John P. Boyd

Dynamics of the Equatorial Ocean

Springer

Dynamics of the Equatorial Ocean

John P. Boyd

Dynamics of the Equatorial Ocean

Springer

John P. Boyd
CLaSP Department
University of Michigan College
of Engineering
Ann Arbor, MI
USA

ISBN 978-3-662-55474-6 ISBN 978-3-662-55476-0 (eBook)
https://doi.org/10.1007/978-3-662-55476-0

Library of Congress Control Number: 2017949138

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Germany
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

“Don’t you know what a headache children are to a good teacher?”
Saint Barsanuphius of Thavatha, first half of the 6th century
To my students,
Shun der Ko, Sue Haupt, Hong Ma, Beth Wingate, Guan-yu Chen, Natasha Flyer, Andrei Natarov, Cheng Zhou and Jianping Xiao and as co-chair Mark Schumack, Javad Abdollahi-Alibeik, Hengchu Han, Wan-li Wu, Laila Guessous, Zhengjie Xu, Lei Wang and Burhan Sadig and as collaborator Zaphiris Christidis, Nam Young Lee, Brett Sanders and Defei Deng and visiting doctoral students and collaborators Shan Li, Zhu Huang and Xiaolong Zhang.
and as always to Marilyn, Ian and Emma

Preface

> Writing projects was always an agony for me … I always tried to live up to Leo Szilard's commandment, 'don't lie if you don't have to do'. I had to. I filled up pages with words and plans I knew I would not follow. When I go home from my laboratory in the next afternoon, I often do not know what I am going to do the next day. I expect to think that up during the night. How could I tell then, what I would do a year hence?
>
> —Albert Szent-Gyorgi (1893–1986), Nobel Laureate, quoted on pg. 382 of *Discovering* by Robert S. Root-Bernstein [1]

This book is literature, not journalism.

Although it is unusual to apply labels like "literature" and "journalism" in science and mathematics, it is nevertheless true that oceanographic observations and contour plots of supercomputer flow are as ephemeral as yesterday's newsprint. The lonely oceanographic research vessels of the twentieth century, laboriously deploying an instrumented mooring here, a tide guage there, painfully struggling with leaking batteries, failed anchor releases, storms and heavy seas, were crewed by heroes. Their data, though, has already been forgotten. Who in his right mind will consult a field experiment in which the whole of the tropical ocean was inferred from a diamond of five moorings when thousands of data points are available from the Argos drifters?

In contrast, good theory is literature. To replace awe with understanding, we must walk the theory road through increasingly complicated approximations. The stepping stones remain unchanged even as the road is extended. In this sense, most of the analysis presented here will be as long-lived as Shakespeare, though without the swordfights and iambic pentameter.

This book reflects my biases. First, most oceanographers flock to oceanographic conferences, exchange the latest oceanographic gossip, and follow each hint of new trends and programs from ONR and NSF with the zeal of the Enigma codebreakers, and talk only to other "water people". They are connoisseurs of five hundred coastal

currents, refer to vortices as "Ring Bob" and "TIV Samantha", and nod sagely at every wiggle in a CTD sounding. Even the theorists share war stories, with rogue waves and forty-degree rolls replacing artillery and charges up the hill, earned by many months at sea. Their job, as they see it, was expressed eloquently by a student whose response was the title of Dallas Murphy's book [2]: *To "Follow the Water"*.

Yes, but a physical and mathematical science is healthy only if some of its citizens follow mathematics and physics, too.

Second, although I have published 69 papers that employ some form of singular perturbation theory, I also have written two long books on Chebyshev, Hermite, and Fourier numerical methods and nearly two hundred articles that employed them. As visualized in Fig. 1, analysis and perturbation expansions ("chirugery") and spectral methods and other numerical algorithms ("arithmurgy") are not separate worlds, but rather the view from the left eye and the right eye fused into a single scene.

This book also bears the scars of its long and difficult genesis. The text began as a graduate course I first taught in the early 1980s. Unfortunately, the audience for my advanced classes dried up, our undergraduate oceanography major went away, and finally my department changed its name to entirely eliminate "oceanic". I told myself that I would return to the book when I was older. My father died at 97, active and independent for almost all of a very long life. Plenty of time, I thought.

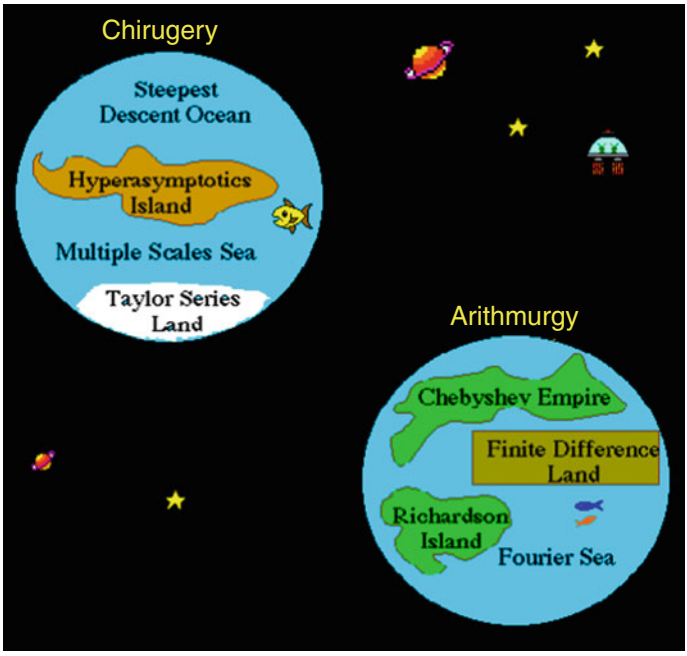

Fig. 1 The Unity of Chirugery and Arithmurgy

Unfortunately, Parkinson's disease intervened. My arms and legs are always sore ("cogwheel rigidity"). My handwriting trails off into tiny random squiggles; three tries and 30 seconds to write a single word that I might perhaps read tomorrow but not in a week ("micrographia"). I limp due to a semi-permanent cramp in my left foot ("dystonia"). Double-clicking usually takes several tries because nothing happens at first ("freezing" or "festination"). I freeze when walking, too. I often move with tiny, shuffling steps with my left arm motionless, though I can sometimes walk almost normally if I consciously take long steps and swing my bad arm. My soft voice is even softer now ("hypophonia") and a little slurred; my wife is urging speech therapy. My typing is slow and error-prone, three or four typos per sentence ("bradykinesia"). With heavy use of dictation software and a student scribe hired by my department, I manage. I have twice reviewed the final manuscript of this book, line by line, for typos, but my typing is so compromised that the errors were twice as numerous as the stars in our galaxy. I ask the reader's forgiveness for the mistakes that remain.

There are some consolations. An incurable, steadily progressive disease is a great foe for procrastination, and otherwise perhaps this book would never have been finished.

Illness has released me from grant-chasing, and that, too, is a blessing.

My department, with few students and great dollops of NASA space money, always expected every faculty member to provide academic year salary support. This became rather awkward after OCE changed its policy to reject such support for tenure track faculty. For 30 years, my proposals had huge budgets for page charges and sundries, but I published only in "free" journals and diverted the funds to pay the department chair's tax.

After many years of failure, I finally got a grant from the Math Directorate! It all went in tax, and I never spent a dime.

My project descriptions were even more fantastic than my budgets, alas. Three years of milestones and detailed plans, focused on a single project, dummied-down to what a graduate student can do. Actually, I averaged 20 to 30 journal articles every 3 years, and hardly ever did I know that I was going to do a problem 6 months in advance of submitting the paper solving same. Year after year, reviewers and program managers scolded me: One project per proposal! Three years! Work plans! Deliverables!

In my other life, I have published twenty science fiction and fantasy short stories. (A list can be found on the Internet Speculative Fiction Database at www.isfdb.org) It is wonderful that my future fantasies will all be literary!

Still, I mourn the terrible waste of time spent preparing and reviewing proposals. Only one of my last fourteen proposals was funded before I was finally cut off by NSF physical oceanography after nearly 40 years. "Not really an oceanographer" was said at the review panels. "Not enough face time at oceanography meetings". A real oceanographer must "follow the water", apparently, even if the water is Powerpoint fakery at a giant conference center in the Rose City. If I still had a face, instead of a Parkinsonian mask, I would even so wish to be judged on what I had written instead of my skill in networking.

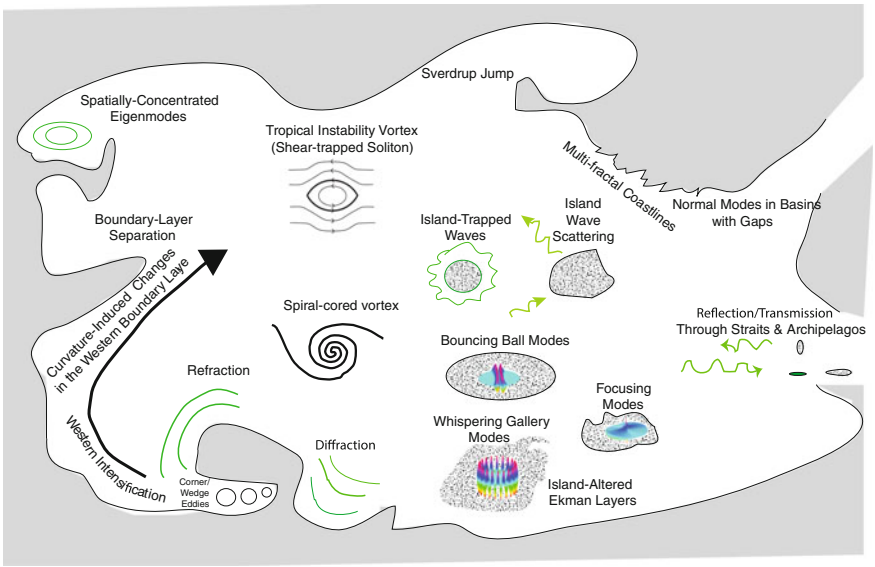

Fig. 2 Geometric Oceanography

An oceanography cutoff from mathematics, insular, is ill-prepared for parallel-in-time reduced basis models. As visualized in Fig. 2, the future of oceanography is as much in geometry as in the connoisseurship of currents.

My 8-year-old laptop can run the LINPACK benchmark four hundred million times faster (50 floating operations per second versus 20 gigaflops) than the HP personal computer I used as a grad student and has a million times more memory (eight kilobytes versus eight gigabytes), but oceanographers do not have a million times more insight. Rather, number-crunching is bound by the following:

First Law of Arithmurgy:
Insight grows no faster than logarithmically with resolution
—J.P. Boyd [3]

Supercomputers shall not accelerate us into a wiser future unless numerical algorithms and oceanographic theory push and push and jump in the bobsled to descend the turns together.

Of course, deans and department chairs were indifferent to such philosophy-of-science issues but just as frowny as the civil servants.

The horror! I program.

The horror! I expand, derive and integrate.

The horror! Didn't I know I was supposed to leave all that to the graduate students? Science is quantized and the quanta are [grad] student triennia.

Where were my big grants? A sausage factory is supposed to turn out sausage in bulk, and where was all my fine doctoral bratwurst? And awards? Surely in a time

when departments and professional societies spend enormous energy on self-congratulation, why was I not a fellow of AGU and AMS? (I was nominated but not elected AMS fellow three years running.) In my own reckoning, to be blessed with the time and opportunity to write four books and 250 journal articles and mentor more than twenty graduate students is prize enough. And I am not yet done.

STEM education is focused entirely on inspiring the young with the joy of science, but many students see professors as Not-Scientists. Faculty at a teaching institution have such heavy teaching loads that they have no time to *do science*, as opposed to merely sharing the classics. Professors at a research university are managers, a hundred parsecs from the lab bench or computer simulation, and are Not-Scientists, too. I suffered for being a Not-Not-Scientist and don't know what to say to students who see no real career in science except an endless succession of postdocs, moving as often as military families, always working on someone's else agenda to impress a review panel, scientific *ronin*.

It is doubtless a great relief in Arlington that the agencies are no longer tortured by someone who consistently Colored Outside the Lines. My chair no longer needs fret about my academic year tax. Our research administrators are happy that one who paid little attention to budgets and balances has been replaced by junior faculty who have been drilled since candidacy in the importance of good management.

I was an awful manager, accountant, and pitchman, but mostly by ignoring those roles entirely, I was mostly, honestly, a scientist, and had a lot of fun. And through blatant dereliction of duty, as defined by deans and federal agencies, I had time to write this book.

I thank Dennis Moore for his review. Thanks also to Peter Gent, James Luyten, Julian P. McCreary, Jr., Mark Cane, and Edward Sarachik for permission to reproduce figures and for helpful comments.

I also express appreciation for permission to reproduce figures to the American Meteorological Society, the Royal Society of London, and the Journal of Marine Research.

I thank Kristina Neal, Temitope Akinlua, Miesha Williamson, and Ninad Ramachandra Naik for help with illustrations.

I thank Leah Lindsey, my scribe, for her assistance, especially with the innumerable corrections.

I acknowledge a little support for more than 35 years from the physical oceanography section of the NSF and also small, non-renewed grants from NASA and the Department of Energy. For many years, my one and only little grant was split-funded between physical oceanography, atmospheric sciences, and computational mathematics, but the latter two dropped out and then physical oceanography refused repeated renewal proposals a few years later.

In view of the huge amount of time lost in reviewing, review panels, and proposal writing, whether the agencies helped or hindered my research is known but to God.

The funds did support some of my students. I was blessed to be chair or co-chair for seventeen doctoral students and two master's students, collaborated on papers with four other students though not an official advisor and mentored three doctoral students who visited from China. I dedicate this book to them.

I am fond of funny words like "arithmurgy", "chirugery", and "hydroarithmomancy" [etymologies in the glossary] not only because I studied Latin and classical Greek and never recovered, but also, I have thought that the spirit of St. Theresa of Avila's advice to her novices was sound for science, too: "God deliver me from frowning saints". It is possible to combine conviction with humility and humor, and I do not apologize for trying my best to do so.

Future work will likely find some of my most charming insights are wrong and some wrong-headed. AIDS was caused only by homosexual contact until it was caused also by heterosexual contact, drug needles, blood transfusions, and mother's milk. All cholesterol was bad until some of it was good. And 30 years of string theory has yield zero testable propositions. Error is not merely a broken formula; sometimes an entire field can chase the ether.

But true scientific understanding is not a drawer of index cards, but rather a structured network which is fault-tolerant to cracks and corrections. And it is also good, both for science and for reading, to reach beyond facts and formulas to a poetic sensibility where number-crunching is arithmurgy and "all of loneliness and grief" is "an empty doorway and a falling leaf".[1]

The poet sees the world on many levels simultaneously, and so must we. Remember, as you read, that John Keats was multiscaled and fractal long before Benoit Mandelbrot.

Ann Arbor, USA John P. Boyd
May 2017

References

1. Root-Bernstein RS (1989) Discovering: inventing and solving problems on the fronters of scientific knowledge. Harvard University Press, Cambridge, Massachusetts
2. Murphy D (2008) To follow the water. Basic Books, New York, p 296
3. Boyd JP (2008) Challenges and controversies in multiscale numerical algorithms for weather forecasting and climate modeling. SIAM News 41(9):1

[1]Archibald MacLeish in "Ars Poetica".

Contents

Symbols & Notation

To conform with existing conventions and previous literature, it is sometimes necessary, to borrow a term from computer science, to "overload" a symbol so that it represents more than one entity. Such overloaded symbols are clearly marked in the table.

f	Coriolis parameter
A	NLS solution, the envelope of a nonlinear wavepacket (Chaps. 2 & 10)
A(Overloaded)	NLS solution, the envelope of a nonlinear wavepacket (Chaps. 16 & 17)
D	Difference variable $\phi - u$
E	Energy
$f_{M/N}$	Padè approximant to $f(x)$, the ratio of a polynomial of degree M divided by a polynomial of degree N
F	Zonal wind stress
g	Gravitational constant
G	Latitudinal wind stress
h	Height field [proportional to pressure]
$H(x)$	Heaviside step function: $H = 1, x > 0$, $H = 0, x < 0$
$H_n(y)$	Unnormalized Hermite polynomials
$\tilde{H}$	Buoyancy flux (forcing for the height/continuity eq.)
k	Zonal (x) wavenumber
k_{ell}	Elliptic modulus, $k_{ell}^2 = m_{ell}$
K	Complete elliptic integral
L	Depth of sponge layer
m	Baroclinic mode number
m (Overloaded)	Latitudinal wavenumber in Sec. 6.2
m_{ell}	Elliptic modulus
m_i	Latitudinal wavenumber of incident wave
m_r	Latitudinal wavenumber of reflected wave
n	Latitudinal mode number

N^2 (Overloaded)	Brunt–Vaisala frequency (squared)
N (Overloaded)	Number of grid points or numerical degrees of freedom
p	Pressure
s	Integer zonal (longitudinal) wavenumber
S	Sum variable $\phi + u$
T	Temperature
$U\,(y)$ *or* $U\,(y,z)$	Longitudinally averaged east–west current
v	North–south velocity
w	Vertical velocity
W	Wavelength
$\tilde{y}$	Mercator latitudinal coordinate, $\tilde{y} \equiv a\,\mathrm{arctanh}\,(\sin(latitude))$
y_t	Turning point [second latitudinal derivative is zero]
z	Vertical coordinate (depth)
$\hat{z}$	WKB vertical coordinate
$\mathfrak{h}_i$	Deviation of top of i-th layer from the mean in multilayer models
$\mathfrak{a}$	Latitudinal scale for different baroclinic modes
$\mathfrak{A}$	Envelope or "modulation" of a wavepacket
$\mathscr{C}$	Carrier wave factor of a wavepacket, $\mathscr{C} = \exp(ik_s[x - c_p(k_s)t])$
$\mathfrak{d}$	sine/cosine switch in basis functions
$\mathfrak{F}$	Inhomogenous term in Bjerknes flow
$\mathfrak{d}$	Modified perturbation parameter for Euler series acceleration
$\mathscr{A}$	NLS solution, the envelope of a nonlinear wavepacket
$\mathscr{B}$	Modified Bernoulli Function
$\mathscr{E}$	Energy density
$\mathscr{F}$	Mass flux $\int_{-\infty}^{\infty} \phi u dy$
$\mathscr{F}$	Zonal mass flux, $\int_{-\infty}^{\infty} \phi u dy$
$\mathfrak{H}$	$d^2/dy^2 - y^2$, Hermite eigenoperator
$\mathscr{L}$	Generic linear differential operator
$\mathfrak{L}$	$(d/dy + y)$, lowering operator
$\mathscr{M}$	Zonal momentum density
$\mathfrak{P}$	$\int_0^Z dx\sqrt{1-x^2} = (1/2)Z\sqrt{1-Z^2} + (1/2)\arcsin(Z)$
$\mathscr{P}$	Zonal pseudomomentum
$\mathfrak{Q}$	Coefficient of ODE in McCreary's linear steady mean current theories
$\mathfrak{R}$	$(d/dy - y)$, raising operator
$\mathfrak{s}$	Vertical mean wind shear perturbation parameter
$\mathfrak{S}_n(\mathfrak{x})$	Hermite function series coefficients for S
$\mathfrak{U}$	Vector Hough-Hermite eigenfunction
$\mathfrak{Y}$	Constant amplitude of homogeneous part of Yoshida jet
$\mathfrak{z}_m(\mathfrak{z})$	Baroclinic mode
α	Thermal expansion coefficient (Chap. 2)
β	$df/dy\big\|_{y=y_0}$, a constant, usual beta-plane meaning

γ	$\tan(\theta)$ where θ is the angle between the local tangent to the coast and due north
δ_{jk}	Kronecker delta; $\delta_{jk} = 0$ if $j \neq k$, 1 if $j = k$
Δ	Divergence (Chap.3)
Δ (Overloaded)	Scale factor for latitudinal tapering (Appendix A)
ε	Lamb's parameter (nondimensional) [Everywhere except in nonlinear chapter]
ε (Overloaded)	Perturbation parameter (Chap. 16)
ζ	Relative vorticity
ζ (Overloaded)	$\varepsilon(x - c_g(k)t)$ in NLS perturbation theory (Chap. 12)
θ	Angle between the local tangent to the coast and a meridian.
λ	Longitude
λ_m	Eigenvalue of vertical structure equation
Λ	Dilation parameter for triad solutions
Λ_n	Auxiliary array defined in (6.35)
Λ	$\frac{\partial}{\partial z}\left\{N^2 \frac{\partial}{\partial z}\right\}$
ν	$2n + 1$ where n is the latitudinal mode number
ξ	Shifted latitudinal coordinate in the gamma-plane approximation
Ξ	Height-varying nondimensional latitudinal coordinate
π	3.14159
ρ	Density
σ	Frequency nondimensionalized by 2Ω
ς	Dilation factor
τ (Overloaded)	Wind stress
τ (Overloaded)	Slow time variable in multiple scales perturbation theory (Chap. 12)
ϕ	Height [pressure/density]
ϕ (Overloaded)	Phase constant for triad solutions in Chap. 17
φ_n	Phase constants
Φ	Phase factor for quasi-normal modes in the long wave approximation
χ	Velocity potential
$\psi_n(y)$ (always with subscript)	Normalized Hermite functions
ψ (without subscript n)]	Streamfunction or mass-weighted streamfunction
Ψ	Streakfunction (streamfunction in a moving coordinate system)
ω	Frequency, nondimensionalized by 2Ω
$\hat{\omega}(y,z)$	$k(U(y,z) - c)$
Ω	Angular frequency of the earth's rotation in radians/second $= 2\pi/86,400\,s^{-1}$
$\aleph$	Parameter with values 0 or 1, used to switch on/off terms in the latitudinal structure equation

א (Overloaded)	$\int^X v_1$ in Boyd 4-mode model (Chap. 6)
ב	Constant of integration for a Bjerknes flow solution
ב	Proportionality constant for radiation in nonlocal solitary waves whose amplitude is proportional to $\exp(-\beth/\varepsilon)$

Chapter 1
An Observational Overview of the Equatorial Ocean

Abstract Latitudinal variations of the Coriolis force refract planetary waves to create a waveguide that confines important wave motions to a narrow belt around the equator. Jets parallel to, rather than perpendicular to, the mean wind arise at low latitudes. Surface jets in opposite directions alternate in latitude. Vertically, jets of alternating signs are stacked one atop the other in the abyssal equatorial ocean like a stack of pancakes. The equatorial oceans have their own waves, their own jets and an ornery indifference to the rules and balances of extratropical dynamics. Some of the key observational phenomena are reviewed in this introductory chapter to lay a foundation for what follows. The alternating surface jets, deep internal jets, seasonal reversals of the Somali Current with the monsoons, upwelling in the Gulf of Guinea and the coupled hemispheric and oceanic oscillation of huge climatic impact, ENSO, are all briefly surveyed.

There is danger that computation becomes a substitute for a nonexistent theory.

Peter Lax, in an address at ICIAM87.

1.1 The Thermocline: The Tropical Ocean as a Two-Layer Model

As a first approximation, the ocean may be divided into two layers: a warm, well-mixed surface layer and a deeper layer of cold water. The boundary between the two is known as the "thermocline". The thermocline can be loosely defined as the middle of the thin layer in which the temperature varies sharply with height from very warm above to very cold below. Figure 1.1 shows a vertical sounding of temperature at the equator. Although almost absent in the polar water, the thermocline is well-defined both in the middle latitudes and in the tropics, particularly the latter. Since the density of sea water is a function of temperature, one can approximate the density of the ocean in a similar idealized way: a layer of uniform density (light!) above the thermocline and a heavier layer of uniform density below.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_1

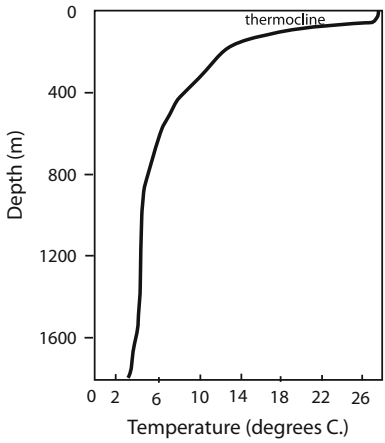

Fig. 1.1 A typical temperature versus depth profile at 0.5° N. The thermocline is missing in polar waters, but is very sharp and well-defined around the equator

These observational facts explain why two-layer models have been a traditional first approximation in oceanography. Experimentally, it is easy to create a two-layer configuration simply by pouring ordinary cooking oil on top of water. [Try it!] This can be done in your own kitchen. It is quite easy to make internal gravity waves on the interface between the two fluids, i.e., on the "thermocline". Analytically, two-layer models give equations for baroclinic instability, etc., which are very similar—in some cases identical—to the two-level models so popular among meteorologists.

In reality, there is some density stratification within the deep ocean and internal waves are possible below and above the thermocline, not merely along the thermocline. However, when the flow is decomposed into vertical modes, most energy goes into the "barotropic mode", which we can identify with wave motions along the air-sea interface, and the "first baroclinic mode", which we can identify with waves along the warm water-cold water interface, i.e., along the thermocline. Furthermore, the linearized equations for each vertical mode are identical except that the depth parameter must be replaced by a different "equivalent depth" for each different vertical mode. Consequently, much theoretical work on the equatorial sea and much numerical modeling, too, has concentrated upon two-layer models, and upon the motion of the thermocline in particular.

The thermocline has a mean depth of about 75–100 m along the equator, but this mean is not very meaningful because the time-averaged depth in the Pacific varies from perhaps 200 m in the west to only about 50 m in the east. (The Atlantic thermocline slopes from about 120 m down off Brazil to only about 20 m in the Gulf of Guinea.) The trade winds piles up water in the west, thus creating a sea surface that slopes from highest in the west to lowest in the east; hydrostatic balance forces the thermocline to tilt in the opposite direction so that it is deepest where the surface of the sea is highest. (The opposite slopes of the sea surface and thermocline are necessary so that the horizontal pressure gradient in the abyss is zero.)

1.2 Equatorial Currents

Figure 1.2 shows a simplified picture of the surface currents. The North and South Equatorial Currents (NEC, SEC) are forced directly by the wind stress exerted on the sea by the trade winds. One can create exactly the same sort of wind-driven currents by blowing a fan over the water. The North Equatorial Counter-Current (NECC) is a return flow towards the east sandwiched between the other two. A major part of the return flow, however, is also carried by the major subsurface current, which is known as the Equatorial Undercurrent (EUC) or as the "Cromwell Current".

The EUC/Cromwell jet is approximately centered on the equator as one would expect, but the system of three surface currents is badly skewed so that the center jet, the North Equatorial Counter-Current (NECC), is a maximum around 6–7° N. The South Equatorial Current, in defiance of its name, is roughly centered on the equator and is, if anything, a little stronger in the northern hemisphere. The North Equatorial Current (NEC) is a puny little weakling in comparison to the other currents, barely attaining peak velocities of 15 cm/s.

The reason for all these asymmetries is the hemispheric asymmetry of the trade wind/Hadley cell circulation that drives it all: the mean location of the Intertropical Convergence Zone (ITCZ), which is the boundary between the southerly and northerly components of the trade winds, the confluence zone where these branches converge and then ascend, is around 7° N., too. (In the mean; the latitude of the ITCZ varies with the season.) Thus, the middle surface jet, the NECC, is located directly at the dynamical equator of the winds, not at the geographical equator. In contrast, the major subsurface flow, the Equatorial Undercurrent, is centered on the geographical

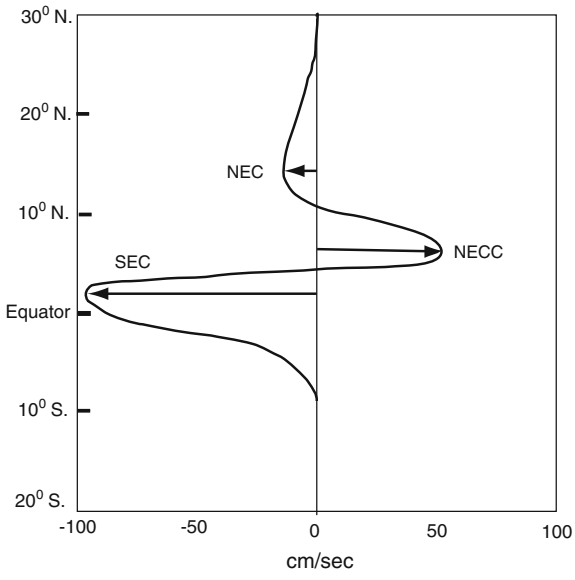

Fig. 1.2 A schematic of the surface equatorial currents

equator: it, too, is ultimately an offspring of the wind, but at a further remove so that where the Coriolis parameter vanishes is more significant to it than where the meridional wind does.

The North and South Equatorial Currents thus are directly underneath the Northeast and Southeast Trade winds so that the direction of these ocean currents is roughly parallel to the wind stress of the Trade Winds. This is rather odd in comparison with the middle latitudes where the turning of the currents by the Coriolis force tends to create currents (wind-driven Ekman layers) in which the vertically-averaged ocean transport is *perpendicular* to the *wind stress*. The NEC and SEC serve as reminders that it is dangerous to blindly apply mid-latitude intuition in the vicinity of the equator. The NEC and SEC are also the westward flowing jets of the great anticyclonic gyres that occupy the major ocean basins. The NEC in the Atlantic is the southern side of the Atlantic gyre whose western side is the Gulf Stream.

In addition, Subsurface Countercurrents (SSCC's), also known as the "Tsuchiya Jets"[1], occur on the flanks of the thermostad. The thermostad is a 150 m-thick layer of water of very uniform temperature (about 13° C.) and salinity (34.9 parts per thousand), which is found in the western Pacific between 150 m and 300 m underneath the surface in the band between 5° N. and S. [2–5]. Their speed is O(20 cm/s) and the transport is 0(5–10 sverdrups). The velocity and transport are weak compared to the

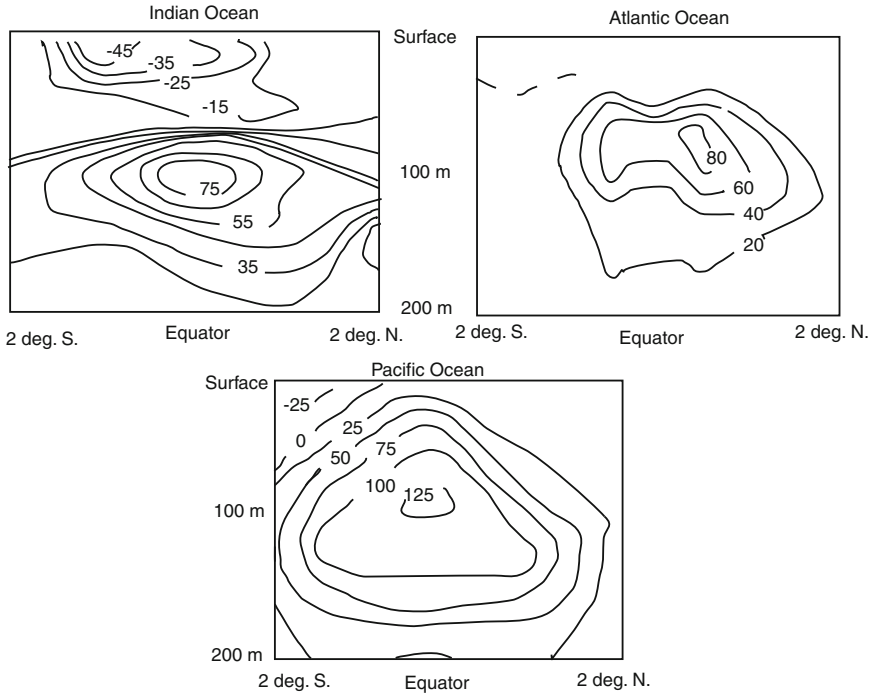

Fig. 1.3 Meridional cross-sections of the zonal currents in cm/s for the three major oceans

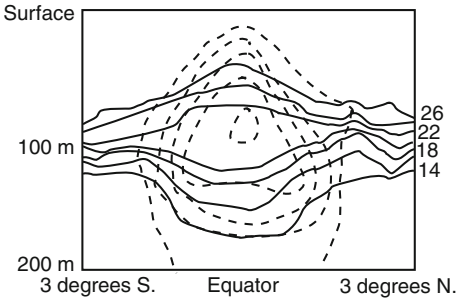

Fig. 1.4 Typical Pacific Ocean velocity contours (*dashed*) superimposed on temperature contours (*solid*) from the same section

0(1 m/s) flow and 20–50 sverdrup transports of the EUC.[1] Even so, the SSCC's are stable and well-defined flows and currents similar to them have also been identified in the Atlantic.

Since the surface currents are directly driven by the wind, the major theoretical challenge has been explaining the Equatorial Undercurrent. We will review and classify the rather bewildering variety of theories later on. In the meantime, Fig. 1.3 shows meridional sections of the zonal current for three different oceans. Because of the monsoon, the Undercurrent in present only part of the year in the Indian Ocean, but when it is, it can still attain peak velocities of 75 cm/s – very fast for the sea. In the Pacific, the EUC can achieve velocities as high as 150 cm/s (Fig. 1.4).

1.3 The Somali Current and the Monsoon

The idealized picture of the equatorial currents given in the preceding section must be modified in the Indian Ocean. Because of the monsoon, the winds over the Indian Ocean blow predominantly north-south rather than east-west. During the Northeast (NE) Monsoon from November to March, the winds blow from the north and east — like the usual NE trades but with a much stronger northerly component than over the other oceans. From May to September, however, the Southwest (5W) Monsoon sends southerly winds over the ocean. (Fig. 1.5).

There is always a South Equatorial Current (SEC), but it is more to the south than in the other oceans, usually coming no closer to the equator than 5° S. latitude. It is stronger during the SW Monsoon than during the NE Monsoon, but even at its time of maximum strength, it is not very conspicuous in Fig. 1.5. because the drawing extends only to 10° S.

In contrast, a westward-flowing surface current north of the equator is present only during the NE Monsoon–in the summer months, it reverses, and the area of the NEC and NECC becomes a single broad current flowing towards the east. Different references use different terminology about this. Knauss [6] describes this as a merging

[1] A sverdrup is one million cubic meters/second.

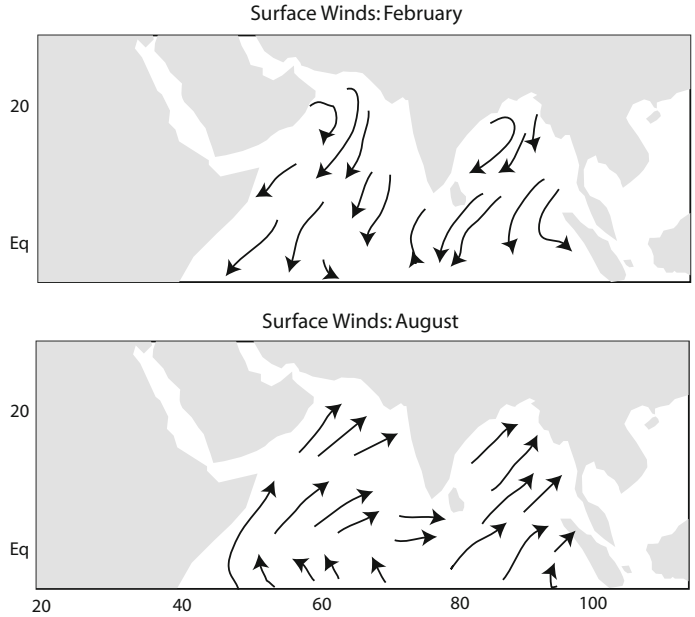

Fig. 1.5 Monsoon surface winds. (*Top*) Northwest Monsoon (February). (*Bottom*) Southwest Monsoon (August)

of the NEC with the Equatorial Countercurrent. Neumann and Pierson (1966) [7] state that the NEC and ECC exist only during the N. Hemisphere winter, and disappear in the summer to replaced by an eastward flowing Monsoon Current. Düing (1978) [8] replaces the NEC throughout the year by a Northeast Monsoon Current (winter) and a Southwest Monsoon Current (summer) but retains an Equatorial Countercurrent (ECC) during the winter months only. We shall generally follow Düing's terminology in these notes. It is worth noting that the NECC is simply the ECC in the Indian Ocean; because the wind patterns are shaped so strongly by the monsoon, the Equatorial Current is strongest on—or even a little to the south—of the equator instead of being confined to the Northern Hemisphere as it is in the other two oceans.

Along the coast of Somaliland flows an intense, seasonally-reversing jet known as the Somali Current. It always flows with the monsoon winds–towards the south in winter and towards the north in summer. Like the Gulf Stream, it is a very intense boundary current because of beta-induced "western intensification". Again like the Gulf Stream, the Somali current separates from the coast and can form meanders and throw off gyres, especially when it separates off Cape Gardafui near the inlet to the Red Sea at about 12° north. However, the Somali Current is more complex than the Gulf Stream because it straddles the equator and reverses with the seasons.

The major theoretical problems have been to (i) generalize Gulf Stream theory to low latitudes (ii) understand coastal separation and formation of gyres and (iii) comprehend the timing of the seasonal reversal. The latter question turns out to

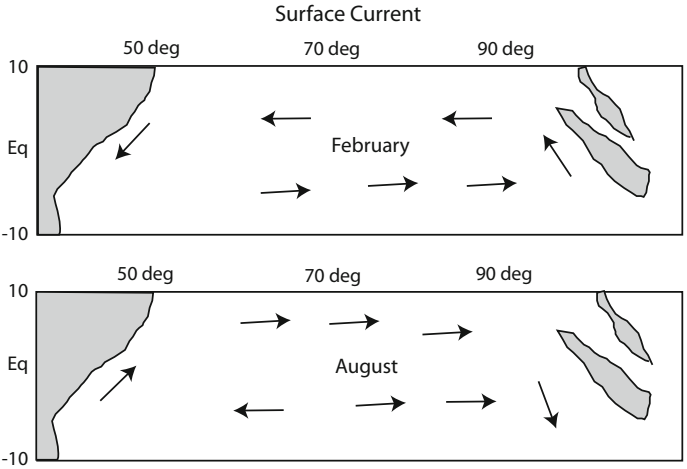

Fig. 1.6 Monsoon-driven surface currents.The *top* panel is at the time of the maximum of the NE Monsoon; the *bottom* during the SW Monsoon

be more complicated than at first sight because the great escarpment of East Africa creates an intense, low-level boundary jet in the atmosphere, too. There is still debate as to whether local wind forcing, i.e., the atmospheric jet, or remote wind forcing, i.e., the changes in wind stress occurring less violently over the whole basin of the Indian Ocean, are more important in making the flow reverse (Fig. 1.6).

1.4 Deep Internal Jets

In 1976, Luyten and Swallow discovered that beneath the well-known upper level jets of the Indian Ocean were a series of weaker but still pronounced alternating jets extending from just beneath the thermocline to nearly the ocean bottom [9]. Although commonly described by the somewhat vague term of "deep internal jet", they are not necessarily axisymmetric, longitude-independent currents like the Equatorial Undercurrent and the surface jets, but may in fact be merely the eddy velocities produced by waves with vertical wavelengths on the order of 300 m. Although their nature is still subject to some argument, later work has shown that these alternating deep jets are a ubiquitous feature of the equatorial ocean. The principle tool in these early observations is the acoustic dropsonde, in which the currents are inferred from the trajectory of a freely falling probe. CTD (Conductivity-Temperature-Depth) profilers have also provided useful information even though the associated temperature perturbation is very small (on the order of a few hundredths of a degree C.). These can be combined with the acoustic dropsonde in the composite instrument called

the Whitehorse, which was developed by Luyten and others at Woods Hole, and the TOPS dropsonde, which is used in Hayes and Milburn (1980) [10].

The work of Luyten and Swallow (1976) [9], Hayes and Milburn (1980) [10], Leetmaa and Spain (1981) [11], Muench, Kunze, and Firing (1994) [12], and Eriksen (1981) [13], and Firing (1987) [14], [9–11, 13] has revealed a number of characteristics of the deep jets including the following:

1. The deep flow is predominantly zonal and has a very different structure (and usually much greater coherence) than the meridional velocity, if the latter is large enough to be accurately measured.
2. The amplitude of the zonal velocity is 0(5–20) cm/s with the larger figure applying just below the thermocline and the smaller figure close to the ocean bottom. In some profiles, the amplitude decreases only slightly as one goes towards the bottom so that 0(10 cm/s) near the bottom is not unheard of.
3. The vertical wavelength is on the order of a few hundred meters; the equivalent depth of such a wavelength is on the order of 0.1 cm.
4. There is usually great zonal coherence, or in other words, the jets seem to have horizontal wavelengths on the order of 1000 km or more.
5. The jets are very highly geostrophic, even to within a fraction of a degree of the equator.
6. Inertial instability of deep currents is now considered a likely suspect for creating the alternating stacked jets.
7. Surface-forced waves ("vertical beams") play a role also.
8. Similar stacked jets in the middle atmosphere, which erupt during periods of intense cross-equatorial latitudinal shear, are known as "pancake instabilities".
9. Similar deep jets are found in the Atlantic, too [15, 16].

1.5 The El Niño/Southern Oscillation (ENSO)

The Southern Oscillation in the atmosphere was discovered by Sir Gilbert Walker in the 1920s. El Niño, the episodic replacement of the nutrient-rich cold, upwelled water off the coast of Peru by a layer of very warm, nutrient-poor water, has been known to Central and South American fishermen for many centuries. It was not until the late 60's, however, that Jacob Bjerknes, already a legend for his contributions as much as forty years earlier, pointed out that the two were merely the atmospheric and the oceanic manifestations of a single, coupled oscillation of both wind and water, a view that is now universally accepted.

The name "Southern Oscillation" is rather unfortunate because it seems to convey the impression of a phenomenon of the Southern Hemisphere whereas in reality the SO is centered on the equator. Walker used the name to distinguish this equatorial oscillation from another cycle he found involving the Icelandic High and the Aleutian High. It must be admitted that the equator is far to the south of Iceland. Although ENSO is a very complex phenomenon, Walker discovered that it could

be characterized by a rather simple "Southern Oscillation Index". The SO Index is simply the monthly mean pressure difference between Port Darwin (15° S., 130° E.) on the northern coast of Australia and another station many thousands of kilometers to the east, usually either Tahiti (15° S., 150° W.) or Santiago, Chile (33° S., 70° W.). Since Darwin and Santiago are roughly 15,000 km apart, almost on opposite sides of the globe, the SO Index is mute testimony to the global scale of ENSO.

A low value of the Tahiti-Darwin difference is accompanied by warm water off Peru (El Niño), little SST (sea surface temperature) difference between the western and eastern Pacific, a deep ocean thermocline (60–120 m) throughout the Pacific, westerly wind anomalies in the western Pacific, heavy rain both on the Line Islands (equatorial, at 160° W.) and in Ecuador, and sinking flow in the western Pacific accompanied by compensating rising motion in the eastern Pacific. This side-to-side sloshing of air – perpendicular to the predominantly north-south, axisymmetric Hadley flow – is known as the Walker circulation. ENSO has been a hot topic for a couple of decades: a monster El Niño occurred in the middle of 1982, and the following winter was mild in the Midwest but terrible in California and the Rockies. Australia endured the "Great Dry", the worst drought that country has known in a generation. The water in the eastern Pacific became so warm over so large an area that it wreaked havoc with global circulation patterns. The relationship between the ENSO and the changes in the position of the mid-latitude jet stream, for example, are not well understood, but this is precisely why a deeper understanding of the ENSO has become an urgent priority. Major El Niño episodes occur at intervals of five years or so, but the phenomenon is highly irregular and intermittent.

In spite of its great importance, the discussion of ENSO here and in the rest of the book will be somewhat limited. The reason is that these topics are covered by excellent monographs by Clarke (2008) [17], Sarachik and Cane (2010) [18] and Philander (1990) [19].

1.6 Upwelling in the Gulf of Guinea

Upwelling, i.e., a positive vertical velocity in the upper layer of the sea, occurs in the Gulf of Guinea only between June and October even though the local winds are favorable for upwelling (i.e., tend to drive upwelling by creating a surface Ekman layer) throughout the whole year and have no seasonal variability. The French have long maintained an oceanographic research station at Abidjan in the Republic of the Ivory Coast, right on the Gulf of Guinea, so this seasonal dependence of the upwelling and the invariability of the winds are both well-documented. Although this phenomenon is not in and of itself profoundly important, it is interesting that it was one of the earliest physical events to be explained in terms of equatorial waves in the note of Moore et al.(1978) [20]. The theory of these eight authors is based on the fact that while the winds are steady in the eastern Atlantic in the Gulf of Guinea, there is a sudden seasonal onset of winds in the western part of the ocean off the

coast of Brazil. This in turn will excite a Kelvin wave which will propagate from west to east to create conditions favorable for upwelling.

There are dissenting views in the literature; see, for example, Philander (1979) [21]. However, Verstraete's lengthy (1992) review concludes that remote forcing is important [22].

The Gulf of Guinea upwelling is a simple and clear-cut example of how remote events can drastically alter local dynamics through the propagation of planetary waves across the ocean.

1.7 Seasonal Variations of the Thermocline

The reversal of the Somali Current and the upwelling in the Gulf of Guinea are but two examples of many seasonal events that are important in the ocean. Another is the complex pivoting of the thermocline in the Atlantic with the seasons. The thermocline normally is tilted to be deeper in the west in consequence of the wind stress of the trade winds. Merle (1983) [23] shows that the thermocline simultaneously pivots about two perpendicular axes.

One pivot is a see-saw in an east-west direction with a fulcrum at about 25° west longitude. Most of the variation occurs in the eastern half of the ocean, so the pivoting is not symmetric about the midpoint of the ocean. It seems to be a fairly straightforward response of the thermocline to the seasonal changes in the strength of the trade winds. The other pivot is about a fulcrum located somewhere between 3 and 8° north latitude, which through no coincidence whatsoever is roughly the mean location of the Intertropical Convergence Zone (ITCZ). It is well-known that the thin line of clouds which is the ITCZ moves north and south with the seasons with a mean position several degrees into the northern hemisphere. The north-south wind stresses created by flow converging into the ITCZ are necessarily affected by the ITCZ's cyclic north-south movements.

The numerical models of Cane and Sarachik (1981) [24] and Busalacchi and Picaut (1983) [25] have been remarkably successful in explaining this double pivot of the thermocline. In turn, these observations of seasonal motion have been the impetus for a rather large number of both theoretical and numerical efforts focused on a common goal: How does the equatorial ocean respond to periodic variations in the applied wind stress, especially the annual cycle?

1.8 Summary

The observed phenomenon described above have been deliberately sketched in rather limited detail because we will return to them, and to the theories and models that attempt to explain them, in later chapters. The purpose of this once-over-lightly

Table 1.1 Selected reviews on equatorial dynamics

Reference	Notes
Moore and Philander (1977) [26]	Both observations and theory
McCreary (1980) [27]	Systematic exposition of 3D separable models
Philander (1980) [28]	Equatorial Undercurrent
Leetmaa, McCreary and Moore (1981) [29]	Brief treatment of both observations and theory
McCreary (1985) [30]	Modeling
McPhaden and Ripa(1990) [31]	Wave mean flow interactions
Verstraete (1992) [22]	Seasonal upwellings in the Gulf of Guinea
Neelin, Latif and Jin (1994) [32]	Coupled ocean-atmosphere models
Battisti and Sarachik (1995) [33]	ENSO: Description and prediction
Hamilton (1998) [34]	Equatorial middle atmosphere; quasibiennial oscillation
McPhaden et al. (1998) [35]	TOGA observing system
Dijkstra and Burgers (2002) [36]	El Niño
Chelton, Schlax and Lyman (2003) [37]	Equatorial Rossby in latitudinal mean shear
Philander and Fedorov (2003) [38]	El Niño
Kessler (2006) [39]	Eastern tropical Pacific
Zeitlin (2007) [40]	Nonlinear waves and numerical geostrophic adjustment
Fedorov and Brown (2009) [41]	Encyclopedia article
Khoulder, Majda and Stechmann (2013) [42]	Climate in the tropics, PDE theory, mostly atmosphere

is to give a feel for the diversity of phenomenon in the equatorial ocean. Further background can be found in the review articles collected in Table 1.1.

References

1. Tsuchiya M (1975) Subsurface countercurrents in eastern equatorial Pacific Ocean. J Marine Res 33(Supplement S):145–175
2. McPhaden MJ, Semtner AJ (1984) On the dynamics of equatorial subsurface countercurrents. J Phys Oceanogr 14(7):1216–1225
3. Donohue K, Firing E, Rowe GD, Ishida A, Mitsudera H (2002) Equatorial Pacific subsurface countercurrents: a model/data comparison in extreme coordinates. J Phys Oceanogr 32:1252–1264
4. Rowe GD, Firing E, Johnson GC (2000) Pacific equatorial subsurface countercurrent philosophy, transport, and potential vorticity. J Phys Oceanogr 30:1172–1187
5. Firing E, Wijffels S, Hacker P (1998) Equatorial subthermocline currents across the Pacific. J Geophys Res 103(21):413–423
6. Knauss JA (2005) Introduction to physical oceanography, 2nd edn. Waveland Press, Portland, p 320

7. Neumann G, Pierson WJ Jr (1966) Principles of physical oceanography, 2nd edn. Prentice-Hall, New Jersey, p 545
8. Düing W (1978) The Somali Current: past and recent observations. In: Moore DW, Witte J (eds) Review papers of equatorial oceanography FINE workshop proceedings. FGGE/Index/Norpax Equatorial, Nova University/N. Press, Dania, Florida, Y. I. T, pp 1–31
9. Luyten JR, Swallow JC (1976) Equatorial undercurrents. Deep Sea Res 23:999–1001
10. Hayes SP, Milburn HB (1980) On the vertical structure of velocity in the eastern equatorial Pacific. J Phys Oceanogr 10:633–635
11. Leetmaa A, Spain PF (1981) Results from a velocity transect along the equator from 125-degrees-w to 159-degrees-w. J Phys Oceangr 11(7):1030–1033
12. Muench J, Kunze E, Firing E (1994) The potential vorticity structure of equatorial deep jets. J Phys Oceanogr 24:418–428
13. Eriksen CC (1981) Deep currents and their interpretation as equatorial waves in the western Pacific Ocean. J Phys Oceangr 11(1):48–70
14. Firing E (1987) Deep zonal currents in the central equatorial Pacific. J Marine Res 45:971–982
15. Johnson GC, Kunze E, McTaggart KE, Moore DW (2002) Temporal and spatial structure of the equatorial deep jets in the Pacific Ocean. J Phys Oceanogr 32(12):3396–3407
16. Claus M, Greatbatch RJ, Brandt P (2014) Influence of the barotropic mean flow on the width and the structure of the Atlantic equatorial deep jets. J Phys Oceangr 44(9):2485–2497
17. Clarke AJ (2008) An Introduction to the dynamics of El Niño and the southern oscillation. Academic Press, New York, p 324
18. Sarachik ES, Cane MA (2010) The El Niño-southern oscillation phenomenon. Cambridge University Press, Cambridge, p 384
19. Philander SGH (1990) El Niño, La Niña and the southern oscillation. Academic Press, New York, p 293
20. Moore DW, Hisard P, McCreary JP Jr, Merle J, O'Brien JJ, Picaut J, Verstraete J, Wunsch C (1978) Equatorial adjustment in the eastern Atlantic. Geophys Res Lett 5:637–640
21. Philander SGH (1979) Upwelling in the Gulf of Guinea. J Marine Res 37:23–33
22. Verstraete J (1992) The seasonal upwellings in the Gulf of Guinea. Prog Oceanogr 29(1):1–60
23. Merle J (1983) Seasonal variability of subsurface thermal structure in the tropical Atlantic Ocean. In: Nihoul JCJ (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 31–50
24. Cane MA, Sarachik ES (1981) The response of a linear baroclinic equatorial ocean to periodic forcing. J Marine Res 39:652–693
25. Busalacchi AJ, Picaut J (1983) Seasonal variability from a model of the tropical Atlantic Ocean. J Phys Oceangr 13(9):1564–1588
26. Moore DW, Philander SGH (1977) Modelling of the tropical oceanic circulation. In: Goldberg ED (ed) The sea. Wiley, New York, pp 319–361
27. McCreary, Jr JP (1980) Modelling wind-driven ocean circulation. Technical report, Hawaii Institute of Geophysics/JIMAR, Honolulu, School of Ocean and Earth Science and Technology, p 64
28. Philander SGH (1980) The equatorial undercurrent revisited. Annu Rev Earth Planet Sci 8:191–204
29. Leetmaa A, McCreary JP Jr, Moore DW (1981) Equatorial currents: observations and theory. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge, pp 184–234
30. McCreary JP (1985) Modeling equatorial ocean circulation. Ann Rev Fluid Mech 17(1):359–409
31. McPhaden MJ, Ripa P (1990) Wave-mean flow interactions in the equatorial ocean. Annu Rev Fluid Mech 22:167–205
32. Neelin JD, Latif M, Jin FF (1994) Dynamics of coupled ocean-atmosphere models - the tropical problem. Annu Rev Fluid Mech 26(1):617–659
33. Battisti DS, Sarachik ES (1995) Understanding and predicting ENSO. Rev Geophys 33(Part 2, S):1367–1376

34. Hamilton K (1998) Dynamics of the tropical middle atmosphere: a tutorial review. Atmosphere 36(4):319–354
35. McPhaden MJ, Busalacchi AJ, Cheney R, Donguy JR, Gage KS, Halpern D, Ji M, Julian P, Meyers G, Mitchum GT, Niiler PP, Picaut J, Reynolds RW, Smith N, Takeuchi K (1998) The tropical ocean global atmosphere observing system: a decade of progress. J Geophys Res Oceans 103(C7):14,169–14,240
36. Dijkstra HA, Burgers G (2002) Fluid dynamics of El Nino variability. Annu Rev Fluid Mech 34:531–558
37. Chelton DB, Schlax MG, Lyman JM, Johnson SC (2003) Equatorially trapped Rossby waves in the presence of meridionally sheared baroclinic flow in the Pacific Ocean. Prog Oceanogr 56(2):323–380
38. Philander SG, Fedorov A (2003) Is El Nino sporadic or cyclic? Annu Rev Earth Planet Sci 31:579–594
39. Kessler WS (2006) The circulation of the eastern tropical Pacific: a review. Progr Oceanogr 69(2–4):181–217
40. Zeitlin V (2007) Nonlinear wave phenomena in rotating shallow water with applications to geostrophic adjustment. In: Zeitlin V (ed) Nonlinear dynamics of rotating shallow water: methods and advances. Elsevier, Amsterdam, pp 257–321
41. Fedorov AV, Brown JN (2009) Equatorial waves. In: Steele J (ed) Encyclopedia of Ocean sciences. Academic Press, San Diego, pp 3679–3695
42. Khoulde B, Majda AJ, Stechmann SN (2013) Climate science in the tropics: waves, vortices and PDEs. Nonlinearity 26(1):R1R68

Chapter 2
Basic Equations and Normal Modes

Abstract This chapter presents a systematic linearization of the equations of motion about a state of rest, followed by additional approximations, each carefully analyzed for estimated error, culminating in a full four-dimensional separation of variables. The general solution to these approximate equations is a superposition of normal modes. Each mode is the product of four one-dimensional factors. The latitudinal structure of the modes is a Hermite function or the sum of a pair of Hermite functions. The vertical structure functions are the eigenmodes of the vertical structure equation. The normal mode is the product of a factor of latitude multiplied by a function of depth z only and lastly multiplied by $\exp(ik(x - ct))$.

> A traveler who refuses to pass over a bridge until he has personally tested the soundness of every part of it is not likely to go far; something must be risked, even in mathematics.
>
> Sir Horace Lamb

2.1 Model

The fundamental assumption for studying fluid dynamics at low latitude is the "equatorial beta-plane". In this approximation, the factors of sine and cosine of latitude are approximated by power series about zero latitude, i.e., the cosine is replaced by one and the sine is replaced by latitude itself. The rapid increase of the Coriolis force as one goes away from the equator confines a broad class of waves and jets to very low latitudes where these geometric approximations are quite good. Typically, the atmospheric equatorial waves have little amplitude beyond 20° — 1/3 of a radian — north and south of the equator while the ocean motions are much narrower still. Since $\cos\theta = 1 - (1/2)\theta^2 + ...$ and $\sin(\theta) = \theta - \theta^3/6$, we see that the relative error in retaining only the lowest order terms is no worse than 5% for either function as long as $\theta <= 1/3$.

By geophysical standards, this is a very small error: more important still, it is a purely numerical error that does not change the qualitative behavior of the equations of motion. The great advantage of these trigonometric approximations is that they permit one to replace *spherical* coordinates latitude and longitude by *Cartesian*

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_2

coordinates x and y, which are simply latitude and longitude multiplied by the radius of the earth.

McCreary (1980) is a very detailed treatment of the assumptions and approximations made in equatorial fluid mechanics [1]. The equatorial beta-plane is so basic that he does not even mention it, assuming correctly that his audience will already be familiar with it! Following his practice of rating assumptions with labels in "all caps", we will supply what he omitted for this one approximation.

Assumption 2.1 Equatorial beta-plane

EXCELLENT for all equatorial motions that are truly equatorial, i.e., confined to low latitudes, provided that the zonal length scale is not too small.[1]

The equations of motion on the equatorial beta-plane then take the form

$$u_t + uu_x + vu_y + wu_z - fv + (1/\rho)p_x = (\nu u_z)_z \tag{2.1}$$

$$v_t + uv_x + vv_y + wv_z + fu + (1/\rho)p_y = (\nu v_z)_z \tag{2.2}$$

$$(1/\rho)p_z = -g \tag{2.3}$$

$$\rho_t + (u\rho)_x + (v\rho)_y + (w\rho)_z = 0 \tag{2.4}$$

$$\rho_t + u\rho_x + v\rho_y + w\rho_z = (\gamma\rho_z)_z \tag{2.5}$$

where subscripts denote differentiation with respect to the subscripted variable and ν is viscosity, γ is the thermal diffusivity and p is pressure. All symbols are defined in the "Symbols and Notation" section of the frontmatter.

On the equatorial beta-plane, the Coriolis parameter $f = 2\Omega y$ where Ω is the angular frequency of the earth's rotation. Parenthetically, note that (2.1)–(2.2) are also the proper equations for the "mid-latitude beta-plane" in which the Taylor expansions of the sine and cosine of latitude are about a middle latitude, usually 45°, rather than the equator. The separation-of-variables procedure is quite unaffected by the approximation we use for f.

Even in these equations, simplified further below, there are implicit a number of additional assumptions including the following:

Assumption 2.2 Neglect of vertical Coriolis force.

These terms are normally extremely small and this assumption is always made even in numerical models. McCreary rates this as GOOD; I would say EXCELLENT.

This is part of the "traditional approximation" in meteorology [2, 3] and is in fact necessary for energetic consistency with other simplifications of the "traditional

[1] In Sect. 3.12, it is shown that the latitudinal scale is controlled by $E + s^2$ where E is the nondimensional parameter known as"Lamb's parameter" (defined below as 2.34) and s is the integer zonal wavenumber. The classical equatorial beta-plane must be generalized by keeping some additional terms in order to capture the zonal wavenumber effect on the north-south width of the waves. Fortunately, because $\varepsilon \sim O(10^5)$ or larger for ocean baroclinic modes, it is only for wavenumbers $s \sim O(300)$, or equivalently for east-west wavelengths of a couple of hundred kilometers or less, that an "extended equatorial beta-plane" is needed.

approximation". For dissenting views, see [4] and the review [5] and also [6–14, 14].

Assumption 2.3 Horizontal viscosity is neglected.

This would be a bad assumption in a pipe flow, but the horizontal scale of large-scale motions in both the atmosphere and ocean is very large (25–100 times) the vertical scale, so the errors in ignoring horizontal viscosity are only $O(10^{-4})$ of the vertical viscosity effects even when the horizontal viscosity coefficient is as large as the vertical viscosity coefficient. McCreary describes this as SENSIBLE; I would say EXCELLENT. (It should be noted, however, that many numerical models keep horizontal viscosity for computational stability.)

Assumption 2.4 Hydrostatic approximation.

This is poor for very small scales (wavelengths of a couple of kilometers or less), but it is a terrific approximation for the large-scale waves of interest here in both the atmosphere and ocean. Another way of putting it is that the vertical acceleration dw/dt is important only for motions whose time scales are on the order of $2\pi/$ the Brunt–Vaisala frequency. McCreary rates hydrostaticity as GOOD.[2]

Assumption 2.5 Incompressibility.

Equation (2.4), which expresses continuity of mass, is almost never used in that form. By subtracting (2.5), which is the heat equation, from it we obtain

$$\rho(u_x + v_y + w_z) = -(\neg\rho_z)_z \tag{2.4*}$$

We now simplify this by ignoring the right-hand side of (2.4*):

$$u_x + v_y + w_z = 0 \tag{2.4**}$$

Veronis [15] has shown that this is VERY GOOD for a typical ocean wind-driven circulation.

An important note: Veronis shows that the right-hand side of the heat equation (2.5) generally cannot be neglected. However, some authors do replace (2.5) by an equation which the temperature rather than the density ρ is the actual unknown. Semtner and Holland [16] use the approximate equation of state

$$\rho = \rho_0[1 - \alpha(T - T_0)] \tag{2.6}$$

$$\alpha = 0.0002/degree \tag{2.7}$$

where ρ_0 and T_0 and the thermal expansion coefficient α are constants. Equation (2.6) ignores salinity, which can be very important in general oceanography, but salinity

[2]The hydrostatic approximation must fail sufficiently close to breaking for a nonlinear Kelvin wave as described in Chap. 16.

differences are largest in northern waters where calving icebergs can overlay the salty seas with relatively fresh water. Salinity differences seem to be of secondary importance for the equatorial ocean, so they have been largely ignored up to now, and we shall do the same. Equations (2.6) and (2.7) make it possible to replace (2.5) with the alternative (but equivalent) heat equation

$$T_t + uT_x + vT_y + wT_z = (\neg \rho_z)_z \tag{2.5*}$$

The system of Eqs. (2.1)–(2.3), (2.4**) and either (2.5) or (2.5*) are the basic set of nonlinear equations that has been the basis for multi-level numerical simulations of the ocean. Atmospheric modellers use an equation of the same form except that (i) z is replaced as a vertical coordinate by pressure or log-pressure (ii) vertical viscosity is usually ignored and (iii) the eddy conductivity term in (2.5*) is normally replaced by an explicit heating function.

To obtain *analytical* models of ocean flow, it is necessary to simplify the equations still further. The first and most important approximation is to drop the nonlinear terms. There are analytical theories that include nonlinear effects, but these are perturbative calculations in which the starting point is the linearized set of equations. Hence, linearize-and-separate-variables is the fundamental first step in all analytical calculations up to the present.

Assumption 2.6 Drop *all* nonlinear terms in the momentum equations.

McCreary rates this QUESTIONABLE; I would add the qualifier NECESSARY to obtain linear, analytically solvable equations. This assumption will be relaxed in Chaps. 16 and 17 on instability and nonlinear waves.

Assumption 2.7 Drop all nonlinear terms in the heat equation *except* for wdT_0/dz (or $wd\rho_0/dz$) where $T_0(z\,\text{only})$ and $\rho_0(z\,\text{only})$ represent the mean temperature /density stratification of the ocean, which is assumed to be a function only of the vertical coordinate.

This also is QUESTIONABLE but NECESSARY. This one nonlinear term cannot be neglected, even in a lowest approximation, because this term is *buoyancy*. If it is missing, then the linearized equations have no resistance to vertical motions and cannot respond to thermal driving. In reality, the density stratification of both the ocean and atmosphere strongly resists vertical motion, which is why the synoptic high and low pressure systems on a daily weather map are typically a hundred times larger in the horizontal dimension than the vertical. Thus, we cannot even come close to the right answer unless density stratification is included.

Fortunately, the mean density/temperature profile is known in both air and sea. It does vary with altitude and longitude, too, but rather slowly; in particular, the ocean and atmosphere are stably stratified over the whole globe. Consequently, inserting a mean temperature or density that is a function only of depth is a satisfactory first approximation.

It should be noted that the mean buoyancy term must also be inserted into the fully nonlinear heat equation, (2.5) or (2.5*), since numerical models usually are

not sufficiently sophisticated and sufficiently high resolution to parameterize the small-scale convection, radiative transfer, etc. well enough to correctly generate the observed stratification from first principles. Thus, even for numerical models, the mean stratification is usually specified rather than internally generated.

Assumption 2.8 Assume special forms for mixing coefficients:

$$\nu = \urcorner = A/N^2 \tag{2.8}$$

where A is a constant and N is the mean Brunt–Vaisala frequency defined by

$$N^2 = -\frac{g}{\rho_0}\frac{d\rho_0}{dz} \tag{2.9}$$

and modify the form of the eddy mixing of heat to

$$(\urcorner\rho_z)_z \longrightarrow (\urcorner\rho)_{zz} \tag{2.10}$$

McCreary does not rate this assumption, but I would describe it as CONVENIENT and AS REASONABLE AS ANYTHING ELSE. The reason that it is convenient is that these forms for the mixing are the only known forms that permit separation of variables. Whatever their lack of realism, these approximations give analytical models with damping, and with a damping that increases with decreasing vertical scale as is true of almost any physically reasonable dissipation.

In point of fact, we need not feel too guilty about using these approximations. For laminar laboratory flow, the viscosities are molecular and have known, measured values, and off-the-wall assumptions are unneeded. In the atmosphere and ocean, however, the dominant mixing is by small-scale turbulence. Since these are known only to within an order-of-magnitude anyway — questionable approximation — (2.8)–(2.10) really are reasonable as any plausible alternative.

The Assumptions (2.8)–(2.10) were first used by Fjelstad and Mork to study internal waves and by McCreary (1980) to explore equatorial dynamics. When small-scale turbulence provides the mixing, it is reasonable that heat and momentum will be mixed with equal efficiency even though the *molecular* viscosity and diffusivity are different. The greater the static stability, the larger the Brunt–Vaisala frequency N; we expect that greater static stability would, by resisting vertical motion, reduce the vertical eddy mixing, and the forms (2.8) explicitly allow for this. However, there is no particularly good physical reason to suppose that ν and $\urcorner$ must decrease precisely as the square of N; the argument merely implies that bigger N will make ν and $\urcorner$ smaller.

Assumption 2.9 The pressure is replaced by a new variable ϕ such that

$$\phi_x = (1/\rho)p_x; \qquad \phi_y = (1/\rho)p_y \tag{2.11}$$

Equation (2.11) is both a definition and an approximation. The definition is the introduction of the new variable ϕ. This is a meteorologist's notation rather than an oceanographer's; the "geopotential" ϕ always replaces the pressure when pressure or its logarithm is used as the vertical coordinate, and no further approximations to justify (2.11) are needed in the atmospheric case except the hydrostatic approximation, which we have already made. In the ocean, taking

$$\phi \equiv p/\rho_M \tag{2.12}$$

where ρ_M is a constant equal to the vertical average of the basic state density, $\rho_0(z)$, is necessary to obtain a simple relation between p and its replacement, ϕ. This approximation of ignoring density variations in the *momentum* equations while leaving ρ *untouched* in the *heat* and *hydrostatic* equations is known as the Boussinesq approximation. Since $(\rho - \rho_0)/\rho$ is $O(10^{-2})$ or smaller in the ocean, the Boussinesq approximation is VERY GOOD.

The replacement of p by ϕ is very convenient. McCreary [1] simply drops the factor of ρ_0 and uses p as the symbol for both the pressure and for the pressure divided by the mean density. It seems less confusing to borrow the meteorologist's symbol.

All these simplifications finally give the linearized set of equations

$$u_t - fv + \phi_x = [(A/N^2)u_z]_z \tag{2.13a}$$

$$v_t + fu + \phi_y = [(A/N^2)v_z]_z \tag{2.13b}$$

$$\rho_M \phi_z = -\rho g \tag{2.13c}$$

$$\rho_t - (\rho_M/g)N^2 w = [(A/N^2)\rho]_z z \tag{2.13d}$$

$$u_x + v_y + w_z = 0 \tag{2.13e}$$

where we have ignored the distinction between ρ_M and $\rho_0(z)$ in the denominator of N^2 in obtaining (2.13d).

2.2 Boundary Conditions

The boundary conditions of the separable model are four at the top and four at the bottom; since the system is eighth order in z (i.e., 8 z-derivatives appear in these equations), one unknown (ϕ, or in other words, the pressure) must be left unconstrained on the boundaries. The imposed conditions are:

SURFACE: (at $z = 0$)

$$\nu u_z = \tau^x \tag{2.14a}$$
$$\nu v_z = \tau^y \tag{2.14b}$$
$$w = 0 \tag{2.14c}$$
$$\rho = \rho_0 (z = 0) \tag{2.14d}$$

where τ^x and τ^y are the surface wind stresses on the ocean in the x and y directions.

The first two boundary conditions are physically correct and are used even in nonlinear numerical models. The difficulty is that wind stress is created by turbulent air-sea interaction which is parameterized by formulas relating wind stress to wind speed. Once the wind stress has been calculated by some formula/incantation/voodoo, the boundary conditions are mathematically correct and induce no further error.

Assumption 2.10 Rigid lid.

Taking $w = 0$ at the top of the ocean is physically equivalent to imposing a rigid lid at $z = 0$ and is obviously unrealistic; an ocean voyager sees only an endless field of waves. However, when we separate variables and analyze the structure of individual vertical modes, we find that setting $w = 0$ at $z = 0$ has a major effect only on the barotropic mode, that is to say, the mode in which the horizontal currents are independent of depth. To be sure, all the baroclinic modes have very small but non-zero vertical velocities at the sea surface. However, setting $w = 0$ at the top for the baroclinic modes creates an error of $O(\Delta\rho/\rho)$ where $\Delta\rho$ is the variation of the density with depth. Since this is $O(1/250)$ for the equatorial ocean, we make only a negligible error for the baroclinic modes by making this approximation.

But what about the barotropic mode? As we shall see below, only the baroclinic modes support equatorially trapped motions, so we must largely exclude the barotropic mode from consideration anyway. McCreary therefore rates this approximation as GOOD.

Assumption 2.11 Density = mean density at surface; equivalently, temperature = mean temperature at the surface.

This assumption is "UNPLEASANT" in McCreary's words because it demands that the atmosphere act as a constant-temperature source of heat, and does not allow the sea surface to change with time, latitude, or anything else. A coupled ocean-atmosphere model would *never* make this assumption because in reality, the situation is the other way around. The ocean surface temperature does change, and this provides a constant-temperature source of heat for the atmosphere. Namias' method of long-range weather forecasting, for example, is based on looking at sea surface temperatures and their effect on the atmosphere. A more realistic condition would be to specify the heat flux at the ocean surface, or better yet, some kind of radiative balance equation involving such complications as evaporation and precipitation, albedo, and so on.

Nonetheless, this assumption is essential to separating variables, so for analytical theory, McCreary rightly labels it as UNFORTUNATE but NECESSARY.

The bottom boundary conditions are

BOTTOM: (at $z = -D$ where D is depth of the ocean, a constant)

$$\nu u_z = 0 \tag{2.15a}$$

$$\nu v_z = 0 \tag{2.15b}$$

$$w = 0 \tag{2.15c}$$

$$\rho = \rho_0(z = 0) \tag{2.15d}$$

Assumption 2.12 Constant bottom depth D

In reality, the bottom topography of the sea is highly irregular and this can give rise to all kinds of complicated effects: topographic Rossby waves, jets that follow the depth contours, topographic scattering of internal waves, etc., etc. However, the important equatorial motions – at least the sort that have been studied up to now – do not extend to the bottom of the abyss. The equatorial ocean is sufficiently deep in most places so that topographic effects are not important; topography is most important in very shallow seas and along the continental shelves. I rate this assumption as NECESSARY, but NOT BAD for the class of motions we are interested in. This assumption of a flat bottom then forces us to impose (2.15c); note over *sloping* topography, the vertical velocity w does not have to be 0.

Assumption 2.13 No stress at the bottom.

A numerical model would use the normal viscous condition of no horizontal velocities at the bottom; the condition of no stress filters out a bottom Ekman layer driven by the geostrophic current just above the bottom. However, for surface trapped motions, there is no bottom Ekman layer anyway. While Ekman spin-down could in principle be an important dissipative mechanism, the turbulent eddy viscosities are so low in the deep ocean in comparison to the surface layer that the bottom Ekman layer can often be ignored when there are bottom currents. McCreary rates this assumption as NECESSARY and NOT BAD.

Assumption 2.14 No temperature changes at the sea bottom.

In view of the approximate equation of state (2.6), requiring that the density equal the mean density at the bottom has multiple implications.It is equivalent to compelling the bottom temperature to equal the mean temperature. This assumption also forces the sea-bottom to act as a constant temperature source of heat to the ocean, i.e., a reservoir of infinite heat capacity and conductivity. This is a little unrealistic, but the lack of realism is not serious for surface-trapped motions. McCreary rates this NECESSARY and NOT BAD.

For clarity, the full list of assumptions is collected in Table 2.1.

As noted by McCreary, the single most restrictive assumption implicit in the boundary conditions is the constant temperature condition at the surface of the sea.

Table 2.1 The assumptions of the basic model and their ratings

Equatorial beta-plane	EXCELLENT
Neglect of vertical Coriolis	GOOD
No horizontal viscosity	SENSIBLE
Hydrostatic approximation	GOOD
Incompressibility (Nondivergence)	VERY GOOD
Drop all nonlinear terms in momentum equation	QUESTIONABLE
Drop all nonlinear terms except wdT_0/dz in heat equation	QUESTIONABLE
Assume special forms for mixing coefficients: $\nu = \urcorner = A/N^2$	NECESSARY
Replace of p by ϕ with Boussinesq approximation	VERY GOOD
Rigid lid upper boundary	GOOD (for baroclinic modes)
Density and temperature equal their mean values at surface	(UNFORTUNATE but NECESSARY)
Constant bottom depth D	NECESSARY but NOT BAD (for surface-trapped motion)
No stress at the sea bottom	NECESSARY and NOT BAD
No temperature changes at sea bottom	NECESSARY but NOT BAD

This is a real problem because a coupled ocean-atmosphere model of the Southern Oscillation/El Niño phenomenon, for example, must have sea surface temperature changes to drive the atmospheric changes of the cycle. The other surface conditions are either correct or very accurate approximations, and we need not worry about them. None of the bottom boundary conditions, in contrast, is realistic, but equatorial flows are not sensitive to the bottom boundary conditions.

2.3 Separation of Variables

In a nonlinear, numerical model, all five of (2.13) would be retained with the local time derivatives in the momentum and heat equations being replaced by total time derivatives, and perhaps the special, stability-dependent viscosity coefficients with more realistic forms. In a linearized treatment, however, we can reduce the number of equations in stages.

The first step is to rewrite the heat equation, using the hydrostatic equation to replace ρ by ϕ_z, to give an expression for the vertical velocity:

$$w = -(1/N^2)[\phi_t - A(\phi_z/N^2)_z]_z \tag{2.16a}$$

The density is given in terms of ϕ by the hydrostatic equation,

$$\rho = -(\rho_M/g)\phi_z \tag{2.16b}$$

We are then left with a system of 3 equations in the three other unknowns:

$$u_t - fv + \phi_x = A\Lambda u \tag{2.16c}$$
$$v_t + fu + \phi_y = A\Lambda v \tag{2.16d}$$
$$-\Lambda\phi_t + u_x + v_y = -A\Lambda^2\phi \tag{2.16e}$$

where Λ is the linear operator defined by

$$\Lambda = \partial_z \frac{1}{N^2}\partial_z \tag{2.17}$$

The key point is that as a result of the careful choice of the form of the viscosity, conductivity, and boundary conditions, the vertical derivatives in (2.16) appear solely in the form of the operator Λ, the resulting spectral equations will uncouple and we will have separated variables. The eigenfunction problem is

$$\frac{d}{dz}\left[\frac{1}{N^2}\frac{d}{dz}\mathfrak{z}_m\right] = -\lambda_m \mathfrak{z}_m \tag{2.18}$$

with boundary conditions

$$\frac{d}{dz}\mathfrak{z}_m = 0 \qquad \text{at } z = 0, -D \tag{2.19}$$

This is a classic Sturm–Liouville eigenproblem – no singularities, no complications of any kind since $N(z)$ is positive definite – but a little care isneeded because different unknowns have different dependence on $\mathfrak{z}_m$. The proper expansions are

$$u = \sum_{m=0}^{\infty} u_m(x, y, t)\mathfrak{z}_m(z) \tag{2.20a}$$
$$v = \sum_{m=0}^{\infty} v_m(x, y, t)\mathfrak{z}_m(z) \tag{2.20b}$$
$$\phi = \sum_{m=0}^{\infty} \phi_m(x, y, t)\mathfrak{z}_m(z) \tag{2.20c}$$
$$\rho = \rho_0(z) - \frac{\rho_M}{g}\sum_{m=0}^{\infty} \phi_m(x, y, t)\frac{d\mathfrak{z}_m}{dz} \tag{2.20d}$$

$$w = \sum_{m=0}^{\infty} w_m(x, y, t) \int_{-D}^{z} \mathfrak{z}_m(z')dz' \tag{2.20e}$$

The three independent variables (u, v, and ϕ) which solve the closed set of equations, (2.16a)–(2.16c), thus depend on the vertical eigenfunctions directly, but ρ and w are different. That ρ depends on the derivatives of the eigenfunctions rather than on the $\mathfrak{z}_m$ themselves is an obvious consequence of the hydrostatic equation – ρ is on one side while the *z-derivative* of ϕ is on the other – which also gives the simple relationship of the expansion coefficients of ρ to those of ϕ which is explicitly displayed in (2.20d).

The vertical dependence of w can be deduce from the equation of continuity in its original form

$$u_x + v_y + w_z = 0 \tag{2.21}$$

The vertical differentiation in (2.21) eliminates the integral in (2.20e) so that u, v, and the z-derivative of w are all proportional to $\mathfrak{z}_m$ for a given vertical mode.

The assumed expansions are all consistent with the boundary conditions, too. One can see why it is necessary to impose somewhat artificial conditions, such as no stress at the bottom, rather than the more natural no-slip: there are four boundary conditions at each surface, but only a single vertical structure function $\mathfrak{z}_m$. The sole boundary condition on $\mathfrak{z}_m$ at $z = 0$ is imposed, like it or not, on all four fields, and similarly at the bottom boundary.

The mechanics of this consistency are a little tricky. The integrals in (2.20e), for example, automatically impose the boundary condition $w = 0$ at the bottom because the lower limit of integration is the bottom, $z = -D$. The top boundary condition of $w = 0$ at $z = 0$ can only be satisfied if

$$\int_{-D}^{0} \mathfrak{z}_m(z')dz' = 0 \tag{2.22}$$

Extra boundary conditions like this cannot be satisfied in general, but (2.22) is fact a direct consequence of the eigenvalue problem which defines the $\mathfrak{z}_m$ as can be seen by rewriting (2.18) and (2.19) in integral equation form. By integrating both sides of (2.18) with respect to z, one obtains

$$\mathfrak{z}_{mz} = -N^2\lambda \int_{-D}^{z} \mathfrak{z}_m(z')dz' \tag{2.23}$$

This is valid for all z. Evaluating (2.23) at $z = 0$ and imposing the boundary condition (2.19), which makes the L.H.S. of (2.23) equal to 0, then gives (2.22).

The top boundary conditions on u and v are more complex: the vanishing of the derivatives of $\mathfrak{z}_m$ implies that each term in the z-derivative of the sums for u and v, (2.20a) and (2.20b), must individually vanish at $z = 0$ — and yet the boundary conditions there demand that u_x and v_x be non-zero to balance the wind stress. This issue is complex enough to deserve further discussion in Appendix B. Here, we will simply go ahead to derive the equations for each baroclinic mode and then invoke the integration-by-parts argument of McCreary [1].

Since (2.18) and (2.19) constitute a regular Sturm–Liouville problem, the eigenfunctions $\mathfrak{z}_m$ are a complete orthogonal set. This justifies the expansions (2.20). It also implies that we can derive spectral equations to compute the coefficients $u_m(x, y, t)$, $v_m(x, y, t)$, etc. through the usual procedure of multiplying the equations of motion by each basis function $\mathfrak{z}_m(z)$ to reduce each infinite sum to a single term.

The x-momentum equation is typical. Since none of the terms involve z-differentiation, the expansion of the L.H.S. of (2.16c) gives only terms proportional to $\mathfrak{z}_m(z)$ for various m. Multiplication by a particular basis function, $\mathfrak{z}_m(z)$, followed by integration then eliminates all terms in the infinite sum except $m = n$. The spectral projection of the R.H.S. of (2.16c) is

$$\int_{-D}^{0} \mathfrak{z}_n A[u_z/N^2]_z dz = A\mathfrak{z}_n(u_z/N^2)|_{-D}^{0} - \int_{-D}^{0} A(\mathfrak{z}_{nz}/N^2)u_z\, dz \tag{2.24}$$

$$= A\mathfrak{z}_n(u_z/N^2)|_{-D}^{0} - Au(\mathfrak{z}_{nz}/N^2)|_{-D}^{0} + \int_{-D}^{0} Au[\mathfrak{z}_{nz}/N^2]_z\, dz \tag{2.25}$$

by twice integrating-by-parts. To evaluate the integral, use the physical boundary conditions on $\partial u/\partial z$ to evaluate the first surface term, the boundary conditions on $\mathfrak{z}_m$ to evaluate the second, and the expansion of u into eigenfunctions and the orthogonality of the eigenfunctions. This gives

$$\int_{-D}^{0} \mathfrak{z}_n A[u_z/N^2]_z dz = \tau^x \mathfrak{z}_n(0) - 0 - A\lambda_n u_n(x, y, t) \int_{-D}^{0} \mathfrak{z}_n^2 dz \tag{2.26}$$

We see from (2.26) that the convenient normalizations for the vertical eigenfunctions are either

$$\mathfrak{z}_n(0) = 1 \qquad \text{[McCreary normalization]}, \tag{2.27}$$

which simplifies the first term in (2.26) and is the choice made by McCreary [1], or

$$\int_{-D}^{0} \mathfrak{z}_n^2 dz = 1, \qquad \text{[orthonormal]} \tag{2.28}$$

which simplifies the last term in (2.26). Either way, the spectral equivalent of (2.2) becomes

$$(\partial_t + A\lambda_n)u_n - fv_n + \phi_{nx} = F_n \tag{2.29a}$$
$$(\partial_t + A\lambda_n)v_n - fu_n + \phi_{ny} = G_n \tag{2.29b}$$
$$\lambda(\partial_t + A\lambda_n)\phi_n + u_{nx} + v_{ny} = 0 \tag{2.29c}$$

and the two dependent equations

$$\rho_n = \phi_n(\rho_M/g) \tag{2.30a}$$
$$w_n = \lambda_n(\partial_t + A\lambda_n)\phi_n \tag{2.30b}$$

where

$$F_n = \tau^x \mathfrak{z}_n(0)/\int_{-D}^{0} \mathfrak{z}_n^2 dz \tag{2.31a}$$
$$G_n = \tau^y \mathfrak{z}_n(0)/\int_{-D}^{0} \mathfrak{z}_n^2 da \tag{2.31b}$$

The closed set of three equations, (2.29), is identical in form with the linearized shallow water wave equations, and will therefore sometimes be referred to as the "shallow water wave equations". The two dependent equations, (2.30), will be needed only rarely.

2.4 Lamb's Parameter, Equivalent Depths, Kelvin Phase Speeds and All that

The reason for the rather lengthy section title is that there are several popular conventions for expressing the eigenvalues of the vertical structure equation. One of the older conventions is based on the analogue between (2.29) and the shallow water equations: the form is identical provided that we write

$$\lambda_n = \frac{1}{(gH_{eq})} \qquad \text{["equivalent depth"]} \tag{2.32}$$

where $H_{eq}(n)$ is the so-called "equivalent depth" for the n-th baroclinic mode, which replaces the actual depth H in the shallow water equations. Because ocean tides were first studied using the model of a single layer of homogeneous fluid, i.e., the shallow water equations, this "equivalent depth" convention is popular in both oceanic and atmospheric tidal theory, among others.

The second convention is based on the fact that $(gH)^{1/2}$ is the speed of a (non-rotating) gravity wave in the shallow water model. This would seem to have little relevance to the large-scale motions that are our primary interest here since these are strongly affected by rotation. However, the equatorial Kelvin wave (and its

mid-latitude counterpart, the coastal Kelvin wave) are in geostrophic balance in one direction while having the structure of a non-rotating gravity wave in the other horizontal coordinate and z; its phase speed is always equal to $(gH)^{1/2}$ or to $(gH_{eq})^{1/2}$ for the baroclinic Kelvin waves that will be discussed at great length in later chapters. Because the Kelvin wave is of such physical importance in its own right as well as providing a convenient scale for other waves–the lowest meridional mode Rossby wave always travels, for long wavelength, at exactly 1/3 the phase speed of the Kelvin wave of the same vertical mode, for example — some equatorial dynamicists use the "Kelvin wave convention":

$$\lambda_n = \frac{1}{c_n^2} \qquad \text{["Kelvin convention"]} \tag{2.33}$$

where c_n is the phase speed of the Kelvin wave for the n-th baroclinic mode. This is the way of writing the eigenvalue which is used in McCreary [1].

The third convention is to define a quantity, usually called "Lamb's parameter", to be the non-dimensional equivalent of the eigenvalue. Letting a = radius of the earth and Ω = angular frequency of the earth's rotation ($= 2\pi/86,400$s),

$$\varepsilon = (4\Omega^2 a^2)\lambda_n = \frac{(4\Omega^2 a^2)}{gH_{eq}} \qquad \text{["Lamb's parameter"]} \tag{2.34}$$

Alas, there is no standard symbol for Lamb's parameter; the famous treatise of Longuet–Higgins [17] used ε. The nonlinear papers of Boyd used E so as to reserve ε for the perturbation parameter.

Fortunately, the issue is not of paramount importance because it is conventional in equatorial oceanography (although there a few diehards) to nondimensionalize the shallow water equations (2.29) in an ε-dependent way so that (nondimensionally) $\lambda_m = 1$. Note that because we have separated variables, we can study each baroclinic mode in isolation and are free to adopt a different nondimensionalization, i.e., one that depends on λ_m, if this is convenient.

The actual eigenvalues, expressed as equivalent depths, for a set of observations and for an idealized numerical model are given in Table 2.2.

Table 2.2 Properties of baroclinic modes for a simple analytical $T(z)$ [top] and observations [bottom]

Equivalent			Length	Time
Vertical mode n	Depth H_0 cm	$(gH_0)^{1/2}$ (cm s^{-1})	Scale L (km)	Scale T (days)
1	100	312	395	1.46
2	22	146	271	2.15
3	9.4	96	219	2.64
4	5.1	71	189	3.09
5	3.3	57	169	3.44
6	2.3	47	154	3.80
7	1.7	41	143	4.04
8	1.3	36	134	4.31
9	1.1	32	126	4.55
10	0.9	29	120	4.80
	H' (cm)	$\sqrt{gH'}$ (cm/s)	L (km)	T (days)
1	60	240	325	1.5
2	20	140	247	2.0
3	8	88	197	2.6
4	4	63	165	3.1
5	2	44	139	3.6

(a) (Model) $T(z) = 4. + 20.e^{z/500\text{m}}$ (in degrees $C°$). Data from Semtner and Holland [16]
(b) (Observed) [Atlantic] E.J. Katz, private communication, [18]

2.5 Vertical Modes and Layer Models

The weak flow below the thermocline in either the layered or continuously stratified case has led to a further idealization popular in equatorial oceanography: the "one-and-a-half layer" model. The vertically integrated horizontal velocities are always zero in the two-layer model. When the lower layer is much deeper than the top, as is always true in the real ocean — the best two-layer fit to the observed density is a bottom layer roughly *40* times thicker (4 km vs 100 m) than the upper — the motion in the lower layer is therefore extremely weak. It is then useful to simplify the model still further by pretending that the depth of the lower layer is infinite, instead of merely very large, so that the *lower layer* is at *rest*. Symbolically,

$$1\frac{1}{2}\text{layer model} \equiv 2\,\text{layer model with lower layer of infinite depth and no motion}$$

This "one-and-a-half layer" model distorts the linear dynamics of the baroclinic mode only a little, but it profoundly alters the nonlinear dynamics by eliminating the coupling between the baroclinic and barotropic modes. The reason is that when

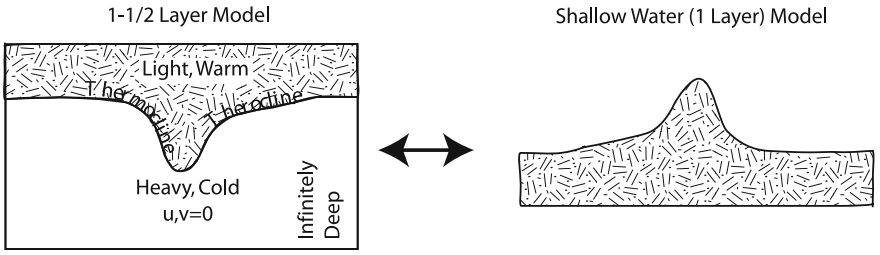

Fig. 2.1 The 1-1/2 layer model (*left*) is mathematically equivalent to the shallow water equations (*right*). Note that a *trough* in the thermocline, which is the boundary between the warm, dynamically-active *upper* layer and the cold, dynamically-passive layer translates into a *crest* in the shallow water model. Note also that the mean depth of the shallow water is the *equivalent* depth (typically half a meter) rather than the actual depth of the layer above the thermocline, which is O(100) m

a horizontal velocity like u is zero in the lower layer, a nonlinear product like uu_x is also zero in the lower layer. Therefore, the nonlinear terms involving the self-interaction of a motion in the baroclinic mode project back only onto the depth-varying baroclinic mode rather than onto the barotropic mode, which is independent of depth. The special cases of (i) equal depth or (ii) an infinite lower layer depth as here are the only circumstances in which the nonlinear barotropic mode/baroclinic mode coupling in the two-layer model is 0.

In reality, however, the coupling to the second baroclinic mode – which does not even exist in the two-layer model – can be as strong as the barotropic mode in the continuously stratified model. Furthermore, the second baroclinic mode motion is more interesting than the barotropic mode because only the baroclinic modes are equatorially trapped. Consequently, it is not possible to do a satisfactory treatment of nonlinear mode-coupling in the two-layer model even when the depth of the lower layer is kept finite, so most equatorial oceanographers have been quite shameless in retreating to the one-and-a-half layer models for nonlinear calculations. (Note that for purely *linear* calculations, the baroclinic and barotropic modes are completely uncoupled, and the difference between the two-layer and one-and-a-half layer models is irrelevant.)

The virtue of the one-and-a-half layer model is that it is identical in form with the *nonlinear* shallow water wave equations except that (i) the actual depth of the upper layer is a factor of a hundred or more larger than the *equivalent* depth, which is (O[0.6 m]) and (ii) the model must be turned upside down (Fig. 2.1). What is meant by this is that the height ϕ is proportional to the total *thickness* of the layer that is in motion (Fig. 2.1). Since the density difference across the thermocline is two orders of magnitude smaller than the difference across the air-water interface, it is the thermocline that moves, and not the sea surface. When the thermocline moves *down*, the total thickness of the upper layer increases. Thus, when the shallow water wave equations are being used to represent the baroclinic mode of the one-and-a-half layer model, a positive value of ϕ corresponds to downward movement of the thermocline, an increased thickness of the upper layer, and downwelling at the thermocline, i.e., negative vertical velocity. The numerical values in the shallow

water model must be multiplied by the ratio of thermocline depth/equivalent depth to obtain the actual thermocline changes: ϕ equivalent to a 30 cm displacement in a shallow water model with an equivalent depth of only 60 cm corresponds to a 50 m change in the thermocline depth.

The one-and-a-half layer model has been described here in some detail because it has been surprisingly effective in explaining many features of the large-scale ocean circulation when supplied with realistic boundaries and wind stresses.

Nonetheless, the formalism of continuously stratified models is well understood, and they are clearly the wave of the future. Figure 2.2 compares the first four baroclinic modes as calculated by Semtner and Holland [16] for the idealized stratification

$$T(z) = 4 + 20e^{z/500m} \quad ^\circ C \tag{2.35}$$

(the eigenvalues are given in Table 2.2) with the lone baroclinic mode of the two-layer model for a case in which the bottom layer is much deeper than the top layer. The structure of the first baroclinic mode for the exponential stratification is similar to that of the baroclinic mode of the two-layer model except for the smearing out of the step-function jump in the corresponding eigenfunction. However, both first baroclinic modes have strong flow in the upper ocean and much weaker flow – in the opposite direction – in the deep ocean. Thus, the layer model gives a reasonable, albeit crude, approximation to the lowest baroclinic mode of a more realistic, continuously stratified model. While the baroclinic modes shown in Fig. 2.2 are intended to be representative only—a more realistic stratification would give less amplitude at the bottom—it is easy to calculate the vertical modes for any specified stratification.

2.6 Nondimensionalization

Letting a = radius of the earth, Ω = angular frequency of the earth's rotation, and

$$\varepsilon = 4\Omega^2 a^2 / g H_{equivalent} \qquad \text{["Lamb's parameter"]} \tag{2.34bis}$$

as before, one can *eliminate* a, Ω, and ε as explicit parameters by choosing vertical mode-dependent scalings. Letting asterisks denote the non-dimensional quantities,

$$x^* = \varepsilon^{1/4} x/a \tag{2.36a}$$
$$y^* = \varepsilon^{1/4} y/a \tag{2.36b}$$
$$t^* = \varepsilon^{-1/4}(2\Omega)t \tag{2.36c}$$
$$u^* = \varepsilon^{1/2} u/(2\Omega a) \tag{2.36d}$$
$$v^* = \varepsilon^{1/2} v/(2\Omega a) \tag{2.36e}$$
$$\phi^* = \varepsilon\phi/(2\Omega a)^2 \tag{2.36f}$$
$$A^* = \varepsilon^{5/4} A/[(2\Omega a)^2 2\Omega] \tag{2.36g}$$

will reduce the shallow water equations to a form with *no explicit parameters* except the viscosity coefficient, which we shall usually set equal to 0 (Fig. 2.2).

Planetary wave theory and middle atmosphere dynamics often uses a nondimensional frequency σ = dimensional frequency$/(2\Omega)$ and the zonal wavenumber is an integer s. The corresponding nondimensional phase speed is $c_{planetary}$ = phase speed$/(2\Omega a)$. These are connected to the nondimensional quantities used in equatorial oceanography by

$$k^* = s/\varepsilon^{1/4} \tag{2.37}$$

$$\omega^* = \varepsilon^{1/4}\sigma \tag{2.38}$$

$$c^* = \varepsilon^{1/2} c_{planetary} \tag{2.39}$$

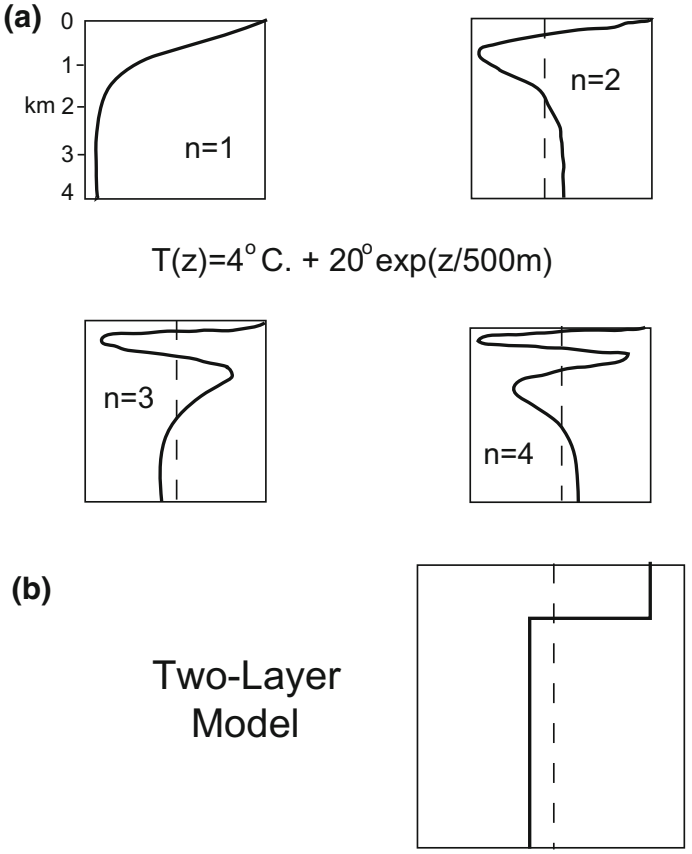

Fig. 2.2 Baroclinic modes versus depth where $\mathfrak{z}$ is the horizontal axis and depth z is the vertical axis. The *dashed vertical lines* are the axis $\mathfrak{z} = 0$. **a** First four modes for an exponential variation of mean temperature with depth. **b** First (and only) baroclinic mode of a two-layer model

$$\varepsilon\sigma^3 - [\varepsilon^{1/2}(2n+1) + s^2]\sigma - s = 0 \qquad \leftrightarrow \qquad \omega^3 - [(2n+1) + k^2]\omega - k = 0$$

The scalings in (2.36) can be summarized by stating that the nondimensional length scale L, time scale T and velocity scale U are given by

$$L = \varepsilon^{-1/4} a \tag{2.40a}$$

$$T = \varepsilon^{1/4}/(2\Omega) \tag{2.40b}$$

$$U = \varepsilon^{1/2}\, 2\Omega\, a \tag{2.40c}$$

where typically $L \sim O(300\text{km})$, $T \sim O(2\text{days})$ and U is $O(2\text{m/s})$.

References

1. McCreary JP Jr (1980) Modelling wind-driven ocean circulation. Technical report, Hawaii Institute of Geophysics/JIMAR, Honolulu, School of Ocean and Earth Science and Technology, p 64
2. Phillips NA (1966) The equations of motion for a shallow rotating atmosphere and the 'traditional approximation'. J Atmos Sci 23:626–628
3. Phillips NA (1968) Reply to comments on "the equations of motion for a shallow rotating atmosphere and the 'traditional approximation"'. J Atmos Sci 25:1155–1157
4. Fruman MD (2009) Equatorially bounded zonally propagating linear waves on a generalized β plane. J Atmos Sci 66:29372945
5. Gerkema T, Zimmerman J, Maas L, van Haren H (2008) Geophysical and astrophysical fluid dynamics beyond the traditional approximation. Rev Geophys 46:RG2004
6. Fruman MD, Shepherd TG (2008) Symmetric stability of compressible zonal flows on a generalized equatorial β plane. J Atmos Sci 65(6):1927–1940
7. Ripa P (1997) "Inertial" oscillations and the beta-plane approximation(s). J Phys Oceangr 27(3):633–647
8. Harlander U, Maas LRM (2007) Internal boundary layers in a well-mixed equatorial atmosphere/ocean. Dyn Atmos Oceans 44:1–28
9. Müller P (2006) The equations of oceanic motions. Cambridge University Press, Cambridge, p 302
10. Mueller R (1989) A note on the relation between the traditional approximation and the metric of the primitive equations. Tellus A 41A:175–178
11. Thuburn JN, Wood N, Staniforth A (2002) Normal modes of deep atmospheres. II: f-plane geometry. Q J Z Meteorol Soc 128:1793–1806
12. Grimshaw RHJ (1975) A note on the β-plane approximation. Tellus 27:351–357
13. Kasahara A (2003) The roles of the horizontal component of the earths angular velocity in nonhydrostatic linear models. J Atmos Sci 60:1085–1095
14. Gerkema T, Shrira VI (2005) Near-inertial waves in the ocean: Beyond the "traditional approximation". J Fluid Mech 529:195–219
15. Veronis G (1973) Model of world ocean circulation. I. Wind-driven, two-layer. J Mar Res 31:228–287
16. Semtner AJ, Holland WR (1980) Numerical simulation of equatorial ocean circulation, Part 1, a basic case in turbulent equilibrium. J Phys Oceangr 10:667–693

17. Longuet-Higgins MS (1968) The eigenfunctions of Laplace's tidal equations over a sphere. Philos Trans R Soc London A 262:511–607
18. Moore DW, Philander SGH (1977) Modelling of the tropical oceanic circulation. In: Goldberg ED (ed) The sea. Wiley, New York, pp 319–361

Chapter 3
Kelvin, Yanai, Rossby and Gravity Waves

Abstract The Hough-Hermite functions are the normal modes of the shallow water equations. Each eigenfunction is a vector with different latitudinal and vertical structure functions for the three fields (u, v, ϕ). Geophysicists commonly classify the normal modes into different species: Kelvin wave, Rossby waves, gravity waves and the Yanai wave, also known as the mixed Rossby-gravity mode. In this chapter, we catalogue the properties of these normal modes. The latitudinal structure functions are given explicitly as Hermite functions. Derived fields such as the vorticity are also cataloged. The eigenvalues are the frequencies, which are analyzed both through explicit approximations and a plethora of graphs. Hough functions also appear when Lamb's parameter ε is the eigenvalue and the frequency is a parameter as in tidal theory, and also when the zonal wavenumber k is the eigenvalue as needed for tides in a land-bounded basin and certain applications in meteorology.

Eyesight should learn from reason.

Johannes Kepler

3.1 Latitudinal Wave Modes: An Overview

The normal modes of the linearized shallow water wave equations, henceforth the "Hough-Hermite" functions, play a dual role in equatorial dynamics. First, they are the idealized solutions of the simplest rational model. Mean flows, transient forcing, irregular coastlines, and dissipation stretch and distort both the spatial and temporal structure of the normal modes, but Rossby waves retain many characteristics of the Rossby Hough-Hermite modes and distorted Kelvin waves are still Kelvin waves.

Second, the Hough-Hermite functions are a complete orthogonal set which can be used as the basis set in applying the spectral or pseudospectral method, even in strongly nonlinear or damped problems. It is possible to eliminate some or all of the gravity wave basis functions to create a restricted basis set that captures the physics of low frequency motions and allows a long time step.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_3

In the previous chapter, we derived the three shallow water wave equations for each vertical mode in the most general possible form for a resting mean state (Eq. 2.29a,b and c). For free oscillations these are simplified by the assumptions that

(i) The wind stresses $\tau^x,\ \tau^y = 0$.

(ii) The eddy viscosity coefficient $A = 0$.

With these two assumptions, the shallow water wave equations will be equally appropriate for either the ocean or the atmosphere. The Hough-Hermite functions are the vector solutions for the *undamped*, *free* oscillations of the equatorial ocean;

In addition, we shall use the vertical mode-dependent nondimensionalization of the previous chapter, which effectively replaces the vertical eigenvalue λ by 1. We shall also drop the subscripts denoting the vertical mode number for notational simplicity. The equations we must solve are then

$$u_t - fv + \phi_x = 0 \tag{3.1a}$$

$$v_t + fu + \phi_y = 0 \tag{3.1b}$$

$$\phi_t + u_x + v_y = 0 \tag{3.1c}$$

where

$$f = \begin{cases} f_0 & \text{[midlatitude f-plane]}, \\ f_0 + \beta y & \text{[midlatitude beta-plane]}, \\ y & \text{[equatorial beta-plane]} \end{cases} . \tag{3.2}$$

where f_0 and β are constants.

We are of course primarily interested in the last case $f = y$, but it is worth nothing that the equations are equally valid for the midlatitude f-plane (or, forming a vorticity equation from the momentum equations in the shallow water trio, the midlatitude beta-plane) because these simpler equations — whose coefficients are constants — will often be useful in later sections for understanding (a) the analogy between coastal and equatorial motions and (b) wind-driven motions in general.

There are two distinct methods for preceding farther. One is to make the usual assumption that all the waves depend on x and t in the form

$$\exp[ik(x - ct)] \tag{3.3}$$

and then reduce the three shallow water equations down to a single equation for one of the unknowns to determine the latitudinal structure. It is best to make v the unknown since this gives the simplest differential equation, the so-called "parabolic cylinder" equation. This has the great virtue that its eigenfunctions can be written down in simple closed form as the Hermite functions, which are described fully in Appendix A — but only if we impose the boundary conditions

$$u(y = \pm\infty) = v(y = \pm\infty) = \phi(y = \pm\infty) = 0 \tag{3.4}$$

Since the equatorial beta-plane cannot be qualitatively correct at high latitudes, it is foolish to impose boundary conditions at some finite value of y merely because that y corresponds to $\pm\pi/2$ radians. The whole justification for the equatorial beta-plane is that its dynamics is dominated by motions which are equatorially trapped and decay *exponentially* fast beyond some finite turning latitude. So long as we choose the boundaries to be beyond these turning points, it matters little where we place them. The boundary conditions at $\pm\infty$ have therefore become a canonical part of the equatorial beta-plane approximation.

One exception is the Gulf of Guinea where the coast of West Africa runs roughly parallel to the equator for a distance of several hundred kilometers at only 4 °N. Here the correct eigenfunctions are general parabolic cylinder functions—which cannot be written down in analytical form except as infinite power series—chosen to vanish at a finite northern boundary (and an infinite southern boundary). For the low order modes, i.e., Kelvin and mixed-Rossby gravity waves, which are the ones of greatest physical importance, the effects of the finite boundary are still negligible. Consequently, we shall use only infinite boundary conditions and Hermite functions.

The disadvantages of the obvious procedure of obtaining a single equation for v are that (i) it omits the Kelvin wave, which has $v \equiv 0$ and (ii) it spuriously gives three solutions for the zeroth-order Hermite mode when there are in fact only two such wave modes for the original set of three equations. Obtaining a single equation in u or ϕ alone is not a useful alternative because these differential equations have "apparent" singularities and are therefore not standard Sturm–Liouville problems; further, u and ϕ are sums of *two* Hermite functions, unlike the eigenmodes for v, whose solutions are proportional to a single Hermite function. The standard textbook approach in [1], for example, is therefore to reduce the three shallow water equations down to a single equation for v alone and then treat the two exceptions separately.

The other approach is less obvious — indeed, it would not make sense unless one already knew the answer — but it has the virtue of neither omitting nor adding waves. Furthermore, this line of attack has proven very useful in solving the *inhomogeneous* shallow water equations when either wind stresses or lower order terms in perturbation theory replace the zeros on the right-hand sides of the shallow water system. This "sum-difference" method is therefore the derivation given here.

The "sum-difference" method uses two ancient tricks. Observe that while u and ϕ are proportional to sums of two Hermite functions, it is possible, since they both involve the *same* two Hermite functions, to define new variables which are linear combinations of u and ϕ that are proportional to a single Hermite function, just as v is [2]:

$$S \equiv \phi + u \qquad \text{["sum variable"]} \tag{3.5a}$$

$$D \equiv \phi - u \qquad \text{["difference variable"]} \tag{3.5b}$$

The second trick, first applied in quantum mechanics, is that the Hermite functions have simple "raising" and "lowering" operators that transmute one Hermite function into another. The beauty of these two tricks is that all the explicit factors of y and

all the y derivatives can be written in terms of these operators $\mathfrak{R}$ and $\mathfrak{L}$. For the *normalized* Hermite functions

$$\mathfrak{R} \equiv (d/dy - y) \qquad \text{["raising operator"]} \tag{3.6a}$$

$$\mathfrak{R}\psi_n = -[2(n+1)]^{1/2}\psi_{n+1} \tag{3.6b}$$

$$\mathfrak{L} \equiv (d/dy + y) \qquad \text{["lowering operator"]} \tag{3.7a}$$

$$\mathfrak{L}\psi_n = [2n]^{1/2}\psi_{n-1} \tag{3.7b}$$

The first step is to take the sum and difference of the x-momentum equation and the height equation so as to obtain replacements that involve only S alone or D alone:

$$S_t + S_x + (v_y - yv) = 0 \tag{3.8a}$$

$$D_t - D_x + (v_y + yv) = 0 \tag{3.8b}$$

Note that the terms involving v in (3.8) can both be written in the simple form $\mathfrak{R}v$ or $\mathfrak{L}v$ by using the "raising" and "lowering" operators. Similarly substituting S, D, $\mathfrak{L}$, and $\mathfrak{R}$ into the y-momentum equation gives the set

$$S_t + S_x + \mathfrak{R}v = 0 \tag{3.9a}$$

$$v_t + 0.5(\mathfrak{L}S + \mathfrak{R}D) = 0 \tag{3.9b}$$

$$D_t - D_x + \mathfrak{L}v = 0 \tag{3.9c}$$

The next step is to substitute Hermite expansions into (3.9). (Note that the sums are now over *latitudinal* eigenfunctions, unlike the sums over vertical modes of the previous chapter.) The general solution, of course, is a double summation over both the vertical and latitudinal modes. The sums are

$$S(x, y, t) = \sum_{n=0}^{\infty} S_n(x, t)e^{-(1/2)y^2} H_n(y) \tag{3.10a}$$

$$v(x, y, t) = \sum_{n=0}^{\infty} v_n(x, t)e^{-(1/2)y^2} H_n(y) \tag{3.10b}$$

$$D(x, y, t) = \sum_{n=0}^{\infty} D_n(x, t)e^{-(1/2)y^2} H_n(y) \tag{3.10c}$$

Substituting these series into (3.9) and exploiting the properties of the raising and lowering operators, one finds that the Hermite coefficients are determined by

$$S_{0t} + S_{0x} = 0 \qquad \text{["Kelvin singlet"]} \tag{3.11}$$

which completely determines the Kelvin wave, for which $v = D = 0$, plus the pair of equations

$$S_{1t} + S_{1x} - 2^{1/2} v_0 = 0 \qquad \text{["Yanai doublet"]} \tag{3.12a}$$

$$v_{0t} + 2^{-1/2} S_1 = 0 \tag{3.12b}$$

plus the infinite set of triples of equation, one for each n with $n \geq 1$,

$$S_{n+1,t} + S_{n+1,x} - [2(n+1)]^{1/2} v_n = 0 \tag{3.13a}$$

$$v_{n,t} + [(n+1)/2]^{1/2} S_{n+1} - [n/2]^{1/2} D_{n-1} = 0 \tag{3.13b}$$

$$D_{n-1,t} - D_{n-1,x} + [2n]^{1/2} v_n = 0 \tag{3.13c}$$

["General Triplet"]

Thus, S_{n+1} is coupled only to v_n and D_{n-1} so that the triplet set of Eq. (3.13) may be solved for a given n independently of all the other triplets.

Because the route of reducing the shallow water wave equations down to a single equation for v was the historical route, it is now a firm convention to define the latitudinal mode number for equatorial waves to be the degree of the Hermite function which describes v. It would be natural to use the Hermite degree of the sum variable S instead; for the Kelvin wave, which has $v \equiv 0$, it is necessary to define $n = -1$ in the usual convention. However, the convention is now too firmly entrenched to be changed.

Since the Eqs. (3.11)–(3.13) are constant coefficient, it is trivial to assume the usual $\exp[ik(x-ct)]$ dependence for all the variables and obtain a dispersion relation. To avoid complex arithmetic, define

$$\begin{vmatrix} S_{n+1}(x,t) \\ v_n(x,t) \\ D_{n-1}(x,t) \end{vmatrix} = \begin{vmatrix} -\tilde{S}\cos(kx-\omega t) \\ \tilde{v}\sin(kx-\omega t) \\ -\tilde{D}\cos(kx-\omega t) \end{vmatrix} \tag{3.14}$$

Substituting into the general triplet gives the 3×3 *algebraic* eigenvalue problem

$$\begin{vmatrix} (k-\omega) & -[2(n+1)]^{1/2} & 0 \\ -[(n+1)/2]^{1/2} & -\omega & [n/2]^{1/2} \\ 0 & [2n]^{1/2} & (-k-\omega) \end{vmatrix} \begin{vmatrix} \tilde{S} \\ \tilde{v} \\ \tilde{D} \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \\ 0 \end{vmatrix} \tag{3.15}$$

Because this matrix is homogeneous, this system has nontrivial solutions only if its determinant is zero. This yields the dispersion relation

$$\omega^3 - (2n + 1 + k^2)\,\omega - k = 0 \tag{3.16}$$

which is solved in Sect. 3.3.

Once ω is known, the first and third equations show that

$$\tilde{S} = \left\{[2(n+1)]^{1/2}/(k-\omega)\right\}\tilde{v} \tag{3.17}$$

$$\tilde{D} = \left\{(2n)^{1/2}/(k+\omega)\right\}\tilde{v} \tag{3.18}$$

What follows in the next section is simply an extensive collection of results.

3.2 Latitudinal Wave Modes: Structure and Spatial Symmetries

The explicit linearized plane wave solutions of (3.13), i.e., the results obtained by assuming that the wave depends upon x and t only through a factor of $\exp[ik(x-ct)]$, are collected as Tables 3.1, 3.2 and 3.3. Somewhat inconsistently, the formulas for $n = 1$ and higher are expressed in terms of normalized Hermite functions while the Kelvin wave and mixed-Rossby gravity wave results are displayed explicitly as ordinary polynomials (because of their simplicity). Only u, v, ϕ, the vorticity ζ, and the linearized potential vorticity $q \equiv \zeta - y\phi$ are given because the other quantities of interest are proportional to those listed. For example, the divergence Δ is given by

$$\Delta \equiv u_x + v_y = ikc\phi \tag{3.19}$$

for all wave modes; this is an obvious consequence of the continuity equation [ϕ_t equation] of the linearized shallow water wave set. Similarly,

$$w = -ikc\int_{-D}^{z} \phi(z')dz' \tag{3.20}$$

$$T = -\alpha\rho = [\alpha\rho_M/g]\phi_z \tag{3.21}$$

from Eq. 2.20.

Note that w, T, and ρ are the only quantities which depend upon either derivatives or integrals of the vertical structure functions; the other variables are directly proportional to $\mathfrak{z}_m(z)$ itself.

Figures 3.1, 3.2 and 3.3 illustrate the horizontal winds and pressure for the Kelvin wave ($n = -1$), the mixed Rossby-gravity or Yanai wave ($n = 0$, westward-traveling), and the $n = 1$ Rossby wave. Several general comments can be made.

First, the height ϕ and zonal flow u are always zonally in phase and have the same latitudinal symmetry with respect to the equator. The latitudinal current v, in contrast, is always 90 degree out of east-west phase with u and ϕ and always has the opposite symmetry with respect to the equator.

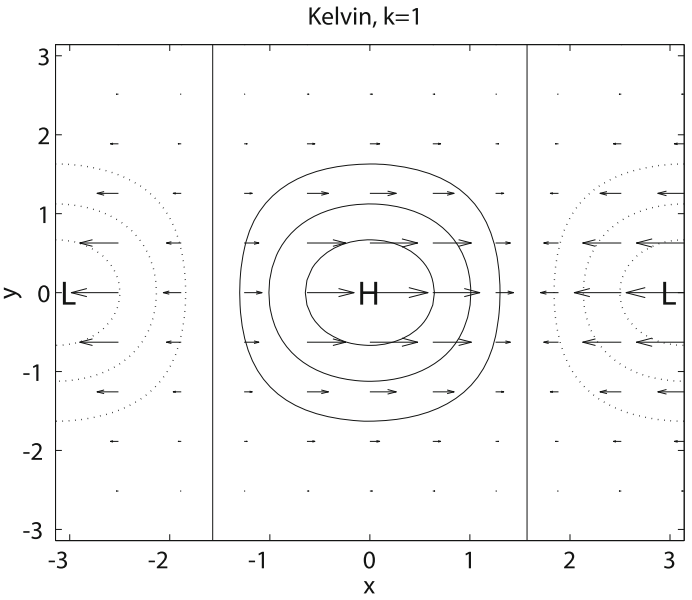

Fig. 3.1 Kelvin wave for east-west wavenumber $k = 1$. Positive isolines of height are *solid* while negative-valued contours are *dotted*. Local maxima and minima of height ϕ are marked with the letters "H" and "L", a common meteorological convention. The *arrows* show the magnitude (vector length) and direction (*arrow* orientation) of the horizontal wave velocity vector whose components are (u, v). No numerical values are given because the amplitude of a linear normal mode is arbitrary

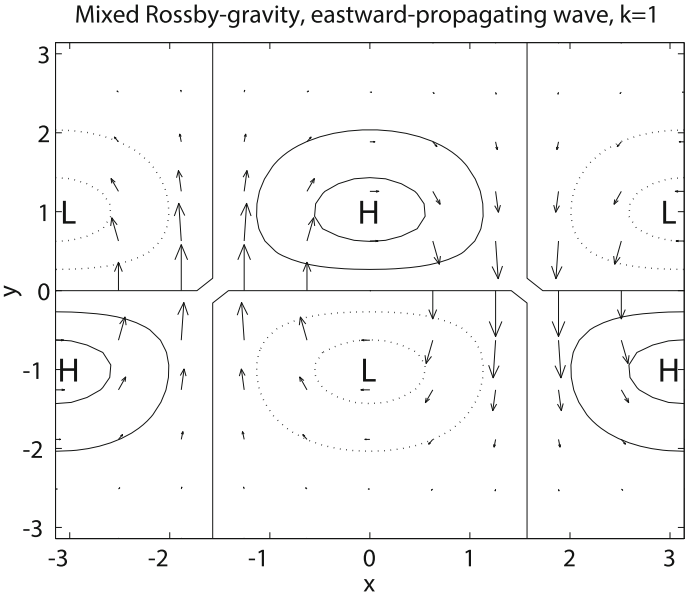

Fig. 3.2 Mixed Rossby-gravity (Yanai) wave for $k = 1$

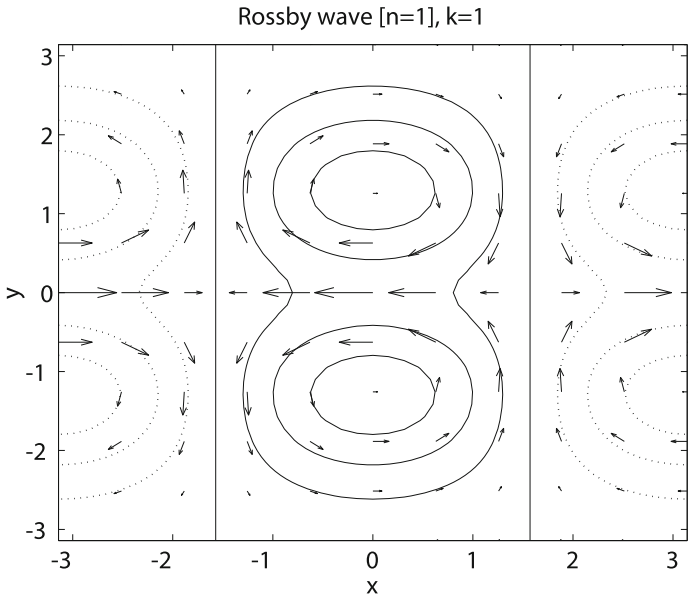

Fig. 3.3 $n = 1$ Rossby mode for $k = 1$

Equations (3.19)–(3.21) show that w, T, and the divergence Δ are all proportional to ϕ (or its *vertical* integral or derivative). Thus, all three of these quantities invariably have the same zonal-latitudinal structure as ϕ.

This explains a rather amusing convention; the equatorial symmetry of a given mode is defined to be the symmetry of u, ϕ, w and T. Thus, the $n = 1$ Rossby mode, for which all these quantities are symmetric with respect to the equator, is called a "symmetric" mode. The irony is that the mode is numbered according to the Hermite degree of v, and this is always opposite in symmetry to the other four fields. One therefore has to remember that while the Kelvin wave and all $n = 1, n = 3, n = 5, \ldots$, *modes* are *symmetric* with the equator, the *Hermite functions* of odd degree n are *antisymmetric* with respect to the equator. Table 3.4 summarizes the latitudinal symmetry of each individual field.

The vorticity ζ and potential vorticity q share the same latitudinal symmetry as v; the potential vorticity is in fact proportional to the same Hermite function as v. However, because of the factor of i in front of v, which is missing from q and ζ, the two vorticities are always 90 degrees out of phase with v and therefore have maxima and minima along the same meridians as u and ϕ.

Zonally speaking, we have that u, ϕ, ζ, and q are in phase while the ridges and troughs of v and the divergence Δ are always a quarter wavelength out of phase with the first four. The zonal phases for w and T are a little more complicated because these depend upon integrals or derivatives of ϕ. When the vertical structure functions represent *standing* waves, unlike those shown in Fig. 3.4, w is zonally out of phase with ϕ while the temperature is in phase with ϕ.

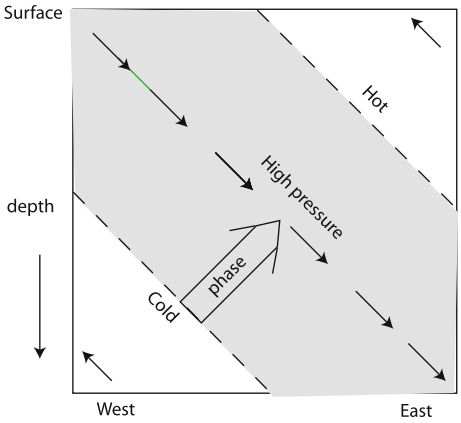

Fig. 3.4 Longitude-depth cross-section for a Kelvin wave with downward energy propagation. Note that the direction of phase propagation is upward, opposite to that of the energy

In the atmospheric case

$$\mathfrak{z}(z) \approx e^{irz} \tag{3.22}$$

which r is the vertical wavenumber. (We have ignored the usual density growth factor of $\exp(z/(2H)$ where H is the scale height of the atmosphere because this factor has no oceanic counterpart, and is omitted anyway in the Boussinesq approximation.) Because the atmosphere is semi-infinite rather than bounded, equatorial waves freely radiate away to space. This implies that free modes are impossible. Instead, one has only forced modes which propagate upward and do not return. We have therefore appropriately omitted a subscript mode number from $\mathfrak{z}(z)$. When one has to vertically integrate the complex exponential in (3.22) one obtains a factor of i, and likewise in differentiating. Therefore, in the atmosphere, the zonal phase relationships of w and T are changed: w is always in phase with ϕ while T is always zonally a quarter wavelength behind.

Because w and T always have the same latitudinal structure as ϕ and because the zonal average of the product of ϕ and w — which is the vertical energy flux — must be positive for an upward propagating wave, we can make the unambiguous statement that the vertical velocity is always maximum upwards where ϕ is a maximum while the strongest downdrafts occur at the troughs of ϕ. Similarly, the temperature maxima always lead the ridges of ϕ by a quarter of a wavelength, regardless of whether the wave is traveling to the east or the west.

Ocean waves that are excited at the surface and transmit energy downward into the abyss may exhibit a similar one-way propagation, only turned upside-down, if there is sufficient damping so that the energy that reflects off the bottom is strongly attenuated before it reaches the surface again. The different vertical propagation direction implies that the oceanic structure is

$$\mathfrak{z}(z) \approx \exp(-irz) \tag{3.23}$$

These zonal phase relationships for a surface-forced, downward-energy-propagating Kelvin wave are shown in Fig. 3.4.

When the vertical wavelength is short in comparison to the mean ocean depth of 4 km, it is therefore reasonable to apply a radiation condition on the eigenfunctions of the vertical structure equation. Then the atmospheric phase relationships apply to the ocean, too, except that for *downward*-propagating waves, maximum *downward* velocities coincide with the ridges of ϕ, and T now lags the ridges of ϕ instead of leading them.

For ocean normal modes in which the $\mathfrak{z}_n(z)$ are real (rather than complex), the vertical integration/differentiation needed to obtain w and T from ϕ does not introduce a factor of i. In this case, w is out of phase with ϕ while the temperature T now is in phase with ϕ. Thus, the distinction between vertical normal modes and vertical traveling waves is very important to the zonal phase relations — which in turn are important in understanding whether observed waves are normal modes, or surface-generated waves on a one-way trip to the bottom.

In the vertical normal modes case, there is no simple rule as to whether ϕ and T are of the same or opposite sign at a given point because both are oscillating with z, but one is proportional to $\mathfrak{z}_n(z)$ and the other is proportional to its *derivative*. Similarly, u and ϕ are always proportional to the same function of depth, but since their latitudinal structure is different, they may be of the same or opposite sign, depending on the latitude.

3.3 Dispersion Relations: Exact and Approximate Frequencies

The dispersion relation for the Kelvin mode is trivial: without approximation

$$c = 1 \qquad \forall k \qquad \leftrightarrow \qquad \omega = k \tag{3.24}$$

The Yanai dispersion relation is the quadratic

$$\omega^2 - k\omega - 1 = 0 \tag{3.25}$$

which has also an algebraic exact solution:

$$\omega = \frac{k}{2} \pm \sqrt{\frac{k^2}{4} + 1} \qquad \text{[Yanai Frequency]} \tag{3.26}$$

For the special case $k = 0$, the frequency simplifies to

$$\omega = \pm 1 \qquad \text{[Yanai Zero-Wavenumber Frequency]} \tag{3.27}$$

The frequencies for the three modes whose north-south current is proportional to $\psi_n(y)$ are the three roots of the cubic equation

$$\omega^3 - (2n + 1 + k^2)\omega - k = 0 \tag{3.28}$$

Note that because the coefficient of ω^2 is zero, the so-called "Vieta Root Identities" (Theorem 11.9 on p. 224 of [3]) show that the sum of the roots is exactly zero:

$$\omega_{Rossby}(n, k) + \omega_{EG}(n, k) + \omega_{WG}(n, k) = 0 \qquad \text{[Vieta Identity]} \tag{3.29}$$

where the three roots are a Rossby wave, an eastward-traveling gravity wave and a westward-traveling gravity wave as discussed further below.

For the special case $k = 0$, the roots are

$$\omega = 0, \pm\sqrt{2n + 1} \qquad \text{[Zero Wavenumber Frequencies]} \tag{3.30}$$

The zero frequency modes are geostrophically-balanced with zero north-south velocity and are often called the "geostrophic modes", although they are also the smooth limits of Rossby waves in the limit $k \to 0$. The frequencies of the other two modes are similarly the well-behaved limits of the frequencies for eastward-traveling and westward-traveling gravity waves

The *exact* roots for general k are given by (Theorem 1.1, p. 199 of [3])

$$\omega_{\text{EG}} = \sqrt{4/3}\,(2n + 1 + k^2)^{1/2} \cos(\Xi/3) \tag{3.31}$$

$$\omega_{\text{WG}} = \sqrt{4/3}\,(2n + 1 + k^2)^{1/2} \cos(\Xi/3 + 2\pi/3) \tag{3.32}$$

$$\omega_{\text{Rossby}} = \sqrt{4/3}\,(2n + 1 + k^2)^{1/2} \cos(\Xi/3 + 4\pi/3) \tag{3.33}$$

where the subscripts EG and WG denote the gravity waves propagating towards the east and towards the west, respectively, and where

$$\Xi \equiv \arccos\left(\frac{3^{3/2}}{2(2n + 1 + k^2)^{3/2}} k\right) \tag{3.34}$$

One can show that Ξ differs form $\pi/2$ by no more than $1/(2n + 1)$ for all n and k; the power series approximations to the arccosine and cosine functions are rapidly convergent and give the *approximate* solutions

$$\omega_{\text{G}} \approx \pm(2n + 1 + k^2)^{1/2} \tag{3.35}$$

where the plus sign gives the eastward-traveling gravity wave and

$$\omega_{\text{Rossby}} \approx -k/(2n + 1 + k^2) \tag{3.36}$$

Boyd [4] proves rigorous error bounds for these approximations as discussed further in the next section. The relative error in the Rossby wave approximation is never worse than 1 part in 60 for all n and k — a far smaller error than that made by ignoring mean currents and other assumptions implicit in (3.28) —so it is rarely necessary to use the exact trigonometric solutions (3.33). The gravity wave approximations may be in error by as much as 10%, but the relative error can be reduced to less than 1% for all n, k by modifying (3.35) to

$$\omega_G \approx \pm(2n + 1 + k^2)^{1/2} - \frac{1}{2}\omega_{Rossby} \tag{3.37}$$

where ω_{Rossby} is given by (3.36). This second approximation for gravity waves is consistent with the Vieta Theorem (3.29) (p. 234 of [3]).

Figures 3.5, 3.6, 3.7 and 3.8 illustrate the dispersion relations for $n = -1$ (Kelvin), $n = 0$ (Yanai), and $n = 1, 2, 3$. The graphs use the common oceanographic convention of plotting eastward and westward traveling gravity waves as a single continuous curve by defining ω to always be positive and allowing k to be either positive (for waves propagating towards the east) or negative (for westward-traveling waves). In

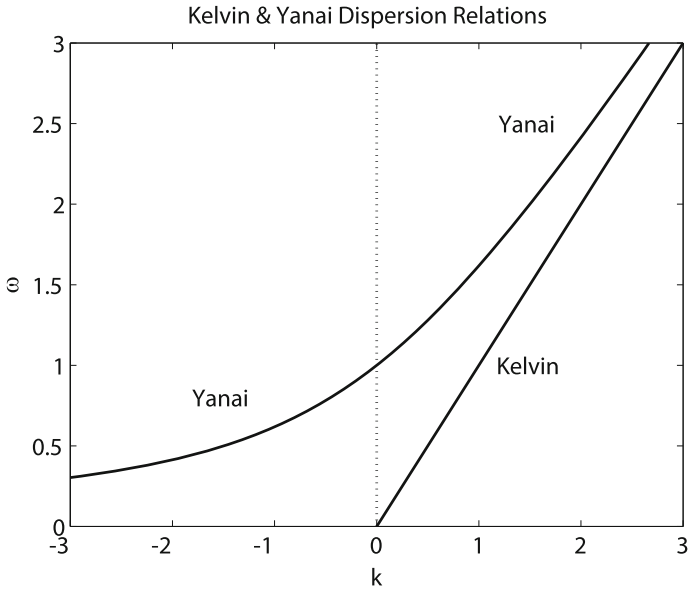

Fig. 3.5 Dispersion relations for the mixed Rossby-gravity (Yanai) wave [latitudinal mode $n = 0$] and for the Kelvin wave [$n = -1$]

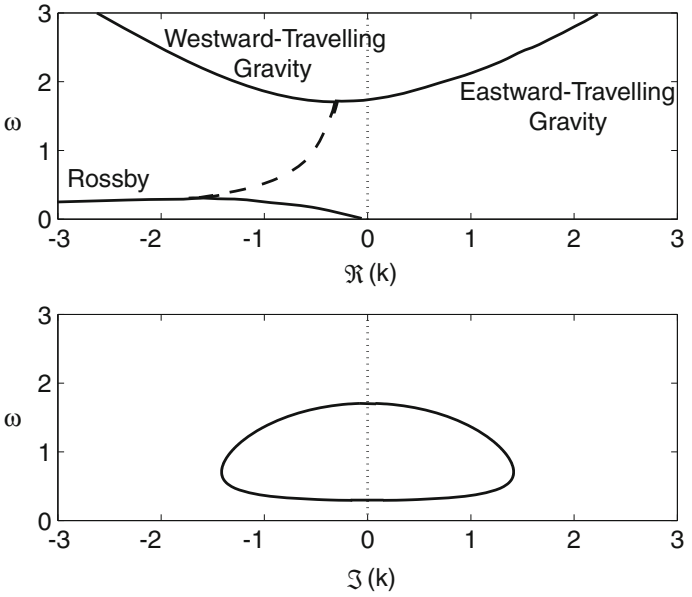

Fig. 3.6 Dispersion relations for the $n = 1$ latitudinal mode. The *dashed line* in the upper diagram shows where k is complex-valued; the imaginary part of the east-west wavenumber k is the oval closed curve in the lower panel. Waves in this intermediate frequency range are coastally trapped, and decay away from shore. The wavenumbers are the two roots of $k^2 - k/\omega + (3 - \omega)$. (Note that the dispersion relation is cubic in frequency ω but only quadratic in zonal wavenumber k for $n \geq 1$.) Complex k is discussed at length in Sect. 3.6

the analytical formulas (3.32), (3.33), (3.35), and (3.37), we have employed the usual meteorological convention of taking $k > 0$ always and allowing the frequency to be either sign. It makes no physical difference so long as we are consistent, but it is important to note that the meteorological and oceanographic conventions are different.

The small positive correction added to the gravity wave frequency for both eastward and westward traveling waves is responsible for the slight asymmetry of the gravity wave branch of the curve; gravity waves have small (but non-zero) group velocities for $k \to 0$.

When the atmosphere drives the ocean at a fixed frequency ω, then we need to solve (3.28) for $k(\omega)$ instead of computing frequency as a function of wavenumber. We will postpone a discussion of this until Sect. 3.6, and first explain in Sect. 3.5 the implications of the large separation between the frequencies of gravity and Rossby waves.

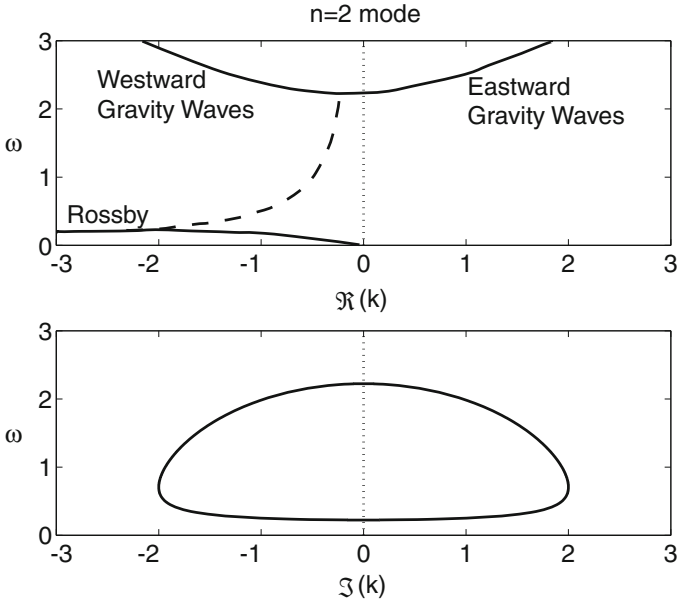

Fig. 3.7 Dispersion relations for the $n = 2$ latitudinal mode, same as the two previous figures but for a different latitudinal mode

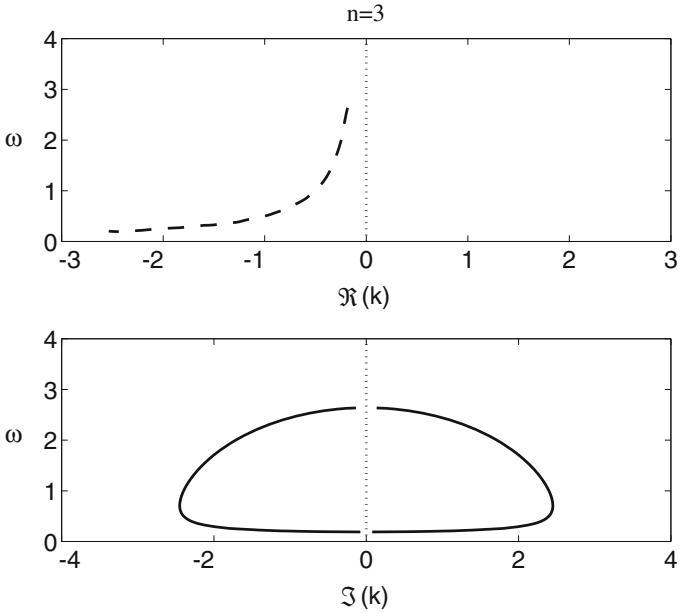

Fig. 3.8 Dispersion relations for the $n = 3$ latitudinal mode, same as the two previous figures but for a different latitudinal mode. The wavenumbers are the two roots of $k^2 - k/\omega + (7 - \omega)$

3.4 Analytic Approximations to Equatorial Wave Frequencies

3.4.1 Explicit Formulas

The exact solutions in terms of the arccosine are precise but much less useful than the algebraic approximations described by the following.

Theorem 3.1 (Approximate Dispersion Relations) *Let w and k denote the nondimensional freuency and wavenumber. Then 1. The dispersion relation for equatorial Rossby waves satisfies the following:*

$$w = \frac{-k}{k^2 + 2n + 1} \{1 + \Delta\} \tag{3.38}$$

for some Δ that satisfies the bound

$$|\Delta| \leq \frac{0.039}{(n + 1/2)^2}, \qquad n \geq 1 \tag{3.39}$$

The relative error is no worse than 1.7% even for the lowest, $n = 1$ mode.

2. For gravity waves,

$$w = \sqrt{k^2 + 2n + 1}\ \{1 + \Delta\} \tag{3.40}$$

$$|\Delta| \leq \frac{0.141}{n + 1/2} \tag{3.41}$$

where the relative error is no worse than 10% even for the lowest, $n = 1$ mode.

3. For gravity waves, the improved approximation and its bound are

$$w = \pm\sqrt{k^2 + 2n + 1}\ \left\{1 \pm \frac{k}{(2\,(k^2 + 2n + 1)^{3/2}}\right\} \{1 + \Delta\} \tag{3.42}$$

$$|\Delta| \leq \frac{0.021}{(n + 1/2)^2} \tag{3.43}$$

the relative error is no worse than 1% even for the lowest, $n = 1$ mode. (From Appendix B of [4].)

3.4.2 Long Wave Series

Long Rossby waves ($|k| \ll c$) have useful Taylor series approximations for the phase speed in powers of the square of the wavenumber. This depends on the mode number through

$$\nu \equiv 2n + 1 \tag{3.44}$$

$$c = -\frac{1}{\nu} + c_2 k^2 + c_4 k^4 + \ldots \tag{3.45}$$

$$c_2 = \frac{-1 + \nu^2}{\nu^4} \tag{3.46}$$

$$c_4 = -\frac{3 - 4\,\nu^2 + \nu^4}{\nu^7} \tag{3.47}$$

The dispersion relation $c^3 - (\nu + k^2)c - 1 = 0$ has a triple point when $k^2 = -\nu$, so the series converges only for $k \le \sqrt{2n+1}$.

The radius of convergence can probably be extended to the entire positive real axis by expanding in the modified perturbation parameter $\eth$ such that

$$k^2 = \eth \frac{\nu}{1 - \eth} \qquad \Leftrightarrow \qquad \eth = \frac{k^2}{k^2 + 2n + 1} \tag{3.48}$$

This is equivalent to Euler acceleration of the power series in k^2. Unfortunately, extensions to quantities other than the phase speed are not yet available.

3.5 Separation of Time Scales

In the middle latitudes, it is well-known that the dominant, low-frequency wave motions are Rossby waves (perhaps modified by shear into baroclinic instability, but still low-frequency) while the gravity waves are merely high-frequency noise. More or the less the same is true in equatorial regions as well, but unfortunately, quasi-geostrophy is not a good approximation for waves at low latitudes. (It becomes better and better for Rossby waves as $n \to \infty$, but is mediocre for the $n = 1$ Rossby wave—which is the most important one.) However, there is still a large gulf between the frequencies of gravity and Rossby waves at the equator.

For simplicity, use the approximate dispersion relations for $n \ge 1$, (3.35) and (3.36). Observe that as the mode number n increases, the frequency (and phase speed) of the Rossby waves becomes smaller and smaller while exactly the reverse happens for the gravity waves. Consequently, the Rossby and gravity modes that are closest in frequency are the $n = 1$ modes. Taking the ratio gives

$$|\omega_G/\omega_R| = (3 + k^2)^{1/2}/[k/(3 + k^2)] \qquad [n = 1 \text{ modes}] \tag{3.49}$$

$$= (3 + k^2)^{3/2}/k \tag{3.50}$$

$$\ge 7.79 \qquad \text{for all } k \tag{3.51}$$

as can be shown by differentiating (3.50) with respect to k, computing the zero of the derivative which is the minimum k_{min}, and lastly evaluating the ratio at this $k = k_{min}$. Thus, even for this worst case, there is a separation of an order of magnitude between the time scales of even the quickest Rossby wave and the slowest pure gravity wave.

The Kelvin wave and mixed-Rossby gravity wave have frequencies intermediate between those of the pure Rossby and gravity waves for $n \geq 1$.[1] Nonetheless, this wide disparity of time scales has made it possible to apply a "two-timing" initialization procedure to replace quasi-geostrophy for global models by an alternative method, still accurate at low levels, which exploits this separation in frequency to filter out high frequency gravity waves (Tribbia, Kasahara, Lynch) [5–8].

3.6 Forced Waves

When we calculate the free oscillations of the ocean, Lamb's parameter is known—it is determined by the particular baroclinic mode we are interested in — and either the phase speed c or frequency ω is the eigenvalue. It is perfectly possible, however, to have waves which are forced at a definite frequency ω. In this case, either ε or the zonal wavenumber k is the eigenvalue.

In the atmosphere, the boundary condition in x is zonal periodicity. It follows that the wavenumber k is always real and is quantized by the requirement that the wave be periodic around the circumference of the earth—in other words, the dimensional zonal wavenumber at the equator must equal an integer divided by the radius of the earth. Lamb's parameter ε, which determines the vertical structure of the waves, is the eigenvalue.

Although ε must always be positive for free oscillations— the free oscillations of a baroclinic mode are analogous to those of a single layer of fluid with the appropriate equivalent depth, and the depth must always be positive—it is perfectly possible that for waves forced at a particular frequency ω, for $\varepsilon(\omega)$ to be negative. (Platzman reviews the seven (!) independent discoveries of negative Lamb's parameter [9].) Physically, $\varepsilon < 0$ simply means that the forced wave is decaying exponentially with height instead of being vertically propagating. For a semi-infinite fluid, i.e., one with no top boundary, this vertical trapping is just as consistent with the upper boundary condition as is the radiating solution where the vertical structure function is proportional to $\exp(irz)$. All atmospheric tidal calculations done before the early 60's were wrong because the modes with negative ε were simply left out of the sum over modes in the mistaken belief that they did not exist.

[1] In the shortwave limit $k \to \infty$, the gravity wave phase speeds converge to one from above creating a shortwave resonance with a steepening, nonlinear Kelvin wave as described in Chap. 16. This is not important for initialization, however.

Wu and Moore (2004) and Wu, Sarachik and Battisti (1999) show that a *continuous spectrum* of modes of negative ε is necessary for some forced problems in meteorology [10, 11].

For *equatorial* waves linearized about a resting mean state, however, $\varepsilon > 0$ always. However, to show this and to do any other useful calculation for forced waves, the ε-dependent nondimensionalization is no longer convenient. To convert the dimensional frequency into a nondimensional frequency, one must know ε— but ε is always 1 in the ε-dependent scaling, and one could not calculate it even if it still appeared as an explicit parameter because one does not know the nondimensional frequency without it.

Consequently, the ε-dependent scaling of this Chapter, which is so ubiquitous in oceanography, is never used in meteorology. Instead, atmospheric scientists use a nondimensionalization in which the length scale is always the radius of the earth and the time scale is always the reciprocal of twice the angular frequency of the earth's rotation, as done in [1]. Let σ denote the dimensional frequency/(2Ω) and s [an integer] denote the dimensional zonal wavenumber divided by the radius of the earth The dispersion relation becomes

$$\varepsilon\sigma^3 - [\varepsilon^{1/2}(2n+1) + s^2]\sigma - s = 0 \tag{3.52}$$

For fixed ω, this equation can be *explicitly* solved as a *quadratic* equation for the square root of ε.

In the ocean, the forced wave problem is different because the sea is bounded both above and below, which forces ε to be quantized, i.e., to assume only a discrete set of values. Coastlines destroy the zonal boundary condition of periodicity that is always correct in the atmosphere. Instead, k becomes the eigenvalue, and it may become *imaginary*, physically representing forced waves that decay exponentially in longitude away from continental boundaries (see O'Connor and Sozou [12–15]). The appropriate dispersion relations in terms of k as the eigenvalue (and with ε nondimensionalized to 1, as always in the ocean) are

$$k = \omega \qquad \text{[Kelvin wave]} \tag{3.53}$$

$$k = \omega - 1/\omega \qquad \text{[Yanai, alias mixed Rossby gravity]} \ \ n = 0 \tag{3.54}$$

$$\text{[also eastward } n = 0 \text{ gravity wave]}$$

$$k = -1/(2\omega) \pm \{\omega^2 + 1/(4\omega^2) - (2n+1)\}^{1/2} \qquad n \geq 1 \tag{3.55}$$

Note that for forced waves, the dispersion relation for general n is a *quadratic* in $\varepsilon^{1/2}$ or k rather than a *cubic* so that (i) all solutions are explicit, requiring nothing more exotic than square roots and (ii) there are at most two forced waves for a given n in contrast to the three free modes for each n and (iii) the eigenvalue can be *complex-valued*.

Figures 3.5, 3.6, 3.7 and 3.8, which illustrate $\omega(k)$, of course also display the forced wave dispersion relations $k(\omega)$ (3.53)–(3.55). The graphs employ the convention, not

used in meteorology, of taking the frequency to always be positive while k can be of either sign. With this convention, waves with *negative* k correspond to *westward* traveling disturbances such as Rossby waves. Note also that the Kelvin and mixed Rossby-gravity waves can never be zonally trapped; it is only when the mode $n \geq 1$ that k can be complex. The dashed line in the upper parts of Figs. 3.6, 3.7 and 3.8 denotes that part of the curve for which k is complex; the solid curves represent the real valued gravity wave [upper curves, opening upward to infinity] and Rossby waves [k negative only and asymptoting to 0 at the origin]. The argument given in Sect. 3.5 to show the large frequency separation of the gravity and Rossby waves applies only when k is real. The branch with k complex connects the real-valued Rossby and gravity wave curves as shown in the figures.

These boundary trapped waves can be readily generated simply by forcing the ocean in a particular baroclinic mode with the structure of a particular latitudinal mode at a frequency between that of the maximum frequency of a free Rossby wave and the minimum frequency gravity wave. For the $n = 1$ mode, this means k will have an imaginary part whenever

$$3^{1/2} > -\omega > 3^{1/2}/6 \tag{3.56}$$

We will see later that such boundary trapped waves are excited not only by wind stress if the stress is oscillating with ω in the necessary range (3.56), but also as part of the reflection of zonally propagating equatorial waves from coastal boundaries.

3.7 How the Mixed-Rossby Gravity Wave Earned Its Name

For $n = 1$, the two westward-traveling free modes that exist for higher n are combined into a single mode. This wave is not described as either a Rossby wave or a gravity wave because in different parametric ranges, it behaves like both — hence its description as a "mixed-Rossby-gravity" wave. Oceanographers often use the shorter (but less descriptive) term "Yanai" wave for this same mode after Michio Yanai of UCLA, who was the first person to detect such waves in the atmosphere in the late 60's. Meteorologists use this name less frequently than oceanographers, in part because the existence of this wave was predicted theoretically by Stanley Rosenthal several years earlier and, independently, by a number of oceanographers (Stern, Bretherton, and Hendershott). Another important note: oceanographers plot both the westward and eastward traveling $n = 0$ waves as a single, continuous function as evident in Fig. 3.5, so they sometimes use the name "Yanai" wave to denote all $n = 0$ waves, regardless of their direction of phase propagation.

When ε is plotted against σ for fixed zonal wavenumber, as meteorologists normally do, where σ denotes a frequency that is allowed to be of either sign, the graphs for the eastward and westward traveling waves cannot be joined into a single, continuous graph, so meteorologists restrict "Yanai" and "mixed Rossby-gravity" to the westward-traveling wave only.

In the atmospheric case where the frequency σ is known and ε is the eigenvalue, the *Rossby* wave-like behavior occurs for *large* σ: ε becomes small and then *negative* as σ increases. No true gravity wave ever has a negative ε! (This is shown in Longuet-Higgin's exhaustive (1968) study of the solutions of Laplace's Tidal Equation on a sphere with equatorial waves representing the $\varepsilon \to \infty$ limit.) In the opposite limit of small σ, ε goes to ∞ as $1/\sigma^4$, like a gravity wave, rather than the slower increase proportional to $1/\sigma^2$. Thus, we have

$$|\sigma| \geq 1/s \to \varepsilon < 0 \qquad \text{[global Rossby wave]} \tag{3.57a}$$

$$|\sigma| \ll 1/s \to \varepsilon \gg 1 \qquad \text{[equatorial gravity wave]} \tag{3.57b}$$

As explained in Longuet-Higgins [16], planetary waves are equatorially trapped only when $\varepsilon + s^2$ is large and positive [assuming no mean current with shear], and waves with negative ε are global and cannot be well approximated by the equatorial beta-plane. (Here s is the integral zonal wavenumber and its role in equatorial trapping is explained in [17, 18] and Sect. 3.12.) For $\varepsilon = 0$, the spherical problem can be solved exactly as done by Haurwitz [19]; this gives $\sigma = -1/s$ for the mixed Rossby-gravity wave. For intermediate σ, the wave is a hybrid of Rossby and gravity wave (hence the name!).

In the oceanic case, $\varepsilon = 1$ (after nondimensionalization) and we are interested in either $\omega(k)$ or, for forced waves, $k(\omega)$. Both gravity and Rossby wave behavior are still manifest in the equatorial beta-plane dispersion relation shown in Fig. 3.5, proving that it is possible for the mixed Rossby-gravity wave to be Rossby-like even when it is equatorially trapped, and not merely when it is a global, vertically trapped ($\varepsilon < 0$) mode. However, an oceanographer would describe the parametric range differently. Comparing Figs. 3.5 and 3.6, observe that the westward-traveling wave ($k < 0$) behaves like a Rossby wave for large negative k in that ω remains bounded (and actually goes to zero, though this cannot be seen from the graph) as $|k| \to \infty$ while ω is a monotonically increasing function of k for gravity waves. For small negative wavenumber, however, the Yanai wave is similar to a gravity wave in that it tends to a finite, non-zero limit as $k \to 0$ whereas all the pure Rossby waves have $\omega \to 0$ in this same limit. Thus, we have

$$-k \gg 1 \qquad \to \omega \ll 1 \qquad \text{[equatorial Rossby wave]} \tag{3.58a}$$

$$-k = O(1) \qquad \to \omega \sim O(1) \qquad \text{[equatorial gravity wave]} \tag{3.58b}$$

3.8 Hough-Hermite Vector Basis

3.8.1 Introduction

In undergraduate physics, trigonometric functions play a dual role. Mathematically, $\cos(kx)$ and $\sin(kx)$ are the *basis* functions for Fourier series. An arbitrary function

$u(x)$, whether it represents waves or not, can always be expanded as a Fourier series. Spectral and pseudospectral methods are schemes for solving differential equations — elliptic, hyperbolic or whatever — that employ a Fourier basis. Physically, however, the structure of infinitesimal amplitude waves in a homogeneous medium is also described by $\cos(kx)$ and $\sin(kx)$. More complicated nonlinear waves are no longer simply proportional to a single trigonometric function; nonlinearity couples different wavenumbers together, and variations in refractive index can modulate the amplitude and wavelength. Nevertheless, this dual role of trigonometric functions is full of insight. Simple, somewhat idealized problems are often solved by expanding the initial condition as a Fourier series. The general solution is that each Fourier component propagates at its own speed. The mathematics of basis functions blends thusly into the physics of waves.

The same is true in more complicated fashion for planetary waves in general and equatorial waves in particular. Hough functions, that is, the normal modes of Laplace's linear tidal equations for a global homogeneous ocean of constant depth, have been sporadically employed as the latitude-and-longitude basis functions for two-dimensional and three-dimensional atmospheric models [7, 20–25].

On the equatorial beta-plane, as noted earlier, the exact solutions to the linearized equations of motion are wave modes that are finite linear combinations of Hermite functions in latitude multiplied by sinusoidal functions in longitude and, in three dimensions, by baroclinic modes in depth. However, the Hermite functions are also a complete set of basis functions for smooth functions on an infinite interval. Even when the ocean is *not* idealized or linear, the Hermite-spectral method is a good numerical method as illustrated in Anderson [26], Holvorcem and Vianna [27, 28], Smith [29], Majda and Khoulder [30] and Tribbia [6] among others. Again, the Hermite functions are both basis function [mathematics] and wave mode [physics].

There is a complication, however: if we expand each of the three unknowns of the shallow water set as Hermite series, then, even using the sum-difference variables, three wave species are mixed together. Thus, $v_n(x, t)$, the n-the Hermite coefficient of the north-south current, is the sum of the n-th Rossby mode plus the eastward-traveling and westward-traveling n-th mode gravity waves. The function $v_n(x, t)$ will have a complicated time evolution as each of its three component waves propagates separately.

To create a one-to-one match between basis functions and waves, we have to use *vector* eigenfunctions that we will dub the "Hough-Hermite" functions. S.S. Hough in the late 19th century showed how to compute the vector eigenfunctions of the shallow water equations on the sphere, which are "Laplace's Tidal Equations" in that context. The eigenfunctions are commonly called "Hough" functions in his honor.[2] The Hough-Hermite functions are asymptotic approximations to the spherical Hough functions in the limit of small equivalent depth, that is, of small mean depth in the shallow water system.

[2] "Hough" is pronounced "Huf"; Sir Frank Dyson, Astronomer Royal, gives a biography of Sydney Samuel Hough, F. R. S., in [31].

Table A.17 is a catalogue of numerical models that use Hermite functions or Hough-Hermite functions. However, Hermite functions and equatorial waves are so entangled that almost every article on equatorial wave theory could be added to the list.

3.8.2 Inner Product and Orthogonality

Theorem 3.2 (Hough-Hermite Orthonormality) *Assume that the zonal wavenumber k is restricted to an integral multiple of the zonal period. Define the scalar inner product by*

$$< f, g > \equiv \int_0^P dx \int_{-\infty}^{\infty} f(x, y)\, g(x, y) dy \tag{3.59}$$

Then the Hough-Hermite vector eigenfunctions are orthonormal with respect to the following vector inner product:

$$< S_{k,n,\mathfrak{q},itype} S_{k',n',\mathfrak{q}',itype'} > + 2 < v_{k,n,\mathfrak{q}} v_{k',n',\mathfrak{q}',itype'} > + < D_{k,n,\mathfrak{q},itype} D_{k',n',\mathfrak{q},itype'} > = \delta_{k,k'} \delta_{n,n'} \delta_{\mathfrak{q},\mathfrak{q}'} \delta_{itype,itype'}, \quad k \neq 0 \tag{3.60}$$

where δ_{ij} is the usual Kronecker δ such that $\delta_{jj} = 1$ whereas $\delta_{ij} = 0, i \neq j$. The reason for the factor of two multiplying the scalar inner product in v, but not the other two terms, will be explained below.

Proof The orthogonality of modes of different latitudinal mode number n is rather trivial. The sum variable S for the n-th mode is always proportional to $\psi_{n+1}(y)$. It follows that the sum variable for a different latitudinal mode number n' will be proportional to a different Hermite function. Integrating the product $S_{k,n,\mathfrak{q},itype} S_{k',n',\mathfrak{q}',itype'}$ will therefore give zero because the Hermite functions are orthonormal. Similarly, when we integrate the products of different zonal wavenumbers k and k' over x, the usual orthogonality of different degrees of a Fourier series will again give zero.

The non-trivial part of the proof is to show that the vector inner product of different wave types of the same n and k are orthogonal.

To prove orthogonality in the case when $k = k'$ and $n = n'$, we need to look carefully at the solutions of the coupled system of three equations in x and t which are solved by the various wave species of a given latitudinal mode number n.

First, substitute

$$\begin{vmatrix} S_{n+1}(x,t) \\ v_n(x,t) \\ D_{n-1}(x,t) \end{vmatrix} = \begin{vmatrix} -\tilde{S}\cos(kx - \omega t) \\ \tilde{v}\sin(kx - \omega t) \\ -\tilde{D}\cos(kx - \omega t) \end{vmatrix} \tag{3.61}$$

into the linearized shallow water equations. This gives the algebraic eigenvalue problem for the three modes of a given n and k as

$$\begin{vmatrix} (k-\omega) & -[2(n+1)]^{1/2} & 0 \\ -[(n+1)/2]^{1/2} & -\omega & [n/2]^{1/2} \\ 0 & [2n]^{1/2} & (-k-\omega) \end{vmatrix} \begin{vmatrix} \tilde{S} \\ \tilde{v} \\ \tilde{D} \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \\ 0 \end{vmatrix} \tag{3.62}$$

The matrix is not symmetric, but can be made so by (i) multiplying the middle equation by $\sqrt{2}$ and (ii) defining the modified unknown

$$\hat{v} \equiv \sqrt{2}\tilde{v} \tag{3.63}$$

The algebraic eigenvalue problem then becomes

$$\begin{vmatrix} (k-\omega) & -(n+1)^{1/2} & 0 \\ -(n+1)^{1/2} & -\omega & n^{1/2} \\ 0 & n^{1/2} & (-k-\omega) \end{vmatrix} \begin{vmatrix} \tilde{S} \\ \hat{v} \\ \tilde{D} \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \\ 0 \end{vmatrix} \tag{3.64}$$

We can then invoke the well-known matrix theorem that the eigenvalues of a symmetric, real-valued matrix are real and the eigenvectors are orthogonal. It follows that the eigenvectors of different modes of the same n and k will be orthogonal with respect to the usual vector inner product. This in turn confirms the orthogonality of the integral inner product in (3.60). Reverting to the original variable $\tilde{v}$ gives the factor of two multiplying $< v, v' >$ in (3.60). ■

It is straightforward to choose a normalization constant, denoted below by $\beth_{k,n,itype}$, so the vector basis functions have unit vector inner product.

3.8.3 Orthonormal Basis Functions

The Hough-Hermite functions can be written in terms of either $\exp(ikx)$, where k must take both positive and negative values, or equivalently as real-valued functions using $\cos(kx)$ and $\sin(kx)$. Since the complex-valued forms are given in Tables 3.1, 3.2 and 3.3, we shall display the real-valued forms here.

The Kelvin basis function is, with q restricted to either $q = 0$ or $q = 1$,

$$\mathfrak{u}_{k,-1,\mathfrak{q}} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{-1} \begin{vmatrix} \cos(kx - \mathfrak{q}\pi/2)\psi_0(y) \\ 0 \\ 0 \end{vmatrix} \tag{3.65}$$

where

$$\beth_{-1} = 1/\sqrt{P/2} \tag{3.66}$$

The Yanai basis functions are, restricting $itype$ to only two modes,

$$\mathfrak{u}_{k,0,\mathfrak{q},itype} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{k,0,itype} \begin{vmatrix} \frac{\sqrt{2}}{\omega - k} \cos(kx - \mathfrak{q}\pi/2)\psi_1(y) \\ \sin(kx - \mathfrak{q}\pi/2)\psi_0(y) \\ 0 \end{vmatrix}, \qquad k \neq 0 \quad (3.67)$$

where

$$\beth_{k,0,itype} = 1/\sqrt{P\left\{1/(\omega - k)^2 + 1\right\}} \quad (3.68)$$

The Yanai basis functions for zero wavenumber are

$$\mathfrak{u}_{0,0,itype,\mathfrak{q}} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{0,0,itype} \begin{vmatrix} \pm\sqrt{2}\,\psi_1(y) \\ \psi_0(y) \\ 0 \end{vmatrix} \quad (3.69)$$

where $\omega = 1$ (and the plus sign applies) for $itype = 1$ while $\omega = -1$ and the negative sign applies for $itype = 2$.

$$\beth_{0,0,itype} = \frac{1}{2\sqrt{P}}, \qquad k = 0 \quad (3.70)$$

where the factor of 2 arises because the average of the constant over $x \in [0, P]$ is P, double that of $\cos^2(kx)$ for any nonzero wavenumber. That is $\int_0^P \cos^2((2\pi\kappa/P)x)\,dx = P/2$ where κ is a nonzero integer and $\int_0^P \cos^2(0)dx = P$.

The basis functions for latitudinal mode $n \geq 1$ are

$$\mathfrak{u}_{k,n,itype,\mathfrak{q}} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{k,n,itype} \begin{vmatrix} \cos(kx - \mathfrak{q}\pi/2)\frac{\sqrt{2(n+1)}}{\omega(k,n,itype)-k}\psi_{n+1}(y) \\ \sin(kx - \mathfrak{q}\pi/2)\,\psi_n(y) \\ \cos(kx - \mathfrak{q}\pi/2)(-1)\frac{\sqrt{2n}}{\omega(k,n,itype)+k}\psi_{n-1}(y) \end{vmatrix} \quad (3.71)$$

where k is the zonal wavenumber, n is the latitudinal mode number ($n \geq 1$), $itype$ is a triple-valued index (one Rossby and two gravity modes) and $\mathfrak{q} = 0, 1$. $\beth$ is the normalization constant,

$$\beth = \begin{cases} 1/\sqrt{P\left\{\frac{(n+1)}{(k-\omega)^2} + 1 + \frac{n}{(k+\omega)^2}\right\}} & k \neq 0 \\ \frac{1}{\sqrt{P}}\,\frac{1}{\sqrt{\frac{2(n+1)}{\omega^2}+2+\frac{2n}{\omega^2}}} & k = 0, itype = 2 \text{ or } 3 \end{cases} \quad (3.72)$$

where P denotes the spatial period in the east-west direction. (The limit $P \to \infty$, which corresponds to a zonally-unbounded ocean, causes no particular difficulties.)

For zero wavenumber, the geostrophic modes [limits of Rossby waves as $k \to 0$] are

$$\mathfrak{u}_{0,n,1} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{0,n,1} \begin{vmatrix} \psi_{n+1}(y) \\ 0 \\ \sqrt{\frac{n+1}{n}}\,\psi_{n-1}(y) \end{vmatrix} \quad (3.73)$$

$$\beth_{0,n,1} = \sqrt{\frac{n}{P\,(2n+1)}} \tag{3.74}$$

For zero wavenumber, the gravitational modes are

$$u_{0,n,itype} = \begin{vmatrix} S \\ v \\ D \end{vmatrix} = \beth_{0,n,itype} \begin{vmatrix} (-1)\,\frac{\sqrt{2(n+1)}}{\omega}\,\psi_{n+1}(y) \\ \psi_n(y) \\ \frac{\sqrt{2n}}{\omega}\,\psi_{n-1}(y) \end{vmatrix}, \qquad itype = 2, 3 \tag{3.75}$$

$$\beth_{0,n,itype} = \frac{1}{\sqrt{P}}\,\frac{1}{\sqrt{\frac{2(n+1)}{\omega^2} + 2 + \frac{2n}{\omega^2}}}, \qquad itype = 2, 3 \tag{3.76}$$

3.9 Applications of the Hough-Hermite Basis: Linear Initial-Value Problems

The Hough-Hermite functions, in their dual role as both waves and basis functions, provide an almost trivial solution to the general unforced, inviscid initial-value problem for the linearized shallow water equations on the equatorial beta-plane. Suppose that the initial conditions are $u^{init}(x, y)$, $v^{init}(x, y)$ and $\phi^{init}(x, y)$. The corresponding initial conditions for the sum and difference variables are $S^{init}(x, y) \equiv \phi^{init}(x, y) + u^{init}(x, y)$ and $D^{init}(x, y) \equiv \phi^{init}(x, y) - u^{init}(x, y)$. Define the integer zonal wavenumber κ by

$$\kappa \equiv (P/[2\pi])\,k \tag{3.77}$$

where P is the east-west spatial period and k is the zonal wavenumber, which is generally not an integer. Further define

$$\left.\begin{matrix} S^{init}_{\kappa,n,\mathfrak{q}} \\ D^{init}_{\kappa,n,\mathfrak{q}} \end{matrix}\right\} \equiv \int_0^P \cos(kx - \mathfrak{q}\pi/2)dx \int_{-\infty}^{\infty} dy \left\{\begin{matrix} \psi_{n+1}(y)\,S^{init}(x, y) \\ \psi_{n-1}(y)\,D^{init}(x, y) \end{matrix}\right. \tag{3.78a}$$

$$v^{init}_{\kappa,n,\mathfrak{q}} \equiv \int_0^P \sin(kx - \mathfrak{q}\pi/2)dx \int_{-\infty}^{\infty} dy\,\psi_n(y) v^{init}(x, y) \tag{3.78b}$$

and similarly for the other fields.

The coefficient of the Hough-Hermite basis function is

$$a_{k,n,\mathfrak{q},itype} = \beth_{k,n,itype} \left\{ S^{init}_{\kappa,n,\mathfrak{q}}\,\frac{\sqrt{2(n+1)}}{\omega(k, n, itype) - k} + 2v^{init}_{\kappa,n,\mathfrak{q}} - \frac{\sqrt{2n}}{\omega(k, n, itype) + k}\,D_{\kappa,n,\mathfrak{q}} \right\}$$
$$\kappa = 0, 1, 2, \ldots;\ n = 1, 2, \ldots;\ \mathfrak{q} = 0, 1; \qquad itype = 1, 2, 3 \tag{3.79}$$

where $\beth$ is defined by (3.72).

The Kelvin and Yanai modes require special treatment, but the ideas are the same as for $n \geq 1$. The Kelvin coefficient is

$$a_{\kappa,-1,\mathfrak{q}} = \frac{2}{P}\int_0^P \cos((2\pi/P)\kappa\, x - \mathfrak{q}\pi/2)dx \int_{-\infty}^{\infty} dy\, \psi_0(y) S^{init}(x,\, y) \qquad (3.80)$$

where index $itype$ is omitted since there is only one species of Kelvin wave.

The Yanai coefficients are

$$a_{k,0,\mathfrak{q},itype} = \beth_{k,0,itype}\left\{ S^{init}_{\kappa,0,\mathfrak{q}} \frac{\sqrt{2}}{\omega(k,n,itype) - k} + 2v^{init}_{\kappa,0,\mathfrak{q}} \right\}$$
$$\kappa = 0, 1, 2, \ldots;\ n = 1, 2, \ldots;\ \mathfrak{q} = 0, 1; \qquad itype = 1, 2 \qquad (3.81)$$

The solution to the general initial value problem is then

$$\begin{vmatrix} S(x,\, y,\, t) \\ v(x,\, y,\, t) \\ D(x,\, y,\, t) \end{vmatrix} = \sum_{\kappa=0}^{\infty}\sum_{\mathfrak{q}=0}^{1} a_{\kappa,-1,\mathfrak{q}}\, \beth_{-1}\, , \begin{vmatrix} \cos(kx - \mathfrak{q}\pi/2)\psi_0(y) \\ 0 \\ 0 \end{vmatrix}$$
$$+ \sum_{n=1}^{\infty} a_{0,n,1}\beth_{0,n,1} \begin{vmatrix} \psi_{n+1}(y) \\ 0 \\ \sqrt{\frac{n+1}{n}}\,\psi_{n-1}(y) \end{vmatrix} + a_{0,0,\mathfrak{q},itype}\beth_{0,n,1} \begin{vmatrix} \psi_{n+1}(y) \\ 0 \\ \sqrt{\frac{n+1}{n}}\,\psi_{n-1}(y) \end{vmatrix}$$
$$+ \sum_{\kappa=1}^{\infty}\sum_{\mathfrak{q}=0}^{1}\sum_{itype=2}^{3} a_{\kappa,0,\mathfrak{q},itype}\, \beth_{\kappa,0,itype} \begin{vmatrix} \frac{\sqrt{2}}{\omega - k}\cos(\kappa x - \mathfrak{q}\pi/2)\psi_1(y) \\ \sin(\kappa x - \mathfrak{q}\pi/2)\psi_0(y) \\ 0 \end{vmatrix},$$
$$+ \sum_{\kappa=1}^{\infty}\sum_{n=1}^{\infty}\sum_{\mathfrak{q}=0}^{1}\sum_{itype=1}^{3} a_{\kappa,n,\mathfrak{q},itype}\, \beth_{\kappa,n,itype} \times$$
$$\begin{vmatrix} \{\cos(\kappa x - \mathfrak{q}\pi/2)\cos(\omega t) + \sin(kx - \mathfrak{q}\pi/2)\ \sin(\omega t)\}\ \frac{\sqrt{2(n+1)}}{\omega(\kappa,n,itype) - \kappa}\psi_{n+1}(y) \\ \{\sin(\kappa x - \mathfrak{q}\pi/2)\cos(\omega t) - \cos(kx - \mathfrak{q}\pi/2)\ \sin(\omega t)\}\ \psi_n(y) \\ \{\cos(\kappa x - \mathfrak{q}\pi/2)\cos(\omega t) + \sin(\kappa x - \mathfrak{q}\pi/2)\ \sin(\omega t)\}\,(-1)\frac{\sqrt{2n}}{\omega(\kappa,n,itype) + \kappa}\psi_{n-1}(y) \end{vmatrix} \qquad (3.82)$$

It is only a modest extension to solved forced problems for the linearized shallow water equations.

3.10 Initialization Through Hough-Hermite Expansion

For nonlinear numerical models, the vector basis functions are equally important. Numerical weather forecasts are awful when raw observations define the initial condition: the small, random observational errors excite unrealistically large amounts of

high frequency gravity waves. The forecast is then poisoned by spurious, rapid oscillations. Consequently, all operational forecasting models apply an "initialization" step which adjusts the data by suppressing the gravity waves. The goal is to adjust the raw data onto the nearest point on the hypothetical "slow manifold", a subspace of the phase space of the model which consists of purely low frequency motion.

As reviewed by Tribbia [5, 6], a popular strategy for initialization of spherical models is to expand the initial conditions in Hough functions. One can suppress the high frequency gravity waves by setting the coefficients of those modes equal to zero.

The same strategy is a natural way to initialize models on the equatorial beta-plane. The coefficients of Hough-Hermite modes with large ω — how large is "large" is up to the modeler — are set equal to zero. The crucial point is that the Kelvin wave should usually *not* be modified in initialization: its phase speed is only three times that of the fastest Rossby wave, and it is the simplest and likely the most energetic wave in the tropical ocean.

Nonlinear interactions usually excite the "initialized" high frequency modes as the flow evolves. Refined nonlinear initializations are described by Tribbia and the articles by Baer and Tribbia [32, 33]. The Baer–Tribbia perturbation series is asymptotic but divergent; a subtlety is that the slow manifold is existent to all orders in powers of the small parameter but nonexistent "beyond all orders" due to a tiny, unavoidable residual of high-frequency motion whose amplitude is an exponential function of the reciprocal of the small parameter [34–44]. The crucial point is that normal mode initialization, even when extended to higher order, is still all about expanding the initial conditions in Hough or Hough-Hermite functions.

3.11 Energy Relationships

The relative amounts of meridional and zonal kinetic energy and potential energy are often as important in identifying waves modes in observations as is the detailed structure collected in Tables 3.1, 3.2, 3.3, 3.4, 3.5 and 3.6, so it is useful to discuss energetics here. If we define

$$MKE \equiv \iint (1/2)v^2\,dx\,dy \qquad \text{["Meridional Kinetic Energy"]} \tag{3.83}$$

Table 3.1 Kelvin mode, $n = -1$

$c \equiv 1$ (dispersion relation)
$\phi = e^{-(1/2)y^2}$
$u \equiv \phi$
$v \equiv 0$
$\psi \equiv y\phi$
$q \equiv 0$

Table 3.2 Yanai wave, $n = 0$. In the dispersion relation, the + sign is the eastward gravity wave; the − sign gives the mixed-Rossby gravity or "Yanai" wave. The expressions for the fields apply to both the eastward and westward traveling $n = 0$ modes with use of the appropriate phase speed c

$c = (1/2) \pm [1/4 + 1/k^2]^{1/2}$
$\phi = \frac{1}{(1-c)} y e^{-(1/2)y^2}$
$u \equiv \phi$
$v = ike^{-(1/2)y^2}$
$q = [-k^2 - 1/(1-c)]e^{-(1/2)y^2}$

Table 3.3 Higher Rossby and gravity waves $n \geq 1$. The phase speeds c are the three solutions of the cubic equation $\omega^3 - (2n + 1 + k^2)\omega - k = 0$ where $\omega \equiv kc$ is the frequency. The dispersion relation is a cubic because the linearized shallow water equations are a set of *three* equations for each positive integer with a first order time derivative appearing in each one. Physically, the three roots correspond to an eastward traveling gravity wave, a westward traveling gravity wave, and a westward traveling Rossby wave. Although the cubic cannot be *exactly* factorized, the roots are so well separated in frequency —the gravity wave roots are of roughly equal magnitude but opposite sign while the Rossby root is always negative and of very small magnitude — that highly accurate approximations are available as given below

$c_{Rossby} = \frac{-1}{2n+1+k^2}[1 + E_R], \|E_R\| \leq \frac{0.039}{(n+1/2)^2}$ $\omega_{gravity} = \pm(2n + 1 + k^2)^{1/2}[1 + E_G]$, $\|E_G\| \leq \frac{0.141}{(n+1/2)}$
$\phi = \frac{1}{(1-c)}[(n+1)/2]^{1/2}\psi_{n+1} + \frac{1}{(1+c)}[n/2]^{1/2}\psi_{n-1}$
$u = \frac{1}{(1-c)}[(n+1)/2]^{1/2}\psi_{n+1} - \frac{1}{(1+c)}[n/2]^{1/2}\psi_{n-1}$
$v = ik\psi_n$
$\zeta = \frac{[(n+1)(n+2)]^{1/2}}{2(1-c)}\psi_{n+2} + \frac{[n(n-1)]^{1/2}}{2(1+c)}\psi_{n-2} - \left[\frac{(n+1)}{2(1-c)} + \frac{n}{2(1+c)} + k^2\right]\psi_n$
$q = -\psi_n\left[\frac{(n+1)}{(1-c)} + \frac{n}{(1+c)} + k^2\right]$

$$ZKE \equiv \iint (1/2)u^2\, dx\, dy \qquad \text{["Zonal Kinetic Energy"]} \tag{3.84}$$

$$PE \equiv \iint (1/2)\phi^2\, dx\, dy \qquad \text{["Potential Energy"]} \tag{3.85}$$

then it is easy to calculate these three terms for each linear mode by performing the integrations using the explicit solutions of Tables 3.1, 3.2 and 3.3. We find that

$$ZKE \equiv PE \qquad \text{[all modes, all wavenumbers]} \tag{3.86}$$

$$MKE/ZKE = \begin{cases} 0 & \text{[Kelvin waves]} \\ = 2k^2(1-c^2)^2/\{(2n+1)(1+c^2)+2c\} & n \geq 0 \end{cases} \tag{3.87}$$

Table 3.4 Latitudinal symmetries and Hermite expansions. Neglecting mean currents and linearizing about a state of rest, all equatorial fields are (neglecting mean currents) either symmetric or antisymmetric about the equator. The middle column lists the symmetry with respect to $y = 0$ for the fields of a *symmetric* mode, that is, one with n odd. The third column lists the Hermite functions which appear in the Hermite series for the function. Mean currents do not break these symmetries if and only if each mean component matches the symmetry of that component for a symmetric wave mode; thus, the mean zonal velocity and height fields must be symmetric about the equator

Variable		Equatorial symmetry	Hermite
North-south current	v	Antisymmetric	ψ_n
Zonal current	u	Symmmetric	$\psi_{n+1},\ \psi_{n-1}$
Height (pressure)	ϕ	Symmetric	$\psi_{n+1},\ \psi_{n-1}$
Vorticity	ζ	Antisymmetric	$\psi_{n+2},\ \psi_n,\ \psi_{n-2}$
Potential vorticity	q	Antisymmetric	ψ_n
Divergence	Δ	Symmetric	$\psi_{n+1},\ \psi_{n-1}$
Temperature	T	Symmetric	$\psi_{n+1},\ \psi_{n-1}$
Vertical velocity	w	Symmetric	$\psi_{n+1},\ \psi_{n-1}$
Perturbation density	ρ	Symmetric	$\psi_{n+1},\ \psi_{n-1}$
Streamfunction	ψ	Antisymmetric	Infinite series
Velocity potential	χ	Symmetric	Infinite series

Table 3.5 Zonal structure groups

Pressure group:	ϕ, u, ζ, q, T
North-south velocity group:	v, Δ, w

Table 3.6 Vertical structure groups

$\mathfrak{z}_m(z)$:	$u, v, \phi, \zeta, q, \Delta$
$\int^z \mathfrak{z}_m(z')\, dz'$:	w
$(d/dz)\, \mathfrak{z}_m(z)$:	T, ρ

For Rossby waves, $|c| \le 1/3$ for all n and k, so (3.87) simplifies to

$$MKE/ZKE \approx 2k^2/(2n+1)\{1 + O[1/(2n+1)^2]\} \quad \text{[Rossby; all } k \,\&\, n\text{]} \tag{3.88}$$

For gravity waves, the approximation $\omega \approx \pm(2n + 1 + k^2)$ gives

$$MKE/ZKE \approx 2(2n+1)^2/\{(2n+1)(2n+1+k^2) \pm 2k(2n+1+k^2)^{1/2}\} \quad [n \le 1] \tag{3.89}$$

The exact dispersion relation for the $n = 0$ mode gives

$$MKE/ZKE = 2 + k^2 \mp 2k(1 + k^2/4)^{1/2} \qquad [n = 0] \tag{3.90}$$

where the negative sign applies to the eastward-traveling gravity wave and the (+) to the mixed Rossby-gravity wave. By considering the limits of these expressions,

we arrive at the following. Longuet-Higgins [16] has already given the long-wave limits.

Theorem 3.3 (Ratio of Meridional Kinetic Energy to Zonal Kinetic Energy)

$$(i) \quad MKE/ZKE \equiv 0 \qquad \textit{[Kelvin wave]}$$

$$(ii) \quad MKE/ZKE \to \begin{cases} 2 & as\ k \to 0 \\ \infty & as\ k \to \infty \end{cases} \qquad \textit{[Mixed Rossby-gravity]}$$

$$(iii) \quad MKE/ZKE \to \begin{cases} 2 & as\ k \to 0 \\ 0 & as\ k \to \infty \end{cases} \qquad [n = 0\ \textit{Eastward-traveling}]$$

$$(iv) \quad MKE/ZKE \to \begin{cases} 0 & as\ k \to 0 \\ \infty & as\ k \to \infty. \end{cases} \qquad \textit{[Rossby; } n \geq 1]$$

$$(v) \quad MKE/ZKE \to \begin{cases} 2 & as\ k \to 0 \\ 0 & as\ k \to \infty \end{cases} \qquad \textit{[Gravity; } n \geq 1]$$

3.12 The Equatorial Beta-Plane as the Thin Limit of the Nonlinear Shallow Water Equations on the Sphere

The *nonlinear* shallow water equations on the equatorial beta-plane are the limit $\varepsilon \to \infty$ of the nonlinear shallow water equations on the sphere. This "equatorial beta-plane" limit corresponds to the dynamics of a global ocean in the limit that the depth (or equivalent depth, for a baroclinic mode) goes to zero. In the opposite limit that $\varepsilon \to 0$, either the velocity potential χ [for gravity and Kelvin waves] or the streamfunction and relative vorticity [for Rossby waves, usually called "Rossby-Haurwitz waves in this limit] are proportional to a single spherical harmonic of latitudinal index n'. Unfortunately, Longuet-Higgins in his magisterial hundred page article on Hough functions (1968) observes that although each normal mode is a single branch $\omega(\varepsilon)$ without bifurcations on the entire range $\varepsilon \in [0, \infty]$, there are complications in cataloging the modes. A common way to distinguish eigenfunctions is to define the mode number to be the number of interior zeros of the eigenfunction, but this changes at some finite ε for some Hough functions. The latitudinal mode numbers n on the equatorial beta-plane [the degree of the Hermite function that is the north-south structure of the latitudinal velocity v] is not generally equal to the

subscript n' of the spherical harmonic that describes a mode in the opposite limit. The Yanai wave is often called the "mixed-gravity wave" because this mode is Rossby-like in part of parameter space and gravity wave-like in the rest of the parameter space as already described earlier. The equatorial Kelvin wave asymptotes to the lowest symmetric gravity wave as $\varepsilon \to 0$. Longuet-Higgins [16] does a thorough job of sorting out these complexities.

An important subtlety omitted from Longuet-Higgins is that the degree of equatorial confinement of a mode, and therefore the accuracy of the equatorial beta plane in describing that particular Hough function, is not controlled by Lamb's parameter ε alone. Rather, if the longitudinal factor is $\exp(is\lambda)$ where λ is the longitude in radians and s is an integer, then latitudinal structure is proportional to $\exp(-\sqrt{\varepsilon + s^2}\,\vartheta^2)$. Thus, it is possible for both Rossby waves [17] and Kelvin waves [18] to be equatorially trapped and therefore well approximated by the equatorial beta plane even when ε is zero, so long as the zonal wavenumber s is sufficiently large.

It has been known for half a century that Rossby Hough functions for finite ε are often well-approximated by simpler transcendentals known as the "prolate spheroidal wave functions" [45–51]. The prolate approximation smoothly merges into the equatorial beta-plane formalism because the Hermite functions are asymptotic approximations to both prolate functions and also to the latitudinal structure of spherical harmonics of large zonal wavenumber s [52–54].

As pointed out by Longuet-Higgins [16], the Kelvin wave on the sphere is always weakly dispersive even in the absence of mean currents. Boyd and Zhou studied the Kelvin solitary waves and cnoidal waves that are a balance between this weak dispersion and the frontogenetic tendencies of weak nonlinearity [55]. Their work is discussed in Chap. 16.

DeLeon, Erlick and Paldor [56] derive various approximations to analyze the equatorial beta-plane/sphere for waves as summarized in Paldor's book [57].[3]

Lindzen's article, though ancient, is still a very good analysis of the relationship of the equatorial beta-plane to the midlatitude beta-plane [58–61].

References

1. Andrews DG, Leovy CB, Holton JR (1987) Fluid Dynamics of the Middle Atmosphere, vol 40. International Geophysics. Academic Press, New York, p 489
2. Gill AE, Clarke AJ (1974) The equatorial current in a homogeneous ocean. Deep Sea Res Oceanogr Abstr 21(5):325–345
3. Boyd JP (2014) Solving Transcendental Equations: The Chebyshev Polynomial Proxy and Other Numerical Rootfinders, Perturbation Series and Oracles. SIAM, Philadelphia
4. Boyd JP (1983) Equatorial solitary waves, part II: envelope solitons. J Phys Oceanogr 13:428–449
5. Tribbia JJ (1979) Nonlinear initialization on an equatorial beta-plane. Mon Weather Rev 107:704–713

[3]The reader is cautioned that Paldor is fond of statements like "the Kelvin wave does not exist on a sphere" which is incorrect.

6. Tribbia JJ (1984b) A simple scheme for high-order nonlinear normal initialization. Mon Weather Rev 112:278–284
7. Kasahara A (1977) Numerical integration of the global barotropic primitive equations with Hough harmonic expansions. J Atmos Sci 34:687–701
8. Lynch P (1992) Richardson's barotropic forecast – a reappraisal. Bull Amer Meteorol Soc 73(1):35–47
9. Platzman GW (1996) The S-1 chronicle: a tribute to Bernard Haurwitz. Bull Amer Meteorol Soc 77(7):1569–1576
10. Wu Z, Moore DW (2004) The completeness of eigenfunctions of the tidal equation on an equatorial beta plane. J Atmos Sci 61(6):769–774
11. Wu Z, Sarachik ES, Battisti D (1999) Thermally driven surface winds on an equatorial beta plane. J Atmos Sci 56(6):2029–2037
12. O'Connor WP (1995) The complex wavenumber eigenvalues of Laplace's tidal equations for ocean bounded by meridians. Proc R Soc Lond A 449:51–64
13. O'Connor WP (1992) On the application of the spheroidal wave equation to the dynamical theory of the long-period zonal tides in a global ocean. Proc R Soc Lond A 439:189–196
14. Sozou C (1997) On azimuthal eigenwavenumbers associated with Laplace's tidal equation. Geophys J Int 130:151–156
15. Boyd JP (1996) Traps and snares in eigenvalue calculations with application to pseudospectral computations of ocean tides in a basin bounded by meridians. J Comput Phys 126:11–20
16. Longuet-Higgins MS (1968) The eigenfunctions of Laplace's tidal equations over a sphere. Philos Trans R Soc Lond A 262:511–607
17. Boyd JP (1985a) Barotropic equatorial waves: the non-uniformity of the equatorial beta-plane. J Atmos Sci 42:1965–1967
18. Boyd JP, Zhou C (2008b) Uniform asymptotics for the linear Kelvin wave in spherical geometry. J Atmos Sci 65(2):655–660
19. Haurwitz B (1940) The motion of atmospheric disturbances on the spherical earth. J Mar Res 3(3):254–267
20. Kasahara A, Puri K (1981) Spectral representation of three-dimensional global data by expansion in normal modes. Mon Weather Rev 109:37–51
21. Kasahara A (1978) Further studies on a spectral model of the global barotropic primitive equations with Hough harmonic expansions. J Atmos Sci 35:2043–2051
22. Callaghan P, Fusco A, Francis G, Salby M (1999) A Hough spectral model for three-dimensional studies of the middle atmosphere. J Atmos Sci 56:1461–1480
23. Dee DP, Moraes Da Silva A (1986) Using Hough harmonics to validate and assess nonlinear shallow-water models. Mon Weather Rev 114:2191–2196
24. Wang H, Boyd JP, Akmaev RA (2016) On computation of Hough functions. Geosci Model Dev 9:1477–1488
25. Zagar N, Boyd JP, Kasahara A, Tribbia J, Kallen E, Tanaka H, Yano J (2016) Normal modes of atmospheric variability in observations, numerical weather prediction and climate models. Bull Amer Meteorol Soc 97(6):ES 125–ES 128
26. Anderson DLT (1973) A low latitude spectral model using Chebyshev-parabolic cylinder functions. Report 7, GARP Programme on Numerical Experimentation W. G. N. E., World Meteorological Organization, Geneva
27. Holvorcem PR, Vianna ML (1992) Integral-equation approach to tropical ocean dynamics. 2. Rossby-wave scattering from the equatorial Atlantic western boundary. J Mar Res 50(1):33–61
28. Vianna ML, Holvorcem PR (1992) Integral-equation approach to tropical ocean dynamics. 1. theory and computational methods. J Mar Res 50(1):1–31
29. Smith NR (1988) A truncated oceanic spectral model for equatorial thermodynamic studies. Dyn Atmos Ocean 12:313–337
30. Majda AJ, Khoulder B (2001) A numerical strategy for efficient modeling of the equatorial wave guide. Proc Nat Acad Sci 98(4):1341–1346
31. Dyson SFW, Hough SS (1923) The Observatory 46(592):269–272

32. Baer F (1977) Adjustment of initial conditions required to suppress gravity oscillations in non-linear flows. Contrib Atmos Phys 50:350–366
33. Baer F, Tribbia JJ (1977) On complete filtering of gravity modes through nonlinear initialization. Mon Weather Rev 105:1536–1539
34. Boyd JP (1998) Weakly Nonlocal Solitary Waves and Beyond- All-Orders Asymptotics: Generalized Solitons and Hyperasymptotic Perturbation Theory, Mathematics and Its Applications, vol 442. Kluwer, Amsterdam
35. Boyd JP (1995) Eight definitions of the slow manifold: Seiches, pseudoseiches and exponential smallness. Dyn Atmos Ocean 22:49–75
36. Boyd JP (1994b) The slow manifold of a five mode model. J Atmos Sci 51:1057–1064
37. Lorenz EN (1992) The slow manifold – what is it? J Atmos Sci 49:2449–2451
38. Lovegrove AF, Read PL, Richards CJ (2000) On the nonexistence of a slow manifold. Q J R Meteorol Soc, Part B 126(570):3233–3254
39. Lorenz EN (1986) On the existence of a slow manifold. J Atmos Sci 43:1547–1557
40. Vanneste J (2008) Asymptotics of a slow manifold. SIAM J Appl Dyn Syst 7(4):1163–1190
41. Lorenz EN, Krishnamurthy V (1987) On the nonexistence of a slow manifold. J Atmos Sci 44:2940–2950
42. Bokhove O, Shepherd TG (1996) On Hamiltonian balanced dynamics and the slowest invariant manifold. J Atmos Sci 53:276–297
43. Camassa R, Tin SK (1996) The global geometry of the slow manifold in the Lorenz-Krishnamurthy model. J Atmos Sci 53:3251–3264
44. Fowler AC, Kember G (1996) On the Lorenz-Krishnamurthy slow manifold. J Atmos Sci 53:1433–1437
45. Dikii LA (1967) On asymptotic behavior of the solutions of the Laplace tidal equation. Sov Phys Dokl 11(9):772–775
46. Dickinson RE (1968a) On exact and approximate linear theory of vertically propagating planetary Rossby waves forced at a spherical lower boundary. Mon Weather Rev 96(7):405–415
47. Chunchuzov YP (1969) The asymptote for eigenfunctions of Laplace's tidal operator. Atmos Ocean Phys 236–238
48. Müller D, Kelly B, O'Brien JJ (1994) Spheroidal eigenfunctions of the tidal equation. Phys Rev Lett 73:1557–1560
49. Müller D, O'Brien JJ (1995) Shallow water waves on the rotating sphere. Phys Rev E 51(2):4418–4431
50. Müller D (1997) The tidal problem for simple circulations. Int J Modern Phys B 11(3):223–254
51. Paldor N, Sigalov A (2011) An invariant theory of the linearized shallow water equations with rotation and its application to a sphere and a plane. Dyn Atmos Ocean 51(1–2):26–44
52. Olver FWJ, Lozier DW, Boisvert RF, Clark CW (eds) (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, New York
53. Boyd JP (2001) Chebyshev and Fourier Spectral Methods. Dover, New York, p 680
54. Huang Z, Xiao J, Boyd JP (2015) Adaptive radial basis function and Hermite function pseudospectral methods for computing eigenvalues of the prolate spheroidal wave equation for very large bandwidth parameter. J Comput Phys 281:269–284
55. Boyd JP, Zhou C (2008) Kelvin waves in the nonlinear shallow water equations on the sphere: nonlinear traveling waves and the corner wave bifurcation. J Fluid Mech 617:185–205
56. De-Leon Y, Erlick C, Paldor N (2010) The eigenvalue equations of equatorial waves on a sphere. Tellus A 62(1):62–70
57. Paldor N (2015) Shallow Water Waves on the Rotating Earth. Sringer Briefs in Earth System Sciences. Springer, New York
58. Hough SS (1898) On the application of harmonic analysis to the dynamic theory of the tides, II. on the general integration of Laplace's tidal equations. Philos Trans R Soc A 191:139–185
59. Hough SS (1897) On the application of harmonic analysis to the dynamic theory of the tides, I. on Laplace's "oscillations of the first species" and on the dynamics of ocean currents. Philos Trans R Soc A 189:201–257
60. Lindzen R (1967) Planetary waves on beta-planes. Mon Weather Rev 95:441–451
61. Chunchuzov YP (1971) Dynamics of atmospheric wave motions in equatorial region. Izvestiya Akademii Nauk SSSR Fizika Atmosfery i Okeana 7(1):69

Chapter 4
The "Long Wave" Approximation & Geostrophy

Abstract The quasigeostrophic approximation, so useful in the middle latitudes, is notoriously inaccurate near the equator. This chapter describes a replacement that has been extremely useful in low latitude dynamics. The "long wave" approximation, which might equally well be dubbed the "low frequency" or "meridional geostrophy" approximation, filters gravity waves while retaining Kelvin and Rossby waves. The key assumption is that there is geostrophic balance in the latitudinal momentum equation only. This chapter also discusses the separation between the high and low frequency waves and the usefulness of this separation for initialization of equatorial numerical models.

> Scientia is knowledge. It is only in the popular mind that it is equated with facts. This is, of course, flattering since facts are incontrovertible. But it is also demeaning, since facts are meaningless. They contain no narrative. Science, by contrast, is story-telling. That is evident in the way we use our primary scientific instrument, the eye. The eye searches for shapes. It searches for a beginning, a middle, and an end.
>
> – John Polanyi in "Science, scientists and society", *Queens Quarterly, 107*, 31–36 (2000).

4.1 Introduction

Quasi-geostrophy has been so useful in understanding the dynamics of the middle latitudes that there is an almost irresistible temptation to extend it to low latitudes, too. In this chapter, we explain why a simple-minded extension of quasi-geostrophy does not work.

4.2 Quasi-Geostrophy

Sufficiently far from the equator and for sufficiently small amplitude, it must be possible to derive the quasi-geostrophic streamfunction equation from the shallow water equations on the equatorial beta-plane. But what is the relationship? In this section, we shall answer that question for infinitesimal amplitude waves.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_4

The midlatitude, linearized streamfunction equation for a traveling wave of phase speed C and zonal wavenumber K is

$$\psi_{YY} + \left\{ \frac{1}{C} - K^2 - 1 \right\} \psi = 0 \tag{4.1}$$

with the standard *midlatitude* nondimensionalization. (The standard midlatitude length scale, the so-called "Rossby radius of deformation", is nondimensionally $O(1)$.) Capital letters are used for the phase speed, zonal wavenumber and latitude because the standard scalings for the equatorial region need not be the same.

For a sinusoidal wave (longitude), $v = ik\psi$, so the equation for ψ must match the equatorial equation for the north-south velocity:

$$v_{yy} + \left\{ \frac{1}{c} - k^2 - y^2 \right\} v = 0 \tag{4.2}$$

Note that β has been nondimensionalized to one in both cases; prior to the nondimensionalization, $1/c$ was the dimensional term β/c. We want the equatorial equation to be as close as possible to its midlatitude, quasi-geostrophic counterpart after rescaling.

Let y_0 denote a reference latitude where, in the equatorial nondimensionalization, $|y_0| \gg 1$. Define

$$y = y_0 + \frac{1}{y_0} Y, \qquad x \equiv \frac{1}{y_0} X \tag{4.3}$$

The equatorial equation becomes

$$v_{YY} + \left\{ \frac{1}{cy_0^2} - \frac{k^2}{y_0^2} - (1 - \frac{2}{y_0^2} Y + \frac{Y^2}{y_0^4}) \right\} v = 0 \tag{4.4}$$

If we assume that the latitudinal scale of the midlatitude disturbance is small compared to y_0 — the usual *midlatitude* beta-plane approximation — then the Y-varying terms in the coefficient of v can be neglected. The equatorial equation for v then becomes identical in form with its midlatitude counterpart if

$$C \equiv cy_0^2, \qquad K \equiv \frac{k}{y_0} \tag{4.5}$$

In the middle latitudes, a phase speed c which is small compared to unity in equatorial scaling, as would be associated with a high order Rossby mode, corresponds to roughly unit phase speed C in terms of the midlatitude coordinates. Thus, $C \gg c$ as expected. Note that fluid velocities will scale in the same way as phase velocities. Thus

$$\bar{U}(Y) = y_0^2 \, \bar{u}(y) \tag{4.6}$$

where $\bar{u}$ is the zonally-averaged current in the equatorial coordinates.

Unit length in Y, the midlatitude coordinate, corresponds to a small change in the equatorial coordinate y. This is in agreement with the principle that the Rossby radius of deformation becomes smaller and smaller as the latitude increases. Indeed, the relationship $(y - y_0) = Y/y_0$ together with the $O(1)$ length scale in Y can be taken as a statement that $1/y_0$ is, in the equatorial nondimensionalization, the Rossby radius at $y = y_0$.

With these scalings, we can, as it were, translate midlatitude solutions — even nonlinear ones — to the standard equatorial beta-plane to examine, either numerically or perturbatively, how propinquity to the equator alters them. However, surprisingly little has been done except through the explicit computation of the linear waves as described in the preceding chapter and through comparisons between equatorial solitons and midlatitude modons as described in Chap. 16.

4.3 The "Meridional Geostrophy", "Low Frequency" or "Long Wave" Approximation

Although a *naive* application of quasi-geostrophy must fail, there is a very useful approximation which has many of the same virtues. In particular, this approximation filters out gravity waves and allows one to numerically integrate the equations of motion with a long time step. More important, all the basic formulas and relationships of equatorial dynamics are greatly simplified, and many important problems, such as the reflection from the eastern boundary of an ocean, can be solved in closed form with — but only with — this approximation.

The requirement for the validity of the approximation is that either (i) the zonal scale is very long or (ii) the frequency is very low. (For Rossby waves of sufficiently large zonal scale to have a westward group velocity, the validity of either of these conditions automatically implies that the other is true, too.) For this reason, the "long wave" approximation can be legitimately dubbed the "low frequency" approximation as well. The key is "meridional geostrophy", which therefore is a third valid name for the approximation. Low frequency phenomenon include the annual cycle in the ocean as well as the waves such as those observed by Lukas and Firing [1].

This simplification was apparently first employed in a (non-equatorial) paper by Pedlosky [2] and introduced into equatorial oceanography by Cane and Sarachik [3].

Definition 4.1 (*Meridional Geostrophy/Long Wave Approximation*) This approximation consists of dropping the local time derivative of the north-south velocity in the latitudinal momentum equation.

Theorem 4.1 (Long Wave/Meridional Geostrophy Approximation) *The neglect of the local time derivative of v in the north-south momentum equation has the following consequences for flow that has been linearized about a state of rest:*

1. *The Kelvin wave is unaffected.*

2. *The Yanai or mixed Rossby-gravity wave is filtered out in both its gravity-like and Rossby-like regimes.*
3. *Gravity waves are filtered out from the approximate shallow water equations.*
4. *The Rossby wave dispersion relation is simplified:*

$$\omega_{Rossby} = -k/(2n+1+k^2) \Rightarrow \omega_{long} \equiv -k/(2n+1) \qquad \textit{[long Rossby]} \tag{4.7}$$

with a relative error of

$$\omega_{Rossby} = \omega_{long}\left\{1 + O(k^2/[2n+1])\right\} \tag{4.8}$$

5. *The magnitude and structures of u, v and ϕ are not altered except through the $O(k^2)$ changes in the frequency ω.*

Proof The Kelvin wave has a zero north-south velocity, so obviously an approximation that modifies only a term in v is irrelevant to this mode.

The Yanai wave is described by the coupled pair of equations

$$S_{1t} + S_{1x} - 2^{1/2} v_0 = 0 \qquad \text{["Yanai doublet"]} \tag{4.9a}$$

$$\boxed{v_{0t}} + 2^{-1/2} S_1 = 0 \tag{4.9b}$$

where the term in the box is neglected in the meridional geostrophy approximation, and $S_1(x,t)$ and $v_0(x,t)$ are the coefficients of $\psi_1(y)$ in the sum variable S and of $\psi_0(y)$ in the north-south velocity v for this mode. However, this immediately implies $S_1(x,t) \equiv 0$, and substituting this in the first equation demands that $v_0 \equiv 0$, too. (This also shows that the mixed Rossby-gravity wave is *never* latitudinally geostrophic even in its Rossby-like regime.)

To demonstrate that (i) all gravity waves with latitudinal mode $n \geq 1$ are filtered and (ii) confirm the modifications to the Rossby wave, it is necessary to look at the "general triplet" (3.13) which arises from Hermite expansion of the shallow water equations in their sum-difference form:

$$S_{n+1,t} + S_{n+1,x} - [2(n+1)]^{1/2} v_n = 0 \tag{4.10a}$$

$$\boxed{v_{n,t}} + [(n+1)/2]^{1/2} S_{n+1} - [n/2]^{1/2} D_{n-1} = 0 \tag{4.10b}$$

$$D_{n-1,t} - D_{n-1,x} + [2n]^{1/2} v_n = 0 \tag{4.10c}$$

The boxed term is the only term which is dropped in the long wave approximation. One might suppose that dropping only one of three time derivatives would reduce the dispersion relation from a cubic (i.e., number of branches of dispersion relation

equal to the number of time derivatives) to a quadratic. However, the (simplified) latitudinal momentum equation shows that

$$S_{n+1} = \sqrt{n/(n+1)}\, D_{n-1} \tag{4.11}$$

for all x, t. Thus, there is only one *independent* time derivative after meridional geostrophy is enforced.

To confirm this explicitly, assume that $S_{n+1}(x,t)$ depends on x, t as $\tilde{S}_{n+1} \exp(ikx - i\omega t)$ where $\tilde{S}_{n+1}$ is a constant, and similarly for the other unknowns. The triplet becomes the matrix problem

$$\begin{vmatrix} i(k-\omega) & -\sqrt{2(n+1)} & 0 \\ \sqrt{(n+1)/2} & \boxed{0} & -\sqrt{n/2} \\ 0 & \sqrt{2n}^{1/2} & -i(k+\omega) \end{vmatrix} \begin{vmatrix} \tilde{S}_{n+1} \\ \tilde{v}_n \\ \tilde{D}_{n-1} \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \\ 0 \end{vmatrix} \tag{4.12}$$

This homogeneous matrix equation has a solution if and only if its determinant is zero. However, the zero in the box, which would otherwise be $-i\omega$, eliminates the ω^3 term from the determinant of the shallow water triplet. The long wave dispersion relation has only the single root:

$$\omega_{longwave} \equiv \frac{-k}{2n+1} \qquad \leftrightarrow \qquad c_{longwave} = \frac{-1}{2n+1}, \qquad n = 1, 2, \ldots \tag{4.13}$$

The deletion of the two roots of the dispersion relation that correspond to gravity waves implies that the long wave approximation has filtered these modes from the shallow water equations.

The error estimate in the long wave Rossby approximation follows from comparing it with the highly accurate approximation (3.36) $\omega_{Rossby} \approx -k/(2n+1+k^2)$: the long wave approximation is the result of expanding the denominator of ω_{Rossby} as a power series in powers of k.

The structural formulas are *unchanged* by the long wave approximation: it is still true that

$$\phi = \frac{1}{(1-c)}[(n+1)/2]^{1/2}\psi_{n+1} + \frac{1}{(1+c)}[n/2]^{1/2}\psi_{n-1} \tag{4.14}$$

$$u = \frac{1}{(1-c)}[(n+1)/2]^{1/2}\psi_{n+1} - \frac{1}{(1+c)}[n/2]^{1/2}\psi_{n-1} \tag{4.15}$$

$$v = ik\psi_n, \tag{4.16}$$

exactly as in Table 3.3. The reason is that the first and third equations of the triplet are unaltered by the long wave approximation, and contain only v_n and *either* [but not both] of the sum and difference variables. This implies that the relationship between v_n and S_{n+1}, as expressed by the first equation, is unaltered. Similarly, the

third equation gives precisely the same proportionality between v_n and D_{n-1} with or without the assumption of latitudinal geostrophy.

However, because the phase speed for each long wave mode is independent of k, it is possible to rewrite the modes alternatively as

$$S_0(x, y, t) = A_0(x - t)\,\psi_0(y) \tag{4.17}$$

and for the Rossby modes with mode number $n \geq 1$:

$$v_n(x - ct) \equiv \frac{dA_n}{dx}(x - ct) \tag{4.18}$$

$$S_{n+1} = \frac{\sqrt{2\,(n+1)}}{1-c}\,A_n(x - ct)\,\psi_{n+1}(y) \tag{4.19}$$

$$D_{n-1} = \sqrt{\frac{2}{n}}\,\frac{n+1}{(1-c)}\,A_n(x - ct)\,\psi_{n+1}(y) \tag{4.20}$$

where A_n is an arbitrary function of a single argument.

However, these arguments that demonstrate the mathematical soundness and consequences of the long wave approximation do not explain: Why does it work?

Equation (4.14) shows that

$$v \sim O(k/n^{1/2}) \begin{Bmatrix} u \\ \phi \end{Bmatrix} \tag{4.21}$$

Since the frequency $\omega \sim O(k/[2n+1])$ and the y-derivative of ϕ is $O(n^{1/2}\phi)$, as can be shown from the identities for the derivatives of the Hermite functions, it follows that

$$-i\omega v \sim O(k^2/n^{1/2}) \begin{Bmatrix} yu \\ \phi_y \end{Bmatrix} \tag{4.22}$$

Thus, the long wave approximation is successful for two reasons. First, the low frequency makes the time derivative in the latitudinal momentum equation small by one factor of k, because $\omega \sim O(k)$. Second, v is itself small compared to the other fields for Rossby waves, which contributes another factor of $O(k)$ to the smallness of the term which is neglected in the long wave approximation.

Thus, the long wave approximation is highly accurate for Rossby waves when $k \ll 1$. The error is always small even for the lowest ($n = 1$) Rossby mode for small k, and becomes smaller still as the mode number increases.

The meridional geostrophy approximation has been justified with such care because of its extreme usefulness. Boyd and Christidis [4–6] have exploited it with great success in interpreting complex mode-mixing in the theory of equatorial instabilities. It is even more useful in obtaining *analytical* solutions for (i) the quasi-free modes of the equatorial ocean and (ii) for the oceanic response to periodic wind stress.

4.4 Boundary Conditions

Reflection from a western boundary would seem to immediately torpedo the long wave approximation because meridional geostrophy filters out the short gravity waves which are part of the reflection of a long Rossby wave from a western boundary. The long wave dispersion relation is,

$$c = -1/(2n+1) \tag{4.23}$$

i.e., c is independent of k so that (4.23) is really a non-dispersion relation. For nondispersive waves, the group and phase velocities are equal so that the long wave approximation forces all Rossby waves to move westward.

Pedlosky (1965) first showed the way out of this conundrum: the short Rossby waves that form part of the reflection from a western boundary have very low group velocities when the wavenumber of the incident Rossby wave is small [2, 7, 8]. To be precise, the full dispersion for equatorial waves in terms of k is

$$k = -\frac{1}{2\omega} \pm \sqrt{\omega^2 + \frac{1}{4\omega^2} - (2n+1)} \tag{4.24}$$

Taking the (+) in (4.24) and carefully expanding the radical for $\omega \ll 1$ shows that this root is $O(\omega)$ and physically corresponds to the incident long Rossby wave packet. The other root is approximately

$$k \approx -1/\omega + O(1) \tag{4.25}$$

(assuming that the mode number n is not too large). By differentiating the dispersion relation with respect to k and solving for $c_{gx} = \partial\omega/\partial k$, the group velocity for Rossby waves can be written in terms of ω, k, and n as

$$c_{gx} = [1 + 2k\omega]/[3\omega^2 - (k^2 + 2n + 1)] \tag{4.26}$$

Eq. (4.25) allows us to substitute $k \to -1/\omega$ which transforms the group velocity into

$$c_{gx} \sim \omega^2 \ (1 + O(\omega)\,) \tag{4.27}$$

For waves of annual period, this is an astonishingly slow rate of travel. For the first baroclinic mode, $\omega^2 \approx 0.00068$; c_{gx} has a dimensional speed of about $(1/7)$ cm/s. Put another way, it would take the reflected short waves about 250 years to cross the width of the Pacific!

In practice, these slowly-traveling waves are destroyed by dissipation before traveling very far away from the coast. For very low frequency incident wave packets, it therefore follows that the short reflected waves give only a *boundary layer* response that is effectively *trapped* near the western side of the ocean. If we are prepared to

forgo a knowledge of the detailed dynamics of this boundary layer, then the long wave approximation does give useful results in the rest of the ocean for $k \ll 1$ or $\omega \ll 1$.

Without the short reflected Rossby waves, however, we still cannot hope to satisfy the true boundary condition $u = 0$. Cane and Sarachik showed that in the long wave approximation, the correct condition to impose at the western boundary is

$$\boxed{\int_{-\infty}^{\infty} u\, dy = 0 \qquad \text{at western boundary of ocean}} \tag{4.28}$$

The justification of (4.28) begins with the observation that when $k \sim O(-1/\omega)$, as is true for the short, reflected Rossby waves, the explicit formulas for u, v, and ϕ) show that

$$v \sim O(-1/\omega) \begin{Bmatrix} u \\ \phi \end{Bmatrix} \tag{4.29}$$

i.e., v is very *large* in comparison to u and ϕ for the short, reflected Rossby waves by exactly the same factor by which it is very *small* in comparison to u and ϕ for the incident *long* Rossby waves. In the equation of continuity, this implies that the $(-i\omega\,\phi)$ term is small in comparison to the other two terms by $O(\omega^2)$:

$$iku + v_y = 0 + O(\omega^2); \qquad k \gg 1; \qquad \omega \ll 1 \tag{4.30}$$

If we integrate (4.30) from $-\infty$ to ∞ in y, then the derivative of v gives only the boundary terms, $[v(\infty) - v(-\infty)] = 0$, implying (4.28), that is

$$\int_{-\infty}^{\infty} u dy = 0, \quad \text{short, reflected Rossby waves} \tag{4.31}$$

This equation is true for all x for the reflected Rossby wave and not merely at the boundary, is (4.28). Since this integral vanishes for each of the short reflected waves individually, it follows that the only way that the integral of the *total* u can be 0 is if

$$\int_{-\infty}^{\infty} (u_{Kelvin} + u_i)\, dy = 0, \quad \text{at the western boundary} \tag{4.32}$$

i.e., the sum of the latitudinal integral of the incident Rossby wave and the reflected Kelvin wave must be zero.

This boundary condition (4.32) is useful mathematically because it makes it possible to treat the long Rossby-cum-Kelvin waves that survive the long wave approximation as a closed set; one can calculate the reflected Kelvin wave from (4.32) without the bother of having to calculate the coefficients of all those irrelevant short Rossby waves through the recurrence formula of Sect. 3. It is useful physically because it shows that short Rossby waves can only serve to redistribute mass in the zonal boundary layer. One could, for example, have a westward flow in the northern hemisphere

and an eastward flow in the southern hemisphere created by these reflected Rossby waves, but the latitudinal average of such zonal flows in the western boundary layer must vanish.

It is in these latter contexts that we can alternatively think of the long wave approximation as a low frequency approximation. For a forcing of annual period, for example, $\omega = 0.026$, and the relative error in the dispersion relation $\omega_{longwave}$ is

$$k^2/[2n+1] \sim O([2n+1]\omega^2) \tag{4.33}$$

which implies that the approximation breaks down only for $n \sim O(700)$, i.e., only for modes whose turning points are well into the middle latitudes so that they cannot properly be considered equatorial waves anyway.

4.5 Frequency Separation of Slow [Rossby/Kelvin] and Fast [Gravity] Waves

Implicit in the arguments in the previous section is that gravity waves are unimportant: otherwise, an approximation that completely neglects gravitywaves would obviously be unacceptable. But why are gravity waves unimportant?

A partial answer is that in the tropical ocean, as in the midlatitude atmosphere and ocean, there is a huge frequency gap between the slow modes (Kelvin and Rossby) and the fast modes (gravity waves) as already discussed for general k in Sect. 3.5. The Kelvin and Rossby modes have frequencies which are $O(k)$ as already seen in the previous section. For gravity waves, including the Yanai mode, the frequency tends to a fixed limit $|\omega_{gravity}| = (2n+1)^{1/2}$ as $k \to 0$. This implies that the frequency separation between the high frequency gravity waves and low frequency Kelvin and Rossby waves, which is always large even for moderate k, becomes enormous for small k.

Just as it does in the middle latitudes, this frequency separation suggests that the high frequency gravity wave motions are just *noise*. If the ocean is perturbed by a small k change in the wind stress — in other words, by a change in the winds with a large zonal scale — the gravity waves will rapidly propagate away from the region of the perturbation and adjust almost instantaneously tothe perturbation in comparison to the much, much slower response of the excited Rossby and Kelvin waves. As in the middle latitudes, the low frequency modes are the only ones that count, and one would like to find an approximation that *filters* out the gravity waves— and yet still accurately reproduces the Kelvin and Rossby waves even at the equator. The long wave approximation does this.

4.6 Initial Value Problems in an Unbounded Ocean, Linearized About a State of Rest, in the Long Wave Approximation

Let $S_{init}(x, y) = u_{init}(x, y, 0) + \phi_{init}(x, y, 0)$ denote the initial condition as expressed in terms of the sum variable. Expand

$$S_{init}(x, y) = \sum_{n=-1}^{\infty} \mathfrak{S}_{n+1}(x)\, \psi_{n+1}(y) \tag{4.34}$$

$$\mathfrak{S}_{n+1}(x) = \int_{-\infty}^{\infty} dy\, S_{init}(x, y) \psi_{n+1}(y) \tag{4.35}$$

The linearized-about-rest long wave solution at a later time t is

$$S(x, y, t) = \mathfrak{S}_0(x - t)\psi_0(y) + \sum_{n=1}^{\infty} \mathfrak{S}_{n+1}(x + [1/(2n+1)]\, t)\, \psi_{n+1}(y) \tag{4.36}$$

Note that there is no $n = 0$ longwave mode, and thus the coefficient of $\psi_1(y)$ in the initial height field must be dropped. (For the full shallow water approximation, this term would generate a very fast-moving Yanai transient.)

The north-south velocity and difference field cannot be specified independently: as soon as one field is specified, the other two are completely determined. (Independent specification of (S, v, D) will excite high-frequency transients which cannot be described by the meridional geostrophy approximation.) In particular, recalling that $c(n) = -1/(2n + 1)$,

$$v_n(x, t) = \frac{(1 + 1/(2n+1))}{\sqrt{2\,(n+1)}} \frac{d\mathfrak{S}_{n+1}(x + [1/(2n+1)]\, t)}{dx} \tag{4.37}$$

$$D_{n-1}(x, t) = \sqrt{\frac{n+1}{n}}\, S_{n+1}(x, t) \tag{4.38}$$

to conclude that

$$v(x, y, t) = \sum_{n=1}^{\infty} \frac{(1 + 1/(2n+1))}{\sqrt{2\,(n+1)}} \frac{d\mathfrak{S}_{n+1}(x + [1/(2n+1)]\, t)}{dx}\, \psi_n(y) \tag{4.39}$$

$$D(x, y, t) = \sum_{n=1}^{\infty} \sqrt{\frac{n+1}{n}}\, \mathfrak{S}_{n+1}(x + [1/(2n+1)]\, t)\, \psi_{n-1}(y) \tag{4.40}$$

Figure 4.1 illustrates dispersion in the long wave approximation.

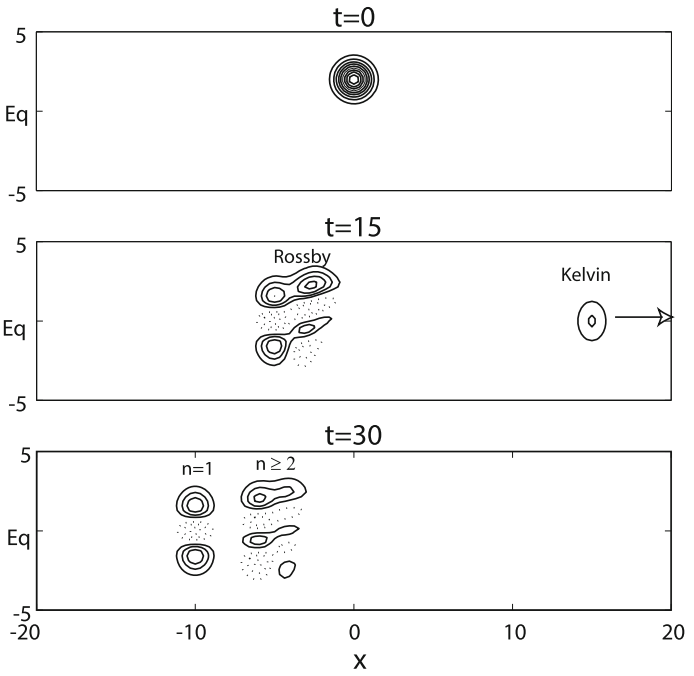

Fig. 4.1 Dispersion of an equatorial pulse in the linearized long wave approximation as seen through contour plots of the "sum" variable $S \equiv \phi + u$; negative values are dashed. The contour interval is 1/10. $S(x, y, 0) = \exp(-x^2)\exp(-y^2)$. The Kelvin wave moves eastward ($c_{Kelvin} = 1$) so rapidly that it is beyond the *right* edge of the graph in the lower panel. The Rossby modes gradually segregate by mode number; in the lower panel, the $n = 1$ mode, which has the fastest nondimensional speed, $c = -1/3$, has separated from the slower and higher modes dispersing to its east. There is a rather strong $n = 0$ pulse which is filtered by the long wave approximation, and therefore not shown. Note that mean currents and all forms of nonlinearity are neglected

4.7 Reflection from an Eastern Boundary in the Long Wave Approximation

4.7.1 The Method of Images

Every function can be decomposed into its symmetric and antisymmetric parts:

$$f(x) = f^S + f^A; \tag{4.41}$$

$$f^S(x) \equiv \frac{1}{2}\left(f(x) + f(-x)\right) \;\&\; f^A(x) \equiv \frac{1}{2}\left(f(x) - f(-x)\right) \tag{4.42}$$

where by construction $f^S(x)$ is symmetric about the origin, that is, $f^S(-x) = f^S(x)$ for all x and similarly f^A is antisymmetric in the sense that $f^A(-x) = -f^A(x)$

for all x. Note that all antisymmetric functions are always zero at the origin since $f^A(0) = -f^A(0)$, and the only number which is its own inverse is zero.

Lord Kelvin realized that such symmetry, and the inflation of a semi-infinite interval to the whole real axis could furnish easy solutions to boundary value problems. His technique is "The Method of Images".

For example, suppose that the problem on the semi-infinite interval is

$$u_{tt} = u_{xx}, \qquad u(x,0) = Q(x), u_t(x,0) = -Q_x;\ u(0,t) = 0, x \in [-\infty, 0] \otimes t \in [0, \infty]$$

If the homogeneous Dirichlet boundary condition at $x = 0$ is ignored, the solution on the real axis is the rightward-moving pulse

$$u(x,t) = Q(x-t) \tag{4.43}$$

If the domain was extended to the right half-interval and $u(x,t)$ was antisymmetrically continued to positive x as the "image" $-Q(-x-t)$, this image pulse is automatically leftward-moving. Furthermore, the sum of leftward and rightward moving pulses is antisymmetric for all time and therefore zero at $x = 0$ for all time. However, if both pulses initially extend individually over the whole real axis, the left-of-the-origin continuation of the right image pulse will mess up the initial conditions. It is therefore necessary to employ the Heaviside step function

$$H(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{4.44}$$

Then the solution on the infinite interval,

$$u(x,t) = Q(x-t)H(-x+t) - Q(-x-t)\,H(x+t) \tag{4.45}$$

satisfies the initial conditions for negative x and is always antisymmetric about the origin and therefore always zero at $x = 0$ (Fig. 4.2).

4.7.2 Dilated Images

"Dilation" is the symmetry operation of scaling a function by a factor ς as in $f(x) \to f(\varsigma x)$. It is useful when left-moving and right-moving waves travel at different speeds. For example, suppose the general wave solution is

$$u(x,0) = Q(x-t) + R(x+t/3) \tag{4.46}$$

How do we create a solution that is antisymmetric about the origin for all time? The answer is to apply dilation as well as reflection and write (Fig. 4.3)

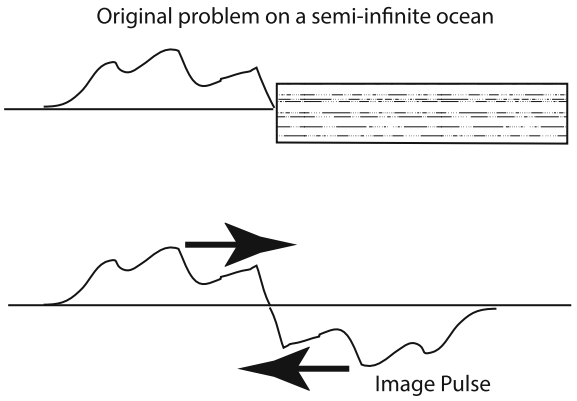

Fig. 4.2 *Top* the original problem on $x \in [-\infty, 0]$. The continent to the *right* of the origin is shaded. *Bottom* the equivalent problem on an infinite interval. For all t, the rightward-moving pulse and its image sum to a function which is antisymmetric about $x = 0$ for all t and therefore $u(0, t) =$ for all t

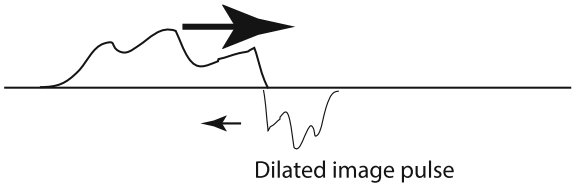

Fig. 4.3 Same as the *bottom* panel of the previous graph except that the image pulse has been dilated by a factor of three [narrowed]. The superposition of leftward and rightward moving pulses still sums to zero at $x = 0$ for all time even though the leftward-moving wave is traveling only one-third as fast as the rightward-moving pulse

$$u(x,t) = Q(x-t)H(-x+t) - Q(3[x+t/3])H(3[x+t/3]) \tag{4.47}$$

4.7.3 Zonal Velocity

The east-west velocity u is

$$u(x,t) = \frac{1}{2}(S-D) \tag{4.48}$$

$$= \frac{1}{2}\sum_{n=1}^{\infty}\left\{\mathfrak{S}_{n-1}(x,t) - \sqrt{\frac{n+1}{n}}\,\mathfrak{S}_{n+1}(x,t)\right\}\psi_{n-1}(y) \tag{4.49}$$

It follows that in the long wave approximation, satisfying the eastern boundary condition at $x = 0$ requires that

$$\mathfrak{S}_{n+2}(0,t) = \sqrt{\frac{n+1}{n+2}}\mathfrak{S}_n(0,t) \qquad n = 0, 1, 2, \ldots \tag{4.50}$$

Thus, if the Kelvin wave has the initial condition

$$S_0(x,0) = Q(x) \qquad \Rightarrow \qquad \mathfrak{S}_2(x,0) = \sqrt{\frac{1}{2}}\, Q\,(3x) \tag{4.51}$$

$$\mathfrak{S}_2(x,t) = \sqrt{\frac{1}{2}}\, Q\,(3[x + t/3]) \tag{4.52}$$

$$\mathfrak{S}_4(0,t) = \sqrt{\frac{3}{4}}\mathfrak{S}_2(0,t) \tag{4.53}$$

$$= \sqrt{\frac{3}{4}}\sqrt{\frac{1}{2}}\, Q\,(t) \tag{4.54}$$

which requires that

$$\mathfrak{S}_4(x,0) = \sqrt{\frac{3}{4}}\sqrt{\frac{1}{2}}\, Q\,(7x) \tag{4.55}$$

$$\mathfrak{S}_4(x,t) = \sqrt{\frac{3}{4}}\sqrt{\frac{1}{2}}\, Q\,(7[x + t/7]) \tag{4.56}$$

The general solution is then, recalling Eq. 4.37, the relationships expressing v and D in terms of the sum variable S,

$$\mathfrak{S}_0 = Q(x - t) \tag{4.57}$$

$$\mathfrak{S}_{2n}(x,t) = \left\{\prod_{k=1}^{n}\sqrt{\frac{(2k-1)}{2k}}\right\} Q\left((4n-1)\left\{x + [1/(4n-1)]\, t\right\}\right) \tag{4.58}$$

$$\begin{aligned} S(x,y,t) &= Q(x-t)\psi_0(y) + \sum_{n=1}^{\infty} \mathfrak{S}_{2n}(x,t)\, \psi_{2n}(y) \\ v(x,y,t) &= \sum_{n=1}^{\infty} \frac{(1 + 1/(2n+1))}{\sqrt{2\,(n+1)}} \frac{d\mathfrak{S}_{n+1}(x + [1/(2n+1)]\, t)}{dx}\, \psi_{2n-1}(y) \\ D(x,y,t) &= \sum_{n=1}^{\infty} \sqrt{\frac{n+1}{n}}\, \mathfrak{S}_{2n}(x,t)\, \psi_{2n-2}(y) \end{aligned} \tag{4.59}$$

This solution will be discussed at much greater length in Sects. 8.2 and 6.4.

4.8 Forced Problems in the Long Wave Approximation

Forced motion will be discussed at much greater length in later chapters. Here, we show that such calculations are easy, at least in the long wave approximation, and also obtain the forced long wave equations that will be needed in the next section in this chapter and in the later discussion of solitary waves.

In sum and difference form, the equatorial beta-plane equations in the long wave approximation can be written

$$S_t + S_x + \Re v = (F + \tilde{H}) \tag{4.60a}$$

$$(1/2)[\mathfrak{L}S + \Re D] = G \tag{4.60b}$$

$$D_t - D_x + \mathfrak{L}\, v = (\tilde{H} - F) \tag{4.60c}$$

where $S \equiv \phi + u$, $D \equiv \phi - u$, $\mathfrak{L}$ and $\Re$ are the lowering and raising operators for the Hermite functions, and where F, G, and $\tilde{H}$ are the forcing terms for x-momentum equation, y-momentum equation and for the height equation, respectively.

Expanding everything in sight in terms of normalized Hermite functions gives the equivalent algebraic set:

$$S_{0,t} + S_{0,x} = \left(F_0 + \tilde{H}_0\right) \tag{4.61}$$

and the triplets, one for each $n \geq 1$

$$S_{n+1,t} + S_{n+1,x} - [2(n+1)]^{1/2}\, v_n = \left(F_{n+1} + \tilde{H}_{n+1}\right) \tag{4.62}$$

$$[(n+1)/2]^{1/2}\, S_{n+1} - [n/2]^{1/2} D_{n-1} = G_n \tag{4.63}$$

$$D_{n-1,t} - D_{n-1,x} + [2n]^{1/2} v_n = -\left(F_{n-1} - \tilde{H}_{n-1}\right) = 0 \tag{4.64}$$

where A_n denotes the n-th normalized Hermite coefficient of the quantity A (wind stress or nonlinear) for any A.

The latitudinal momentum equation, simplified by the neglect of v_t in the long wave approximation, implies that the difference variable can be expressed in terms of the sum variable as

$$D_{n-1} = [(n+1)/n]^{1/2} S_{n+1} - \sqrt{2/n}\, G_n \tag{4.65}$$

The height equation becomes

$$\sqrt{\frac{n+1}{n}} S_{n+1,t} - \sqrt{\frac{n+1}{n}} S_{n+1,x} + \sqrt{2n} v_n = -F_{n-1} - \tilde{H}_{n-1} + \sqrt{\frac{2}{n}} G_{n,t} - \sqrt{\frac{2}{n}} G_{n,x} \tag{4.66}$$

This can be rewritten as an equation for the north-south velocity in terms of the sum variable

$$v_n = \frac{1}{n}\sqrt{\frac{n+1}{2}}\left\{S_{n+1,x} - S_{n+1,t}\right\} - \frac{1}{\sqrt{2n}}F_{n-1} - \frac{1}{\sqrt{2n}}\tilde{H}_{n-1} + \frac{1}{n}G_{n,t} - \frac{1}{n}G_{n,x}$$

where the subscripts on G denote differentiation with respect to t and x, respectively. We can eliminate v_n to obtain an equation in the sum variable only by multiplying the first, Eq. (4.62), by the square root of n and the second, (4.66), by the square root of $(n+1)$. Adding, and then multiplying the result by a common factor of the square root of n gives

$$\begin{aligned}(2n+1)S_{n+1,t} - S_{n+1,x} = n\left(F_{n+1} + \tilde{H}_{n+1}\right) - \sqrt{n(n+1)}\left(F_{n-1} - \tilde{H}_{n-1}\right) \\ +\sqrt{2(n+1)}\left(G_{n,t} - G_{n,x}\right)\end{aligned} \tag{4.67}$$

The homogeneous solution to (4.61)–(4.64) plus (4.67) has already been given in the previous section.

The crucial advantage is that through the long wave approximation, a system of three time-dependent equations has been reduced to a single first-order time-dependent equation — a vast simplification.

To illustrate this simplicity, note that for a zonal wind stress which is (i) independent of x and (ii) has a time dependence of $\exp[-i\omega t]$, it is trivial to obtain the x-independent particular solution as ($F \neq 0$, $G = 0$, $\tilde{H} = 0$)

$$S_0 = -F_0/(i\omega) \tag{4.68}$$

$$S_{n+1} = -[nF_{n+1} - (n[n+1])^{1/2}F_{n-1}]/(i\omega) \tag{4.69}$$

We shall study more complicated forced problems in later chapters.

References

1. Lukas R, Firing E (1985) An annual Rossby wave in the central equatorial Pacific. J Phys Oceanogr 15(1):55–67
2. Pedlosky J (1965) A study of the time dependent ocean circulation. J Atmos Sci 22(3):267–272
3. Cane MA, Sarachik ES (1977) Forced baroclinic ocean motions, II: the linear equatorial bounded case. J Mar Res 35(2):395–432
4. Boyd JP, Christidis ZD (1982) Low wavenumber instability on the equatorial beta- plane. Geophys Res Lett 9:769–772
5. Boyd JP (1983) The continuous spectrum of linear Couette flow with the beta effect. J Atmos Sci 40:2304–2308

6. Boyd JP, Christidis ZD (1987) The continuous spectrum of equatorial Rossby waves in a shear flow. Dyn Atmos Ocean 11:139–151
7. Pedlosky J (2003) Waves in the Ocean and Atmosphere: Introduction to Wave Dynamics. Springer, New York
8. Pedlosky J (1987) Geophysical Fluid Dynamics, 2nd edn. Springer, New York

Chapter 5
The Equator as Wall: Coastally Trapped Waves and Ray-Tracing

Abstract There are many analogies between coastal dynamics and equatorial dynamics. The first three sections therefore discuss coastally-trapped waves, ray-tracing in general, and ray-tracing for coastal waves in particular. The rest of this chapter extends this ray-tracing analysis to equatorial waves.

"A theorist is allowed to introduce one tooth fairy into his theory — but only one."

— David Schramm, quoted in Cosmic Adventures by Bob Berman, p. 217.

5.1 Introduction

The purpose of this chapter is to explore the very close analogy that exists between equatorial dynamics and coastal dynamics. As suggested in the title, the equator does indeed behave very much like a rigid wall.

In particular, the Yoshida jet—parallel to the equator and very strong, driven by the wind stress parallel to the flow in contrast to the weak, perpendicular currents driven by the winds at higher latitudes—is mirrored in the "coastal jet", an intense flow parallel to the coast which is driven by the wind component parallel to the coast. The Yoshida jet is described in Chap. 9.

The equatorial Kelvin wave is mirrored by the coastal Kelvin wave. Indeed, the connection between the two waves is so intimate that when an equatorial Kelvin wave hits a coastline, some of its energy is converted into coastal Kelvin waves that then run north and south away from the equator.

The full spectrum of higher order waves — gravity waves traveling in both directions along the equator/coast and Rossby waves propagating much more slowly than the gravity waves and in just one direction — are present in the shallow water wave equations both in the vicinity of a coast and at low latitudes.

The differences are mostly in the e-folding scale. Equatorial phenomena decay like Gaussians, i.e., as $\exp(-(1/2)\, y^2)$. Coastal jets, coastal Kelvin waves, and coastally-trapped gravity and Rossby waves decay as $\exp[-x]$ for a coast running due north-south. This difference has little or no significance, however. Bottom topography gives different and more complicated decay away from the coast, but the important point is still that the coastal jets and waves are coastally-trapped and exponentially

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_5

small at large distances from the coast. The topography of the ocean varies from bay and basin to bay and basin, but the equatorial case is simpler —always a Gaussian –because the Coriolis force depends only on the shape of the earth, which is always a sphere.

In the next section, the spectrum of coastally trapped waves will be briefly reviewed. In the third section, we will wrestle with a deeper question: why *are* there these similarities between coastal and equatorial phenomena? For the waves with mode number $n \geq 0$, we will at least partly answer this question through ray-tracing and the ideas of geometrical optics.

5.2 Coastally-Trapped Waves

Model assumptions:

(i) linear shallow water wave equations
(ii) infinitely long, straight coast running due north-south
(iii) variable depth $H(x)$ with land for $-x$ and water for $x \geq 0$
(iv) No variations in bottom depth in either y or t
(v) boundary condition as $x \to \infty : u \to 0$
(vi) boundary condition at $x = 0 : u = 0$
(vii) f = constant [neglect *rotational* beta-effect but not *topographic* beta-effect]

The linearized shallow water wave equations are

$$u_t - fv + g\phi_x = 0 \tag{5.1}$$

$$v_t + fv + g\phi_y = 0 \tag{5.2}$$

$$\phi_t + (uH)_x + (vH)_y = 0 \tag{5.3}$$

where ϕ is the displacement of the surface relative to mean sea level, as in earlier chapters, and where g is the gravitational constant. (In discussing coastal phenomenon, we shall retain g as the gravitational constant and use dimensional unknowns to conform to the common practice.) The coordinate system is illustrated schematically in Fig. 5.1.

In equatorial wave theory, it is most convenient to reduce the equations down to a single equation for v alone, but here, it is easier to solve for ϕ instead:

$$\mathscr{L}u = -g(\phi_{xt} + f\,\phi_y); \qquad \mathscr{L}v = -g(\phi_{yt} - f\phi_x) \tag{5.4}$$

where

$$\mathscr{L} \equiv \frac{\partial^2}{\partial t^2} + f^2 \tag{5.5}$$

The single equation for ϕ is then

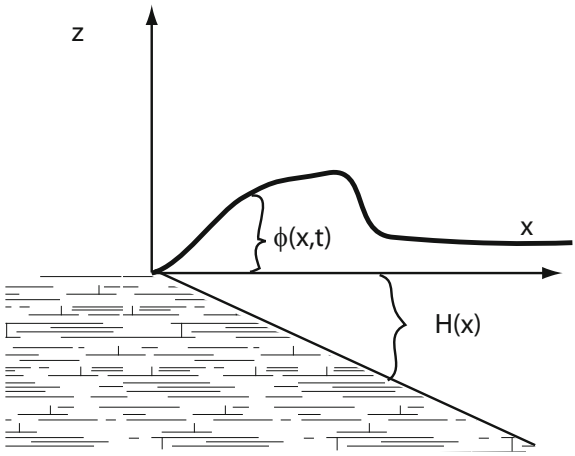

Fig. 5.1 The coordinate system for coastal flow theories. The y-axis is out of the plane of the paper and parallel to the coast

$$H\Delta\phi_t + H_x\phi_{xt} + fH_x\phi_y - (1/g)\mathscr{L}\phi_t = 0 \tag{5.6}$$

where Δ is the usual two-dimensional Laplacian operator and $\mathscr{L}$ is the operator defined by (5.5).

Since (5.6) has coefficients which vary only with x, we can, without approximation, separate variables by assuming

$$\phi = F(x)e^{i(ky-\omega t)} \tag{5.7}$$

The fact that both the y-wavenumber k and frequency ω are constants turns out to be of crucial importance in the ray-tracing discussed in the next section. Equation (5.7) is an exact solution (with the proper $F(x)$).

The x-structure function satisfies the ordinary differential equation

$$(HF_x)_x + \left[-\frac{k}{\omega}\{fH_x\} - k^2H - \frac{(f^2-\omega^2)}{g}\right]F = 0 \tag{5.8}$$

If we treat $H(x)$ in the same way as f is treated by the mid-latitude beta-plane, i.e., replacing it by a constant except where differentiated, then we can divide through by H to obtain an equation which is identical in form to its analogue in the equatorial beta-plane,

$$v_{xx} + \left[-\frac{k}{\omega}\{\beta\} - k^2 - \frac{(f^2-\omega^2)}{gH}\right]v = 0 \tag{5.9}$$

This is written in dimensional form, in contrast to the nondimensional form usually employed elsewhere in this book.

The differences between the two equations are that (i) the trapping is produced by variations in $H(x)$ in one case and by the increase of f with latitude in the equatorial

situation and (ii) the beta term, which is in curly brackets in both (5.8) and (5.9), is equal to

$$\beta = df/dy \quad \text{[rotation]}; \quad \beta = fH_x/H \quad \text{[topographic]} \tag{5.10}$$

For typical continental shelves, the topographic β defined by (5.10) is usually $O(50)$ times larger than df/dy, so an f-plane is a safe approximation as assumed above. The topographic beta effect is important because, just like its rotational counterpart, it creates Rossby waves.

The boundary conditions on u become, using (5.4)

$$F \to 0 \qquad \text{as} \qquad x \to \infty \tag{5.11}$$

and at $x = 0$ either

$$fkF(0) - \omega \frac{dF}{dx}(0) = 0 \qquad H(0) \neq 0 \quad \text{[“cliff”]} \tag{5.12}$$

or

$$\left| \frac{dF}{dx}(0) \right| < \infty \qquad H(0) = 0 \qquad \text{[“beach”].} \tag{5.13}$$

Equation (5.7) and the boundary conditions (5.11)–(5.13) constitute a standard one-dimensional Sturm–Liouville eigenvalue problem. Explicit solutions are possible only for special cases, alas, but numerical solutions are easy. A general statement can be made, however: if $H(x) \gg H(0)$, then all the terms in (5.7) will disappear except the $-k^2 H(x)$. (One can see this by dividing (5.7) by $H(x)$ and then estimating the magnitude of each term as $H(x)$ becomes very large with H' bounded.) Asymptotically, (5.7) thus reduces to

$$F_{xx} - k^2 F = 0 \tag{5.14}$$

with the solution

$$F \sim e^{-kx} \tag{5.15}$$

This asymptotic behavior is explicit in the case of a linear slope, i.e.,

$$H(x) = \alpha x \tag{5.16}$$

since the eigensolutions of (5.7), (5.10) and (5.11) are simply the Laguerre functions

$$F_n(x) = e^{-kx} L_n(2kx) \qquad n = 0, 1, 2, \ldots \tag{5.17}$$

where

$$L_n(x) \equiv n\text{-th Laguerre polynomial} \tag{5.18}$$

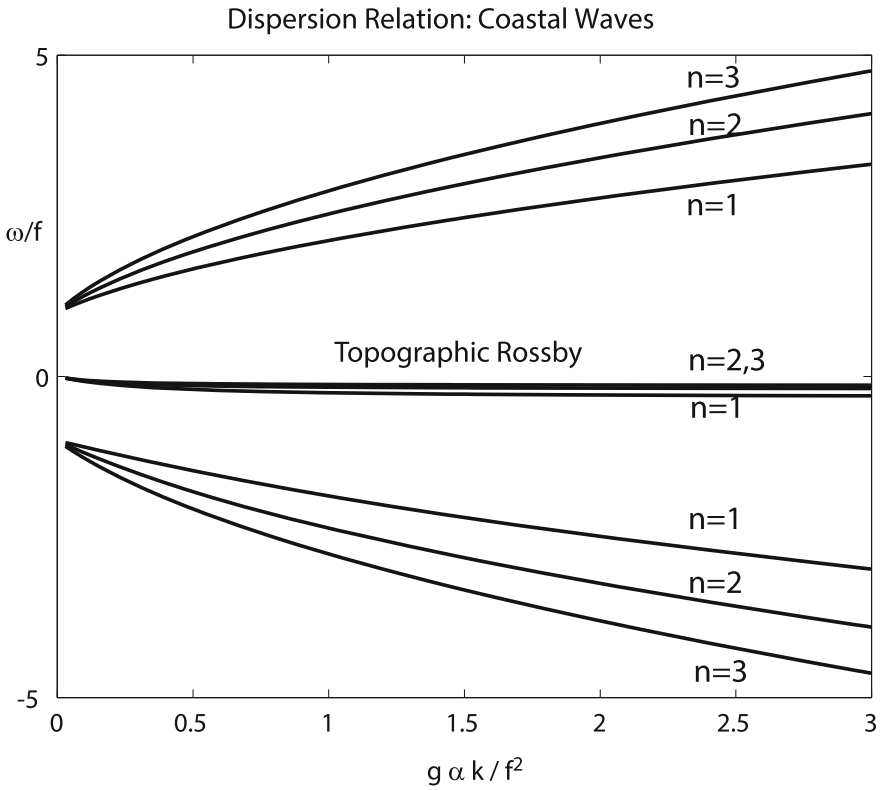

Fig. 5.2 Dispersion relations for the $n = 1$, $n = 2$ and $n = 3$ waves on a beach of constant slope α. The Rossby frequencies, all negative, are of such small magnitude that the *curves* are hard to distinguish on this graph. The $n = 2$ and $n = 3$ Rossby *curves* are thus given a single label, even though their frequencies are different. The $n = 1$ Rossby wave has the largest magnitude of frequency while higher modes travel more and more slowly. In contrast, the magnitude of gravity wave frequencies increases with n. The dimensional dispersion relation can be transformed into a cubic equation for the nondimensional frequency, ω/f versus the nondimensional parallel-to-the-coast wavenumber, $(g\alpha/f^2)k$, so these quantities are the axis labels

with the dispersion relation

$$\omega^3 - [f^2 + (2n + 1)g\alpha k]\omega - fg\alpha k = 0 \tag{5.19}$$

Laguerre polynomials are close cousins of Hermite polynomials. Like them, the subscript is the degree of the polynomial. The differences are that the exponential has a quadratic argument for equatorial waves, but a linear argument for the coastally trapped waves, and that the Laguerre functions are orthogonal on $x \in [0, \infty]$ whereas the Hermite functions are orthogonal on $[-\infty, \infty]$.

The cubic dispersion relation (5.19) is a good imitation of the cubic dispersion relation on the equatorial beta-plane. It implies that, just as for equatorial waves, there

are three distinct wave modes for each integer $n \geq 1$, physically corresponding to gravity waves traveling both up and down the coast and topographic Rossby waves — usually called "continental shelf waves" in this context — traveling at much slower speed in one direction only. Just as for equatorial waves, there are only two waves for $n = 0$, both corresponding to gravity waves; the third root has physical meaning only for $k = 0$ and is often called the "inertial mode" in the literature. Just as for equatorial waves, the gravity wave frequencies increase with increasing wavenumber while the Rossby wave has a maximum frequency at some finite wavenumber (Fig. 5.2).

One importance difference from the equatorial case is that the e-folding scale depends on the wavenumber parallel to the coast, k, whereas for equatorial waves, it is independent of wavenumber. This implies that very long coastally trapped waves will not in fact be coastally trapped. We shall explore the reasons for this when we examine ray-tracing in the next section.

5.3 Ray-Tracing For Coastal Waves

One universal truth about waves is that they propagate in a straight line only when the properties of the medium, i.e., those parameters that enter into the wave's dispersion relation, are independent of space. When the medium varies, waves are refracted. Light waves are refracted when they propagate into a glass lens because the speed of the waves is slower in glass than in air. When a water wave propagates from deep water into shallow, its phase speed changes, and so it usually is refracted, too. It matters not whether the speed of the medium jumps discontinuously, as at the air-to-glass transition of eyeglasses, or smoothly and continuously, as for a water wave above a gently-shoaling bottom: when the speed of the wave varies, the wave path will bend.

The key point here is that neither the type of wave nor the parameter is important: *any* spatial variation in the dispersion relation will refract the waves.

It is this generality of refraction that produces the kinship between equatorially trapped waves and coastally trapped waves. For the former, it is the variation of the Coriolis parameter with latitude that generates the refraction that traps the waves at low latitudes. Wave rays that initially propagate away from the equator are bent so strongly as they travel that they eventually bend around and turn towards the equator again, and so never escape the neighborhood of the equator.

For coastally-trapped waves, the strong refractive bending that traps the waves near the coast is created by variations in the depth of the water. The end result is the same: where the varying property, Coriolis/depth, goes to 0, the bending is so strong that the rays are trapped eternally in that neighborhood.

Ray-tracing is a multi-dimensional generalization of the one-dimensional WKB method; the conditions for validity of the generalization are the same as for the WKB approximation: the length scale on which the phase speed of the wave is varying must be large in comparison to the phase scale of the wave, which is the wavelength divided by 2π. In practice, as shown in Bender and Orszag [1], WKB, ray-tracing, and other

"multiple scale" approximations are quantitatively accurate even when the separation of scales is only a factor of 2. Ray-tracing may still give the correct *qualitative* answer even when the numerical error is rather large. Consequently, we are justified in using ray-tracing as a conceptual tool without fretting about its numerical accuracy.

The basic idea behind ray-tracing is that because the depth of water is varying "slowly" with x, the dynamics of the waves in the vicinity of a certain point x will be more or less the same as those for a flat-bottomed ocean whose depth is equal to the instantaneous depth $H(x)$. Put another way, we assume that the dispersion relation for constant H is still a legitimate description of waves on varying depth provided that we allow $H(x)$ to be variable, and thus obtain a variable $c(x)$ from the flat-bottom dispersion formula.

To clarify the idea, it is sufficient [since we will not perform any numerical calculations here] to restrict ourselves to long coastal gravity waves, only weakly affected by the earth's rotation, so that the variable $c(x)$ is given by the usual formula for ultra-long, non-dispersive, non-rotating gravity waves

$$c(x) = \sqrt{g\,H(x)} \tag{5.20}$$

Because the mean depth varies only in x, the frequency ω and y-wavenumber k are rigorous invariants. Consider a wave whose wavevector is making an instantaneous angle of θ_0 with respect to the perpendicular to the coast at a distance from the coast. (The angle θ is defined in Fig. 5.3) Let $\mathbf{k}$ denote the vector wavenumber. The two invariants are then

$$\omega = |\mathbf{k}|c(x) \tag{5.21}$$

$$k = |\mathbf{k}|\sin\theta(x) \tag{5.22}$$

Eliminating $|\mathbf{k}|$ between (5.21) and (5.22) shows that

$$\sin\theta(x)/c(x) = k/\omega = \text{constant} \tag{5.23}$$

which is equivalent to the more familiar form

$$\frac{\sin\theta(x)}{c(x)} = \frac{\sin\theta_0(x)}{c_0} \qquad \text{["Snell's Law"]} \tag{5.24}$$

Snell's Law, as this is usually called in optics, is completely general, but to actually trace a ray one needs a specific formula for $c(x)$ such as that provided by (5.20). In reality, $c(x) = \sqrt{gH(x)}$ will be slightly modified by the Coriolis force, etc., but nor enough to change the qualitative behavior of the waves: as the wave propagates into *deeper* water, the phase speed $c(x)$ *increases*.

For simplicity, suppose $H(0)$ is finite and take $x_0 = 0$. If the ray leaves the coast almost perpendicular to the coast, i.e., $\theta_0 \ll 1$, then Snell's law shows that as it travels into deeper water so $c(x)$ becomes large, $\sin\theta(x)$ must increase also. This is possible only if θ is increasing, i.e., the wave is bending back towards the coastline.

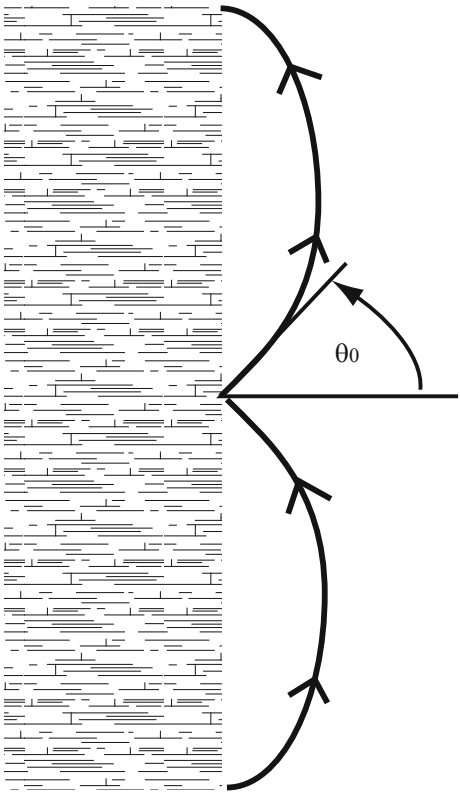

Fig. 5.3 Schematic of gravity wave refraction on a sloping *bottom* which descends to the *right*. A ray that leaves the coast at an angle θ_0, relative to the *normal* to the coast, will bend until it is turned back towards the coast. The wave is thus *coastally trapped*

Eventually, if the depth of the sea continues to increase, $\theta \to \pi/2$, and the wave is propagating *parallel* to the coast. It can go no farther from the coast than that maximum value of x defined by the equation

$$c(x_{max})/c_0 = \text{cosec}\,\theta_0 \tag{5.25}$$

and must then bend back towards the coast as illustrated schematicaly in Fig. 5.3.[1]

There is a kind of Catch-22 in the ray-tracing idea: a wave with $\theta_0 = 0$, i.e. one propagating straight away from land, will not be refracted but instead will travel straight out into the sea. Equation (5.23) shows that such waves have $k = 0$, i.e.,

[1] Strictly speaking, Snell's Law does not tell us what happens after $\theta = \pi/2$; it is consistent with (5.24) for the ray to turn back towards the coast, but there is nothing in Snell's Law that insists upon it. In reality, the multiple scales arguments break down in the vicinity of the turning point, which is also called a "caustic" in optics, and a special analysis is needed which *does* predict that the ray will reflect from the caustic.

are independent of y, and since the exponential decay factor in (5.15) and (5.17) is $\exp(-kx)$, it follows that the analytic solutions predict no decay for such waves also. Thus, it is still consistent to state the conceptual equation

$$\text{refraction} = \text{trapping} \tag{5.26}$$

This is a rather different situation from the equatorial beta-plane where all waves are trapped by the variations of the Coriolis parameter with turning points that are *independent* of the zonal wavenumber k. Qualitatively, however, the behavior is very similar.

5.4 Ray-Tracing on the Equatorial Beta-Plane

The ray-tracing method can also be applied to planetary waves at low latitudes. In this section, we will be more formal than in the discussion of coastal waves above.

The key mathematical assumption is that the waves are approximately sinusoidal in spite of the variable Coriolis parameter so that we may write

$$v = A(x, y, t) \exp\{i\Phi(x, y, t)\} \tag{5.27}$$

where $A(x, y, t)$, the amplitude or envelope, is slowly varying in comparison with the phase function Φ. If the medium is slowly varying, we may still define local wavenumbers and frequencies. One can after all speak of a child as being 3'11" in spite of the fact that growth is a continuous, ever-varying process: one simply applies a yardstick to the little fellow, and then adds the qualifier "is his height at the age of 5 years and 4 months". In the same way, we can define local wavenumbers by simply measuring the distance between adjacent crests of the wave disturbance, and then adding the qualifier "are the x and y wavenumbers at the point (x, y, t)".

The formal mathematical definitions of the local wavenumbers and frequency are

$$k \equiv \Phi_x; \qquad m \equiv \Phi_y; \qquad \omega \equiv -\Phi_t \tag{5.28}$$

and as long as the amplitude satisfies the conditions that

$$(1/A)A_{xx} \ll k^2 + m^2; \quad (1/A)A_{yy} \ll k^2 + m^2; \quad (1/A)A_{tt} \ll \omega^2 \tag{5.29}$$

then k, m, and ω obey the usual dispersion relation for equatorial waves with the factor of $2n + 1$ replaced by $m^2 + y^2$:

$$\omega^3 - (k^2 + m^2 + y^2)\,\omega - k = 0 \tag{5.30}$$

Equation (5.30) can be justified in either of two equivalent ways. First, note that WKB treatment of the parabolic cylinder equation satisfied by the n-th Hermite

function,

$$v_{yy} + [(2n+1) - y^2]v = 0 \tag{5.31}$$

gives

$$m^2 = 2n + 1 - y^2 \tag{5.32}$$

where m is the local latitudinal wavenumber. The second route to (5.32) is to simply derive a dispersion relation from the linearized shallow water wave equations with y-derivatives represented in the form

$$\phi_y = im\phi \tag{5.33}$$

and similarly for u and v.

Since the medium is still independent of x and t, both the zonal wavenumber k and frequency ω are constants, as true for coastally-trapped waves. It is only the wavenumber in the direction of variations of the equations, i.e., the latitudinal wavenumber m, which changes as the wave propagates.

The velocity of a wave packet, i.e., a wave solution whose Fourier spectrum is sharply peaked about particular *initial* values of k and m, is given by the vector group velocity. Formally, we have the following equations from Schopf, Anderson and Smith [2], which are proven in general form in the books by Whitham, Yang and Pedlosky [3–5]:

$$\frac{Dx}{Dt} = \frac{\partial \omega}{\partial k} = c_{gx} \tag{5.34}$$

$$\frac{Dy}{Dt} = \frac{\partial \omega}{\partial m} = c_{gy} \tag{5.35}$$

$$\frac{Dm}{Dt} = -\frac{\partial \omega}{\partial y}\bigg|_{m=\text{constant}} \tag{5.36}$$

The last equation requires a comment since it has just been explained above that ω is constant as the ray or wave packet travels, and the derivative of a constant is usually 0! However, ω is constant because m and y vary *simultaneously* as the packet propagates to make it so. If we take the derivative of ω with fixed m, as shown explicitly in (5.36), and replace ω by the usual approximate expressions for frequency:

$$\omega = -k/(k^2 + m^2 + y^2) \qquad \text{[Rossby waves]} \tag{5.37}$$

$$\omega = \pm[k^2 + m^2 + y^2]^{1/2}, \qquad \text{[gravity waves]} \tag{5.38}$$

we do indeed obtain a non-zero value for $\partial\omega/\partial y$.

The time derivatives have been written using D/Dt to denote that they are the advective derivatives following the motion of the ray or wave packet:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + c_{gx}\frac{\partial}{\partial x} + c_{gy}\frac{\partial}{\partial y} \tag{5.39}$$

We would seem to have a little trouble in evaluating the partial derivatives that define the two group velocities and the advective derivative of the latitudinal wavenumber because the full dispersion relation (5.30) is *implicit* rather than explicit. However, if we differentiate (5.30) with respect to k, m and y respectively, we obtain the three equations

$$3\omega^2[\partial\omega/\partial k] - (k^2 + m^2 + y^2)[\partial\omega/\partial k] - 2k\omega - 1 = 0 \tag{5.40}$$

$$3\omega^2[\partial\omega/\partial m] - (k^2 + m^2 + y^2)[\partial\omega/\partial m] - 2m\omega = 0 \tag{5.41}$$

$$3\omega^2[\partial\omega/\partial y] - (k^2 + m^2 + y^2)[\partial\omega/\partial y] - 2y\omega = 0 \tag{5.42}$$

It is trivial to solve these to obtain the necessary partial derivatives in terms of ω, k, m and y without approximation:

$$c_{gx} = \frac{1 + 2k\omega}{[3\omega^2 - (k^2 + m^2 + y^2)]} \tag{5.43}$$

$$c_{gy} = \frac{2m\omega}{[3\omega^2 - (k^2 + m^2 + y^2)]} \tag{5.44}$$

$$\frac{Dm}{D_t} = -\frac{2y\omega}{[3\omega^2 - (k^2 + m^2 + y^2)]} \tag{5.45}$$

All these expressions are exact, and apply to either gravity or Rossby waves with the choice of the appropriate branch of the solution of (5.30) for ω.

The paths of the rays or wave packets in the $x - y$ plane can be obtained by integrating the usual trajectory equation

$$\frac{dy}{dx} = \frac{c_{gx}}{c_{gy}} = \frac{m}{(k + 1/[2\omega])} \tag{5.46}$$

Recalling that $m = (2n + 1 - y^2)^{1/2}$ for the n-th Hermite function, (5.46) can be solved via separation of variables, i.e.,

$$\frac{dy}{(2n + 1 - y^2)^{1/2}} = \frac{dx}{(k + 1/[2\omega])} \tag{5.47}$$

Integrating each side of (5.47) gives an arcsin function on the left and $x/(k + 1/2\omega)$ on the right, which implies

$$y = (2n+1)^{1/2} \sin\left[\frac{x}{(k+1/[2\omega])} + \Xi\right] \tag{5.48}$$

where Ξ is an arbitrary phase constant. The book by Gill [6] discusses ray-tracing also and gives (5.48), but the context is one of equatorial gravity waves only. All the results in this section, however, apply to both Rossby and gravity waves if the correct ω is substituted into them.

Equation (5.48) shows explicitly that the rays for equatorial waves — any type of equatorial wave, for (5.48) applies to either gravity waves or Rossby waves — are refracted by the latitudinal variations of the Coriolis parameter so that they turn back (at $y_t = \pm(2n+1)^{1/2}$, which are the usual turning points for the n-th mode) and cross and recross the equator as the wave propagates zonally. Strictly speaking, (5.48) is not self-consistent because WKB theory and its generalizations always break down at turning points (which in ray theory are also known as "caustics"); one must do a local treatment in terms of Airy functions and then match that to the solutions away from the caustics given by the generalized WKB treatment. The Airy analysis shows, however, that the wave packets do simply reflect off the turning latitudes and return towards the equator so that (5.48) is the correct trajectory for the wave packets.

In contrast to a wave packet, the Hermite functions that describe the n-th latitudinal eigenfunctions have a cosine-like structure between the turning points with no propagation whatsoever. Inside the turning points, the Hermite functions have the WKB representation [n even]

$$\psi_n(y) \sim 0.5[\exp(imy) + \exp(-imy)] \tag{5.49}$$

where we have suppressed the slowly-varying amplitude factor on the right-hand side of (5.49). A standing wave in y can always be decomposed into two traveling waves, propagating in opposite directions in latitude. A Hermite function is the superposition of two wave packets, and reflection of the waves from the turning points is the key to a physical (as opposed to merely mathematical) understanding of the Hermite functions.

Ripa (1994) also provides a ray-tracing analysis for equatorial waves [7], and one of his themes is specifically the latitudinal modes/ray tracing relationship. He notes in his abstract, speaking also of radiation from a point source on the equator, that, "The full solution…has a structure that would be hard to find in a normal modes expansion". See also [8].

Refraction is the key for coastal waves, too: exactly the same kind of trapping occurs for coastal waves even though the Coriolis parameter is kept constant and it is quite a different physical mechanism — the water depth — which is varying. The point is that variations of almost anything — f, H, the mean shear, the static stability —will cause refraction of the waves. If the refraction is strong enough, refraction will trap waves within certain regions of space.

Figures 5.4 and 5.5 show this refraction-sum-trapping actually occurs for packets of equatorial waves even when no approximations are made, and the equations are solved numerically. The formal condition for the accuracy of WKB and ray-tracing-methods is

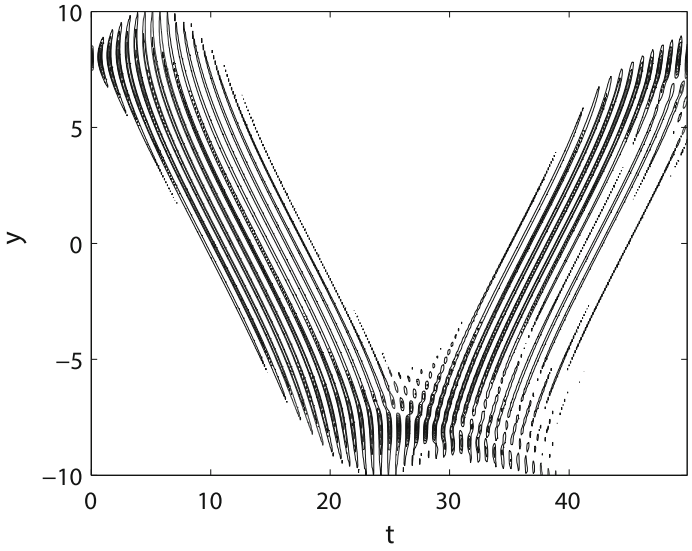

Fig. 5.4 Dispersion and refraction of gravity waves on the equatorial beta-plane. Note that the horizontal axis is time, not longitude, while the vertical axis is latitude. The packet is excited by the initial conditions $S(x, y, 0) = \exp(-8[y - 8]^2)$, $v(x, y, 0) \equiv 0$, $D(x, y, 0) = 0$. The packet is independent of x for all time. Contours of the sum variable, $S = u + \phi$, are plotted with a contour spacing of 1/10 and omission of the zero contour and all negative contours. A similar case with different parameter values is Fig. 6 of [9], which is reprinted in Gill (1982)

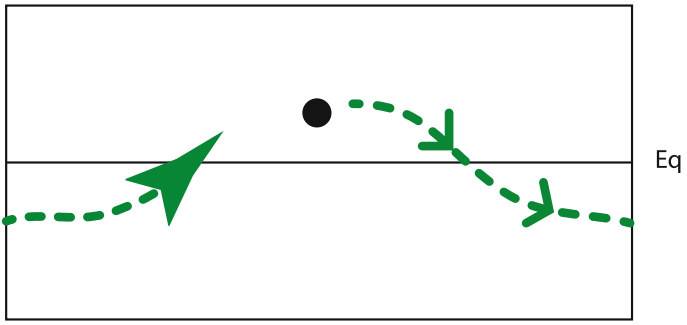

Fig. 5.5 The *dashed line* is the ray of a Rossby wave-train radiating from a circular mountain at 30° N., which disturbs a flow that would otherwise be in solid-body rotation. The *black* disk marks the position of the mountain. As the Rossby wave is refracted towards the equator, and then eventually from a turning latitude in the southern hemisphere back towards the equator, it decays in amplitude because of dissipation. The isolines of a similar experiment can be seen in Figs. 3a, c of Grose and Hoskins [10]

$$n \gg 1 \tag{5.50}$$

i.e., the wave is oscillating with y on a scale small in comparison to 1 (which is the scale on which f is varying) so that the oscillation scale may be considered a "fast" scale in the in sense of the method of multiple scales in comparison to the "slow" scale on which Coriolis parameter or depth varies. In practice, WKB usually gives *qualitatively* correct results even when this criterion of disparity of scales of variations is but poorly satisfied. Equatorial waves are a good example: even the lowly Kelvin wave is equatorially trapped despite the fact that WKB and ray-tracing cannot be rigorously justified for this mode.

Although our first look at ray-tracing has been confined to explaining why coastal and equatorial trapping are possible, it is also a perfect tool in understanding the shadow zones and foci that occur when the equatorial ocean is forced by the annual and semiannual cycles in the atmospheric wind stress. Before we can move on to this relatively advanced topic, we must first consider a more basic issue: how equatorial waves reflect off coastal boundaries, as described in the next chapter.

5.5 Coastal and Equatorial Kelvin Waves

The coastal Kelvin wave is illustrated in Fig. 5.6. In the off-shore direction, there is perfect geostrophic balance while the wave is otherwise a normal gravity wave propagating at the usual gravity wave speed parallel to the coast. Like the equatorial Kelvin wave, one of the horizontal velocity components is identically equal to zero—for the coastal Kelvin wave, the component perpendicular to the coast.

Unlike other coastally-trapped waves, the Kelvin mode can exist in a flat-bottomed channel with no variations in the depth of the sea at all. What is essential is the presence of the shore so that the pressure gradients created by the piling up of the water along the cliff can balance the off-shore Coriolis force. If the coast were suddenly removed, as could be done in a laboratory experiment, the crests of the waves would be free to slop over what had been land, shoved sideways by the now-unopposed Coriolis force.

The coastal Kelvin wave, linearized about a state of rest, is

$$\phi = e^{-x/R} e^{ik(y-ct)} \tag{5.51}$$

$$v = -(g/H)^{1/2} e^{ik(y-ct)} \tag{5.52}$$

where

$$R = (gH)^{1/2}/f \qquad \text{["Rossby radius of deformation"]} \tag{5.53}$$

$$c = -(gH)^{1/2} \qquad \text{[Coastal Kelvin dispersion relation]} \tag{5.54}$$

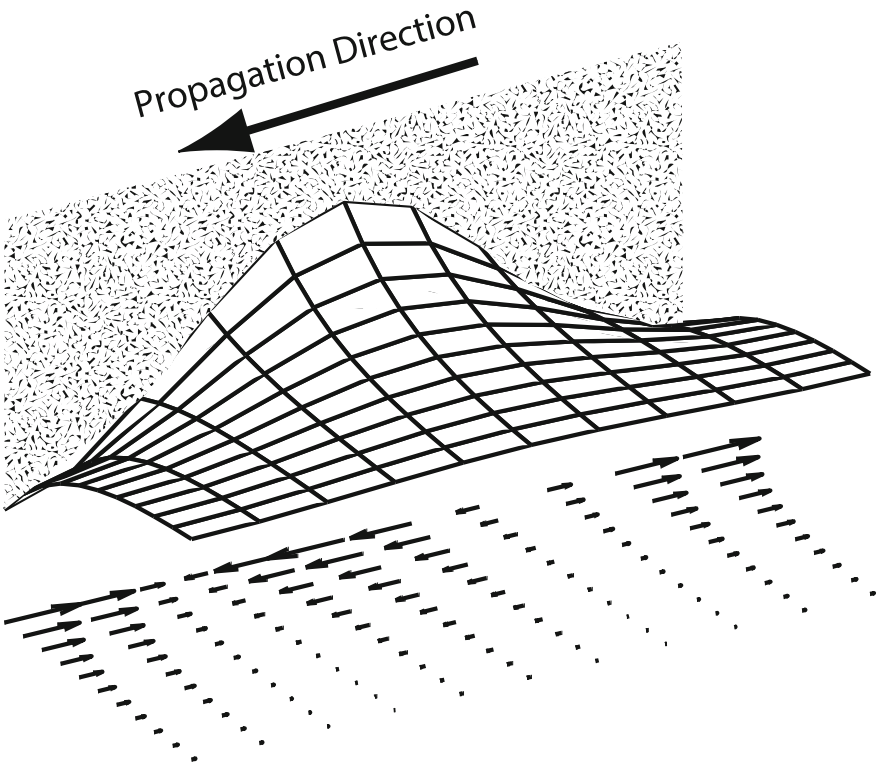

Fig. 5.6 Schematic of a coastal Kelvin wave. The shaded wall is the coast. The height ϕ is depicted by a mesh diagram. The *arrows* show the magnitude and sign of the horizontal velocity, which is everywhere parallel to the coast

The e-folding scale is the usual Rossby radius of deformation. The Kelvin wave is nondispersive and the minus sign in (5.54) is a reminder that coastal Kelvin waves, like coastally trapped topographic Rossby waves, are always "right-bounded", that is to say, keep the land on the right as viewed by an observer traveling with the wave (in the northern hemisphere).

5.6 Topographic and Rotational Rossby Waves and Potential Vorticity

The potential vorticity for a barotropic fluid is given by

$$q = (\zeta + f)/H \tag{5.55}$$

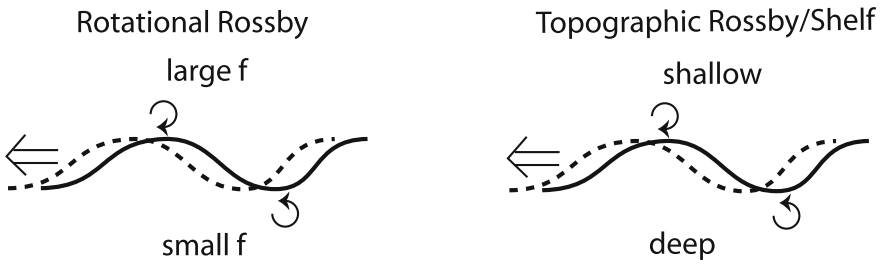

Fig. 5.7 *Left* the mechanism of planetary wave propagation. Conservation of potential vorticity implies that a particle which is perturbed equatorward acquires cyclonic vorticity (*arrow*). The *broad arrow* shows the motion induced by this change in vorticity, and is equivalent to westward propagation. *Right* same but for topographic Rossby waves, also known as shelf waves

The quintessential point is that variations in either the Coriolis parameter or the depth will force changes in the relative vorticity ζ. In particular, if the particles of a fluid are perturbed from their equilibrium positions by being moved to where either f or H is different from its original value, the particles will acquire a relative vorticity so that q is conserved. This vorticity in turn can push the particles back to their original positions while sending the wave propagating in a direction perpendicular to the gradient of f or H as shown schematically in Fig. 5.7.

In either case, it is not the *mean* value of f or H that is important, but rather the fact that f and H *vary*. Without gradients of f or H, the particles would acquire no relative vorticity when displaced, so there would be no horizontal restoring force to push the particles back to their original positions.

Both species of waves are now commonly called "Rossby" waves because the dynamical mechanism is the same, even though Rossby's original paper was limited to variations in f alone. The topographic Rossby waves are somewhat confusingly also called "continental shelf waves" and "edge waves", but it is more sensible to name them "topographic Rossby waves" since these waves can occur in the deep ocean because of gradients of topography along the sea floor, thousands and thousands of kilometers from the continental shelves or the edges of the sea.

An elementary power series expansion shows that if Δy is the northward displacement of the particle, then it acquires a relative vorticity $\Delta\zeta$ given by

$$\Delta\zeta = -(df/dy)\Delta y \tag{5.56}$$

A similar expansion about a point (x_0, y_0) gives

$$\Delta\zeta = (dH/dx)(f_0/H_0)\Delta x \tag{5.57}$$

assuming for simplicity that the topography varies in the x-direction. The identification of a "topographic beta" in (5.57) is obvious.

As described in Sect. 5.2, topographic Rossby waves comprise one of the three modes for each off-shore mode number $n \geq 1$, but they have been observed not only along the continental shelf, but above prominent topographic features in the interior of ocean basins.

References

1. Bender CM, Orszag SA (1978) Advanced mathematical methods for scientists and engineers. McGraw-Hill, New York, p 594
2. Schopf PS, Anderson DLT, Smith R (1981) Beta dispersion of low-frequency Rossby waves. Dyn Atmos Ocean 5(3):187–214
3. Whitham GB (1974) Linear and nonlinear waves. John Wiley & Sons, New York
4. Yang H (1991) Wave packets and their bifurcations in geophysical fluid dynamics, applied and mathematical sciences, vol 85. Springer, New York
5. Pedlosky J (2003) Waves in the ocean and atmosphere: introduction to wave dynamics. Springer, New York, p 260
6. Gill AE (1982) Atmosphere-ocean dynamics. Academic Press, New York
7. Ripa P (1994) Horizontal wave-propagation in the equatorial wave-guide. J Fluid Mech 271:267–284
8. Han W, Huang SX (2004) Equatorial trapped waves in the mean state with slow variation, inhomogeneus equatorial duct. J Hydrodyn 16(2):182–185
9. Anderson DL, Gill AE (1979) Beta Dispersion of Inertial Waves. J Geophys Research-Oceans and Atmospheres 84(NC4):1836–1842
10. Grose WL, Hoskins BJ (1979) Influence of orography on large-scale atmospheric flow. J Atmos Sci 36(2):223–234

Chapter 6
Reflections and Boundaries

Abstract Coastal boundaries reflect ocean waves and in so doing generate new waves, jets and transient currents. The first third is a comprehensive discussion of the reflection of equatorial waves from straight boundaries aligned with a meridian. Although the equatorial oceans are effectively unbounded north and south, normal modes and quasi-normal modes do exist and are analyzed in the middle of the chapter. The final part is reflection from irregular boundaries and scattering from islands.

> Mathematical analysis appears to be a dying or lost art, and I would argue for a better balance between analytical and numerical methods.
>
> Philip D. Thompson on p. 636 of his article, "The Maturing of the Science", *Bull. Amer. Met. Soc., 68*, 631– 637 (1987). Thompson was both a brilliant theorist and an adept number-cruncher; he wrote the first American book on numerical weather prediction.

6.1 Introduction

The equatorial ocean is chopped up into several pieces by the continents, and these zonal boundaries greatly complicate the dynamics. The Yoshida jet solution given in Chap. 9 is still a solution to the linear equations of the equatorial beta-plane, but additional terms — free oscillations – must be added in order to satisfy the zonal boundary conditions. Thus, the first role of coasts is to act as wavemakers.

Even when waves are excited directly in the open ocean by wind stresses that vary with x and t, the coasts play a role because such waves must eventually collide with land. For nonrotating gravity waves, the reflection problem is simple: the wave simply bounces and then propagates with the same speed but opposite direction. The rotation of the earth, however, introduces strong asymmetries in the propagation of planetary waves. For example, only long Rossby waves have a westward group velocity while only short Rossby waves have an eastward group velocity. It follows that even if a Rossby wave reflects into another Rossby wave, we must expect a drastic change in zonal wavenumber. On the equatorial beta-plane, however, the situation is still more complicated.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_6

Nonetheless, boundary effects are essential to synthesizing complete solutions. Because of the complexities involved, we shall begin with Rossby wave reflection on the midlatitude beta-plane where some of the difficulties are suppressed. We shall then proceed to the relatively easy case of reflections of equatorial waves from a western boundary and then to the much more difficult problem of eastern boundary reflections where coastal Kelvin waves are an annoying but essential complication. Finally, we will turn to reflections from slanted boundaries, i.e., those not oriented due north–south, which is an approximation employed in early sections of this chapter.

6.2 Reflection of Midlatitude Rossby Waves from a Zonal Boundary

The problem of this section is that of the reflection of a quasi-geostrophic Rossby wave (or more accurately, a wave packet) whose streamfunction is in the form

$$\psi_i = A(x, y, t) \exp(i[k_i x - m_i y - \omega_i t]) \tag{6.1}$$

from a zonal boundary at $x = 0$. When the incident wave (6.1) is a wave packet, the "envelope function" $A(x, y, t)$ will vary slowly with position and time, but the shape of the envelope has no bearing on the reflection of the packet. In other words, to lowest order in perturbation theory, the precise mathematical form of $A(x, y, t)$ is irrelevant, and we may take A to be a constant for purposes of calculating the reflected wave. However, unless the wave is really a wave packet, the concepts of "incident" and "reflected" would be meaningless because both the incident and reflected would extend to infinity in x, and could not be meaningfully separated. Hence, it is *conceptually* important that A varies with x and t even though this "envelope" or "modulation" factor will not be discussed in detail until Chap. 16.

The fundamental physical condition at a zonal boundary which is parallel to the y-axis is

$$u = 0 \tag{6.2}$$

for any exact or approximate system. For quasi-geostrophic waves, this becomes

$$\psi_y = 0 \tag{6.3}$$

If we assume that the reflected wave is of the same general form as (6.1) — we must check a posteriori to verify that we have indeed obtained a solution to both the differential equation and boundary conditions — then (6.3) implies

$$m_i A_i \exp(i(m_i y - \omega_i t)) + m_r A_r \exp(i(m_r y - \omega_r t)) = 0 \tag{6.4}$$

at $x = 0$. Now the only way that (6.4) can be satisfied for all t is if

$$\omega_i = \omega_r \tag{6.5}$$

i.e., the frequency is invariant to reflection. (This would still be true even for reflection from an oblique coast, and not merely one aligned north–south.) The only way that (6.4) can be satisfied for all y is if

$$m_i = m_r \tag{6.6}$$

(This is *not* true for reflection from an *oblique* coast). Thus, for a straight coastline, both the frequency and y-wavenumber are conserved, and only the changes in k and A remain to be determined. At this point, it is necessary to invoke the explicit dispersion relation for Rossby waves in the form

$$\omega = -\beta k/(k^2 + m^2 + F) \tag{6.7}$$

where F is the Froude number, f^2/gH.[1]

Many years ago, Michael Longuet-Higgins noted that the Rossby dispersion relation (6.7) can be rearranged into

$$\left(k + \frac{\beta}{2\omega}\right)^2 + m^2 = \frac{\beta^2}{4\omega^2} - F \tag{6.8}$$

This is the equation of a circle in the $k - m$ plane with a center at $k = -\beta/(2\omega)$ and a radius of $\sqrt{\beta^2/(4\omega^2) - F}$ as shown in Fig. 6.1. This is precisely what is needed to determine the reflected wavenumber k_r since the total wavevectors for both the incident and reflected wave must (since ω is invariant) touch the same circle. Further, since the y-wavenumber m is conserved, both the incident and reflected wavenumber must lie on the straight line, parallel to the k-axis, for which m is the equal to the y-wavenumber of the initial wave. The dashed line intersects the circle only in two points.

To determine which point is which, recall that a wave packet incident on a western boundary from the east must obviously have a westward group velocity. Only long Rossby waves (small k) have this property. On the other hand, the reflected packet must have an eastward group velocity, which is possible if and only if k is sufficiently large. It follows that the leftmost point on the circle and the line through y-wavenumber m must be the tip of the wavevector for the incident wave whereas the right point — which has much larger k —is the tip of the wavevector of the reflected packet.

This is indeed the correct answer: the x-wavenumber k is signficantly increased by the reflection whereas m and ω are unchanged. To verify that we have obtained

[1] Notation warning: m, which is elsewhere the index of the baroclinic vertical modes, is here used for the latitudinal wavenumber in this section only.

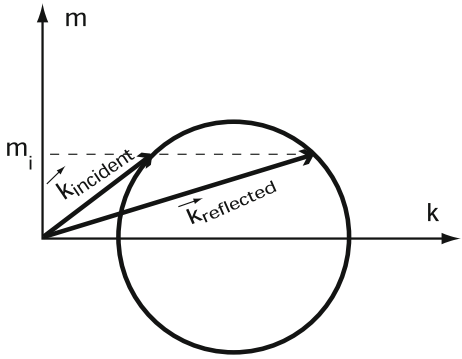

Fig. 6.1 The dispersion relation for Rossby waves, which can be written by clearing denominators as $(k + \beta/(2\omega))^2 + m^2 = \beta^2/(4\omega^2) - F^2$, is the implicit representation of a *circle*. All pairs of longitudinal and latitudinal wavenumbers (k, m) such that the frequency is equal to the common frequency $\omega = \omega_i = \omega_r$ of the incoming and reflected Rossby waves must lie on this *circle*. Since the latitudinal wavenumber m is unchanged by the reflection, i.e., $m_i = m_r$ as marked by the *dashed line*, both the incoming and reflected wavevectors (rays with *arrows*) must lie on the *dashed line* $m = m_i$ as well as on the *circle*. Based on a concept of M.S. Longuet-Higgins

a true solution, we need to check that the waves whose wavevectors lie in the right half of the circle in Fig. 6.1 do indeed have eastward group velocities while those that lie in the left half of the circle travel in the opposite direction. The equations for the two points with fixed m that lie on the circle are given by

$$k = -\beta/2\omega \pm \sqrt{\beta^2/[4\omega^2] - (F + m^2)} \tag{6.9}$$

where the $(-)$ corresponds to the incident wave and the $(+)$ to the reflected wave; the difference of the two roots gives the change in wavenumber due to reflection. It is straightforward to show that the condition that the radical vanish – which is the point of demarcation between the left and right half of the circle at the top of the circle – is the condition for the group velocity $c_{gx} = \partial\omega/\partial k = 0$. (Differentiate both sides of the dispersion relation with respect to k, set $\partial\omega/\partial k = 0$, and this becomes $k = -\beta/(2\omega)$, that is k must lie on a straight vetical line in the k-m plane through the center of the circle.) It follows that wave packets of larger k will have eastward group velocities while waves of smaller k have westward group velocity so that our initial assumption of a reflected wave in the same form as the incident wave always yields a true and consistent solution. Figure 6.2, which shows a graph of frequency versus zonal wavenumber for fixed m, makes this even clearer and also illustrates the increase in k.

Equation (6.4) then, in view of (6.5) and (6.6), implies that

$$A_i(0, y, t) = -A_r(0, y, t) \tag{6.10}$$

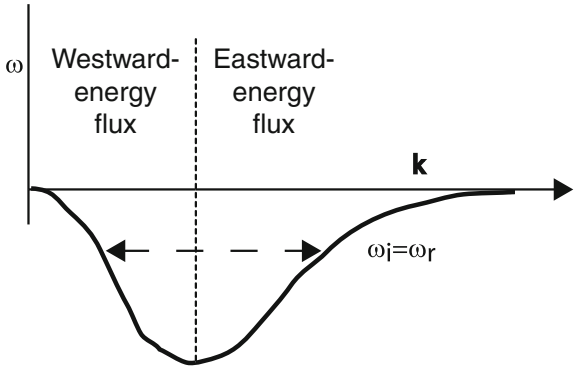

Fig. 6.2 The *thick curve* is the dispersion relation $\omega(k)$ for a Rossby mode where k is the zonal wavenumber. The double-headed arrow connects wavenumbers that have the same frequency. The *vertical dotted line*, which marks the zonal wavenumber k where the group velocity $c_g = d\omega/dk$ is zero, is the boundary between waves of positive and negative group velocity as labeled

In words, the amplitude of the wave is not changed by reflection, but the phase is reversed by 180 degrees.

The zonal velocity, which is given by (minus) the y-derivative of ψ, is unchanged (except in phase) because the y-wavenumber and the amplitude A are unchanged. The latitudinal velocity, however, is greatly increased by the reflection because it is the x-derivative of ψ, and is therefore proportional to k, which is greatly increased by reflection.

It can be shown, as done in Pedlosky ([1], p. 127), that the energy density (energy per unit area) of the reflected wave is greater than that of the incident wave to compensate for the slower group velocity of the reflected packet; thus, the total energy flux stays the same. Because of the change of wavenumber, the reflected packet is squeezed into a narrower region traveling more slowly from the boundary, but the total energy of the packet stays the same.

Pedlosky [1] also discusses reflection from a latitudinal boundary, i.e., one running due east–west at a fixed value of y. For this special case, everything is conserved except that m suffers a change of sign. However, this case has little direct connection with equatorial oceanography, so we will say no more here.

6.3 Reflection of Equatorial Waves from a Western Boundary

On the equatorial beta-plane, the reflection of a wave from a zonal boundary is more complicated because the zonal velocity u is the sum of *two* Hermite functions with different coefficients, i.e.,

$$u = \frac{1}{k-\omega}\sqrt{\frac{n+1}{2}}\,\psi_{n+1} - \frac{1}{k+\omega}\sqrt{\frac{n}{2}}\,\psi_{n-1} \tag{6.11}$$

It would be a blessing if an incoming wave packet of meridional mode number N would reflect as a wave packet of the same meridional wavenumber. However, as in the midlatitude Rossby wave reflection problem solved in the previous section, the east–west wavenumber k is changed by reflection even though the frequency ω is not. Because the coefficients in (6.11) have different dependence upon k, it is not possible to choose the amplitude of the N-th reflected mode so as to simultaneously cancel both Hermite terms in (6.11) for the incoming wave so as to make u vanish at the boundary.

To derive the explicit solution, write the incoming wave as

$$v_i = iA_i(x, y, t)\exp[i(k_i x - i\omega t)]\psi_N(y) \tag{6.12}$$

(Note that we have divided out the wavenumber k from the formulas of Chap. 3).

Assume that the reflected wave is the *finite* series

$$u_r = i\sum_{n=1}^{N} A_r \exp[i(k_n x - \omega t)]\left\{\frac{\sqrt{(n+l)/2}}{k-\omega}\psi_{n+1} - \frac{\sqrt{n/2}}{k+\omega}\psi_{n-1}\right\}$$
$$-[\omega/\sqrt{2}]\,A_0\,\exp[i(\omega - 1/\omega)x - \omega t]\,\psi_1 + A_{Kel}\,\exp[i\omega(x-t)]\,\psi_0 \tag{6.13}$$

Note that both the Yanai and Kelvin waves appear in (6.13) because both have *group* velocities towards the east, even though the Yanai wave has a phase velocity in the opposite direction.

To make (6.13) satisfy the condition

$$u_i + u_r = 0 \quad \text{at } x = 0, \tag{6.14}$$

choose the coefficient A_N so as to match the $(N+1)$-st Hermite coefficient of the u_i. Unfortunately, this leaves the $(N-1)$-st Hermite coefficient of mode N uncancelled, so we need the $N-2$ mode in the reflected packet, too. By choosing A_{N-2} appropriately, we can cancel the second part of u_i — but this leaves uncancelled the degree $N-3$ term in the reflected $(N-2)$ mode. To get rid of this, we are forced to choose A_{N-4} to balance it, and so it goes until we come to modes which have zonal velocities that are proportional to but a *single* Hermite function — the Yanai and Kelvin waves. Thus, these modes play an essential role in making a finite, closed form solution possible for reflection from a western boundary.

The results are

$$A_N = (\omega - k_N)/(\omega - k_i)\mathrm{A}_i \tag{6.15}$$

$$A_{N-2} = \sqrt{N/(N-1)}\,(k_{N-2} - \omega)\left\{\mathrm{A}_i\frac{1}{\omega + k_i} + A_N\frac{1}{\omega + k_N}\right\} \tag{6.16}$$

$$A_{n-2} = A_n \sqrt{n/(n-1)}\ \{(k_{n-2} - \omega)/(k_n + \omega)\} \tag{6.17}$$

$$A_0 = -A_2 \sqrt{2}\ /\ \{\omega(k_2 + \omega)\} \tag{6.18}$$

$$A_{Kel} = A_1 / \left\{\sqrt{2}(k_1 + \omega)\right\} \tag{6.19}$$

Note that the recursion advances by jumps of 2: only those equatorial waves with (i) mode numbers less than or equal to that of the incoming wave packet N and (ii) the same symmetry with respect to the equator will enter into the reflected wave packet. Thus, if the incoming wave is symmetric about the equator, i.e., u_i is symmetric about the equator, then the Yanai wave will not be part of the reflected wave packet. The wavenumbers that appear in (6.16)–(6.19) are easily determined from the dispersion relation

$$k = -1/(2\omega) \pm \sqrt{\omega^2 + 1/(4\omega^2) - (2n+1)} \tag{6.20}$$

which is just the usual dispersion relation for equatorial waves, written so as to solve for k instead of ω. (Note that ω in this case is the same for all waves, both incoming and reflected, and is the frequency of the incoming wave.) For an incoming gravity wave, the negative sign in (6.20) would be appropriate; the (+) sign would be used for all the reflected waves. For an incoming Rossby wave, one would apply the (+) sign in (6.20) to that, and the (−) sign to all the reflected waves.

In either event, note the following:

1. The reflection problem is solved in explicit, closed form and consists of a finite number of waves.
2. All the Hermite functions which enter into the reflected packet have a degree less than or equal to that of the incoming packet. Since the turning points of the waves are proportional to the mode number, it follows that the reflected wave packet is confined to the same latitudinal band as the incoming wave packet.
3. All the wavenumbers k are real so that the reflected wave packets will propagate into the interior of the ocean.[2]

One final note: when the zonal wavenumber k and latitudinal mode number n are both small, as is appropriate to low frequency (annual and semiannual) long wave motions, the Kelvin wave will tend to dominate the response since the denominator in (6.19) is then very small. The terms in the general recursion (6.17) on the other hand, will be of $O(1)$ so that the higher order modes will be of comparable size. Even when n and k are not both small, the Kelvin wave will still play a special role in the reflection of a wave from the western boundary – any wave – because it has a much higher group velocity than any Rossby wave.

[2]Note that the radical in (6.20) must be positive for the incoming, N-th mode so that it can be a propagating (as opposed to a zonally-trapped) wave. Since all the waves in the reflected packet have mode numbers less than or equal to N, it follows that the radical is the same or larger for them. Hence, k is real for all waves in the reflected packet.

6.4 Reflection from an Eastern Boundary

If we attempt to apply the recursive procedure of the preceding section to determine the mode-by-mode components of the reflected wave which is created when a wave of mode number N is incident on a boundary on the eastern side of the ocean, we run into fatal problems because the lowest two modes, the $n = 0$ Yanai and $n = -1$ Kelvin waves, always have group velocities towards the east. Consequently, they cannot be part of the wave ensemble which reflects westward off an eastern boundary. The downward recursion therefore fails because we need the Kelvin and Yanai modes to cancel the components of zonal velocity u which are proportional to $\psi_1(y)$ [for the $n = 2$ mode] and $\psi_0(y)$ [for the $n = -1$].

Instead, we must use an *upward* recursion (in latitudinal mode number), and this generates an infinite series of modes instead of a finite sum. There are, however, three different sets of formulas because the reflection of the Kelvin and Yanai waves must be dealt with separately from the higher modes. (This is unnecessary in discussing western boundary reflection because these two exceptional, eastward-moving wave modes can never reflect from a western boundary).

For the general case of an incoming wave of latitudinal mode number N with $N \geq 1$, the reflected wave takes the form

$$u_r = \sum_{n=N}^{\infty} A_n \exp[i(k_n x - \omega t)] \left\{ \frac{[(n+1)/2]^{1/2}}{k_n - \omega} \psi_{n+1} - \frac{[n/2]^{1/2}}{k_n + \omega} \psi_{n-1} \right\} \quad (6.21)$$

where N is the mode number of the incoming, eastward-travelllng wave packet.

To cancel the term in the sum $(u_r + u_i)$ which is proportional to $N - 1$, we must have

$$A_i \sqrt{N/2}/(k_i + \omega) + A_N \sqrt{N/2}/(k_N + \omega) = 0 \quad (6.22)$$

which implies

$$A_N = -A_i (k_N + \omega)/(k_i + \omega), \qquad N \geq 1 \quad (6.23)$$

Unfortunately, the zonal velocity component of the reflected N-th mode proportional to ψ_{N+1} generally does not exactly cancel the corresponding component of the incoming wave, so the "low" component of mode $(N + 2)$ — that is to say, the part of u which is proportional to the Hermite function of degree *lower* than the mode number — must be pressed into service to cancel the "high" components of both the incoming and N-th reflected mode, demanding

$$\begin{aligned} &- A_{N+2} [(N+2)/2]^{1/2}/(k_{N+2} + \omega) \\ &\quad + A_i [(N+1)/2]^{1/2}/(k_i - \omega) + A_N [(N+1)/2]^{1/2}/(k_N - \omega) = 0 \end{aligned} \quad (6.24)$$

which gives, using (6.23),

$$A_{N+2} = A_i\,[(N+1)/(N+2)]^{1/2}(k_{N+2}+\omega) \times \left\{ \frac{1}{k_i-\omega} - \frac{(k_N+\omega)}{(k_i+\omega)\,(k_N-\omega)} \right\} \tag{6.25}$$

In a similar way, we must cancel the "high" part of the zonal velocity of the n-th mode with the "low" degree component of the $(n+2)$-d mode, which gives the general recursion

$$A_{n+2} = A_n \sqrt{(n+1)/(n+2)}\,\frac{k_{n+2}+\omega}{k_n-\omega}, \qquad n \geq N+2 \tag{6.26}$$

The general recursion (6.26) applies when the incident wave is a Kelvin or a Yanai wave, too, but the starting values are different. For the Kelvin wave

$$A_1 = A_{Kel}\,2^{1/2}\,(k_1+\omega) \qquad \text{[incident Kelvin]} \tag{6.27}$$

with the recursion (6.26) determing A_3 and all higher A_n's. For the Yanai ($n=0$) incoming wave, the lowest mode of the reflected wave disturbance is

$$A_2 = -A_0\,2^{-1/2}\,\omega\,(k_2+\omega) \qquad \text{[incident Yanai]} \tag{6.28}$$

where the general recursion (6.26) again applies to determine all higher coefficients. A number of remarks are in order. First, the reflected wave disturbance always has the same *symmetry with respect to the equator* that the incident wave does, regardless of whether the reflection is from an eastern or western boundary. Thus, when the mode number of the incoming wave packet is odd, all the even number A_n's are automatically 0; when the incident wave is a Yanai wave or other mode of odd symmetry, then $A_1, A_3, \ldots$ are all 0.

Second, one should properly speak of a reflected wave disturbance rather than a reflected wave because the reflection of a single packet invariably consists of a number, perhaps an infinite number, of latitudinal modes, each with its own characteristic group velocity. The reflection of a localized wave packet will therefore inevitably segregate into many wave packets as each mode goes at its own speed.

Third, although the disturbance reflected from an eastern boundary is inevitably an infinite series, only a finite number of the modes of the reflection will be able to propagate westward into the interior of the ocean. The proof follows from recalling that the dispersion relation for k in terms of latitudinal mode number n and frequency ω) (which is the same for all modes, being invariant under reflection) is

$$k = -1/(2\omega) \pm \sqrt{\omega^2 + 1/(4\omega^2) - (2n+1)} \tag{6.29}$$

As n increases for fixed frequency, the radical must inevitably become imaginary, implying that *all* the *higher modes* are *zonally trapped*.

In a brilliant piece of classical mathematical analysis too complicated to repeat here, Moore [2] showed in his thesis that the zonally-trapped terms in the series could be approximately summed for large y, and that the sum is a slowly varying

coastal Kelvin wave propagating poleward along the eastern boundary of the ocean — actually *two* Kelvin waves of identical form propagating in opposite directions. The two Kelvin waves are in phase if the incident wave is symmetric about the equator, but the Kelvin waves are 180 degrees out of phase, i.e., are of opposite sign, when the incident wave packet is antisymmetric. The adjective "slowly varying' is necessary because the classic coastal Kelvin wave is usually derived on an f-plane, but the series of Hermite functions is a solution on the equatorial beta-plane where the Coriolis parameter is varying with y. The correct approximate form for the coastal Kelvin wave is therefore

$$v \sim [\text{constant}]\sqrt{y}\,\exp[i\,\{\omega t - \omega y + x/(2\omega)\} + yx]\,\{1 + O(1/y)\} \tag{6.30}$$

where it is assumed that the coordinate system is chosen so that $x = 0$ corresponds to the eastern boundary of the ocean. The e-folding scale (the Rossby radius of deformation) becomes smaller as one goes away from the equator (linearly with y), so the coastal Kelvin wave must narrow as it propagates poleward. Conservation of energy then forces the amplitude of the wave to grow as the square root of y so as to compensate for the narrowing of the wave in latitude as y increases. The reason for the square root dependence is that the energy is a quadratic function of the wave amplitude. This narrowing also implies that the phase lines must tilt with respect to the coast, which is why a factor of x appears multiplied by i in the exponential in (6.30).

These coastal Kelvin waves have all kinds of profound implications, both physical and mathematical. The major physical consequence is that the equatorial ocean cannot be considered a bounded system — energy always leaks away to high latitudes through coastal Kelvin waves running poleward along the east coast of the ocean. The mathematical implication is that one cannot simply choose a basin of fixed size, and then demand that the solutions vanish at the meridional sidewalls; the coastal Kelvin waves can make the amplitude as large on the northern and southern walls of the model as it is at the equator. In the remainder of this chapter, we will explore these consequences and approximations that are useful in understanding them.

6.5 The Meridional Geostrophy/Long Wave Approximation and Boundaries

The formulas of the preceding two sections are general but messy. When the frequency is small or when the zonal wavenumber k is small — these statements are equivalent for Rossby waves with westward group velocity— it is possible to greatly simplify them. This so-called "long wave","low frequency", or "meridional geostrophy" approximation, was derived and extensively discussed in Chap. 4. It is the equatorial equivalent of the quasi-geostrophic approximation.

The western boundary condition for flow with the long wave approximation is

$$\int_{-\infty}^{\infty} (u_{Kelvin} + u_i)\, dy = 0, \quad \text{at the western boundary} \tag{6.31}$$

i.e., the sum of the latitudinal integral of the incident Rossby wave and the reflected Kelvin wave must be zero.

This boundary condition (6.31) is convenient because, unfiltered by the assumption of meridional geostrophy, the long Rossby waves and the Kelvin waves become a closed set. The reflected Kelvin wave can be computed from (6.31) without computing the irrelevant short Rossby waves through the recurrence formula used earlier. Meridional geostrophy is conceptually valuable because it shows that short Rossby waves can only redistribute mass in the coastal boundary layer.

In the next section, we will give a practical example of the usefulness of the long wave approximation in obtaining simple, closed form solutions to an otherwise almost intractable problem.

6.6 Quasi-normal Modes: Definition and Other Weakly Non-existent Phenomena

In his thesis, Moore [2] calculated the eigenfrequencies and eigenfunctions of the free oscillations of an ocean-sized, rectangular basin that spanned the equator [2]. He showed that the poleward-traveling coastal Kelvin waves could turn the corner and run back towards the west as zonally-propagating coastal Kelvin waves along the northern boundary, then turn the corner again to travel towards the equator along the western boundary. Moore also solved the simpler problem of the free oscillations of a "strip" ocean which is unbounded in y; this is much less realistic because some [fictitious] energy source at high latitudes along the western boundary must pump energy into equatorward-traveling Kelvin waves to make up for what is being lost to higher latitudes on the eastern side of the sea via the poleward traveling Kelvin waves (Figs. 6.3 and 6.4).

In truth, it is thought that the closed basin problem is no more relevant to the actual ocean than the strip model. While the ocean basins are closed — the earth is a sphere with a finite surface area — dissipation will almost certainly clobber the coastal Kelvin waves as they make their long journey from Peru to Alaska and then run back down the coast of Siberia towards the equator again. Further, coastal Kelvin waves will be disrupted by the highly irregular twists and turns of real-life coasts and by local (estaurine and gulf) circulations, which are ignored in the idealized, inviscid, straight coastline model of Moore [2] [and the preceding section]. In reality, the energy taken off along the eastern boundary of the ocean by the poleward-traveling Kelvin waves is lost to the equatorial seas forever. Any waves excited in the equatorial region, therefore, will eventually damp out as the coastal Kelvin waves extract a certain percentage of the remaining energy on each trip of the wave pulse back-and-forth along the equator.

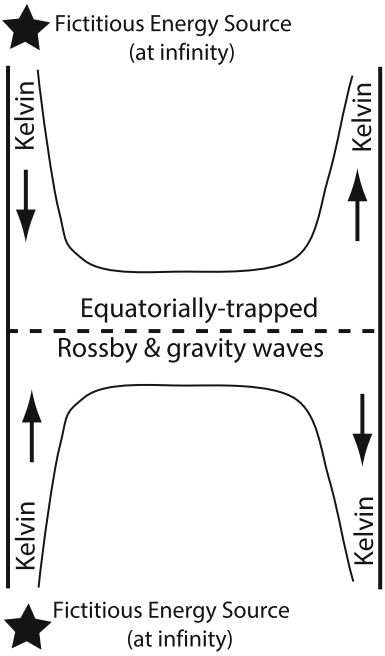

Fig. 6.3 Free oscillations in a strip ocean. To supply the losses from the equatorial zone due to Kelvin waves radiating away from the equator (*dashed line*) on the eastern side of the ocean, one must have fictitious energy sources at $y = \pm\infty$ (*stars*) to excite Kelvin waves moving towards the equator along the western coast

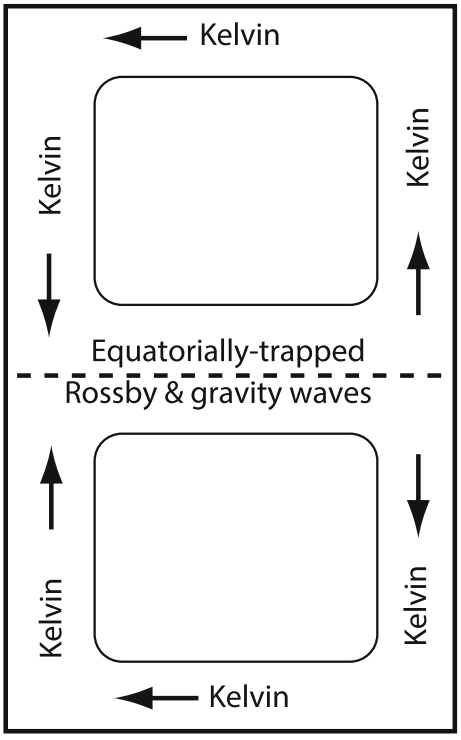

Fig. 6.4 Free oscillations in an ocean basin. The equator is *dashed*. *Arrows* show the direction of propagation of the Kelvin waves

However, there are never true free oscillations in the sea (or air) because every wave will lose energy to dissipation even if spared the radiative loss that depletes equatorial waves. The physical interest in inviscid, free oscillations is that when the rate of energy loss is sufficiently low — to whatever causes — a free mode is a near-resonant response of the ocean to external forcing. Thus, although an immortal equatorial wave pulse is impossible even in the limit of vanishing viscosity, it is still true that after being forced by changes in the wind stress, the equatorial ocean may well be dominated by quasl-normal modes, that is to say, by wave pulses whose frequencies and wavelengths are such that the rate of energy loss to coastal Kelvin waves is very low.

Such thinking motivates the following:

Definition 6.1 *(Quasi-normal Mode/Basin Resonance)* A solution to the inviscid shallow water wave equations is a true normal mode if it oscillates forever. A solution is a quasi-normal mode if it decays by irreversibly radiating energy out of the equatorial region to higher latitudes so slowly that the decay time scale is large compared to the period of oscillation.

"Basin resonance" is often used today in place of "quasi-normal mode". Because of dissipation, permanent free oscillations never exist. However, the response to forcing with a broad frequency spectrum may exhibit tall, narrow peaks at the a discrete set of "resonance frequencies". These match the eigenfrequencies of the normal modes (or quasi normal modes) as computed in an inviscid model. A forced boundary value problem is not, to a mathematician, the same as an eigenvalue problem, but the underlying physics is the same.

Cane and Moore found analytic quasi-normal modes of *low frequency* composed entirely of Rossby and Kelvin waves in the long wave approximation. Gent in collaboration with Semtner found *high frequency* quasi-normal modes. We will discuss each in turn in the following sections.

First, though, it is useful to set quasi-normal modes in a larger context. Similar slowly-decaying phenomenon are ubiquitous in nature including radioactive nuclei, externally-destabilized atoms, breathers in the ϕ^4 field theory and other weakly nonlocal solitary waves [3]. In addition, there are other phenomena which do not decay but share the property of being "beyond-all-orders nonexistent". This property means that the phenomenon such as the slow manifold is predicted by all orders of a power series in the small parameter ε, but the slow manifold is nonexistent because of effects which are exponentially small in $1/\varepsilon$. The slow manifold, for example, is nonexistent because there is an unavoidable fast gravity wave motion which is $O\ (\exp(-1/\varepsilon))$. It may offend the purist to generalize the concept of "normal mode" from an oscillation that is immortal in the absence of dissipation to a "quasi" normal mode that is ever-weakening with time even when viscosity and friction are completely neglected, but that is the way the world is.

A proton is an extreme example. Modern quantum theories assert that the proton is not a permanent state, but rather its intrinsic nature is to decay as surely as any quasi-normal mode of the ocean. On the other hand, the average lifetime of the proton

is larger than the expected lifetime of the universe! Measurement of proton decay would confirm some theoretical ideas, but otherwise is otherwise utterly irrelevant to human affairs.

Equatorial Rossby solitons of latitudinal mode number $n \geq 3$ are "weakly nonlocal" in the sense that they radiatively decay by emission of small amplitude sinusoidal equatorial Rossby waves of lower latitudinal mode number [3–6]. (See Sect. 16.18.) If the soliton amplitude is $O(\varepsilon^2)$, it is possible to compute a power series in ε to arbitrarily high order to capture the structure of the apparently nondecaying solitary wave. The decay is missed by the power series because it lies "beyond-all-orders", proportional to $\exp(-\beth/\varepsilon)$ where $\beth > 0$ is a constant. The decay rate is absurdly small — exponentially small — for small ε. Given that there is always dissipation, it seems silly turn a blind eye to viscosity while asserting that Radiative Decay Changes Everything, as some in the nonlinear waves community do, even now.

The quasi-normal modes discussed in the rest of the this chapter are considerably more "quasi" in the sense that the decay rates are not, alas, exponential functions of anything. Nevertheless, basin modes whose decay scales are long compared to the oscillation period are normal modes in a practical sense, and we must deal with them.

6.7 Quasi-normal Modes in the Long Wave Approximation: Derivation

Using the long wave approximation, Mark Cane and Dennis Moore independently derived explicit solutions for such quasi-free modes and published their results in a joint paper (Cane and Moore, [7]). In the remainder of this section and the next, we will derive and discuss their results. In Sect. 6.9, we will describe other efforts to explore quasi-free oscillations by Peter Gent and others.

The derivation begins by calculating the eastern boundary reflection of a Kelvin wave in the long wave approximation using the Rossby wave dispersion relation

$$k_n = -(2n+1) \tag{6.32}$$

The coefficients of the reflected long Rossby waves simplify to

$$\begin{aligned} A_1 &= A_{Kelvin} \sqrt{2}\,(k_1 + \omega) \\ &= -2^{3/2}\,\omega\, A_{Kelvin} \end{aligned} \tag{6.33}$$

and the general recurrence

$$\begin{aligned} A_{n+2} &= A_n \sqrt{(n+1)/(n+2)}\,(k_{n+2} + \omega)/(k_n - \omega) \\ &= A_n \sqrt{(n+2)/(n+1)} \end{aligned} \tag{6.34}$$

where we have made no approximations in (6.33) or (6.34) except the longwave approximation (6.32).

A convenient notational trick is to introduce the coefficients Λ_n defined by

$$\Lambda_n = \begin{cases} \sqrt{\frac{3\cdot5\cdot7\cdots n}{2\cdot4\cdot6\cdots(n-1)}}\, n \text{ odd}, n \geq 3 \\ \\ 0, \quad n \text{ even} \\ 1, \quad n = 1 \end{cases} \tag{6.35}$$

The Λ_n, which Cane and Moore denote by the more cumbersome notation κ_{1n}, obviously satisfy the recurrence (6.34). These coefficients will appear often in equatorial calculations because they are proportional to the expansion coefficients of y, i.e.,

$$y = 2\,\pi^{1/4} \sum_{n=1, n \text{ odd}}^{\infty} \Lambda_n \psi_n(y) \tag{6.36}$$

$$= \sum_{n=1, n \text{ odd}}^{\infty} a_{2n+1} \psi_{2n+1}(y) \tag{6.37}$$

$$a_{2n+1} = \pi^{1/4} \frac{\sqrt{(2n+1)!}}{2^{n-1}\, n!} \tag{6.38}$$

$$\sim 2.378\, (2n)^{1/4} \qquad [n \to \infty] \tag{6.39}$$

In terms of these coefficients Λ_n,

$$A_n \equiv -2^{3/2} \omega\, \Lambda_n, \qquad n \geq 1 \tag{6.40}$$

The reflection of the Kelvin wave is now given by (top sign refers to u, bottom to ϕ)

$$\left.\begin{matrix} u_r \\ \phi_r \end{matrix}\right\} = -2^{3/2}\, \omega\, A_{Kelvin} \sum_{n=1; n \text{ odd}}^{\infty} \Lambda_n \exp(-i\omega t) \exp(-i\omega(2n+1)x) \times$$
$$\left\{ \frac{\sqrt{(n+1)/2}}{k_n - \omega} \psi_{n+1} \mp \frac{\sqrt{n/2}}{k_n + \omega} \psi_{n-1} \right\} \tag{6.41}$$

where we have already employed (6.32) to replace k_n in the exponential in (6.41).

Introduce the new "stretched" x coordinate

$$s \equiv \omega\,(x - x_E) \tag{6.42}$$

where x_E is the location of the eastern boundary. This definition is convenient because all the wavenumbers k in the long wave approximation are proportional to ω, which

can then be absorbed into the new coordinate. In (6.41), it is implicitly assumed, as in earlier sections, that the eastern boundary is at $x = 0$, but for purposes of applying the western boundary condition below, it is convenient to shift $x = 0$ to correspond to the *western* boundary. Since the phase of the Kelvin wave and its reflection is arbitrary anyway, we can replace x by $(x - x_E)$ without altering anything else in (6.41).

Use of (6.32) for the k_n in the second line of (6.41) shows

$$\begin{aligned}\sqrt{(n+1)/2}/(k_n - \omega) &= \sqrt{(n+1)/2}/(-(2n+1)\,\omega) \\ &= 1/[-\omega\, 2^{3/2}\,\sqrt{n+1}]\end{aligned} \tag{6.43}$$

$$\begin{aligned}\sqrt{n/2}/(k_n + \omega) &= \sqrt{n+/2}/(-2n\,\omega) \\ &= 1/[-\omega\, 2^{3/2}\,\sqrt{n}]\end{aligned} \tag{6.44}$$

The factors of $[-\omega\, 2^{3/2}]$ in (6.43) and (6.44) cancel the same factor which appears in (6.41). The solution for u and ϕ (including the Kelvin wave) becomes

$$\left.\begin{matrix} u \\ \phi \end{matrix}\right\} = A_{Kelvin}\exp(-i\omega t)\,\{\psi_0(y)\,\exp(is) + \sum_{n=1;n\ \text{odd}}^{\infty} \Lambda_n\left[(n+1)^{-1/2}\psi_{n+1}(y) \mp n^{-1/2}\,\psi_{n-1}(y)\right]\exp(-is(2n+1)\right\} \tag{6.45}$$

What is remarkable about (6.45) is that with a couple of further transformation, the infinite series can be summed in closed form. First, we collect degrees of Hermite functions to obtain

$$\left.\begin{matrix} u \\ \phi \end{matrix}\right\} = A_{Kelvin}\exp(-i\omega t)\,\{\psi_0(y)\,[\exp(is) \mp \Lambda_0\exp(-i3s)] + \sum_{n=2;n\ \text{even}}^{\infty} \psi_n(y)[\Lambda_{n-1}\,n^{-1/2}\exp(-is(2n-1)) \mp \Lambda_{n+1}\,(n+1)^{-1/2}\exp(-is(2n+3))\,]\right\} \tag{6.46}$$

By using the recurrence (6.34), which implies

$$\Lambda_{n-1}\,n^{-1/2} = \Lambda_{n+1}\,(n+1)^{-1/2} \tag{6.47}$$

we can rewrite (6.46) as

$$\left.\begin{matrix} u \\ \phi \end{matrix}\right\} = A_{Kelvin} \exp(-i\omega t) \sum_{n=0;n \text{ even}}^{\infty} (n+1)^{-1/2} \; \Lambda_{n+1} \, \psi_n \, \exp(-is(2n+1)) \; \times \\ \{\exp(2is) \mp \exp(-2is)\} \tag{6.48}$$

The crucial point about the form (6.48) is that the sums for u and ϕ differ only in a term which is *independent* of n and therefore can be taken *outside* the sum. Taking the ratio of u/ϕ eliminates all the y dependence to give

$$u = i \tan(2s)\, \phi, \tag{6.49}$$

which is equivalent to (10) of Cane and Moore [7] except for some differences of notation. [Cane and Moore assume the mode is proportional to $\exp(i\omega t)$ whereas we have kept the usual $\exp(-i\omega t)$. This reverses the sign of all factors of s that appear in (6.48) because s is proportional to $(x - x_E)$ times ω, and thus is sign-flipped by a different convention about the sign of ω. Also, Cane and Moore use h in place of our ϕ.]

Equation (6.49) is a philosopher's stone for this problem, turning brass into gold, because all the modes in the previous equation — Kelvin and long Rossby — satisfy the long wave approximation

$$yu + \phi_y = 0 \tag{6.50}$$

Multiplying (6.49) by y and subtracting it from (6.50) eliminates u to give the first order ordinary differential equation for ϕ,

$$\phi_y + \{\, i \tan(2s)\, y\}\, \phi = 0 \tag{6.51}$$

with the general solution

$$\phi = \eta(s) \, \exp(-i \tan(2s)\, y^2/2) \tag{6.52}$$

Equation (6.51) is a differential equation in y alone, so its solution is proportional to an arbitrary function of s, $\eta(s)$. It may be determined by evaluating the x-momentum equation at $y = 0$, which is

$$-\, i\, \omega\, u \, + \, \omega\, \phi_s = 0 \tag{6.53}$$

where we have used $\partial_x = \omega \partial_s$ as a result of the definition of s. Replacing u by $[i \tan(2s)\phi]$ as given by (6.49) converts (6.53) into a differential equation in ϕ alone. Dividing out the common factor of ω, we have

$$-\, i \tan(2s)\phi + \, \phi_s = 0 \tag{6.54}$$

which is really an equation for $\eta(s)$ since $\phi(s, y = 0) = \eta(s)$. The solution is (to within an arbitrary multiplicative constant)

$$\eta(s) = \sqrt{\cos(2s)} \tag{6.55}$$

A solution for v can be obtained by evaluating the x-momentum equation for non-zero values of y. Let us write the height in the form

$$\phi = \eta(s) \exp(-i\,\Phi(s, y)\,) \tag{6.56}$$

The x-momentum equation becomes

$$yv = \omega\,\{-iu + \phi_s\} \tag{6.57}$$

$$= \omega\,\{-i[i\tan(2s)\,(\,\eta\,\exp(-i\Phi)\,)\,] + \exp(-i\Phi)\,\eta_s + \eta\,(\exp(-i\,\Phi)\,)_s\,\} \tag{6.58}$$

$$= \omega\,\exp(-i\Phi)\,\{[\tan(2s)\,\eta + \eta_s] - i\,\eta\,\Phi_s\} \tag{6.59}$$

Now the two terms in square brackets [ ] in (6.59) are the only surviving terms at $y = 0$ where $\Phi = \Phi_s = 0$, so they must sum to 0. (The vanishing of their sum is in fact the differential equation that determined $\eta(s)$.) Therefore, since $\Phi = \tan(2s)y^2/2$,

$$yv = -i\omega\,\eta\,\exp(-i\Phi)\,(\Phi_s) = -\,i\omega\phi\,\left\{(1/2)y^2\,[\tan(2s)]_s\right\} \tag{6.60}$$

where we have used the definition of ϕ in terms of η and s in the second half of (6.60). Dividing out the common factor of y gives finally

$$\phi = \sqrt{\cos(2s)}\,\exp\left\{-i\tan(2s)\,y^2/2\right\} \tag{6.61}$$

$$u = i\,\tan(2s)\phi \tag{6.62}$$

$$v = -i\omega y\,\sec^2(s)\,\phi \tag{6.63}$$

where we recall that $s = \omega(x - x_E)$.

Equations (6.61)–(6.63) give the complete so1ution for an equatorial Kelvin wave and for the long Rossby waves that are its reflection from the eastern boundary. Thus, it automatically satisfies the boundary condition at $x = x_E$. To be a normal mode of the ocean basin, however, it must also satisfy the boundary condition at $x = 0$, which is the western side of the ocean. The long wave boundary normal condition is

$$\int_{-\infty}^{\infty} u(x = 0, y)dy = 0, \tag{6.64}$$

but because of the special form of (6.61)–(6.63), the only way that (6.64) can be true is if u satisfies the exact boundary condition

$$u(x = 0, y) = 0 \tag{6.65}$$

(There are other problems — those with forcing, for example — where we will have to fall back to (6.64) when using the long wave approximation, but we will come to those at a later time). Equation (6.65) demands $\tan(2s) = 0$, or in other words, since $s(x = 0) = -\omega x_E$,

$$\omega x_E = m\pi/2, \qquad m = \text{integer} \tag{6.66}$$

or equivalently in terms of temporal period P

$$P = 4x_E/m, \qquad m = 1, 2, 3, \ldots \tag{6.67}$$

Since the Kelvin wave has unit nondimensional speed and the width of the basin is x_E, it follows that the mode with the lowest frequency ($m = 1$) has a period equal to four times the length of time it takes a Kelvin wave to cross the basin. Equiva1ently, the $m = 1$ mode has a period equal to the length of time required for a Kelvin wave to cross the basin from west to east plus the time for the lowest, $n = 1$ Rossby mode to return. The careful reader is probably bothered by all sorts of questions at this time.

How could we write the eastern boundary reflection of the Kelvin wave as an infinite sum of long, westward propagating Rossby wave when we showed in Sect. 6.4 that all the high order reflected waves are zonally trapped, and form the coastal Kelvin wave? How are we to interpret the singularities of u and ϕ for $s = \pi/4 + j\pi/2$ where j is any integer? How did we make the "leakiness" of the equatorial ocean to higher latitudes disappear? 1n the next section, we will try to answer these questions, and discuss the Cane-Moore normal modes in detail.

6.8 Quasi-normal Modes in the Long Wave Approximation: Discussion

The longwave approximation falsifies the Kelvin wave's eastern boundary reflection by forcing it to consist solely of long, eastward-propagating Rossby waves. In reality, the full dispersion relation

$$k = -1/(2\omega) + \sqrt{\omega^2 + 1/[4\omega^2] - (2n + 1)} \tag{6.68}$$

shows that the radical always becomes imaginary for

$$n \gtrsim N(\omega) \equiv 1/(8\omega^2) \tag{6.69}$$

The sum of the zonally trapped waves with complex k was shown by Moore [2] to be a coastal Kelvin wave propagating away from the equator. Equation (6.68)

implies that while the breakdown must always occur, it happens only for very large n if $\omega << 1$. This makes it at least plausible that the amount of energy which leaks poleward in the form of coastal Kelvin waves is very small for $\omega << 1$. Cane and Moore [7] quote the loss rate per period P as

$$\frac{\Delta \text{ Energy}}{\text{Energy}} \approx (N+1)\, A_N^2 \approx \frac{2}{\sqrt{2\pi N}}, \tag{6.70}$$

but do not give a detailed derivation. They do, however, convert this into a decay time $1/r$, measured in periods of the oscillation, which is shown for various (inviscid) modes in a table in their paper. The decay time/period ratios for the four lowest vertical modes in each ocean range from 5 to 10 in the Indian Ocean, 6–11 in the Atlantic, and 13–25 in the Pacific where in each case the first baroclinic mode is the leakiest (smallest ratio of decay time scale to oscillation time scale).

The tabulated energy loss rates suggest a couple of important points.

The decay rates are always small in comparison to a period, which implies that neglecting the energy loss to coastal Kelvin waves, as we have done in the analytical long wave modes derived in the previous section, is always a sensible first approximation.

These "radiative" decay rates $1/r$ are comparable with or smaller than the e-folding scales produced by reasonable dissipation in the ocean. In other words, dissipation is just as important and probably more important for these modes. Consequently, it is as reasonable to study these "leaky" normal modes of the equatorial ocean, which in reality lose energy to both dissipation and coastal Kelvin wave radiation, as it is to study the free normal modes of Lake Michigan or other confined bodies of water, which lose energy to dissipation [only].

For equatorial basin modes, the fact that the energy e-folding scale is large in comparison to the period implies that the energy loss rate is small enough so that these modes are physically important.

Another question, however, is what to do about the singularities of the Cane–Moore solutions at $s = \pi/4, 3\pi 4, 5\pi/4$, etc.? Cane and Moore [7] point out that the addition of any form of dissipation will remove these singularities. Mathematically, the simplest choice is linear or "Rayleigh" friction, which involves adding $-ru$ to the x-momentum equation, $-r\phi$ to the height equation and $-rv$ to the y-momentum equation. [The Rayleigh friction $-rv$ in the y-momentum equation is neglected in the long wave approximation.] Although there is no physical justification for this form of the dissipation — it should be some form of turbulence parameterization, but turbulence is after all understood only by God — it is a useful conceptual tool because it is formally equivalent to the replacement

$$\omega \to \hat{\omega} \equiv \omega - i\, r \qquad (r = \text{constant}) \tag{6.71}$$

Recalling that $s \equiv \omega\,(x - x_E)$, this implies that

$$\tan(2s) \to \tan\{2(s - ir(x - x_E))\} = \frac{(1-\mu^2)\tan(2d) + i\,\mu\sec^2(2s)}{1+\mu^2\tan^2(2s)} \quad (6.72)$$

where s on the R.H.S. of (6.72) is to be interpreted as $(x - x_E)$ multiplied by ω (not the complex frequency defined by Eq. (6.71), $\hat{\omega}$), and where

$$\mu \equiv -\tanh\{2\,r\,(x - x_E)\} \quad (6.73)$$

If we assume $r << 1$ so that $\mu << 1$, then away from the caustics where $\tan(2s) = \omega$, the R.H.S. of (6.72) is approximately equal to $\tan(2s)$, i.e., the damping has only a negligible effect. Near these singularities ("near" means $|\tan(2s)| > 1/\mu$), the solution is heavily damped because $\tan\{2(s - ir(x - x_E)\} \approx i/\mu$. In this highly dissapative region, $\exp[i(y^2/2)\tan(2s)]$ becomes a rapidly decaying exponential in y.

The abbreviated version of the argument above is that by effectively making the frequency slightly complex, we have shifted the singularity slightly off the real s axis so that it occurs only for imaginary values of s: $s = \pi/4 + i\mu$, to be precise. This slight shift — and the damping — are only a small correction to the solution except within the small neighborhood $s = \pi/4 + 0(\mu)$.

The inclusion of linear friction is a useful generalization of the theory given in the previous section because it makes it possible to not only remove the unphysical singularities at $s = \pi/4$, etc., but also offers a direct numerical reassurance about the neglect of the coastal Kelvin wave. Figures 4, 5, 6 and 7 of Cane and Moore [7] compare the exact solution for the reflection of a low frequency Kelvin mode (coastal Kelvin wave included) with the corresponding long wave solution at various latitudes and various values of the friction coefficient. The panels show that out to $y = 5$ at least, which is 15 degrees of latitude for the first baroclinic mode, the exact and long wave reflections are qualitatively similar even with no dissipation at all. When a weak dissipation with $r = (1/10)\omega$ is included, the two become identical. Since the damping time scale of the ocean is thought to be only a few years and $\omega = 1/30$, used to make the figures, corresponds to a period slightly longer than a year, $r = 0.1\omega$ is actually likely to be too little damping (compared to the real ocean) rather than too much. Thus, the extremely good agreement between the exact and the long wave graphs in the bottom sections for $y = 0$, $y = 1$, and $y = 5$ implies that we will really sacrifice nothing by making the long wave approximation.

Figures 6.5 and 6.6 show the actual amplitude ϕ and u for the long wave free modes. The height field is bounded, but both u and v (where the latter is not illustrated) have poles that rise to infinity at $s = -\pi/4$, $y = 0$. The phase patterns are illustrated by Fig. 6.7, which shows the real part of ϕ, which is similar to that of the other fields. The dissipation was deliberately chosen to be very weak to bring out the character of the inviscid solution. Although the graphs exhibit only the lowest, $m = 1$ mode explicitly, the periodicity of the solutions with respect to s — note $\phi(s - \pi/2) = i\eta(s)$ and similarly for u and v — implies that the solutions for $m > 1$ can be pictured by repeating these figures with the proper phase shifts.

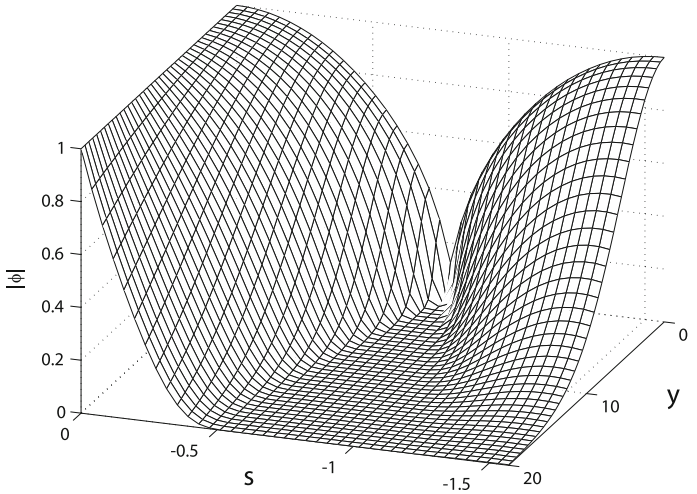

Fig. 6.5 $|\phi|$ for the lowest Cane–Moore free oscillation of the equatorial ocean, viewed looking towards the equator from the northwest. The flat wedge is the shadow zone where none of the rays of the normal mode penetrate; the point of the wedge at the equator is the focus. The longitudinal coordinate is s; the whole width of the ocean is illustrated

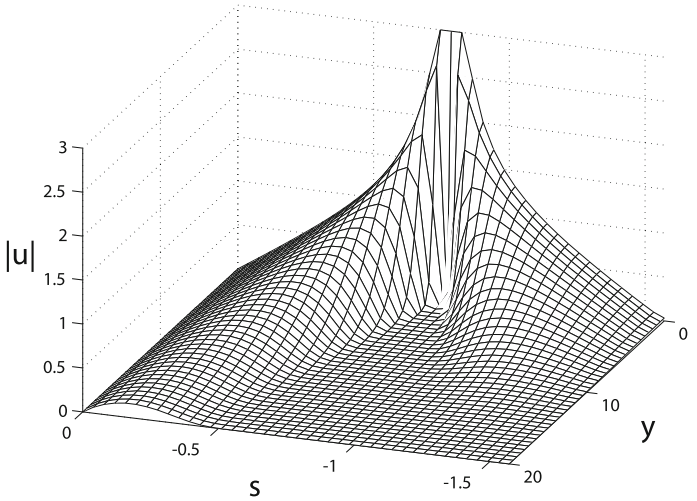

Fig. 6.6 Cane–Moore lowest mode, same as previous figure except the absolute velocity of the east–west velocity is shown. Both u and v (not shown) have poles at the focus, so the amplitude of $|u|$ was truncated at three to avoid infinities

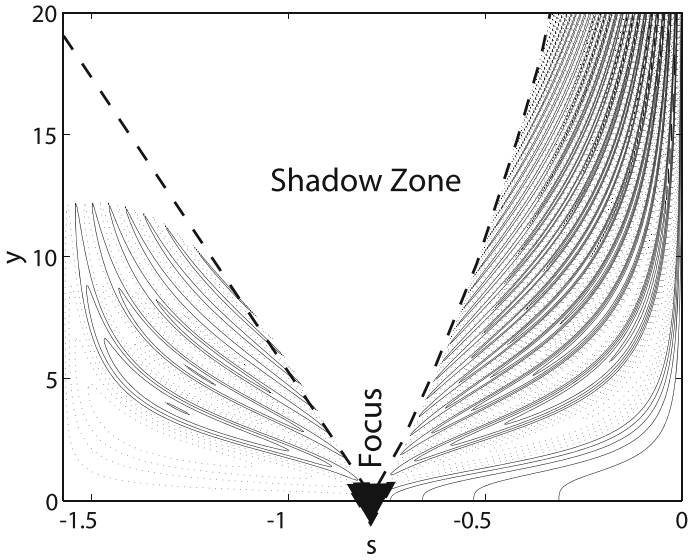

Fig. 6.7 Cane–Moore lowest mode. Same case as previous figure contours of $\Re(\phi)$ are shown with the contour intervals at $[-0.9 : 0.2 : 0.9]$. The *dashed line* bounds the shadow zone where the wave amplitude is zero. The triangle marks the focus, which is *right* on the equator at $s = -\pi/4$, that is, exactly in the middle of the ocean

The ray-tracing theory of Chap. 5 gives the physical reason for both the V-shaped pattern of phase lines seen in the phase plots, radiating like a fan from $s = -\pi/4$, $y = 0$, and also for the very large amplitude of u at that point. Equation (5.48) showed that the trajectory of the rays in the $x - y$ plane is given by

$$y = \sqrt{2n+1}\,\sin[x/\{k + 1/(2\omega)\} + \Phi] \tag{6.74}$$

where Φ is an arbitrary phase factor. In the long wave approximation, we must neglect k in comparison to $1/(2\omega)$ with an error $O(\omega^2)$ to obtain

$$y = \sqrt{2n+1}\,\sin[2\omega x + \Phi] \tag{6.75}$$

Thus, all rays follow a trajectory of the same sinusoidal shape, differing only in magnitude.

If we choose the phase factor Φ to be the same for all n, then all the rays — in the long wave approximation — will cross the equator at the *same point*. The particular choice $\Phi = \pi/2$ causes the focus to occur at $s = -\pi/4$, which is precisely where u is very large and also the point from which the fan of constant phase lines radiates. The trajectories of a few rays are sketched schematically in Fig. 6.8.

Figures 6.9 and 6.10 show the Cane–Moore modes in the space-time plane instead of longitude-latitude plane illustrated in previous figures.

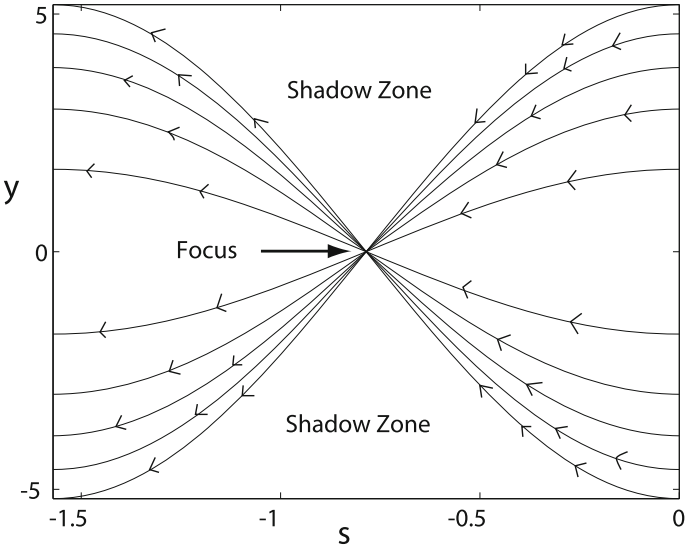

Fig. 6.8 Cane–Moore lowest mode, same as previous figure except showing the rays. The ray-tracing approximation, which gives the ray curves as $y(s) = \sqrt{2n+1}\,\sin(2s+\pi/2)$, is not exact and produces a spurious symmetry with respect to reflection about the middle of the ocean, $s = -\pi/4$. Otherwise the ray pattern closely resembles the Cane–Moore analytical approximation. The wave looks symmetric in longitude with respect of the focus, but the previous figure shows that there should be a slight asymmetry

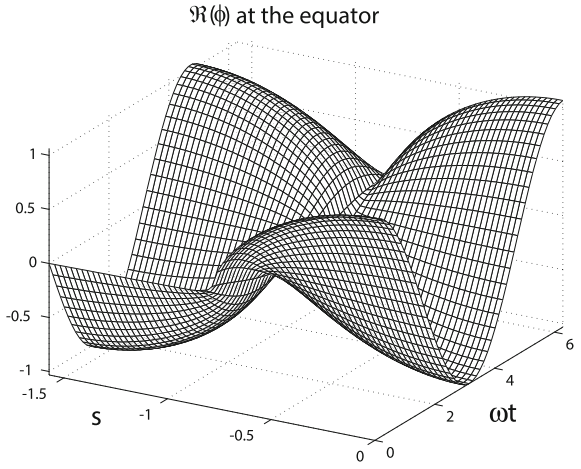

Fig. 6.9 Cane–Moore normal modes in the space-time plane [not longitude-latitude plane as in previous figures]: ϕ at the equator versus s (rescaled longitude) and the product of ω with time

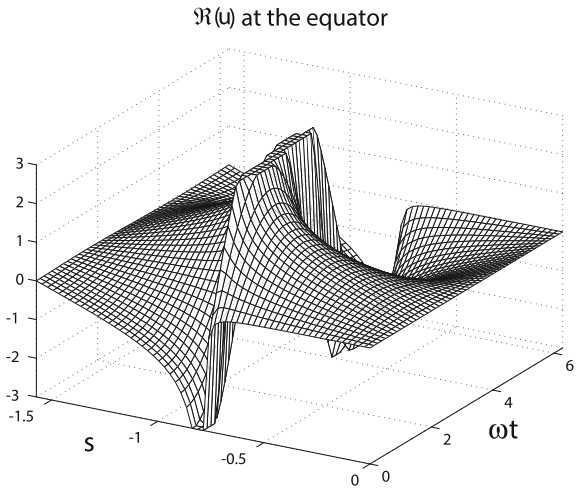

Fig. 6.10 Same as the *left panel* except it shows u

Equation (6.74) and the focus evident in Figs. 6.5, 6.6 and 6.7 show that when the Kelvin wave bounces off the eastern boundary, it effectively converts the boundary into a wavemaker. We could create the same westward pattern of Rossby waves (without the Kelvin wave) if we replaced the boundary by a north–south column of wavemakers, all synchronized to generate wave packets in phase.

What we have not done is explained precisely why all the wave packets generated by the boundary reflection of the Kelvin wave must be in phase. To do this requires the sophisticated ray-tracing study of Schopf, Anderson and Smith [24], and we shall explore the more detailed results of their paper — some of the elementary ray formulas for a single wave packet given earlier were also taken from their work — when we consider the response to periodic wind-forcing in the next chapter.

For now, it is sufficient to note that through the longwave approximation, we have obtained analytical normal modes without ray-tracing.

6.9 High Frequency Quasi-free Equatorial Oscillations

Moore [2] showed that the free oscillations of a rectangular basin or open strip spanning the equator are not equatorially trapped because coastal Kelvin waves carrry energy away from low latitudes [2]. Gent [8] observed that it was possible to impose the condition that the coastal Kelvin waves are zero by allowing the width of the ocean L to be a free eigenparameter (in addition to the frequency ω, which is also kept as an eigenvalue.) In physical terms, Gent's standing modes are a finite sum of equatorial waves whose phases and amplitudes are in such a ratio that the coastal Kelvin waves exactly cancel.

The catch is that the width of the L must be (mathematically!) taken as variable. A geologist would hardly approve! In Moore's scheme, there are always two modes (one symmetric about the equator and one antisymmetric) in the latitudinally unbounded ocean for any value of the frequency ω.[3] Gent's mode exists only for *discrete* set of pairs of the values of (ω, L). As a mathematician would put it, the allowed discrete values of L are a set of "measure zero" in comparison to the continuous intervals of L between the eigen-L's for which no modes exist; expressed differently, there is a zero probability that the width of any of the real oceans is *exactly* equal to one of the free mode values of L.

The obvious rejoinder is that we are interested in the ocean rather than mathematics, and oceanography is no more an "exact" science than medicine or economics. In reality, Gent's approach is just as legitimate as that of Cane and Moore because if the width of the ocean is *close* to that of a free mode — not exactly equal, but only close — the ocean will still exhibit a quasi-resonant response at roughly the frequency of the free oscillation its width is "close to".

For brevity, we will not attempt to derive Gent's dispersion relation in detail, but the method is to apply the eastern boundary reflection recurrence formulas derived earlier in the chapter. The simplest of his standing modes consists of the Kelvin wave and either two $n = 1$ gravity waves or two $n = 1$ Rossby waves. For either, he shows that the condition on L is

$$L = \sqrt{p(p+m)} \qquad \text{for any nonzero integers } m,\ p \tag{6.76}$$

Figure 6.11 shows the values of ω for the two types of Kelvin-plus-two-($n = 1$) modes when the integers m and p are both 1. The integer p is the number of complete wavelengths (across the basin) of the relative phase of the Kelvin wave and the shorter of the two $n = 1$ modes while the integer m is the number of wavelengths of the relative phase between the two $n = 1$ modes.

In physical terms, only one member of each triplet of waves that make up the normal mode is carrying energy westward – the *long* Rossby wave in the case of the lower frequency mode. Two waves, the Kelvin and the short Rossby wave, carry energy eastward. Individually, a Kelvin wave generates an infinite series of coastally trapped modes — the coastal Kelvin wave — when it reflects. Individually, the short Rossby wave does the same. Together, however, the coastal Kelvin wave excited by the reflection of the short Rossby wave exactly cancels that excited by the equatorial Kelvin wave, and the result is that the sole survivor in the reflection of the sum is a single long Rossby wave.

The essential role of the short Rossby wave shows that Gent's approach is complementary to that of Cane and Moore: this short Rossby wave would be filtered by the long wave approximation employed by the latter. For low frequency motions, the long wave approximation is legitimate, which strongly suggests that Gent's modes

[3] In other words, for the "strip" ocean, we have a *continuous* spectrum of eigenmodes. When the north and south of the ocean are walled off so that the ocean becomes a basin, then the spectrum becomes *discrete* as for any other species of water waves in a basin.

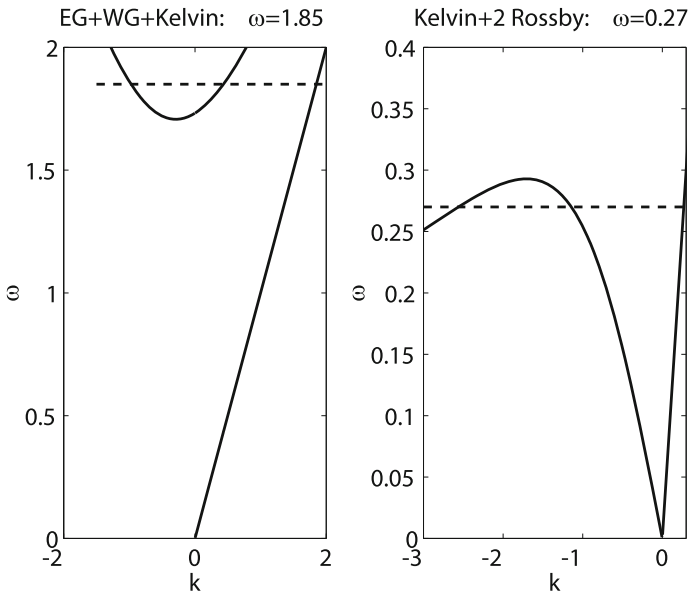

Fig. 6.11 Dispersion relations for the two lowest quasi-normal modes of [8]. *Left* the horizontal dividing line at the mode's frequency of $\omega = 1.85$ intersects the $\omega(k)$ curves of the three participating modes: a westward-traveling $n = 1$ gravity wave, an eastward-traveling gravity wave of the same latitudinal mode, and a Kelvin wave. The *right panel* is the same except for the mode whose frequency is $\omega = 0.27$; this consists of a short (high $|k|$) Rossby wave, a long Rossby wave, and a Kelvin wave. In each panel, the Kelvin wave has the linear dispersion relation, $\omega = k$ for $k > 0$

are of much higher frequency than those found by Cane and Moore. Indeed, the Rossby mode in Gent's Fig. 2 still has a period of only 36 days, which is much shorter than any of the Cane–Moore modes which all have periods of at least 120 days. Gent's gravity wave mode in his Fig. 3 has a period of only 5 days.

The two modes composed of a Kelvin wave plus a pair of $n = 1$ waves are hardly the only types possible; Gent's [8] paper gives examples of many others formed from a Yanai wave plus an $n = 1$ and $n = 3$, etc., and of modes formed from four, five, or many different waves.

Gent's illustrations bring up an interesting point: there is apparent zonal phase propagation in these waves, even though they are standing modes trapped between zonal boundaries. The reason for this is constructive and destructive interference between the three zonally propagating waves, each with its own phase speed, that collectively make up the standing mode. On can follow a node as it propagates from one side of the ocean to the other as shown in his Fig. 3, but the apparent phase velocity so defined is not spatially uniform, being larger in the center of the basin than near the boundaries. The average phase speed is about 0.11 that of the Kelvin wave or equivalently 0.47 of the long Rossby wave — that is to say, it is different from the phase speeds of all the waves which make it up. Gent's mode thus serves

as a warning that sums of waves can mimic the phase behavior of a single wave, and thus lead to totally erroneous identifications of different wave modes in ocean data. If the observations were sufficiently bountiful, of course, it would be possible to avoid such errors by looking carefully at the latitudinal structure of the mode. However, oceanographic data is always sparse, so sometimes apparent phase propagation is all that can be inferred. What appears to move may not be a single wave, but rather the pattern of interference between several different waves.

Gent and Semtner [9] numerically tested Gent's [8] theory [8, 9]. By artificially imposing a large friction along the northern and southern walls of their computational basin, they insured that any energy that escaped the equatorial region in the form of coastal Kelvin waves would be lost forever.

McPhaden and Semtner [10] have added mean zonal flows to the picture, and find that the low-frequency standing modes — those made of Rossby waves plus the Kelvin wave or Yanai wave — become very leaky in the presence of mean shear [10].

Later work has established the existence of significant basin resonances at 90-day and 180-day (semiannual) periods in the equatorial Indian Ocean. Han, McCreary, Masumoto, Vialard and Duncan [11] give a good review of previous work and then apply a hierarchy of models to show that the resonances are robust. The Maldive Islands weaken the semiannual oscillation, but lie in a node of the higher frequency quasi-normal mode and therefore have little effect on the 90-day resonance.

To leak away as a coastal Kelvin wave or not to leak away? That is still a question.

6.10 Scattering and Reflection from Islands

Scattering of waves by islands has a long history that extends from non-equatorial contributions by Meyer and Painter [12] to later contributions like Pedlosky et al. [13] and Spall and Pedlosky [14].

Yoon [31] showed through numerical shallow water simulations that Kelvin waves always get through, even when the island is of unit nondimensional width. Neither the Maldive Islands in the Indian Ocean nor the Gilberts in the Pacific have much effect on Kelvin waves. In contrast, an island on the equator can reduce transmission of Rossby waves to less than one-third the incident energy.

The appendix of Spall and Pedlosky [14] presents a theory that reiterates this point: the fundamental assumption is that because an equatorial Kelvin wave can penetrate a narrow gap so freely, there is no reflected Kelvin wave when a long Rossby wave forces a narrow gap propagating in the opposite direction. All the mass flux forces the gap, too, since the reflected short Rossby waves carry negligible mass flux. The mass flux through the gap is idealized as a Dirac δ-function. (The gap was 75 km or 1/4 nondimensional length units in their extensive numerical calculations.) An incident $n = 1$ latitudinal mode Rossby wave

$$v = A_1 \exp(k_i x - i\omega t) \tag{6.77}$$

yields the mass flux on the far side of the gap

$$\mathscr{F} \equiv \int_{-\infty}^{\infty} \phi u dy \tag{6.78}$$

$$= \sum_{j=1,\, j odd}^{J} \frac{|A_1|^2}{4} \left\{ \frac{j+1}{\omega - k_j)^2} - \frac{j}{\omega + k_j)^2} \right\} \tag{6.79}$$

where J is the largest (odd) mode number for which the dispersion relation gives real-valued $k_J(\omega)$:

$$k_j = -\frac{1}{2\omega} + \frac{1}{2}\sqrt{\left(\frac{1}{\omega} - 2\omega\right)^2 - 8j} \tag{6.80}$$

Analysis of island effects is still a topic of active research for both tropical and extra-tropical latitudes, even today, but theory, numerical studies and laboratory experiments are addicted to circular islands, infinitesimally thin islands, square and rectangular islands. Clarke [15] provides at least a veneer of respectability for such idealizations. His motive is that the "State of the art models of El Niño/Southern Oscillation (ENSO) events …suggest that reflection of low-frequency energy from the western Pacific boundary is crucial to ENSO dynamics: if too little reflection occurs, an interannual oscillation does not develop." This motivates him to consider reflection of low-frequency waves with roughly annual period.

Since the nondimensional timescale of the first baroclinic mode of the ocean is a couple of days, unit nondimensional frequency corresponds to a period of a couple of weeks. It follows that the annual frequency is very small. The dispersion relation for Kelvin waves, $c = k/\omega = 1$ implies $k = \omega \ll 1$ is always very small for Kelvin waves (and by similar reasoning, Rossby waves) of such low-frequency. This justifies his statement in his abstract that "At low frequencies, equatorial Kelvin and Rossby waves have very large East–West scales ….Consequently, these land masses may be treated as islands that are infinitesimally thin in the East–West direction."

His article generalizes previous single-island low frequency theory by Cane and du Penhoat [32] to as many as seven islands. His results are too extensive to be reviewed in detail here, but one noteworthy feature is direct engagement with "Sverdrup jumps", although he does not use this terminology [16].

Classic Sverdrup flow [1, 17–19] satisfies the ordinary differential equation

$$\beta\, \psi_x = \text{curl-of-the-wind-stress} \Leftrightarrow v = \text{curl-of-the-wind-stress} \tag{6.81}$$

where the left is the familiar quasi-geostrophic midlatitude balance and the right is the equivalent nondimensional statement on the equatorial beta-plane. Because the ordinary differential equation for ψ is first-order, only a single boundary condition may be imposed. Ocean texts explain that the correct boundary condition is that

$$\psi = \chi \text{ at } x = x_E(y) \tag{6.82}$$

where χ is a constant, usually set equal to zero on the boundary of the ocean basin, and where $x_E(y)$ is the longitude of the basin boundary at the eastern side of the ocean [1, 17, 18].

An island or peninsula creates a discontinuity in Sverdrup flow. Let y_N denote the northern boundary of an island, and let $x_I(y)$ denote the western boundary of the island. Then the longitude where the boundary condition is applied jumps discontinuously by hundreds or thousands of kilometers at $y = y_N$:

$$\psi(x = x_E(y), y) = 0, \qquad y > y_N \tag{6.83}$$

$$\psi(x = x_I(y), y) = \chi, \qquad y < y_N \tag{6.84}$$

The island boundary condition is $\psi = \chi$ where χ is a constant.

An equatorial beta-plane treatment of Sverdrup flow is given in Sect. 8.3, but the Sverdrup jumps at the northernmost and southernmost tips of an island are similar regardless of latitude.

If a specific model for dissipation is added to the inviscid dynamics of Clarke's model, then the discontinuities can be resolved by diffusive boundary layers. Pedlosky, Pratt, Spall and Helfrich [20] give a good discussion of the boundary layer on the northern and southern sides of the rectangular island, the "zonal boundary layer" in their terminology.

Temam and Jung have written a series of fifteen papers (so far) developing both the analysis and the numerical analysis of similar boundary layers [21, 22]. They have not yet developed a theory for the hardest case, which is the boundary layer at a promontory (as opposed to a bay. The Stommel layer on the Eastern coast of the islands thickens as the northernmost point is approached and then detaches from the coast to erupt westward into the interior of the flow, the diffusive, widening interior layer in which longitude is a time like coordinate and the diffusion is a latitudinal spreading.

In Clarke's model, the Sverdrup jumps simply induce δ-functions in the east–west velocity. This illuminates a very general principle.

Assertion 6.1 (**δ-Functions and Boundary Layers**) Every δ-function in geophysical fluid dynamics is the limit of a boundary layer, a cover-up for a layer of finite thickness that is unanalyzed for the sake of simplifying the theory.

Other examples are discussed in [16, 23]. The proper choice of the dissipation model near the coast of an island is still a matter of dispute, so Clarke's decision to avoid the controversy at the price of δ-function jets was sensible. The reader who is unhappy with δ-functions can take heart that there always is an underlying boundary layer that could, with the use of the method of matched asymptotic expansions, replace the singularity by a thin layer of high gradients.

McAlpin applied (6) of Cane and Gent's low frequency theory, which we repeat as

$$U = \int_{Y_S}^{Y_N} \{u[X(y), y] - \gamma(y)\, v[X(y), y]\}\, dy \tag{6.85}$$

where $X(y)$ denotes the nondimensional longitude of the boundary. Using numerical quadrature of the integral, he calculated reflection from piecewise linear approximations to the western boundaries of all three oceans. Waves of annual period or longer in low meridional and baroclinic modes reflect as if from a straight boundary, but increasing the latitudinal and/or baroclinic mode number increases the effects of complicated geometry, especially for the most complicated boundary, that of the Pacific.

Rowlands [24] is a posthumous article completed by Anderson and Moore. It is unusual in that the low frequency approximation is not made; instead, as in Anderson and Rowlands [25, 26], Laplace transforms in time are applied. East of the island, assumed infinitely thin in longitude as in many other studies, there is an ever-narrowing boundary layer of short Rossby waves. Although neglected by most other authors, the steady decrease of spatial scale in the short Rossby waves blows up the consistency of inviscid, linear models because either dissipation or nonlinearity must become important in this thin, east-of-the-island layer. The same is true of the δ-function jets (the spoor of Sverdrup jumps as noted earlier) at the tips of the islands.

Another exception that goes beyond the long wave approximation is du Penhoat, Cane and Patton [27]. Figure 6.12, original but drawing heavily on their ideas, is a schematic of the Kelvin wave diffracting around a thin island.

Much of the discussion in Rowlands is focused on island perturbations of the Equatorial Undercurrent and other mean jets, as in Hendry and Wunsch [28]. This properly belongs in our treatment of jets and so mean currents will be largely ignored here.

McPhaden and Gill used a 2–1/2 layer model. This has two active layers above an unmoving, infinitely deep abyss. They employed the model to examine the scattering of Kelvin waves and also Rossby waves from undersea ridges [34]. Kelvin waves in the first baroclinic mode are strongly reflected by shallow ridges but indifferent to deep ridges; second baroclinic mode Kelvin waves are just the opposite. Other work on scattering by undersea features has been sparse, in part because understanding reflection from islands, peninsulas and archipelagoes is far from complete.

Clarke [29] extends Moore's theory for reflections from straight north–south coasts by incorporating boundary curvature. He derives analytical formulas for small curvature combined with numerical treatment of large curvature. Cane and Gent [30] attacked straight coastlines rotated by angle γ from due N-S, and gave analytical formulas applicable to curved boundaries, too. Neither paper is about island scattering per se, but was part of the foundation for later work that was.

The territorial boundaries of the intellectual geography of the studies of reflections from curved coasts, gaps, islands and barriers are as artificial as the aftermath of a really bad peace treaty. We have tried to be inclusive.

Table 6.1 catalogs a few illuminating references.

We have omitted vast amounts of detail from the literature on equatorial wave reflection, diffraction and scattering because this is an area where Lord Kelvin's

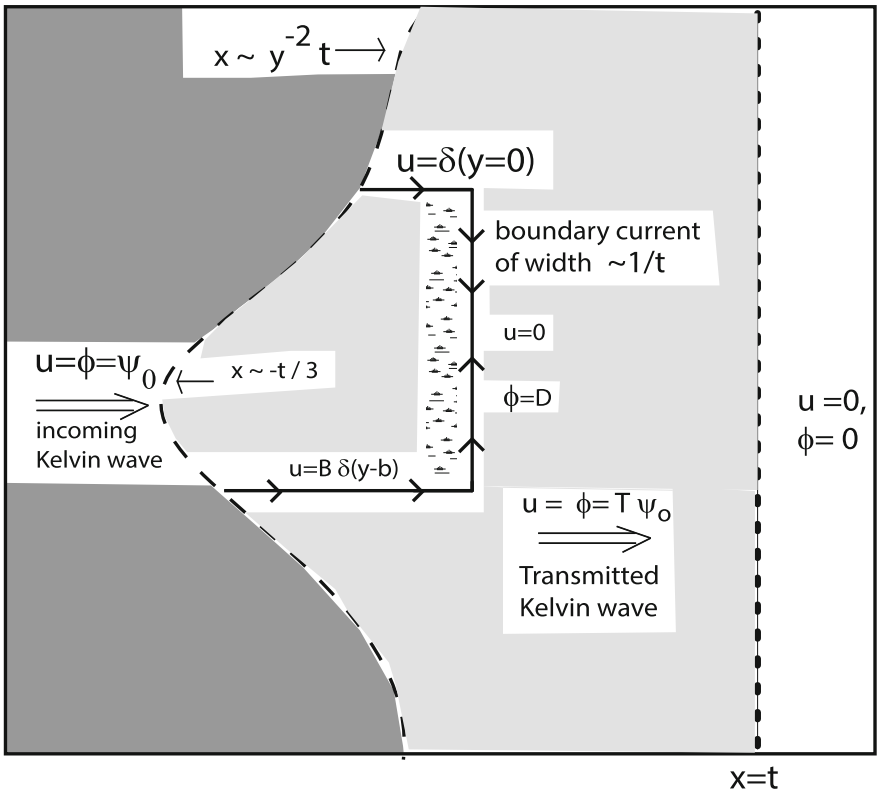

Fig. 6.12 Schematic of the transient solution for a Kelvin wave reflecting/diffracting from a thin island, which is the texture shaded rectangle, narrow in x and extending from $y \in [0, b]$. The incoming Kelvin wave is at unit value everywhere in the *dark shaded area*; this part of the ocean had not yet been effected by the reflected waves. For large time, the asymptotic approximation is accurate out to $x \sim t$. This edge of the front is marked by the *dotted line*; the *white space* to the *right* of this line is where the wave amplitude is zero because the switched on Kelvin wave has not yet reached the unshaded part of the ocean by time t. This is the farthest (eastward) the Kelvin wave travels. The western boundary of the asymptotic approximation is a curving front located at $x \sim -t/3$ at the equator and at $x \sim y^{-2}t$ at high latitudes. East of the western front and west of the eastern front at $x = t$, the amplitude is everywhere T where $T < 1$ as marked by light shading. Theory from du Penhoat, Cane and Patton [27]

famous remark about all of science being "physics and stamp collecting" breaks down. That is to say, the peculiarities of individual islands, undersea ridges, peninsulas, and island groups create detail as individual as the 200,000 species of beetles or the peculiarities of a misprinted Penny Black. But mathematics and physics are ingrained as deeply in physical oceanography as they are irrelevant to botany and stamp collecting.

Table 6.1 Island scattering

Reference	Notes
Reflection from curved boundaries	
Clarke [29]	Analytical formulas for small curvature numerical treatment of large curvature; applications
Cane and Gent [30]	Straight coastlines rotated by angle γ from due N-S
Island and Island Gap Scattering	
Yoon [31]	Purely numerical study; includes Gilbert Is., Maldives Is.
Rowlands, Anderson and Moore [24]	Laplace transforms-in-time for Kelvin waves
Cane and du Penhoat [32]	Single island, low frequency
du Penhoat, Cane sand Patton [27]	Reflection from partial boundaries
McCalpin [33]	Realistic western boundaries of all 3 oceans in low frequency approximation
McPhaden and Gill [34]	Kelvin scattered from submarine ridges in 2.5 layer model
du Penhoat and Cane [35]	Reflection from complicated gaps
Clarke [15]	Irregular Pacific boundary
Cravatte, Boulanger and Picaut [36]	Intraseasonal equatorial Rossby reflected by western Pacific
Spall and Pedlosky [37]	Mix of numerical model and narrow-gap theory for western boundary with gaps
Yuan and Han [38]	Reflection and seasonal Indian Ocean flow

References

1. Pedlosky J (1987) Geophysical fluid dynamics, 2nd edn. Springer, New York
2. Moore DW (1968) Planetary-gravity waves in an equatorial ocean. PhD dissertation, Harvard University, Applied Mathematics
3. Boyd JP (1998b) Weakly nonlocal solitary waves and beyond- all-orders asymptotics: generalized solitons and hyperasymptotic perturbation theory, mathematics and its applications, vol 442. Kluwer, Amsterdam
4. Williams GP, Wilson RJ (1988) The stability and genesis of Rossby vortices. J Atmos Sci 45:207–249
5. Boyd JP (1989b) Non-local equatorial solitary waves. In: Nihoul JCJ, Jamart BM (eds) Mesoscale/synoptic coherent structures in geophysical turbulence: proceedings of the 20th Liege Coll. on Hydrodynamics. Elsevier, Amsterdam, pp 103–112
6. Boyd JP (1991c) Weakly nonlocal solitary waves. In: Osborne AR (ed) Nonlinear topics of ocean physics: Fermi Summer School. Course LIX, North-Holland, Amsterdam, pp 527–556
7. Cane MA, Moore DW (1981) A note on low frequency equatorial basin modes. J Phys Oceangr 11:1578–1584
8. Gent PR (1979) Standing equatorial wave modes in bounded ocean basins. J Phys Oceangr 9(4):653–662

9. Gent PR, Semtner AJ (1980) Energy trapping near the equator in a numerical ocean model. J Phys Oceangr 10(6):823–842
10. McPhaden MJ, Semtner AJ (1982) On the structure of low-frequency equatorial waves. J Phys Oceanogr 12(8):795–804
11. Han W, McCreary JP, Masumoto Y, Vialard J, Duncan B (2011) Basin resonances in the equatorial Indian Ocean. J Phys Oceangr 41(6):1252–1270
12. Meyer RE, Painter JF (1979) Wave trapping with shore absorption. J Eng Math 13:33–45
13. Pedlosky J, Iacono R, Napolitano E, Spall MA (2011) The two-layer skirted island. J Mar Res 69(2–3):435–478
14. Spall MA, Pedlosky J (2013) Interaction of Ekman layers and islands. J Phys Oceangr 43(5):1028–1041
15. Clarke AJ (1991) On the reflection and transmission of low-frequency energy at the irregular Pacific Ocean boundary. J Geophys Res-Oceans 96(Suppl. S, Feb. 28):3289–3305
16. Huang Z, Boyd JP (2017) Smoothing the Sverdrup flow discontinuity. J Phys Oceangr To be submitted
17. Knauss JA (2005) Introduction to physical oceanography, 2nd edn. Waveland Press, Portland, Oregon
18. Vallis GK (2006) Atmospheric and ocean fluid dynamics: fundamentals and large-scale circulation. Cambridge University Press, New York
19. Boyd JP, Sanjaya E (2014) Geometrical effects on western intensification of wind-driven ocean currents: the rotated-channel Stommel model, coastal orientation, and curvature. Dy Atmos Oceans 65(1):17–38
20. Pedlosky J, Pratt LJ, Spall MA, Helfrich KR (1997) Circulation around islands and ridges. J Mar Res 55:1199–1251
21. Jung CY, Temam R (2014) Boundary layer theory for convection-diffusion equations in a circle. Russian Math Surv 69(3):435–480
22. Hong Y, Jung CY, Temam R (2014) On the numerical approximations of stiff convection-diffusion equations in a circle. Numerische Mathemaik 127(2):291–213
23. MacKay MD, Moore GWK (1995) Strongly curved flow profiles in quasigeosstrophic stability theory. J Atmos Sci 52(22):3879–3884
24. Rowlands PG, Anderson DLT, Moore DW (1981) The flow of equatorial Kelvin waves and the equatorial undercurrent around an island. J Mar Res 40(4):915–936
25. Anderson DLT, Rowlands PB (1976a) Role of inertia-gravity and planetary waves in response of a tropical ocean: to incidence of an equatorial Kelvin wave on a meridional boundary. J Mar Res 34(3):295–312
26. Anderson DLT, Rowlands PB (1976b) Somali current response to southwest monsoon – importance of local and remote forcing. J Mar Res 34(3):395–417
27. du Penhoat Y, Cane MA, Patton RJ (1983) Reflections of low frequency equatorial waves on partial boundaries. In: Nihoul JCJ (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 237–258
28. Hendry R, Wunsch C (1973) High Reynolds number flow past an equatorial island. J Fluid Mech 58(1):97–114
29. Clarke AJ (1983) The reflection of equatorial waves from ocean boundaries. J Phys Oceangr 13:1193–1207
30. Cane MA, Gent PR (1984) Reflection of low-frequency equatorial waves at arbitrary western boundaries. J Mar Res 42(3):487–502
31. Yoon JH (1981) Effects of islands on equatorial waves. J Geophys Res 86(NC11):10,913–10,920
32. Cane MA, du Penhoat Y (1982) On the effect of islands on low frequency equatorial motion. J Mar Res 40:937–962
33. McCalpin JD (1987) A note on the reflection of low-frequency Rossby waves from realistic western boundaries. J Phys Oceangr 17:1944–1949
34. McPhaden MJ, Gill AE (1987) Topographic scattering of equatorial Kelvin waves. J Phys Oceanogr 17(1):82–96

35. du Penhoat Y, Cane MA (1991) Effect of low-latitude western boundary gaps on the reflection of equatorial motions. J Geophys Res-Oceans 96(Suppl. S, Feb. 28):3307–3322
36. Cravatte S, Boulanger JP, Picaut J (2004) Reflection of intraseasonal equatorial Rossby waves at the western boundary of the Pacific Ocean. Geophys Res Lett 31(10):Article Number: L10,301
37. Spall MA, Pedlosky J (2005) Reflection and transmission of equatorial Rossby waves. J Phys Oceangr 35(3):363–373
38. Yuan DL, Han WQ (2006) Roles of equatorial waves and western boundary reflection in the seasonal circulation of the equatorial Indian Ocean. J Phys Oceangr 36(5):930–944

Chapter 7
Response of the Equatorial Ocean to Periodic Forcing

Abstract The annual and semiannual cycles in the atmosphere are the result of low-frequency periodic forcing of the ocean. This chapter systematically develops the theory of the ocean response to a periodically-varying wind stress. Special emphasis is given to low frequency forcing and the use of the low frequency/meridional geostrophy approximation to analyze it.

> "Nature is not embarrassed by difficulties of analysis."
>
> Augustin Fresnel

> Lighthill [1] modeled the response of the Indian Ocean by postulating an abrupt onset of the southwest monsoon ... observation suggests that the onset of the southwest moon is far from abrupt, occurring not over a few days, but rather over weeks to perhaps two months. ... at least part of the temporal spectrum is better described as a quasi-periodic process than as a transient.
>
> Carl Wunsch [2]

7.1 Introduction

Historically, the first studies of the response of the equatorial ocean to forcing were driven by the need to test Wyrticki's theory that the rapid event known as El Niño was a reaction to changes in the Trade Winds. For this reason, the papers of 1974–1978 were almost entirely concerned with the response to a wind stress forcing which was *impulsive*, and was usually idealized as a step function in time. The problem of a periodic forcing — the annual and semiannual cycles of the sea, for example — is mathematically much simpler since time derivatives can be replaced by multiplication by $-i\omega$, whereas inversion of Laplace transforms via the steepest descent method is necessary for impulsive forcing. We shall therefore reverse the historical order and consider periodic-in-time forced problems first.

For an annual period, $\omega = 0.026$ for the first baroclinc mode of the ocean, which implies that the "long wave" approximation is just as useful for the seasonal cycles of the sea as it is for the quasi-free modes discussed in the previous chapter. The

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_7

seminal paper of Cane and Sarachik [3] is still the analytical state-of-the-art; in the next section, we discuss various levels of approximation to explain why. We then briefly review the pioneering numerical studies by Kindle, O'Brien, Busalacchi, and their descendants using actual ship wind data and realistic coastal boundaries.

7.2 A Hierarchy of Models for Time-Periodic Forcing

The theoretical studies described in later sections make a lot of approximations. The schematic Fig. 7.1 is a visualization of options with the Cane–Sarachik straight-coasts-with-long-wave framework as the bottom.

Nonlinearity is always a serious complication. When the forcing is confined to a single frequency ω, the "fundamental", nonlinearity generates components whose frequencies are integer multiples of the fundamental, the "harmonics". Denoting the coefficients of $\exp(im\sigma t)$ by $u_m(x, y, z)$ and similarly for the other fields, the numerical task is a single set of nonlinear PDEs coupling all the fields and all the harmonics. Such studies almost always eschew the nonlinear boundary value problem in favor of a general circulation model or another elaborate initial value code. (In nonlinear wave theory, solving for harmonics using perturbation theory is common, and discussed at some length in Chap. 17.)

If the coastlines are curvy but nonlinearity and mean currents are neglected, the forced periodic problem is a set of linear PDEs with coefficients that vary with y. The method of separation of variables does not apply because the boundaries do not coincide with coordinate isolines. It is necessary to attack the linear inhomogeneous PDE system

$$-i\omega u - yv + hx = F \tag{7.1}$$

$$-i\omega v + yu + hy = G \tag{7.2}$$

$$-i\omega h + u_x + v_y = \tilde{H} \tag{7.3}$$

where F and G are the zonal and meridional wind stress and $\tilde{H}$ is a buoyancy flux.

When the coastlines run north and south, it is convenient to rewrite the system above in sum and difference variables, using the lowering and raising operators $\mathfrak{L}$ and $\mathfrak{R}$, as

$$-i\,\omega S + S_x + \mathfrak{R}v = F + \tilde{H} \tag{7.4a}$$

$$-i\,\omega v + (1/2)[\mathfrak{L}S + \mathfrak{R}D] = 0 \tag{7.4b}$$

$$-i\,\omega D - D_x + \mathfrak{L}v = (\tilde{H} - F) \tag{7.4c}$$

A Hermite Galerkin method then replaces the PDE by systems of three ODEs in x, identical in form to the equations for normal modes except that the frequency is now a specified parameter instead of an eigenvalue and the ODE triplets are inhomogeneous.

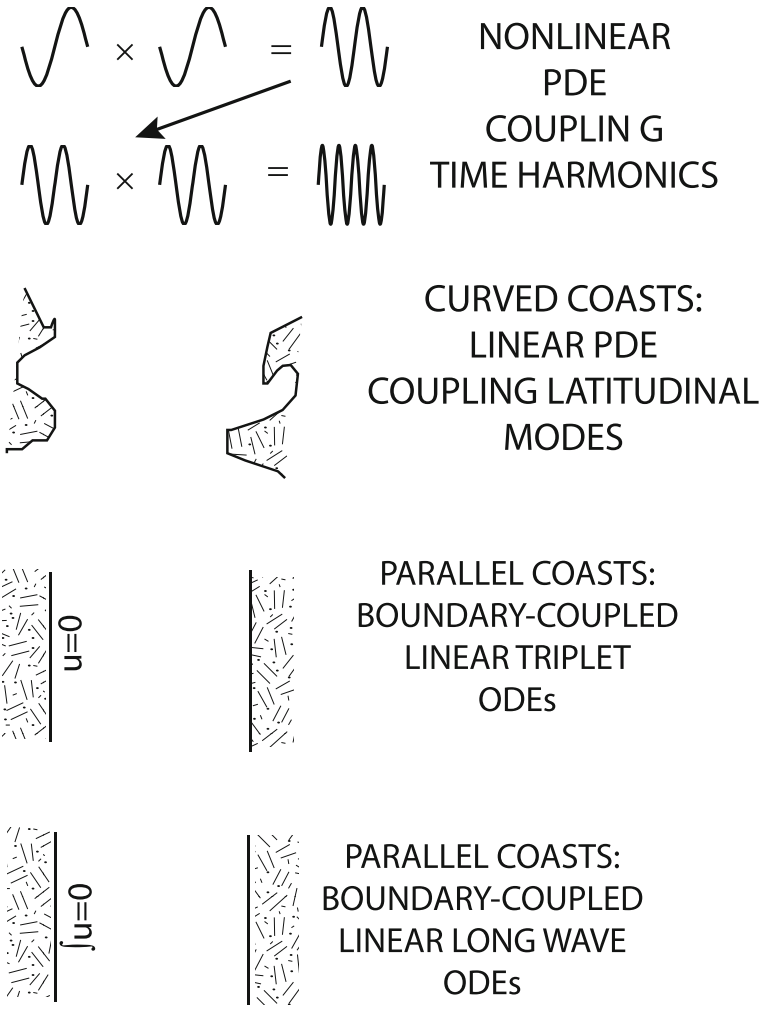

Fig. 7.1 Hierarchy of models for periodic-in-time forced motion. *Top* Quadratic nonlinearity will multiply two copies of the forcing frequency to generate the second harmonic; nonlinear interaction of second harmonics will generate a frequency equal to four times that of the forcing, and so on. A domain bounded by curving coasts [second from the *top*] can be mapped into a rectangle, but the metric factors make the transformed PDEs nonseparable. If the coasts are straight and nonlinearity is omitted [second from the *bottom*], the PDEs are separable and the triplets of (S_{n+1}, v_n, D_{n-1}) are uncoupled by the PDEs, but still chained together by the boundary conditions on the zonal velocity. In the long wave approximation, the triplets become single long wave modes, but are still coupled by the boundary conditions on *u Bottom*

It is convenient to introduce new variables for the inhomogeneous terms:

$$\hat{F} = F + \tilde{H}; \qquad \hat{G} \equiv \tilde{H} - F \tag{7.5}$$

The equations for the Hermite coefficients $\{S_0, S_1, \ldots\}$, $\{v_0, v_1, \ldots\}$, and $\{D_0, D_1, \ldots\}$ are

$$-i\omega S_0 + S_{0,X} = \hat{F}_0 \tag{7.6}$$

$$-i\,\omega S_{n+1} + S_{n+1,x} - \sqrt{n+2}\,v_n = \hat{F}_{n+1} \tag{7.7}$$

$$-i\,\omega v_n + \sqrt{(n+1)/2}S_{n+1} + \sqrt{n/2}D_{n-1} = 0 \tag{7.8}$$

$$-i\,\omega D_{n-1} + D_{n-1,x} + \sqrt{2n}v_n = \hat{G}_{n-1} \tag{7.9}$$

For the annual and semiannual cycles, the long wave approximation is quite accurate. This yields the problem discussed in Sect. 4.8 and the solutions provided in the rest of this chapter.

With x-independent forcing on an acquaplanet, the equations simplify still further to

$$-i\omega S_0 = \hat{F}_0 \tag{7.10}$$

$$-i\,\omega S_{n+1} - \sqrt{n+2}\,v_n = \hat{F}_{n+1} \tag{7.11a}$$

$$-i\,\omega v_n + \sqrt{(n+1)/2}S_{n+1} + \sqrt{n/2}D_{n-1} = 0 \tag{7.11b}$$

$$-i\,\omega D_{n-1} + \sqrt{2n}v_n = \hat{G}_{n-1} \tag{7.11c}$$

7.3 Description of the Model and the Problem

The model makes the following assumptions:

1. hydrostatic approximation
2. linear shallow water wave equations
3. "long wave" approximation
4. zonal boundaries at $x = 0$ and $x = X_E$ (running due N–S)
5. forcing is independent of x
6. y-dependence of forcing is a Gaussian.

(Usually; we will derive solutions for general y-dependence first.)

This model is the most complicated model with periodic forcing that allows analytical theory.

For the first baroclinic mode, the annual and semiannual periods are roughly $\omega = 0.026$ and 0.052, respectively, so both frequencies are small in comparison to 1 and

the long wave approximation is quite accurate. However, atmospheric fluctuations on shorter time scales can always be represented by a Fourier integral over frequencies; if the time scale of the fluctuation is only a few days, then the long wave approximation will be terrible. However, the general solution (without the approximation) is so messy that adopting the long wave approximation seems a small price to pay for a much clearer physical understanding.

The assumption of coasts parallel to the y-axis is made of dire necessity: it is not possible to obtain analytical solutions without it.

The assumption of x-independent forcing is made solely for simplicity. For a zonally bounded ocean, one can synthesize a solution as a Fourier series over discrete wavenumbers in x, but because the solution is not periodic in x, the Fourier series converge poorly, and a numerical model is superior. Since the problem is after all only two-dimensional, numerical solutions are not hard. Personal computers are making messy analytical solutions less and less useful. (By "messy" is meant analytical solutions that cannot be more easily interpreted than a numerical solution.)

The Gaussian dependence in latitude is convenient because one has to choose something or other, and the Hermite coefficients of a Gaussian of arbitrary width are known in analytical closed form (Appendix A). We will derive the solution for general y-dependence first, however, because we do not need to assume any specific form for the y-structure of the forcing until a rather late stage in the analysis. A general property of the solutions to linear, inhomogeneous differential equations — quite independent of the five assumptions given above — is that the general solution is the sum of (a) a particular solution, which satisfies the inhomogeneous equations, but not the forcing and (b) the general homogeneous solution, which contains arbitrary constants that must be chosen so that the sum of the homogeneous solutions plus the particular solution satisfies the boundary conditions. In our problem, it is easy to derive a particular solution which is independent of x (since the forcing is) and has only the trivial ($\exp[-i\omega t]$) dependence upon time. The general homogeneous solution is the sum of all the free oscillations allowed, i.e., the Kelvin wave plus the long Rossby waves (if we make the long wave approximation), and these same waves plus gravity waves and short Rossby waves if we do not. The arbitrary constants are simply the countable infinity of the amplitudes of these modes. Each of the homogeneous solution waves has the same ($\exp[-i\omega t]$) dependence as the forcing.

The forced, particular solution in the long wave approximation is

$$\bar{u} = (i\omega)^{-1} \exp(-i\omega t) \left\{ d_K \bar{M}_K + \sum_n r_n \bar{R}_n \right\} \tag{7.12}$$

where the summation is over all Rossby modes. $\bar{u}$ is the vector (u, ϕ) and

$$\bar{M} = 2^{-1/2} \begin{vmatrix} \psi_0(y) \\ \psi_0(y) \end{vmatrix} \tag{7.13a}$$

$$\bar{R}_n = \frac{1}{4n(n+1)} \begin{vmatrix} -(2n+1)\psi_{n,y} - y\psi_n \\ (2n+1)y\psi_n + \psi_{n,y} \end{vmatrix} \tag{7.13b}$$

$$= 2^{-3/2} \begin{vmatrix} \frac{1}{\sqrt{n+1}}\psi_{n+1} - \frac{1}{\sqrt{n}}\psi_{n-1}, \\ \frac{1}{\sqrt{n+1}}\psi_{n+1} + \frac{1}{\sqrt{n}}\psi_{n-1} \end{vmatrix} \tag{7.13c}$$

where the $\psi_n(y)$ are the usual normalized Hermite functions and (7.13) give the latitudinal dependence of the Kelvin and Rossby waves as evaluated in the long wave approximation. It is noteworthy that these structure functions are *independent of* k in this approximation. These functions also satisfy the (vector) orthogonality relations

$$(\bar{M}_K, \bar{R}_n) = 0, \qquad (\bar{R}_m, \bar{R}_n) = 0 \quad \text{for any } m, n \tag{7.14}$$

where

$$(A, B) \equiv \int_{-\infty}^{\infty} A_u B_u + A_\phi B_\phi \, dy \tag{7.15}$$

for any vectors A, B. These orthogonality relationships are important because they imply that one can use the long wave approximations to the Kelvin and Rossby modes as freely as their exact counterparts so that there is no difficulty in solving for the as yet undetermined constants d_K and r_n. Precisely because this task is relatively simple, we shall postpone it for the moment and instead turn to the determination of the homogeneous part of the solution.

The free modes are

$$\bar{u}_{free} = \left\{ b'_K \overline{M}_K(y) \exp(i\omega x) + \sum_n a'_n \bar{R}_n(y) \exp(-i\omega x[2n+1]) \right\} \tag{7.16}$$

where we have used the fact that (i) $k = \omega$ (exactly) for the Kelvin wave and (ii) $k_n = -\omega(2n+1)$ in the long wave approximation.[1] It is convenient to rewrite the particular solution in a form that explicitly adds a sum of the free modes (7.16) so that the new particular solution automatically satisfies the boundary condition at the eastern boundary. With the definitions

$$\Phi \equiv \omega X_E; \qquad \xi \equiv (x - X_E)/X_E \tag{7.17}$$

— note that the rescaled x-coordinate ξ is different from the variable s employed in the discussion of long wave quasi- free modes — the full solution is

[1] It should be noted that Cane and Sarachik assumed a time dependence of $\exp[+i\omega t]$, which effectively multiplies the exponentials in (7.16) by a minus sign. Their convention means they represent waves in the unorthodox form $\exp[-ikx + i\omega t]$; the usual representation $\exp[ikx - i\omega t]$ will be employed in these notes.

$$\bar{u} = (i\omega)^{-1} \exp(-i\omega t) \left\{ d_K \bar{M}_K \left[1 - \exp(i\Phi\xi)\right] + \sum_n r_n \bar{R}_n [1 - \exp(-i(2n+1)\Phi\xi)] \right.$$
$$\left. + b_K \bar{M}_K \exp(i\Phi\xi) + \sum_n a_n \bar{R}_n \exp(-i(2n+1)\Phi\xi) \right\} \tag{7.18}$$

where the constants b_K and a_n are different from the primed equivalents in (7.16).

The advantage of the form (7.18), which at first glance seems needlessly complicated, is that since the particular solution now satisfies the eastern boundary condition of vanishing zonal velocity, i.e., $u_{particular}(\xi = 0) = 0$, it follows that the sum of the free modes must also separately vanish at the eastern boundary. However, we have already solved the problem of free modes at an eastern boundary in the long wave approximation: the solution consists of the incoming Kelvin wave plus its Rossby wave reflections. It follows from Eq. (6.45) of the previous chapter that

$$\bar{u}_{free} = (i\omega)^{-1} \exp(-i\omega t) \left\{ d_K \bar{M}_K \left[1 - \exp(i\Phi\xi)\right] + \sum_n r_n \bar{R}_n [1 - \exp(-i(2n+1)\Phi\xi)] \right.$$
$$\left. + b_K \bar{M}_K \exp(i\Phi\xi) + \sum_n a_n \bar{R}_n \exp(-i(2n+1)\Phi\xi) \right\} \tag{7.19}$$

The eastern boundary condition thus determines the homogeneous solution completely to within an overall multiplicative constant, which is the still undetermined amplitude b_K.

The coefficients of the particular solution, $\{d_K, r_n\}$, are determined entirely by the wind stress that is driving the motion, so the single factor b_K is the only remaining unknown. What determines it is the boundary condition at the west where $x = 0$ and $\xi = -1$. For the long wave quasi-free oscillations of the previous chapter, it is possible to satisfy the rigorous boundary condition $u = 0$. Here, however, the solution consists of forced as well as traveling waves, and we must fall back upon the general long wave boundary condition derived in Chap. 6:

$$\int_{-\infty}^{\infty} u \, dy = 0 \qquad \text{at } x = 0 \ [\xi = -1] \tag{7.20}$$

Using the known y-integrals of the Hermite functions gives

$$\int_{-\infty}^{\infty} (\bar{M}_K)\, u \, dy = \pi^{1/4}, \qquad \int_{-\infty}^{\infty} (\bar{R}_n)\, u \, dy = \pi^{1/4} \, \Lambda_n / [2n(n+1)] \tag{7.21}$$

where, as before, $\Lambda_1 = 1$ and (6.35) $\Lambda_n = \sqrt{[3 \cdot 5 \cdot 7 \cdots n]/[2 \cdot 4 \cdot 6 \cdots (n-1)]}$ if $n \geq 3$ is odd and $\Lambda_n = 0$ for all even n. Substituting these into (7.20) and using (7.18) and (7.19) gives

$$b_K = \{B(i\Phi) + d_K(\exp(-i\Phi) - 1)\} \,/\, \{1 - \exp(2i\Phi)\, \mathfrak{S}(i\Phi)\} \tag{7.22}$$

where the sums B and $\mathfrak{S}$ are given by

$$B(i\Phi) \equiv \sum_n [2n(n+1)]^{-1} r_n \Lambda_n [1 - \exp(i\Phi(2n+1))] \tag{7.23}$$

$$\mathfrak{S}(i\Phi) \equiv \sum_n [n(n+1)]^{-1} [\Lambda_n]^2 \exp(i2n\Phi) \tag{7.24}$$

At this point, we have obtained the complete solution in the long wave approximation. Several comments are necessary.

First, the limit $N \to \infty$ of the sums over the Rossby modes have deliberately been left unspecified above. As was the case in our treatment of the long wave basin modes in the previous section, it is analytically convenient to extend the sums to ∞. However, it is not possible to obtain closed form exact sums, but only approximate results that are useful in interpreting results in various parameter ranges. For numerical purposes, it is safer to evaluate the series by simply summing them. In this case, the proper truncation is

$$N = 1/(8\omega^2) \qquad \text{[cutoff mode number]} \tag{7.25}$$

since, according to the full dispersion relation for equatorial waves, modes of higher n are zonally trapped. It is senseless to include these higher modes in the sums since the long wave approximation hopelessly falsifies their nature by forcing all Rossby modes to be zonally propagating. For the annual cycle in the first baroclinic mode with an equivalent depth of 0.65 m, $N \approx 180$.

Second, the latitudinal structure functions $\bar{M}_K$ and $\bar{R}_n$ are deliberately restricted to describe only u and ϕ. When $\omega << 1, v << u$ as explained in Chap. 6 in justifying the long wave approximation. The calculation of v is not ill-conditioned since all the terms in the x-momentum and height equations involving u and ϕ are multiplied by ω or k, and thus are as small as those involving v. It is somewhat pointless to compute v, however, since it is so small in comparison to u for low frequency motions that it is below the noise level of observational instruments. Consequently, Cane and Sarachik [3] completely ignore the north-south velocity, and we shall do likewise.

Third, the long wave y-momentum equation, $yu + \phi_y = 0$, demands

$$\phi_y = 0 \quad \text{[at the coastal boundaries]} \tag{7.26}$$

i.e., the height ϕ is constant along the coast. (Equivalently, in the language of the 1–1/2 layer model, the thermocline depth does not vary with latitude along the coast.) Strictly speaking, this is true only for $|y| < \sqrt{(2N+1)}$ where N is the limiting mode number given by (7.25). At higher latitudes, higher modes which sum to the coastal Kelvin wave dominate the solution, and (7.26) is no longer a good approximation. This is rather important because coastal measurements are the easiest to take (for obvious reasons) and (7.26) implies that the signal will have no phase or amplitude variations along the coast to bedevil observationalists. Along the eastern boundary in particular, the particular solution in (7.18) is zero by construction. Therefore, the

eastern boundary height is given entirely by the ϕ component of the sum of free waves in (7.19) — but we already saw that this sum has a closed form solution in the previous chapter. Evaluating the sum at $x = X_E$ gives

$$\phi_E = X_E\, b_K\, \exp(-i\omega t)/[\pi^{1/4}\, i\, \Phi] \tag{7.27}$$

Thus, the coefficient of the free Kelvin wave, b_K, determines the amplitude and phase of the height along the eastern side of the ocean (note that b_K is normally complex), and vice versa.

Fifth, we have multiplied (and divided) by X_E in (7.27) because the overall amplitude of ϕ is, all other things being equal, directly proportional to the width of the ocean basin. The reason for this can be understood by taking the limit of very low frequency motions so that the waves, inviscid in the approximations used here, would be suppressed by the damping of the real ocean. In this limit, the ocean is always close to being in quasi-equilibrium with the wind stress, that is to say, the solution for the ocean is approximately equal to that for a *steady* wind stress equal to the instantaneous value of the (fluctuating) wind stress. For a steady wind , the x-momentum equation is

$$\phi_x = \tau^{(x)}; \tag{7.28}$$

If $\tau^{(x)}$ is oscillating with frequency ω and is independent of x, then (ignoring the waves) ϕ tilts up and down like a seesaw. The important point is that (7.28) shows that the *slope* of ϕ is determined entirely by the wind stress. It follows that if we double the width of the basin, the *amplitude* of the motion of the ends of the seesaw (ϕ_E and its western counnterpart) will have to double also. Consequently, ϕ_E is proportional to X_E, and the graphs of the height fields from Cane and Sarachik [3] are invariably graphs of ϕ/X_E rather than ϕ itself.

Cane and Sarachik [3] contains a number of typographical errors [3]. There is a factor of $\pi^{-1/4} d_K/(i\Phi)$ missing from the unnumbered equation just below their Eq. (24); their Eq. (14) is missing absolute value bars and a factor of ω. The most significant error is that in their Eq.(19), they make the replacement

$$b_K \rightarrow\ d_K\, \rho(i\Phi) \tag{7.29}$$

where their (20) defines ρ (comparing formulas) as

$$\rho = \exp(-i\Phi)\, b_K \tag{7.30}$$

This is inconsistent; since d_K, which is part of the particular solution and is therefore directly proportional to the strength of the wind stress, appears in the formula for b_K, the product of d_K with ρ contains a term proportional to the *square* of d_K and therefore to the square of the wind stress — which is rather remarkable in a linear model.

However, all is not lost. At worst, this mistake gives only an overall amplitude and phase error; the relative variations of amplitude and phase with x, y, and t are

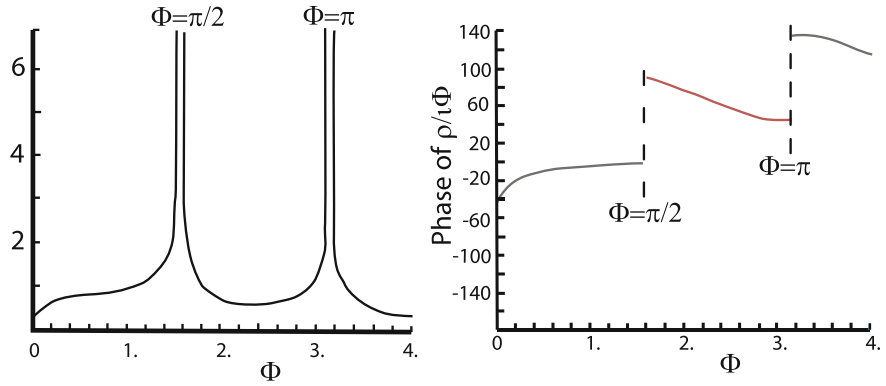

Fig. 7.2 *Left* Amplitude of $\rho(i\Phi)/i\Phi$ for $\mu = 1/5$ (corresponding to a forcing scale of 10 degrees of latitude) as a function of Φ, x and y are scaled by $L_{eq} = \sqrt{c/\beta}$ [=1, nondimensionally]. This is also a graph of $b_K/(i\Phi)$ versus Φ where b_K is the amplitude of the Kelvin wave in $\bar{u}_{free}$, the homogeneous solution. *Right* Phase of $\rho(i\Phi)/i\Phi$ for $\mu = 1/5$ as a function of Φ. Redrawn after Cane and Sarachik [3], Fig. 1, who assume a time dependence of $\exp(i\omega t)$ versus the $\exp(-i\omega t)$ used here; this reverses the sign of the phases

not affected. The mistake is probably only a typo anyway. However, it is not obvious (as it is with most typos) how to correct the mistake.

In the rest of their paper, Cane and Sarachik [3] use only a forcing with a Gaussian dependence upon latitude,[2] i.e.,

$$\tau^x = \exp(-i\omega t)\ \exp(-[1/2]\,\mu\, y^2); \qquad \tau^y = 0 \tag{7.31}$$

The advantage of this form, as mentioned earlier, is that Hermite expansion coefficients of (2.19) are known in analytical closed form as given in Appendix A. Unfortunately, this is not sufficient to obtain simple, analytical answers for the solution as a whole: Eq. (26) of Cane and Sarachik [3] is very intricate and includes integrals which cannot be done in closed form. However, it is possible to obtain closed form approximations to the solution in various parameter ranges, but that for the most interesting and realistic case, $\Phi \sim 0(1)$, is unfortunately quite messy. We therefore will content ourselves with a description of their numerical results.

Figure 7.2 shows the graph of a quantity proportional to ϕ_E as a function of the parameter $\Phi = \omega X_E$. The resonances for $\Phi = m\pi/2$ correspond to the Cane–Moore free oscillations of the basin that were described in Chap. 6. A white noise spectrum of forcing, i.e., one in which there is roughly equal energy over some wide band of frequency, will give a response dominated by those frequencies that correspond to the Cane–Moore free modes for that particular ocean. Figure 7.3 shows that b_K

[2] Strictly speaking, they also consider an antisymmetric zonal wind stress obtained by multiplying (7.31) by y. However, this does not excite a Kelvin wave and does not offer any particularly interesting new features, so we shall only discuss the cases where the stress is symmetric about the equator.

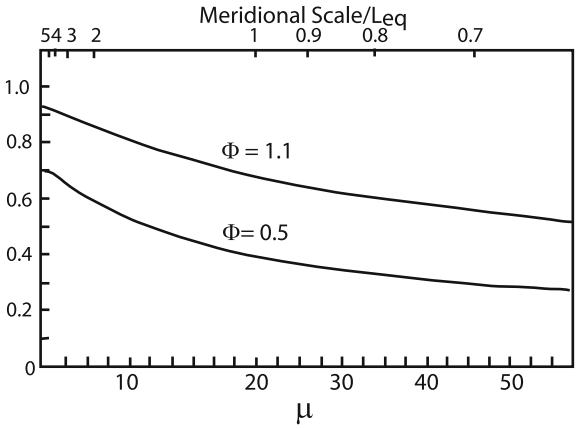

Fig. 7.3 Amplitude of $\rho(i\Phi)/i\Phi$, same as previous figure, but versus μ, the width of the forcing. There is little variation for $\mu < 0.2$ [forcing wider than 10 degrees of latitude]. Figure 2 of Cane and Sarachik [3]

is rather insensitive to μ, i.e., to the width of the forcing, so long as $\mu < 0.2$, but becomes sensitive for narrower forcing. The phase for small y is sensitive to the width of the forcing for all μ. These comments are based on more than the graph, of course; after inspecting many sets of output, Cane and Sarachik concluded that a forcing with a half-width of 10 degrees of latitude or more would give the same features in amplitude as taking the forcing to be independent of y.

The remaining figures from their paper illustrate four different cases: (i) $\Phi << 1$, narrow forcing (Fig. 7.4) (ii) $\Phi << 1$, broad forcing (Fig. 7.5) (iii) $\Phi \approx 0(1)$, annual cycle in the Atlantic Ocean and (iv) $\Phi \approx 0(1)$, annual cycle in the Pacific Ocean.

The first case (Fig. 7.4) — $\Phi = 0.01$ and a narrow forcing — is probably the least realistic since, for an ocean as wide as the Atlantic (5000 km), this Φ implies a period of forty years (far longer than the dissipation time scale) and furthermore the winds are broad rather than narrow. The parameter $P = 2\Phi/\mu$ plays a crucial role in Cane and Sarachik's analytical approximations. The response is strongly concentrated about the equator and the amplitude of the tilt increases linearly from small values $[O(P^{1/2})]$ in the east to rather large values [0(10)] in the west. The tilt of the height field ϕ_x is everywhere in phase with the forcing and is indeed precisely what it would be for the quasi-equilibrium solution discussed above. The wave-like part of the solution, which is approximately independent of both x and y in this parameter range, is complex, however, so that ϕ itself is *not* in phase with the forcing. However, this distinction is important only in a thin boundary layer of width $O(P^{1/2})$ near the eastern boundary because the amplitude of the wave part of ϕ is only $O(P^{1/2})$.

The second case (Fig. 7.5) uses the same unrealistically small value of Φ, but now the forcing is broad and the parameter P is 0(1). The amplitude of ϕ is (except for such large values of y that they are off the graph) approximately independent of y in both amplitude and phase. The height field is seesawing up and down around a pivot point located at the amplitude minimum at $\xi = -0.572$ — not quite the middle of the ocean. In the limit that the forcing becomes so broad that the parameter $P >> 1$

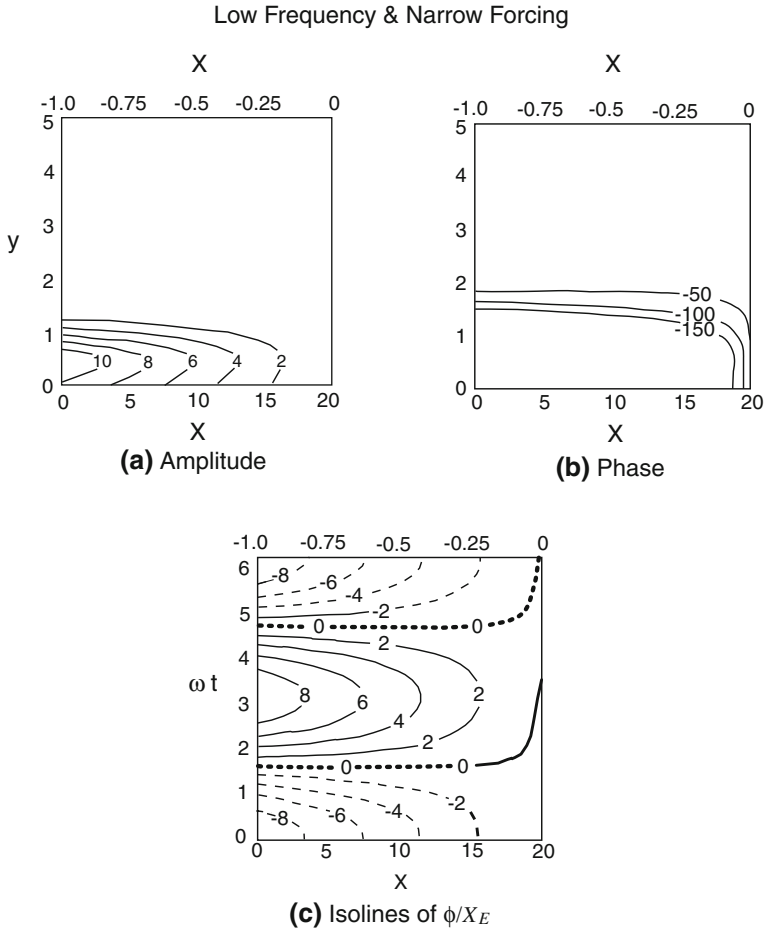

Fig. 7.4 The *low* frequency ($\Phi = 0.01$) narrow symmetric forcing case ($\mu = 4.5$, corresponding to a forcing scale of 3 degrees of latitude). *Left* Amplitude of height field oscillations normalized by X_E. The isoline interval in 0.1. *Right* Phase of height oscillations with respect to the forcing (negative phases *lead* the forcing). Contour interval is 50 degrees. *Bottom* ϕ / X_E versus time on the equator. Contour interval 1/5. Time is scaled by the reciprocal of ω. Redrawn after Fig. 3 of Cane and Sarachik [3]

(not illustrated by a graph), the pivot point moves out to $\xi = -0.6667$ so that there is twice as much ocean on the eastern side of the pivot point as on the western:

$$\phi = (2/3 + \xi)\, X_E \, \exp[i\omega t] \qquad [\Phi << 1; \quad P >> 1] \tag{7.32}$$

for y not too large; the thermocline pivots like a rigid seesaw in phase with the wind to the east of the pivot point and out of phase to the west.

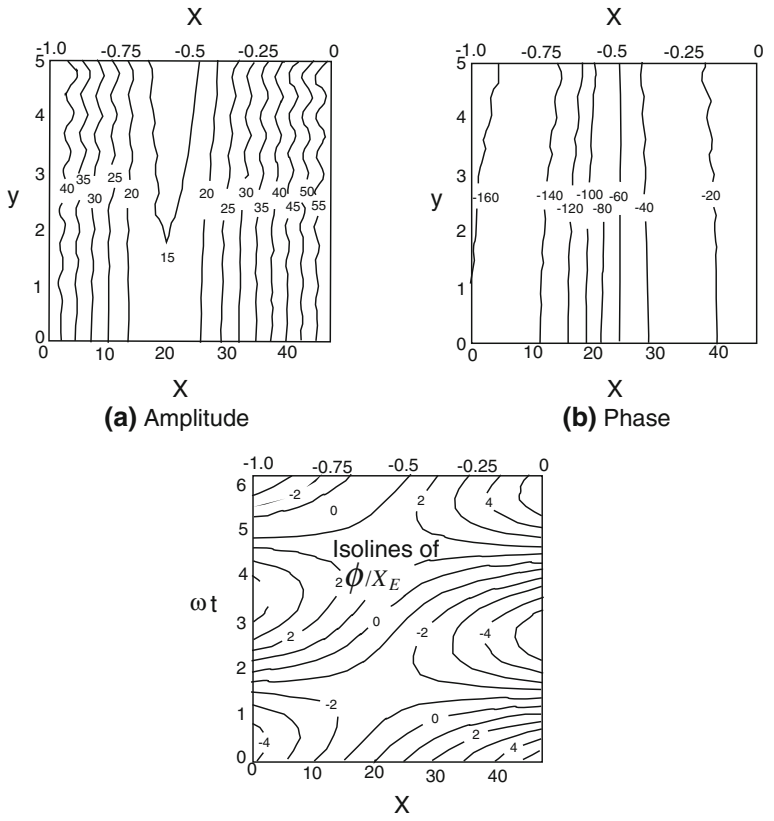

Fig. 7.5 The *low* frequency ($\Phi = 0.01$) wide symmetric forcing case ($\mu = 0.006$, corresponding to a forcing width of 60 degrees of latitude). *Left* Amplitude of ϕ/X_E. Contour interval of 0.05. *Right* Phase of ϕ relative to the forcing (negative phases lead). Contour interval is 20 degrees. *Bottom* ϕ/X_E versus times on the equator. Contour interval 1/10. Figure 4 of Cane and Sarachik [3]

This location of the pivot point, which has certain obvious observational consequences since thermocline tilting on seasonal time scales can be seen with existing intruments, seems to defy common sense, which would predict that the height would tilt about the exact middle of the ocean basin. This common sense result is indeed what happens for a *closed* basin [4] or a basin for which mass is conserved. The unbounded equatorial ocean, however, is free to exchange mass with the reservoir at infinity, gaining and losing during different phases of the cycle. Thus, the generation and repeated reflection of Kelvin and Rossby waves at the boundaries drastically alters the dynamics of the ocean's response to the periodically varying wind.

The phase, however, is varying across the basin (east-to-west), so the pivoting is not occurring as if the thermocline were a rigid surface, but rather as if it were a "writhing snake" to use Cane and Sarachik's rather poetic expression. One can

define a local x-wavenumber as $\partial\chi/\partial x$ where χ is defined by $\phi = \exp[i\chi]$, and the corresponding phase speed as $c_{phase} = \omega/k_{local} = \omega/\partial\chi/\partial x$. Even away from the equator where the Kelvin wave has negligible amplitude and only long Rossby waves are found, there is nonetheless an eastward phase propagation. [Note that with our phase convention of taking the time dependence as $\exp[-i\omega t]$, the phases have opposite signs from what is shown in the Cane and Sarachik diagrams.] This phase propagation in the wrong direction is another reminder that the behavior of a *sum* of waves cannot be directly related to the phase behaviour of any of the waves making up the sum. This seems rather obvious, but because oceanographic data is always so sparse, there is always an almost irresistable temptation to identify any evidence of phase propagation with an individual equatorial mode — which is probably wrong. Many studies have looked hard for evidence of "annual Rossby" waves in the ocean; again the eastward phase propagation of the forced response implies that this sort of search is likely to be futile, at least in most parameter ranges.

When Φ is 0(1), the only analysis that can be done is rather complicated and will not be presented here. Cane and Sarachik give approximations that predict rather well the minima and maxima of amplitude as shown by the dashed lines in Figs. 7.6 and 7.7. The graph for the Atlantic annual cycle is interesting; the structure can be interpreted by using the ray-tracing ideas of Schopf, Anderson and Smith [5]. When $\Phi > \pi/4$, as it is for the Pacific annual cycle, the focusing, caustics, and shadow zones so characteristic of the Cane–Moore quasi-normal modes of the basin reappear. However, the focus for the Cane–Moore mode of longest period was located in the middle of the ocean whereas in Fig. 7.7 it is rather near the western boundary.

The shadow zone and the caustics, which are the regions of large amplitude that bound the shadow zone, are very similar to long wave basin modes because the particular solution, although present here, is overwhelmed by the homogenous solution in (7.18) wherever the latter is singular, or nearly singular.

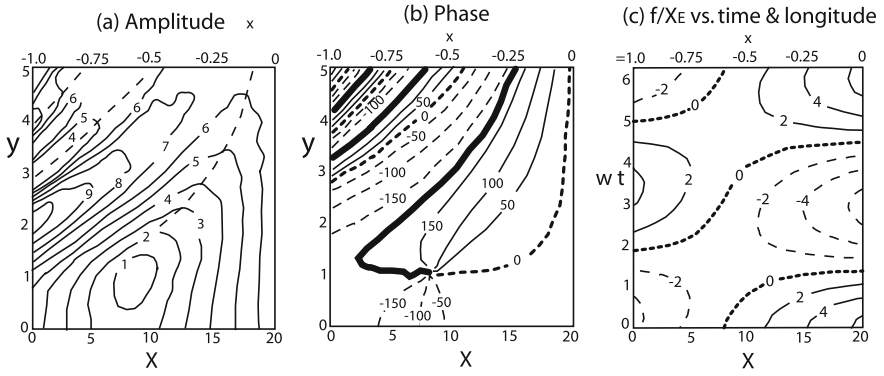

Fig. 7.6 The Atlantic annual ($\Phi = 0.54$) Symmetric Forcing Case ($\mu = 1/5$). *Left* amplitude of ϕ/X_E with a contour interval of 1/10. *Right* Phase relative to forcing (negative phases lead the forcing). The thick multiple lines mark the jump from 180 degrees to –180 degrees. The contour interval is 50 degrees. Redrawn after Fig. 5 of [3]

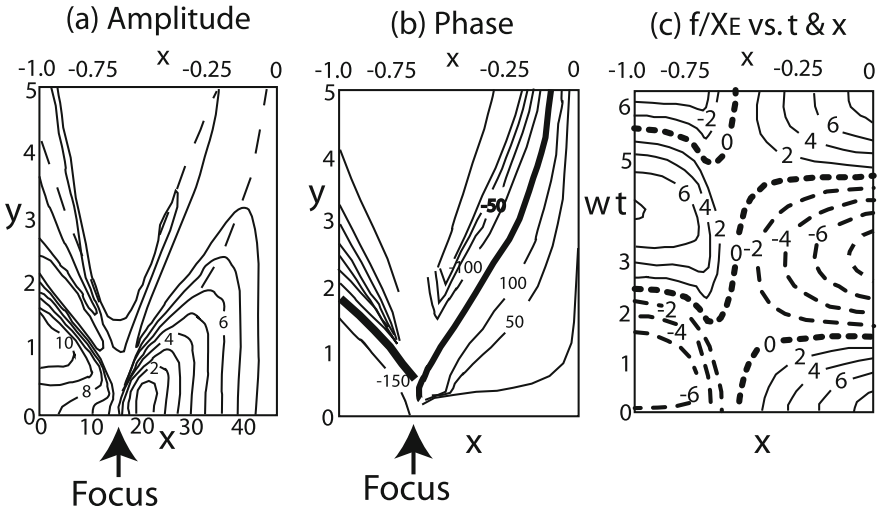

Fig. 7.7 Pacific annual ($\Phi = 1.13$) Symmetric forcing case ($\mu = 1/5$). *Left* amplitude of ϕ / X_E with a contour interval of 1/10. Minima are marked by *dashed lines* while maxima are *thick solid lines*. The contours are in agreement with theory. *Right* Phase relative to forcing (negative phases lead the forcing). The thick multiple lines mark the jump from 180 degrees to –180 degrees. The contour interval is 50 degrees. *Bottom* ϕ / X_E versus t and x. Redrawn after Fig. 6 of [3]

For both of these annual cycles, the phase variation across the basin is about 180 degrees. If we assume that the variation occurs linearly for purposes of scale analysis, then this implies an effective x-wavenumber of about π / X_E. The order-of-magnitude of the apparent zonal phase propagation of the thermocline will be $\omega\pi / X_E$, or about 30 cm/s for the annual signal in the Atlantic. This is just about the value noted by [6].

7.4 Numerical Models: Reflections and "Ringing"

Kindle [7] solved the nonlinear shallow water wave equations (i.e., the nonlinear 1–1/2 layer model) on the equatorial beta-plane using numerical integration for both impulsive and seasonal wind stresses. He also did a simple analytical treatment, which, although far cruder than the long wave theory of [3], nevertheless gives some useful insights.

When the wind is turned on impulsively, the response in the interior of the ocean is identical with that of the Yoshida jet, which is why it is worthwhile to study the Yoshida jet even though the real equatorial oceans have continental boundaries. The Yoshida jet is only a particular solution, however; free Kelvin, gravity, and Rossby waves must be added to it to satisfy the boundary conditions. Since the zonal flow in the Yoshida jet is increasingly linearly varying with time, more and more

waves must be generated at the boundary to make $u = 0$ on the eastern and western boundaries; these coastlines are therefore forced to play the role of continuously operating wavemakers.

The Kelvin and Rossby modes which dominate the long time behavior of the waves take a few weeks to penetrate the interior, but eventually they reach the other side of the sea and reflect. After a few more months, the reflections have crossed the ocean and been transmuted into re-reflections. For large times, the overall solution is rather messy because so many waves are pinging back and forth in the ocean basin. However, after a time scale of about $4X_E$, which is long enough for a Kelvin wave to cross the basin once and make one return as an $n = 1$ Rossby wave,[3] the amplitude of the fluctuations becomes rather small and the flow is in near equilibrium with the wind stress, even though a complicated pattern of small wavy perturbations on top of this equilibrium may persist for many years. (Recall Gent's normal modes). Each time the Kelvin and Rossby waves reflect, they lose energy to coastal Kelvin waves, and the remainder becomes spread out among a large number of modes which then largely cancel each other out through interference effects. Furthermore, in the real ocean, energy is lost to dissipation. Thus, although technically the re-reflections are part of the solution for many years after the turn-on of the winds, in practice only the initial Kelvin and Rossby waves — the ones that begin to move towards the interior of the ocean from the boundaries as soon as the wind stress is turned on — and their first reflections are large enough to worry about.

Kindle exploits these ideas for *seasonal* forcing by assuming that the upwelling rate along the equator can be represented in the form

$$w(x,t) = w_K \sin(\omega t) + w_K \sin[\omega(t - x)/c_K] + w_R \sin[\omega(t + \{X_E - x\}/c_R)] + w_{KR} \sin[\omega(t - X_E/c_K + (X_E - x)/c_R)] + w_{RK} \sin[\omega(t - X_E/c_K + x/c_R)] \tag{7.33}$$

with $c_K = 1$ and $c_R = -1/3$ where w_I is the Yoshida jet (interior) contribution, w_K and w_R are the Kelvin and $n = 1$ Rossby waves excited directly at the boundary, and w_{KR} and w_{RK} are respectively the $n = 1$ Rossby part of the reflection of the directly excited Kelvin wave, and the Kelvin reflection of the directly excited $n = 1$ Rossby wave. His Figs. 7 and 8 compare the analytical solution as given by (7.33) [together with the Kelvin and Rossby re-reflections, which are not written out explicitly in (7.33)] with the numerical solution. The graphs are virtually indistinguishable [7].

Equation (7.33) is remarkable for what it leaves out. First, all higher order Rossby modes are omitted! The reason is that the first few Hermite coefficients of the series solution given in Sect. 7.3 decrease very rapidly; the $n = 1$ mode has much more energy than $n = 3$ and the higher modes. Hence, within a degree or two of the equator, the higher modes can be ignored. At higher latitudes, the situation is reversed because the $n = 1$ mode has turning points at $y = \pm 1.73$, so at high latitudes, it has decayed away to nothing and the modes of larger n completely determine the solution. This

[3] Note that some of the Kelvin wave energy goes into all the higher and slower moving Rossby waves; however, the $n = 1$ mode gets the lion's share.

primacy of the $n = 1$ mode along the equator is still worth noting, however; this mode is highly energetic in comparison to the rest.

Even more notably, (7.33) and the figures together show what has been said in words above: though the basin "rings" with reflections forever in an inviscid model, in practice the numerical solution and the real ocean are both likely to be dominated by just the first reflection or two plus the waves and interior x-independent flow that are directly excited by the wind stress and boundaries.

Kindle calculates the reflection coefficients in his model in two ways: (i) by using the analytical formulas of [8] and (ii) directly from his computer integrations by carefully examining the amplitude in places and times where the upwelling is determined by only a single wave mode, or by one plus others whose amplitude is already known. The results are summarized below:

$$w_K = -1.4\, w_I, \qquad w_R = 0.6\, w_I \tag{7.34}$$

$$w_{KR} = R_E\, w_K, \qquad w_{RK} = R_W\, w_R \tag{7.35}$$

$$R_E = \begin{cases} 0.36 \ \text{[numerical]} \\ 0.5 \ \text{[analytical]} \end{cases}, \qquad R_W = \begin{cases} 0.75 \ \text{[numerical]} \\ 1.0 \ \text{[analytical]} \end{cases} \tag{7.36}$$

Note that these results are expressed in terms of relative upwelling velocities along the equator; since w and ϕ are proportional to one another with a proportionality constant of $-i\omega$ times the ratio of the integral of the vertical structure function to the vertical eigenmode $\mathfrak{z}_n(z)$, the relative upwelling is equivalent to specifying relative values of $\phi(y = 0)$ rather than amplitudes of S_n or $\bar{M}_K$ and $\bar{R}_n$.

It is striking that he underestimates the analytical reflection coefficients by about 25%. Kindle suggests that this is probably due to the viscosity and periodic numerical smoothing built into his computer model, and chooses to use the numerical values in comparing his analytical solution (7.33) with the results of the full numerical integration. Although this pastiche of numerical reflection coefficients with the theoretical solution is a little inconsistent, it is nonetheless true that there is dissipation in the real ocean. While damping will little effect on w_K and w_R, the directly excited waves, it will have some effect on the first reflections, making their amplitude lower than in an inviscid calculation, and will more and drastically suppress the higher order reflections. Kindle's numerical reflection coefficients have this dissipative reduction built into them, but it must be noted that a good model or parameterization for the dissipation of the real ocean is not available, and his viscosity was included more for numerical stability than for physical realism.

The large amplitude of the directly excited Kelvin wave, which is opposite in sign to that of the Yoshida jet, results in a near-cancellation whereas the Rossby wave, though smaller in amplitude, adds its own upwelling to that produced directly by the interior solution. The result is that the initial motion of the thermocline is four times greater at the eastern boundary than at the western boundary (for an impulsive turn on of the wind; for periodic forcing, this simple picture is modified by the reflected waves).

It is also interesting that the eastern boundary reflection coefficient is only half as large as that at the western boundary in both the inviscid analytical and viscous numerical calculations. Kindle notes that Moore's (1968) analysis of the reflection problem, discussed in Chap. 6, showed that when a wave reflects from a *western* boundary, energy is funneled *downwards* in *mode number* so that the solution involves only waves of smaller n. At an eastern boundary, on the other hand, energy is shifted into all higher modes, so the $n = 1$ Rossby mode receives a relatively small share of the energy of the incident Kelvin wave because it must share with $n = 3, 5, 7, \ldots$.

7.5 Atlantic Versus Pacific

Observations show [3] that the Pacific has a weak annual signal in the equatorial thermocline with most of the variation from the annual mean occurring at the eastern end. In marked contrast, the Atlantic has a strong *annual* signal at *all* longitudes. An obvious motive for exploring how the ocean responds to periodic forcing is to determine if the differing widths of the two oceans is responsible for the weakness of the Pacific annual cycle.

Both the numerical work of Kindle (1979) and the analytical model of [3] show quite the opposite: the response becomes *stronger* and more complex as the width of the ocean increases. In the long wave theory, the maximum thermocline variations at the boundary for the Pacific annual [>8 at both coasts] is significantly larger than for the Atlantic annual [∼5 at the western boundary and ∼7 at the eastern boundary] for the same strength of forcing (unity).

Kindle (1979) shows a plot of the upwelling amplitude (proportional to ϕ) versus the basin width-and-frequency parameter Φ (instead of time, as in the Cane–Sarachik diagrams). The pattern of contours shows quite clearly that the amplitude is increasing with Φ and reaches truly large values for $\Phi \sim \pi/2$, which corresponds to the lowest Cane–Moore quasi-normal mode, and is truly enormous for $\Phi = \pi$, which is the second Cane–Moore mode. (Kindle points out that in either case, the condition for resonance is that all waves cross the ocean and return with exactly the same phase as when they first left the boundary. However, the resonance is weaker at $\Phi = \pi/2$ than at $\Phi = \pi$ because waves which leave one boundary in phase with the wind stress forcing arrive at the other boundary exactly 90 degrees out of phase with the forcing. For the first baroclinic mode, the very strong $\Phi = \pi$ resonance occurs only for *semiannual* period, and only when the effective width of the ocean is taken to be ∼15,000 km.)

Still, the picture is clear. Resonances do occur even in a numerical model like Kindle's which allows the coastal Kelvin waves to radiate poleward and also includes dissipation. The Pacific is far closer to resonance than the Atlantic, and should, all things being equal, have a much stronger response.

The only way out of this dilemma is to suppose that annual component of the wind stress forcing is much weaker in the Pacific than the Atlantic. In fact, this seems to be the case. The mean wind stress in both oceans is about 0.4 dynes cm^{-2}, but the

amplitude of the Atlantic annual is about 0.15 dynes cm^{-2} whereas it varies from 0.02 to 0.1 dynes cm^{-2} in the Pacific [9, 10].

Although this is not terribly obvious in their Figures, the smoothness of the contours for $\Phi = 0.5$ (Atlantic) versus the much more highly structured contours for $\Phi = 1.1-1.5$ (Pacific) show that the waves are much less important for the smaller ocean. In the Atlantic, "the solution is essentially in equilibrium with the wind stress" (Kindle, 1979) insofar as the *equatorial* thermocline is concerned; the refractive effects so evident in Fig. 7.6a occur only *away* from the equator. The implications of this wave-free simplicity and the larger signal imply that the Atlantic is a much better laboratory for studying the annual cycle than the Pacific. In contrast, the interannual variations of wind in the Pacific are bigger than in the smaller ocean, and these longer period changes correspond to smaller values of ω such that the response is closer to equilibrium. Thus, the Pacific is likely to be a good place to explore long-term variations (time scale of several years) in the ocean due to the wind.

7.6 Summary

The overall conclusion is that the broad features of the equatorial response to periodic wind forcing are understood. "Ringing" in the form of directly excited and reflected Kelvin and Rossby waves, the Yoshida jet-like interior, the refraction and focusing of Rossby waves, and local wind stress curl all play roles. The mean tilt of the thermocline and its annual and semiannual harmonics and those of the three equatorial surface currents (NEC, NECC, and SEC) are all reasonably well-reproduced by a simple 1–1/2 layer model. There are, however, many discrepancies in detail between theory and observation. It appears that these are due at least as much to the crudity of the wind stress data as to the limitations of the 1–1/2 layer model. The course for the future is largely one of refinement, correction, and the explanation of features of smaller scales (in both time and space).

References

1. Lighthill MJ (1969) Dynamic response of the Indian Ocean to onset of the S. W. monsoon. Phil Trans 265:45–92
2. Wunsch C (1977) Response of an equatorial ocean to a periodic monsoon. J Phys Oc 7(4):497–511
3. Cane MA, Sarachik ES (1981) The response of a linear baroclinic equatorial ocean to periodic forcing. J Marine Res 39:652–693
4. Cane MA, Sarachik ES (1979) Forced baroclinic ocean motions, III: the linear equatorial basin case. J Marine Res 37(2):355–398
5. Schopf PS, Anderson DLT, Smith R (1981) Beta dispersion of low-frequency Rossby waves. Dyn Atmos Ocean 5(3):187–214

6. Merle J (1983) Seasonal variability of subsurface thermal structure in the tropical Atlantic Ocean. In: Nihoul JCJ (ed) Hydrodynamics of the Equatorial Ocean. Elsevier, Amsterdam, pp 31–50
7. Kindle JC (1979) Equatorial Pacific ocean variability — seasonal and El Niño time scales. PhD dissertation, Florida State University, Department of Oceanography, 134 pp
8. Moore DW (1968) Planetary-gravity waves in an equatorial ocean. Applied Mathematics, Harvard University
9. Meyers G (1979a) Annual Rossby wave in the tropical North Pacific Ocean. J Phys Ocean 9(4):663–674
10. Meyers G (1979b) Annual variation in the slope of the 14C isotherm along the equator in the Pacific Ocean. J Phys Ocean 9:885–891

Chapter 8
Impulsive Forcing and Spin-Up

Abstract On the equatorial beta-plane, the response of the flow in the linear shallow water wave equations to wind stress which is a step function in time can be formally solved by Laplace transform integrals (multiplied by the usual Hermite functions). Although these transform integrals must evaluated numerically, it is still possible to analyze the oceanic response to jumps in the wind. The work of Anderson and Rowlands (J Mar Res 34(3):295–312, 1976) [1] and Cane and Sarachik (J Mar Res 35(2):395–432, 1977) [2] is the point of entry to the complex wave patterns that result.

> An analytical solution, if sufficiently complicated, is no more useful, and useful only in the same ways, as a numerical solution.
>
> — J.P. Boyd

8.1 Introduction

In this chapter, we will deal with the response of the linear, 1–1/2 layer ocean to wind stresses which are applied impulsively. Usually, for the sake of simplicity, we shall assume that the wind is turned on as a step function in time. This sort of sudden, rather than periodic, change in the wind is an appropriate means of understanding El Niño events and perhaps the seasonal reversal of the Somali Current off East Africa. To be sure, both are part of quasi-periodic variations of the ocean-atmosphere system, but the data suggest that both events occur on time scales short in comparison to those of the interannual (El Niña) or annual (Somali Current) oscillations in which they are embedded.

In principle, of course, any time dependence of the forcing can be represented as a Fourier integral over time and is therefore just the superposition of the periodic forcing results obtained in the previous chapter. In practice, Fourier methods are a clumsy tool for understanding spin-up and switch-on problems because the Fourier integral may have properties and behavior very different from those of the single-frequency

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_8

solutions that are the integrand. Our purpose is to gain insight, so it is more useful to look at impulsive forcings directly.

The next section is an analysis of Anderson and Rowlands' [1] on the reflection of a switch-on [not periodic] Kelvin wave from the eastern boundary. They do not make the long wave approximation and present some interesting time sequences that clearly show that low latitudes respond faster than high latitudes and also illustrate the generation of the coastal signal.

In Sect. 8.3, we will briefly review the spin-up problem in the absence of boundaries; this is basically the Yoshida jet, but we will also discuss the effects of y-variations in the forcing. In the rest of the chapter, we will then discuss, along the lines of Cane and Sarachik [2], the corresponding solution with boundaries.

8.2 The Reflection of the Switched-On Kelvin Wave

This problem was treated by Anderson and Rowlands [1] as a warm-up to a study of the Somali Current reversal which was published in the same issue [3]. The switched-on Kelvin wave is assumed to be of the form

$$S_0(x,t) \equiv H(x-t) \qquad \Leftrightarrow \qquad S_0 = \begin{cases} 0 & t < x \\ 1 & t > x \end{cases} \tag{8.1}$$

where S_0 is the zeroth-th degree Hermite coefficient of the usual sum variable S. The eastern boundary is at $x = 0$ (not $x = X_E$ as in the Cane-Sarachik work, and elsewhere in this book). Thus, the Kelvin wave is implicitly assumed to be switched on far upstream for some negative time and then propagates until it hits the coast at $t = 0$. Equation (8.1) is assumed to be the complete solution to the Kelvin wave for all time, which is an assumption that can be made without approximation or apology since the Kelvin mode (i.e., S_0) is completely uncoupled from all the others. The other modes are assumed to be 0 until $t = 0$, but when the Kelvin wave hits the boundary, all the higher modes are switched-on so that the boundary condition $u = 0$ at $x = 0$ can be satisfied.

Anderson and Rowlands ignore all effects of a western boundary by implicitly assuming that the ocean is unbounded to the west, north, and south. This allows them to retrieve the general linear solution by taking Laplace transforms in time. The linear partial differential equations in x and t are thus reduced to constant coefficient equations in x alone which are easily solved to give the Laplace transform of each S_n or D_n.[1]

The bad news is that the final integral is very messy and cannot be done in closed form even for D_0, and the complexity rapidly mounts for the corresponding integrals

[1] Notational comment: Anderson and Rowlands use q and r to denote our sum and difference variables S and D, respectively. They work in terms of parabolic cylinder functions (asymptotically proportional to $\exp(-0.25y^2)$ instead of Hermite functions, which are proportional to $\exp(-0.5y^2)$, so there are lots of unexpected factors of 2 in their paper.

for higher modes. However, it is easy to evaluate the Laplace Transform numerically. The results for D_0 are shown in Fig. 4 of Anderson and Rowlands and Fig. 8.1 here. The most striking feature is the sharp spike at the leading edge of the wavefront. This front propagates at the same speed as the Kelvin wave ($c = -1$, nondimensionally) but in the opposite direction, and is due to the fact that Anderson and Rowlands do not filter out the gravity waves by using the long wave approximation. By asymptotically evaluating the integral in the vicinity of the front, Anderson and Rowlands show that (viewed in a coordinate system moving with the front), it can be described by the same $J_0([2xt]^{1/2})$ factor that describes the western boundary layer for the spin-up of a bounded, midlatitude ocean. This similarity exists because gravity waves are indifferent to their direction of propagation (eastward from a western boundary, or westward from an eastern coast) and are the waves responsible for the steep gradients in either case. As is visible in the figure, the width of the first spike shrinks so that it is $O(1/t)$, so for large times, we would expect viscosity would destroy it. Consequently, we should not be too upset when these gravity wave spikes are filtered by the pseudo-long wave approximation employed by Anderson in his earlier solo paper, Anderson [4], and by the long wave approximation employed in our later illustrations below.

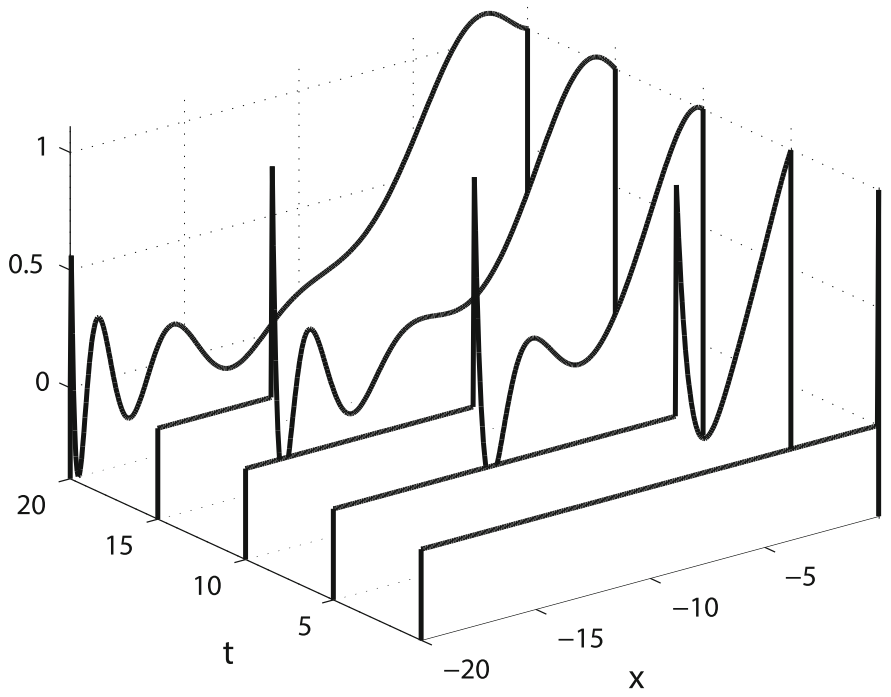

Fig. 8.1 $D_0(x, t)$ in the exact model as a function of x and t as generated by the reflection of a Kelvin wave at $t = 0$ from an eastern boundary at $x = 0$

They interpret the poleward front as the coastal Kelvin wave which is illustrated in their Figs. 5, 6 and 7. The wave front amplifies as it propagates like $y^{1/2}$, just as predicted by Moore's solution. This amplification is a purely geometric effect caused by the narrowing of the Kelvin wave as the stronger Coriolis forces at higher latitudes trap the wave into a narrower and narrower layer along the coast. In the wake of the front, the Kelvin signal subsides to a value that Anderson and Rowlands show that at all latitudes is $1/2^{1/2}$ [1/2 in our convention] of the amplitude of the equatorial Kelvin wave. The coastal Kelvin wave asymptotes to a constant (rather than 0) because the equatorial Kelvin wave is unbounded to the west and therefore keeps pumping a constant amount of energy into the coastal boundary layer for all $t > 0$. This constant is less than 1 because some of the energy of the incoming Kelvin wave (of unit amplitude) is reflected westward as Rossby waves.

The meridional velocity is strongly trapped to the coast for all times as shown at four different times by contour plots in Fig. 11 of Anderson and Rowlands; v drops off to very small amplitudes after the front of coastal Kelvin wave has passed. Part of the reason is that, near the coast but well in back of the head of the coastal Kelvin wave, the solution is a sum of Rossby waves which are not zonally trapped. Cane and Sarachik [2] point out that in a switch-on problem, the frequency component $\omega = 0$ is always present. Thus, all Rossby latitudinal modes can propagate some energy westward, and for moderate to large times when the high frequency transients have propagated off the map, only the low frequency (and zonally unbounded) Rossby waves remain. Cane and Sarachik therefore object to Anderson and Rowland's description of the coastal Kelvin wave asymptoting to $1/2^{1/2}$ because it is actually the sum of zonally propagating Rossby waves that is doing the asymptoting except near the front of the coastal signal.) These long, low-frequency Rossby waves have $v << u$; this observation is the basis for the long wave approximation, and v becomes smaller as the frequency 0 becomes lower. Asymptotic analysis of the Fourier integrals over ω in the switch-on problem show that for large t, the integrals are dominated by smaller and smaller values of ω, so v drops off to nothing rather quickly.

Anderson [4] reduced the system of three equations for each latitudinal mode number n to a single equation for each n [4]. To avoid severe explicit time-stepping restrictions, Anderson dropped the third time derivative from this reduced, third order equation, thus filtering out gravity waves. This is less drastic than the long wave approximation in that the zonal velocity and height are not forced into perfect geostrophic balance in the y-direction, but the gravity waves are filtered out. His Figs. 6 and 7 shows that this "pseudo-long wave" approximation eliminates the gravity wave spikes, but otherwise, since Rossby wave dispersion is still included in a very accurate form, it changes the solution but little.

To elaborate on their ideas, it will suffice here to use the long wave approximation. The reflection of the Kelvin wave derived in Sect. 6.4 above as Eq. 6.27:

$$S(x,y,t) = \overline{H}(x-t)\psi_0(y) + \sum_{n=1}^{\infty}\left\{\prod_{k=1}^{n}\sqrt{\frac{(2k-1)}{2k}}\right\}\overline{H}\left((4n-1)\{x+[1/(4n-1)]t\}\right)\psi_{2n}(y)$$

$$D(x,y,t) = \sum_{n=1}^{\infty}\sqrt{\frac{n+1}{n}}\left\{\prod_{k=1}^{n}\sqrt{\frac{(2k-1)}{2k}}\right\}\overline{H}\left((4n-1)\{x+[1/(4n-1)]t\}\right)\psi_{2n-2}(y) \quad (8.2)$$

Each reflected Rossby mode in the long wave approximation (which is free of dispersion within a given latitudinal mode) is a step-function pulse in x as illustrated in Fig. 8.2. The superposition of many modes, each traveling at its own speed, creates the more complex pattern illustrated by the contour plot in Fig. 8.3.

It is possible to roughly estimate the boundaries of the wave pulse as it expands both poleward and westward. At a time and longitude (x,t), only those latitudinal modes are to be found for which $|ct| > |x|$, that is, $t > (2n+1)(-x)$ where n is the latitudinal mode number. As elaborated in Appendix A, the n-th Hermite function has "turning latitudes" $y_t = \sqrt{2n+1}$ where the function transitions from oscillation (for smaller $|y|$) to very rapid exponential decay (for $|y| > |y_t|$). If follows that the latitudinal extent of the disturbance at (x,t) is roughly bounded by the turning latitudes of the highest mode which has reached longitude x at time t. The highest

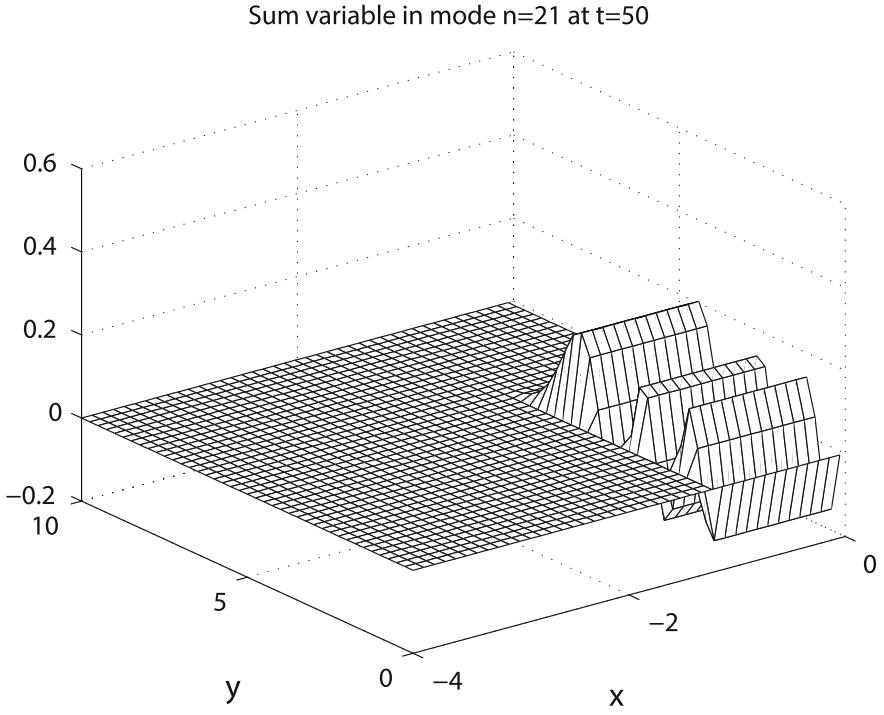

Fig. 8.2 Surface plot of $S_{22}(x, y, t = 50)$, the coefficient of the sum variable in latitudinal mode 21

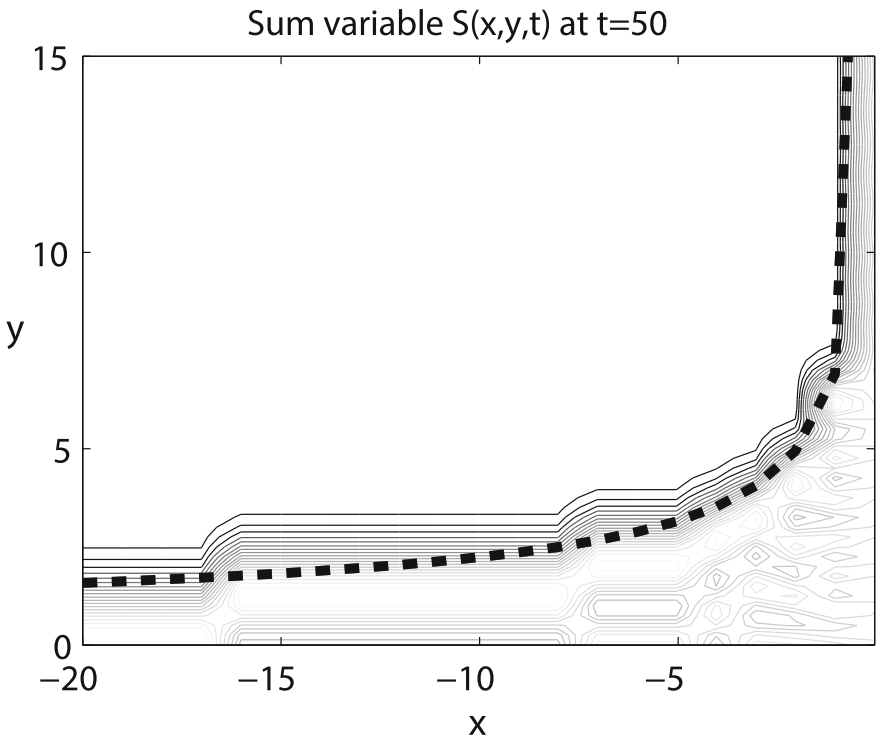

Fig. 8.3 The *thick dashed line* is the wavefront. The *thin gray curves* are the isolines of $S(x, y, t = 50)$, the sum of all modes, as opposed to the single mode illustrated in the previous figure

mode has $t/|x| \approx 2n+1$, but the right-hand side is the square of the turning latitudes for the mode, yielding for the frontal position $y_{front}(x, t)$

$$y_{frontal}^2 = \frac{t}{-x} \tag{8.3}$$

This is graphed as the heavy dashed line in Fig. 8.3 and also in the "zoom plot" for larger t, Fig. 8.4.

Anderson and Rowland's Figs. 9 to 11 are each a quartet of contour plots illustrate ϕ, u, and v at various times as calculated via their numerical model. Initially, sharp gradients of pressure form in a coastal boundary layer, but by $t = 20$, the front has propagated off the map, and the rather flat and much smoother field associated with the Rossby waves is left behind. Notice that the disturbance is spreading westward at $y = -8$ [bottom of graph] throughout all four plots in each of their figures. At this high latitude, only Rossby modes with $n > 8$ have significant amplitude, so the effective rate of propagation of the sum of reflected Rossby waves at this latitude is bounded by $c_8 = -1/17$.

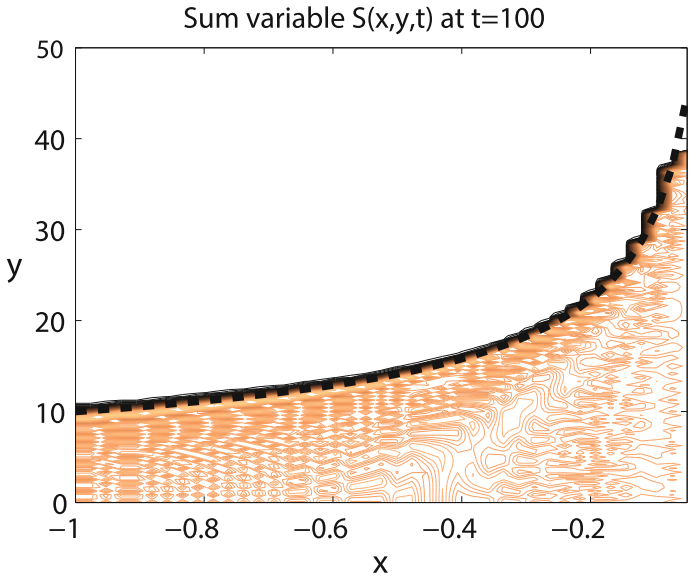

Fig. 8.4 Same as previous figure except that the time is doubled to $t = 100$ and the zonal scale is zoomed in (shrunk) to an interval 20 times smaller. The *thin curves* are the isolines of $S(x, y, t = 100)$

The westward spreading is much more subdued for u than for the height ϕ because at the boundary, the Rossby modes and coastal front must cancel the zonal velocity of the equatorial Kelvin wave so that $u = 0$ everywhere along the coast. It is interesting that for large times a small region of westward flow develops along the equator and near the coast.

8.3 Spin-Up of a Zonally-Bounded Ocean: Overview

In the absence of coastal boundaries, the problem of a switched-on wind stress is that of the Yoshida jet, discussed in Chap. 9. The response to an x-independent zonal stress is a zonal jet that accelerates linearly with time plus inertial waves needed to satisfy the initial conditions of no motion at $t = 0$. The response to a north- south stress is merely a steady jet also accompanied by inertial oscillations.

Boundaries at $x = 0$ and $x = X_E$ modify the response in the following two ways:

(A) The wind stress forcing is cut-off at the boundaries so that an x-independent wind creates a forcing in the shape of a "top hat" or "square well".

(B) The boundaries are barriers to zonally propagating motions so that any energy which is driven onto the boundaries in the Yoshida solution will inevitably generate reflected waves to carry the incident energy away.

Both types of modifications are important.

Overall, the solution consists of five parts: (i) The interior (Yoshida) solution $\bar{u}_I$, which is independent of x.

(ii) The inertial gravity waves (a) needed to satisfy the initial conditions in the Yoshida problem *plus* (b) their *boundary reflections*.

(iii) The eastern boundary response consisting of (a) long Rossby waves needed to bring $\bar{u}_I$ to 0 at the eastern boundary and (b) reflections of Kelvin waves (which are also long Rossby waves plus a coastal signal.)

(iv) The short Rossby wave part of the western boundary response, which involves (a) waves which superimpose to give a thinning boundary layer described by Bessel functions so as to bring $\bar{u}_I$ to zero plus (b) the short Rossby reflections of long Rossby waves. (By "thinning" we mean that its width is $\sim O(1/t)$]

(v) The Kelvin wave part of the western boundary response, which consists of (a) a wave directly excited at $t = 0$ plus (b) reflections of long Rossby waves.

In terms of the two species of Yoshida modifications described above, parts (a) of the last three categories correspond to (A): the zonal truncation of the Yoshida jet. Parts (b) of the last (four) categories above correspond to (B): coastal boundaries-as-wavemakers-via-reflection.

Parts (iii.a) and (iv.a) also appear in the solution to the barotropic vorticity equation for a bounded ocean. The long Rossby waves which propagate westward from $x = X_E$, (iii.a), are the wavefront that spins-up the interior of the sea to the Sverdrup solution. Since the boundary does not modify these long Rossby waves, they are unchanged by the presence or absence of the coast – what is important to them is merely that the *forcing* stops at $x = X_E$. The Bessel function-like, ever-thinning western boundary layer of short Rossby waves is needed to bring the Sverdrup flow to 0 along the western boundary in the solution to the barotropic vorticity equation; the fact that $x = 0$ is a coast, and not merely the edge of the wind stress, is essential to the existence of this thinning boundary layer. The new kid on the block, the one feature which was not found in the solution of the barotropic vorticity equation, is the Kelvin wave. Because of it, one of the waves which is needed to bring u to 0 at the western boundary can now propagate energy eastward. In theories of El Niño, it is this wave which propagates the decisive downwelling signal that Peruvian fishermen have observed – and dreaded – for centuries.

In the next few sections, we shall successively review the interior solution, the inertial-gravity waves, the western boundary layer of short Rossby waves, and then finally the Kelvin and long Rossby waves that dominate the spin-up of the equatorial ocean.

8.4 The Interior (Yoshida) Solution

In Chap. 9, we derive the solution for an x-independent, y-independent wind stress acting in the x-direction only. The methods employed in Sect. 9.3 can be easily generalized to include (a) a meridional wind stress (b) a buoyancy source (i.e., a

heating function which provides forcing for the height equation) and (c) latitudinal variations of the wind stress. Since the method is the same, we shall merely quote results from Cane and Sarachik [2].

Before doing so, it is helpful to define some auxiliary 3-component vectors that give the meridional structure of each mode. Let

$$\bar{w}_n(y) \equiv (y\,\psi_n, 0, -d\psi_n/dy), \qquad (n \geq 0) \tag{8.4}$$

$$\bar{v}_n(y) \equiv (0, \psi_n, 0) \qquad (n \geq 0) \tag{8.5}$$

$$\bar{M}_n(y) \equiv (-d\psi_n/dy, 0, y\,\psi_n) \qquad (n \geq 0) \tag{8.6}$$

$$\bar{R}_n(y) \equiv \left\{(2n+1)\,\bar{M}_n(y) - \bar{w}_n(y)\right\} / [4n(n+1)] \tag{8.7}$$

$$\bar{M}_K(y) \equiv (\psi_0, 0, \psi_0)/1.4141 \tag{8.8}$$

where $\psi_n(y)$ is the n-th *normalized* Hermite function. The solution for a forcing vector that involves only a zonal wind stress and buoyancy source can be written

$$\bar{F} = (F(y), 0, \overline{H}(y)) \qquad [\text{independent of } x, t] \tag{8.9}$$

$$\bar{u}_I = (t\,u^{(1)},\ v^{(1)},\ t\,\phi^{(1)}) \tag{8.10}$$

$$(u^{(1)}, 0, \phi^{(1)}) = d_K\,\bar{M}_K(y) + \sum_{n=1}^{\infty} r_n\,\bar{R}_n(y) \tag{8.11}$$

$$v^{(1)} = -\sum_{n=0}^{\infty} [d_n/(2n+1)]\,\psi_n(y) \tag{8.12}$$

[Notational comment: Cane and Sarachik [2] use $\bar{u}_I$ to denote the inertial-gravity part of the problem, and, unlike us, have no separate symbol to denote the accelerating jet.) We adopt the notation

$$A_n = \int_{-\infty}^{\infty} A\,\psi_n(y)dy = n\text{-th Hermite coefficient of } A \tag{8.13}$$

for any quantity A. The Hermite coefficients in (8.11) and (8.12) are

$$d_K = 2^{-1/2}\,(F + \overline{H})_0 \tag{8.14a}$$

$$d_n = (yF + d\overline{H}/dy)_n \tag{8.14b}$$

$$r_n = (dF/dy + y\overline{H})_n - d_n/(2n+1) \tag{8.14c}$$

The corresponding solution for a purely meridional wind stress is given by

$$\bar{F} = (0, G(y), 0) \qquad \text{[independent of } x, t\text{]} \tag{8.15}$$

$$\bar{u}_I = (u^{(2)}, 0, \phi^{(2)}) \tag{8.16}$$

$$= \sum_{n=0}^{\infty} [g_n/(2n+1)]\, \bar{w}_n(y) \tag{8.17}$$

where

$$g_n = (G)_n = (\psi_n(y), G) \tag{8.18}$$

As will be noted in our discussion of Yoshida's problem, a latitudinally unbounded stress implies a solution which decays only algebraically with y. This in turn implies that the coefficients a_n of the Hermite series decrease only algebraically with n — small, fractional powers of n — so Moore/Hutton/Euler summation, described in Appendix B, must be employed. Cane and Sarachik note that if the forcing is a Gaussian in y, then the exponential decay of the stress will cause the solution to decay exponentially with $|y|$ also and in turn the coefficients of its infinite series to decay exponentially with degree n as well so that the series can be summed without special tricks. If the width of the Gaussian is wide, then the solution will differ only a little from Yoshida's. Figure 8.5 illustrates the structure functions for various types of wind stresses. One does not sacrifice an analytical solution by using a Gaussian since the Hermite expansion of such a function is known explicitly:

$$\exp(-b^2 y^2) = \sum_{n=0}^{\infty} \pi^{1/4} \frac{\sqrt{2}}{\sqrt{1+2b^2}} \frac{\sqrt{(2n)!}}{2^n n!} \left| \frac{1-2b^2}{1+2b^2} \right|^n \psi_{2n}(y) \tag{8.19}$$

where the sum is over the even coefficients only. For small b, the coefficients that satisfy $(2n+1) << 0.5b^2$ are essentially the same as in the Hermite expansion of $f(y) = 1$, i.e., as in Yoshida's problem. Cane and Sarachik make heavy use of such Gaussian wind stresses [2, 5–7].

Although we shall not go through the details here, Cane and Sarachik [5] describe how to calculate the response to forcings with an arbitrary dependence on x. The required Fourier integrals are rather messy, however, so most work has either used brute-force numerical methods or idealized wind stress as a step function or "top hat" in x.

The x-independent solutions $\bar{u}_I$ derived here have been labelled with the subscript "I" to denote that they are the "interior" solution: away from the boundaries, $\bar{u}_I$ dominates the flow for small times because none of the waves – which are entirely generated at the boundaries of the ocean – have had time to penetrate deep into the interior. At later times, the x-independent interior solution is overlaid with waves at all longitudes. In particular, the wave part of the solution *cancels* the secular growth with time implied by (8.10) so that the solution with coastal boundaries spins up to a flow which is bounded in time.

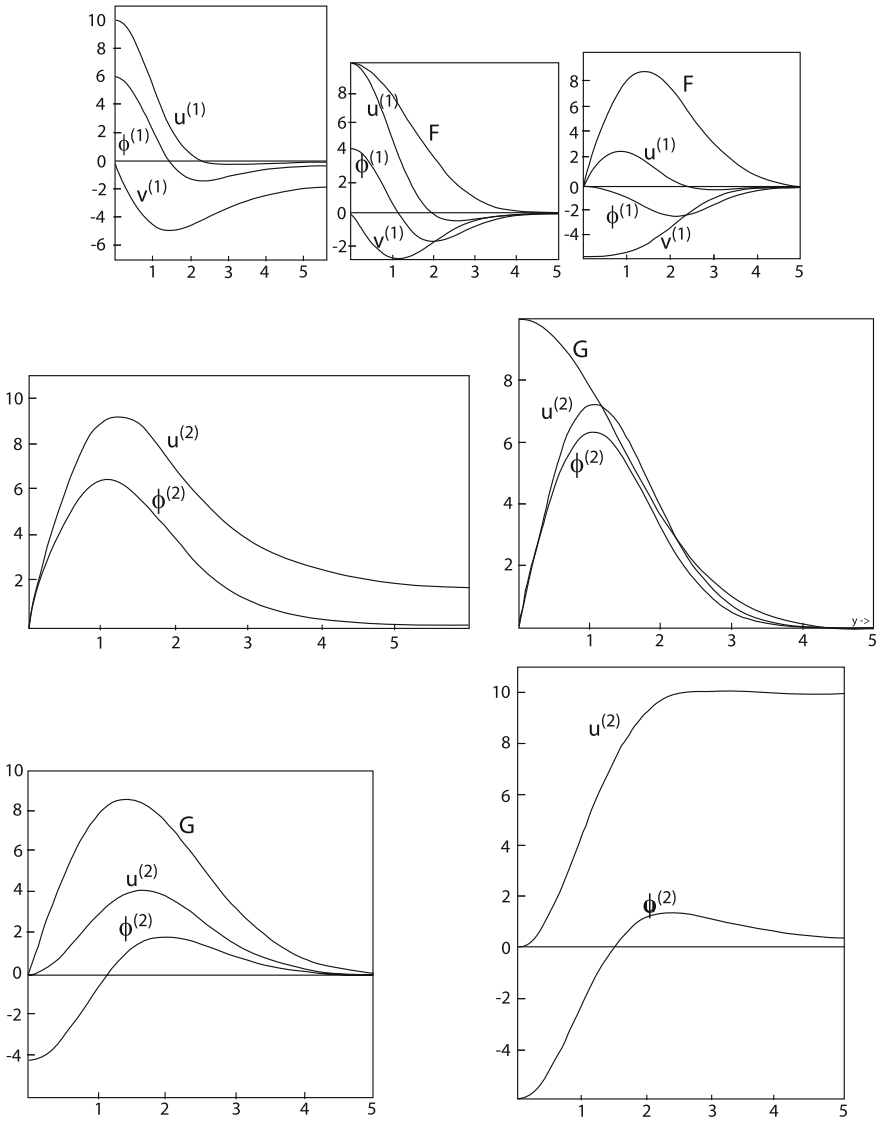

Fig. 8.5 *Top three panels* Response in an unbounded ocean to a westerly wind stress ($F \neq 0$) while the other forcing components are zero ($G = \overline{H} = 0$). *Left top* $F(y) = 1$. *Middle top* $F(y) = \exp(-y^2/4)$. *Right top* $F(y) = y\exp(-y^2/4)$. *Bottom four panels* Response to a southerly wind ($G \neq 0$), also in an unbounded ocean with $F = \overline{H} = 0$. *Middle left* $G(y) = 1$. *Middle right* $G(y) = \exp(-y^2/4)$. *Bottom left* $G(y) = y\exp(-y^2/4)$. *Bottom right* $G(y) = y$. Figures 7 (*top*) and 8 (*bottom four panels*) of Cane and Sarachik [5]

8.5 Inertial-Gravity Waves

Retain the same auxiliary functions as in the previous section. When a forcing is switched on as a step function at $t = 0$, Cane and Sarachik [2, 5] show that, to satisfy the initial conditions of no motion, the general linear solution for the inertial-gravity waves is

$$\bar{u}_{inertial} = \sum_{n=0}^{\infty} m^{-2} \left[\bar{w}_n(y) + \bar{v}_n(y)\, \partial/\partial t\right] \left\{ \frac{1}{\sqrt{2n+1}}\, d_n \, \sin(\sqrt{2n+1}\, t) + g_n \, \cos(\sqrt{2n+1}\, t) \right\} \tag{8.20}$$

As in the Yoshida problem, $\bar{u}_{inertial}$ is independent of x. However, contrary to what one might expect, the inertial- gravity waves in (8.20) have an *eastward* group velocity and this has rather important consequences when zonal boundaries are included. To calculate this group velocity, we note that the usual gravity wave approximation

$$\omega \approx \pm\sqrt{2n+1+k^2} \tag{8.21}$$

predicts a zero x-group velocity for $k = 0$. This immediately tells us that the exact group velocity for $k = 0$, whatever it is, has to be very small, or otherwise it would not have been missed by what is otherwise a good approximation. There are two ways of obtaining the correct group velocity.

The most direct method is to simply take the exact cubic dispersion relation

$$\omega^3 - [2n+1+k^2]\omega - k = 0 \tag{8.22}$$

and, using (8.21) as the zeroth-order approximation, apply Newton's method to generate higher approximations. The next lowest approximation is given by

$$\omega = \pm\sqrt{2n+1+k^2} + k/[2n+1+k^2] \tag{8.23}$$

Differentiating with respect to k and then taking the limit $k \to 0$ gives

$$c_{gx}(k=0) = 0.5/(2n+1) \tag{8.24}$$

for gravity waves of either sign of frequency. Thus, the group velocity of the inertial-gravity waves is (i) always eastward and (ii) of a magnitude equal to exactly half the velocity of the corresponding (same n) long Rossby mode.

When zonal boundaries are added, the situation becomes that shown schematically in Fig. 8.6. Since there are (obviously!) no inertial-gravity waves at $t = 0$ west of the boundary at $x = 0$ — there is no water west of $x = 0$ — it follows that for

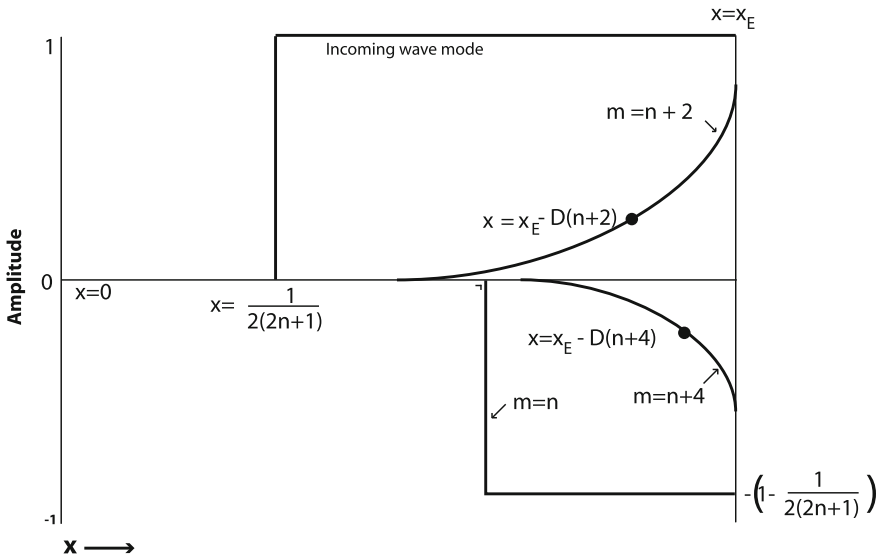

Fig. 8.6 Inertial-gravity equatorial wave of latitudinal mode number n and its boundary responses at time t for the Yoshida problem with boundaries at $x = 0$ and $x = x_E$. The directly excited initial mode reflects as a very long gravity wave of the same mode number and almost the same speed. That is to say, the western boundary cancels the initial wave for $x \leq t/\{2(2n+1)\}$. The eastern boundary reflection is a series of modes of mode number $m, m+2, m+4, \ldots$. Only the n-th reflected mode propagates; the higher reflected modes are trapped with an e-folding scale for the m-th mode given by $D(m) = \sqrt{2(m-n) - 1/4(2n+1)^{-1}}$. A little of the initial energy thus goes into a zonally-trapped eastern boundary layer. Figure 3 of Cane and Sarachik [2]. Note that $D(n+2)$ is plotted for D_{n+2}

$t > 0$, the gravity wave signal propagates eastward towards the interior of the ocean. Conceptually, what happens at the western boundary is that a given inertial-gravity mode merely generates the same mode at the boundary with equal amplitude but opposite phase. As time passes, the superposition of the two – which sum to 0 – propagates eastward at the velocity given by (8.24). One can hardly dignify this as a western boundary "reflection" since all the energy is being swept *away* from the boundary by the intrinsic free group velocity of the waves.

At the eastern boundary, the situation is more complicated. The only reflected wave that is zonally propagating has $m = n$, i.e., the same mode number as the incident wave, and it has very nearly the same energy, too: if the amplitude of the incident wave is normalized to 1, then the amplitude of the $m = n$ reflected wave is $[1 - 1/(2n+1)]$ as shown in Fig. 8.6, which is almost indistinguishable from the initial wave for large n. The small fraction of energy that is left over goes into an infinite number of zonally-trapped modes with mode numbers greater than that of the incident wave (i.e., $m > n$) and the same parity ($m = n \mod 2$). The reflected long wave has $k \approx 1/(2n+1)$, which is certainly a large wavelength, but it is *not* proper to speak of this as a Rossby wave because it – and all the zonally-trapped

modes – oscillate at the same high frequency ($\omega = [2n + 1]^{1/2}$) of the incident wave, and such high frequencies are an order of magnitude too large for Rossby waves. (Thus, there is no coupling at the eastern boundary between the high and low frequency parts of the response, i.e., between the inertial-gravity waves and the long Rossby waves which we will discuss later.)

The long term outcome is that the inertial-gravity waves ping back and forth across the ocean basin for a long time, gradually losing energy into the zonally-trapped motions at the eastern boundary. It should be noted that the time for even one crossing of the ocean is very slow: even the fastest-moving $n = 1$ mode takes a whole year to cross the width of the Pacific. The return mode has *slightly* less energy and a *slightly* lower group velocity than the incident mode, "slightly" meaning $O[1 - 1/(2n + 1)]$ or smaller. When the return mode reflects off the western boundary, a little of the energy goes into a Kelvin mode (among others), but most is put into $k = 0$ waves of the same latitudinal mode number. As a first approximation, however, the mode simply bounces endlessly back and forth.

The inertial-gravity waves therefore never settle down into a steady-state in and of themselves. Their ultimate fate is to be destroyed by viscosity. They play no role whatsoever in the spin-up of the equatorial ocean to a Sverdrup flow, but rather are simply high-frequency noise superimposed on this adjustment process. For this reason, we shall not discuss these waves further. Note, however, that such modes do play a role as constituents of the rather high frequency quasi-free basin modes of Gent which were discussed in Chap. 6.

8.6 Western Boundary Response

The zonal velocity along the western boundary has a time dependence of the form

$$u(y, t) = u(y)t^s, \qquad s = 0 \text{ or } 1; \tag{8.25}$$

A similar equation but with a different shape in y can be written for the eastern boundary. The case $s = 0$ applies to a cross-equatorial wind stress whereas $s = 1$ denotes a steady zonal stress, the accelerating Yoshida jet. Cane and Sarachik [2] show that the western boundary response $\bar{u}^W$ (which excludes the incoming long Rossby waves) is

$$\bar{u}^W(x, y, t) = b_K \, H(t - x) \, (t - x)^s \, \bar{M}_K(y) + \bar{u}^B(x, y, t) \tag{8.26}$$

The Kelvin wave term is exact – note the similarity for $s = 0$ to the form assumed by Anderson and Rowlands [1] – but $\bar{u}^B$ is described by horribly messy integrals which cannot be integrated in closed form. Asymptotic approximations to the short Rossby waves that superpose to give $\bar{u}^B$ can be obtained, however, and the result can be written either of two ways. The first form is

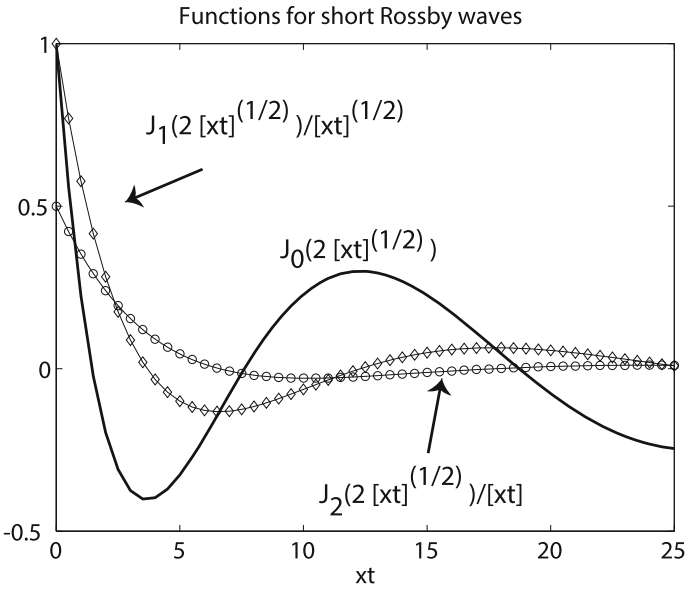

Fig. 8.7 The Bessel functions which describe the short Rossby waves in the western boundary layer. Note that the abscissa is xt, not x: this implies that all these functions shrink in longitudinal scale like $O(1/t)$

$$\bar{u}^B(x, y, t) \sim \left| \begin{array}{l} (t/x)^{s/2}\, J_s(2[xt]^{1/2}) \sum_{n=0}^{\infty} b_n\, H(c_n t - x)\, (d\psi_n/dy, 0, y\,\psi_n) \\ -(t/x)^{(s+1)/2}\, J_{s+1}(2[xt]^{1/2}) \sum_{n=0}^{\infty} b_n\, H(c_n t - x)\, (0, \psi_n, 0) \end{array} \right| \tag{8.27}$$

where the first sum gives the latitudinal structure of the zonal velocity (first component of the vector) and ϕ (third component) whereas the second sum gives the structure of the north-south current. The Bessel functions that appear in (8.27) are illustrated in Fig. 8.7.

The reason for the step functions in (8.27) [omitted from (8.28) below] is to serve as a reminder that (8.27) is merely an asymptotic approximation and is valid only close to the coast. The parameter $c_n = 1/[8(2n+1)]$ is the maximum eastward group velocity of the n-th mode; when $x > c_n t$, the n-th mode is exponentially small and the proper asymptotic approximation has a different form.

The second, alternative form is

$$\begin{aligned} \bar{u}^B(x, y, t) &= (u^B, v^B, \phi^B) \\ &= [-\partial/\partial y, \partial/\partial x, y]\, (t/x)^{s/2}\, J_s(2[xt]^{1/2}) \sum_{n=0}^{\infty} b_n\, \psi_n(y) \end{aligned} \tag{8.28}$$

The simple form of (8.28) is possible because the short Rossby waves are, for very low frequencies (i) nondivergent and (ii) in geostrophic balance in the x-direction,

i.e., v^B is in geostrophic balance with ϕ_x. These two observations remain true whether we are dealing with short Rossby waves at a single frequency $\omega << 1$, or whether we are dealing with those produced by a switch-on forcing, which gives a solution dominated by a mix of frequencies – all very low – for large time.

The condition of nondivergence,

$$u_x^B + v_y^B = 0 \qquad [+ O(\omega^2) \text{ error for a single frequency}] \tag{8.29}$$

can be integrated in y from $y = -\infty$ to $y = \infty$ to give, since $v(\pm\infty) = 0$,

$$\int_{-\infty}^{\infty} u_x^B(x, y, t)\, dy = 0 \qquad \forall\,(x, t) \tag{8.30}$$

In words, since (8.30) is true for *all* x, this equation implies that the y-integral of u is at most a constant independent of x. Since u^B is 0 at $x = \infty$ (or X_E), it follows that this constant must be 0. Thus,

$$\int_{-\infty}^{\infty} u^B(x, y, t)\, dy = 0 \qquad \forall\,(x, t) \tag{8.31}$$

We have, of course, already derived this equation in Chap. 6: the Kelvin wave carries all the mass flux away from the western boundary, and the short Rossby waves only create some local recirculation. It is worth recapitulating the argument here, however, because it gives a quick and efficient method for calculating the Hermite coefficients b_n that appear in (8.27) and (8.28).

The boundary condition at $x = 0$ is

$$-u_i(y)\, t^s = u^B(0, y, t) + 2^{-1/2}\, b_K\, t^s\, \psi_0(y) \tag{8.32}$$

where $u_i(y)$ is the incident wave disturbance (sum of long Rossby waves) and where the second term on the right in (8.32) is the contribution of the Kelvin wave. [The factor of $2^{1/2}$ is due to the normalization of the Kelvin wave employed by Cane and Sarachik [2].] Integrating this equation from $y = -\infty$ to $y = \infty$ and applying (8.31) gives

$$-\int_{-\infty}^{\infty} u_i(y)\, dy = \pi^{1/4}\, b_K \tag{8.33}$$

where we have used (i) $\psi_0(y) = \pi^{-1/4} \exp(-0.5y^2)$ and (ii) the integral from $[-\infty, \infty]$ of $\exp(-0.5y^2)$ is $(2\pi)^{1/2}$. With the Kelvin wave response thus completely determined by the zonal mass flux of the incoming waves, we can then obtain the Hermite coefficients of u^B at the boundary by calculating those of $u(y)$ through the usual method and then correcting that for ψ_0 by using (8.33). Cane and Sarachik [2] also describe a second method based on recurrence relations, but this is less efficient unless the incoming wave consists of but a single long Rossby mode.

Of course, as discussed in Chap. 6, the boundary layer of short Rossby waves has little physical importance, and it is best to ignore it. This is easier to do in theoretical treatments such as our study of Cane-Moore normal modes, than it is in numerical models. The computer codes will inevitably show the thinning, Bessel function-like layer, so it important that one is not confused into thinking the numerics have gone beserk. In what follows, however, we shall generally ignore the western boundary response except for the all-important Kelvin wave.

8.7 Sverdrup Flow on the Equatorial Beta-Plane

The steady-state equations of motion on the equatorial beta- plane take the form

$$-yv + \phi_x = F(y) \tag{8.34a}$$

$$yu + \phi_y = G(y) \tag{8.34b}$$

$$u_x + v_y = \overline{H}(y) \tag{8.34c}$$

Steady-state solutions are possible because of the assumption that the forcing functions are step functions of time, and thus constant in time after the switch-on at $t = 0$. The solution for v can be obtained by taking the curl of (8.34a) and (8.34b) (i.e., differentiating the x-momentum equation with respect to y and the y-momentum equation with respect to x and subtracting) to obtain

$$v = (G_x - F_y) - y\,(u_x + v_y) \tag{8.35}$$

The divergence term can be rewritten in terms of the forcing via (8.34c) to give the final vorticity equation

$$v = (G_x - F_y) - y\,\overline{H} \tag{8.36}$$

In the absence of the buoyancy forcing $\overline{H}$, (8.36) expresses the classical balance $\beta v = \text{curl}(\boldsymbol{\tau})$, where the vector $\boldsymbol{\tau}$, is the horizontal wind stress. The β does not explicitly appear in the left- hand side of (8.36) because the usual equatorial nondimensionalization sets $\beta = 1$, but the left-hand side of (8.36) is nonetheless dimensionally equivalent to βv.

The height equation (8.34c) can be written, using the explicit solution for v, as

$$u_x = -\left\{(G_x - F_y) - y\overline{H}\right\}_y + \overline{H} \tag{8.37}$$

Integrating this with respect to x then gives the Sverdrup solution for u. Note that u is determined by a *first* order equation: this implies that one can impose at most

one boundary condition in x upon u. This is the usual Sverdrup dilemma; its proper resolution is to demand $u = 0$ at $x = X_E$ and to add in a *time-dependent* western boundary layer. We have already discussed this boundary layer, and pointed out that it consists of short Rossby waves which sum to solutions that can be approximated by Bessel functions of argument $(2[xt]^{1/2})$. The fact that the width of the western boundary layer is $O(1/t)$ insures that it is *always* time-dependent and thus can resolve the "Sverdrup paradox" that the steady solution can satisfy only a single boundary condition. Integration gives

$$u = \int_{X_E}^{x} dx \left\{ [G_x - F_y]_y - [2\overline{H} + y\overline{H}_y] \right\} \tag{8.38}$$

The solution for the height field is the most complicated of the lot. Solving the x-momentum equation for ϕ_x and using (8.36) for v gives

$$\phi = \int_{X_E}^{x} dx \left\{ y\,[G_x - F_y] - y^2\,\overline{H} + f \right\} + A(y) \tag{8.39}$$

where $A(y)$ is an arbitrary function of integration. [Since the differential equation for ϕ involves x only, the constant of integration is free to vary with y). For u, the corresponding arbitrary function was 0 so that u could vanish at the eastern boundary $x = X_E$, but ϕ is under no such constraint. To find $A(y)$, we must apply the y-momentum equation at $x = X_E$. Since $u(x = X_E) = 0$, (8.34b) reduces to

$$\phi_y(x = X_E) \equiv A(y) = G(x = X_E, y) \tag{8.40}$$

Integrating with respect to y shows that $A(y)$ is equal to the integral of the boundary value of the north-south wind stress G plus another constant of integration ϕ_0 which is independent of both x and y. The full solution for ϕ is then

$$\phi = \int_{X_E}^{x} dx \left\{ y[G_x - F_y] - y^2\,\overline{H} \right\} + \int_{\infty}^{y} dy\, G(x = X_E, y) + \phi_0 \tag{8.41}$$

The constant ϕ_0 cannot be determined directly from the equations of motion (8.34) since ϕ appears only in differentiated form. Instead, ϕ_0 must be found via a mass conservation condition. If the integral of $\overline{H}$ over the basin is 0, then the integral of ϕ over the basin must vanish also. Cane and Sarachik [2] show that the contribution to the double integral of ϕ from the time-dependent western boundary layer is $O(t^{-1/2})$ so that it can be ignored. (The Sverdrup solution is the steady solution, remember, and is a good approximation only when t is large so that the transients from the switch-on have died away.) If the integral of $\overline{H}$ over the basin is *not* zero, then the integral for ϕ as given by (8.41) must increase linearly in time, but this corresponds to adding mass to the basin and is therefore rather artificial.

Just as on the midlatitude beta-plane, the zonal velocity for the equatorial Sverdrup solution depends upon the curl of the wind stress rather than wind stress itself. However, unlike the midlatitude case, there is a nonzero solution even if the curl is 0, given by (with $\overline{H} = 0$ also for simplicity)

$$\phi_x = F; \qquad \phi_y = G; \qquad u \equiv v \equiv 0 \qquad [\text{curl}](\boldsymbol{\tau}) = \overline{H} = 0] \tag{8.42}$$

i.e., the height field adjusts so that pressure gradients balance the wind stress. It is precisely this adjustment that is responsible for creating the tilt of the equatorial thermocline due to the zonal wind stress of the Trade Winds. Correspondingly, there will also be an adjustment process with radiation of transient Rossby and Kelvin waves when F and G change, and one can in fact model El Niño by changes in a zonal wind stress which is independent of x so that there is no curl at all. On the midlatitude beta-plane, in contrast, when the wind stress has zero curl, there is neither a steady solution nor a transient one.

Note that the two equations for ϕ for the zero-curl case do have a solution, even though there are two equations in but a single unknown. Taking the derivatives of both equations and demanding that the mixed derivatives of ϕ be equal requires that $F_y = G_x$ [mathematicians call this the "compatibility condition" for this particular set of seemingly overdetermined equations] — but this is just the condition that there should be zero curl.

8.8 Spin-Up: General Considerations

Figures 8.8 and 8.9 illustrate the response of the zonally unbounded equatorial ocean to a switched-on wind stress which is a step function to the east of the origin. The patterns for a zonal wind stress in Fig. 8.8 are identical with those for the corresponding solutions of the barotropic vorticity equation except that the solutions for each different mode are superimposed. This shows explicitly what was always implicit in the barotropic vorticity equation (in the dependence on the latitudinal mode number n that we carried around): there is strong modal dispersion for Rossby waves, and the crucial long waves, which propagate westward to spin-up the interior to the Sverdrup solution, travel at different rates. Since this fast-moving $n = 1$ mode is also the one trapped nearest to the equator, the spin-up will occur most rapidly along the equator and much more slowly at higher latitudes. This is evident both in the schematic, Fig. 8.10, and in the later diagrams of a model El Niño from McCreary [8].

For a given mode, the solutions shown in Figs. 8.8 and 8.9 are very similar to those for various Green's functions for the midlatitude barotropic vorticity equation as reviewed in Cane and Sarachik [5].

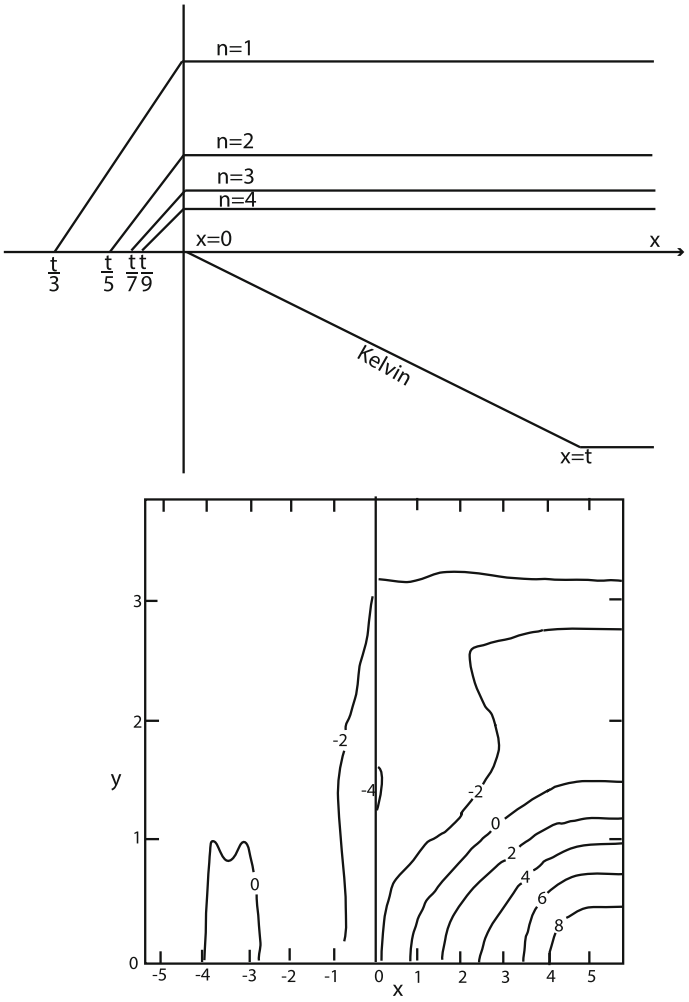

Fig. 8.8 Response in an unbounded ocean to a westerly wind which is impulsively switched-on and is a step function in x. *Top* longitudinal dependence of the low n modes. *Bottom* ϕ at $t = 5$ when $F = H(x)$, $G = \overline{H} = 0$. From Cane and Sarachik [5], Fig. 9

The Rossby waves generated by a zonal wind stress $F = H(x)H(t)$ are identical with $G^{(3)}$ for the barotropic vorticity equation. The solution for the Rossby waves generated by a meridional wind stress $G = H(x)H(t)$ is quite different, however: the solution is zero for all positive x except for a Bessel-function-like, narrowing boundary layer around the jump discontinuities at $x = 0$. The reason is that the

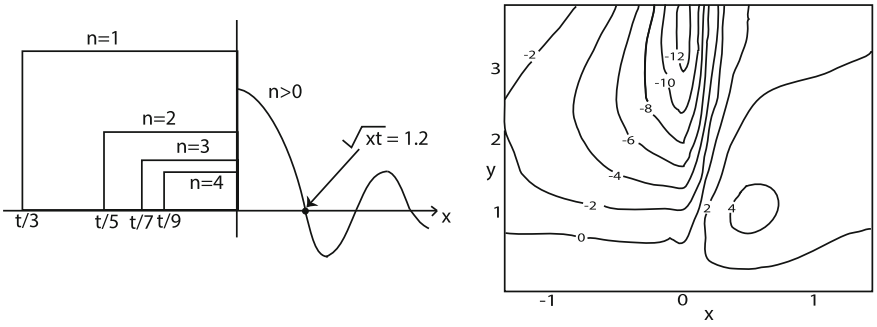

Fig. 8.9 Same as previous figure except that the wind stress is southerly instead of zonal: $G = H(x)$, $G = \overline{H} = 0$. From Cane and Sarachik [5], Fig. 10

meridional stress enters the vorticity equation, the equatorial Sverdrup solution and so on as the *curl* of the wind stress, i.e., as $\partial G/\partial x$. Since the x-derivative of a step function is the δ-function, it follows that the equatorial solution resembles the midlatitude Green's function which is forced by something proportional to $\delta(x)$ rather than, as in Fig. 8.8, something proportional to the smoother function $H(x)$.

The schematic Fig. 8.10 reviews different stages in the equatorial spin-up problem, showing successively: (a) before any wave has completely crossed the Pacific ($t < X_E$) (b) after the directly-excited Kelvin wave has reflected off the east coast, creating a second Rossby front, but before any Rossby waves have completely crossed the ocean ($X_E < t < 3X_E$) (c) after the directly-excited Rossby waves have reflected, creating a second edge to the western boundary layer and also a second front in the Kelvin mode and (d) the final asymptotic solution in which the transient waves have disappeared, leaving only the Sverdrup solution and the nonpropagating, ever-thinning western boundary layer which is needed to satisfy the western boundary condition $u = 0$. The individual components of this diagram and their roles have already been discussed above in some detail, but a comparison of Figs. 8.8 and 8.9 shows that a general schematic like Fig. 8.10, which makes no distinction between different kinds of wind forcing, is glossing over some details: the mathematical form is actually quite different depending upon whether the wind stress is zonal or meridional. It therefore is useful to take a step- by-step approach to the problem, breaking wind stress forcing into four cases depending on (a) direction and (b) symmetry with respect to the equator, and discussing the characteristics of each in turn. This is done in the next section.

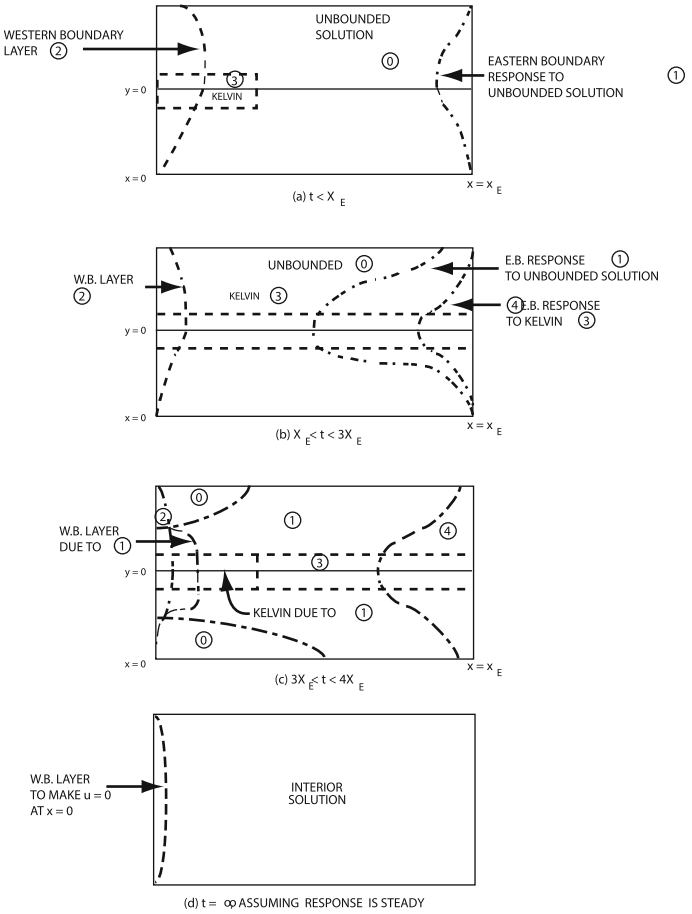

Fig. 8.10 A schematic of the equatorial spin-up in a bounded ocean with walls at $x = 0, x = X_E$. "W.B." is an abbreviation of "Western Boundary Layer". The inertial-gravity oscillations have been suppressed so that the schematic shows only the Kelvin, Rossby, and the unbounded x-dependent interior solution $\bar{u}_I$. *Top* $t < X_E$. *Upper middle* $X_E < t < 3X_E$. *Lower middle* $3X_E < t < 4X_E$. *Bottom* $t = \infty$. From Cane and Sarachik [2], Fig. 4

8.9 Equatorial Spin-Up: Details

Since the problem is linear, the general solution can be constructed by superposing special solutions. It is useful to consider zonal stress separately from latitudinal wind stress. It also useful to make a distinction between stresses that are symmetric about the equator and antisymmetric about the equator since all the individual equatorial modes fall themselves into one or the other of these two symmetry categories. A general function $F(y)$ can be split into its symmetric and antisymmetric components in latitude via

Table 8.1 Characteristics of equatorial spin-up for various stresses. F and G denote purely zonal and purely meridional stresses, respectively. The subscript "S" denotes a wind stress which is symmetric about the equator while "A" denotes one which is antisymmetric

	G_S	F_A	G_A	F_S
Symmetry	Asymm.	Asymm.	Symm.	Symm.
Kelvin waves	No	No	Yes	Yes
Sverdrup flow	$u = v = 0$	Nonzero	$u = v = 0$	Nonzero
Yoshida flow: u, ϕ linear growth with t; $v \neq 0$	–	Yes	–	Yes
Yoshida flow: u, ϕ steady; $v \equiv 0$	Yes	–	Yes	–
Kelvin wave in Yoshida flow	–	–	–	Yes
Total $v \equiv 0$	Yes	–	Yes	–
Western boundary layer decays with t	Yes	No	Yes	No

$$
\begin{aligned}
F_S(y) &= \frac{1}{2}[F(y) + F(-y)]; \\
F_A(y) &= \frac{1}{2}[F(y) - F(-y)]; \\
F(y) &= F_S(y) + F_A(y)
\end{aligned}
\tag{8.43}
$$

so we lose no generality through this approach. Some of the general characteristics of the four cases we will look at are summarized in Table 8.1.

Meridional wind stress is easier to discuss than zonal because $F = 0$ produces a number of useful simplifications. First, the Yoshida (x-independent) part of the solution, regardless of the symmetry of the north-south wind, always has (i) $v \equiv 0$ with (ii) u and ϕ that are independent of time and (iii) no Kelvin wave. The final steady state Sverdrup solution is also simplified because the currents depend upon the latitudinal wind stress only in the form of $\partial G/\partial x$, which is zero for an x-independent stress. Thus,

$$
u^{Sverdrup} = v^{Sverdrup} = 0; \qquad \phi_y^{Sverdrup} = G \qquad [F \equiv 0; G_x \equiv 0] \tag{8.44}
$$

In contrast, the zonal wind stress appears both undifferentiated and as its y-derivative, so its x- independence does not prevent a non-zero Sverdrup flow. For realistic winds, of course, G is generally a function of x and there are some steady currents associated with it also.

It is also true for a latitudinal wind stress that $v \equiv 0$ for the *total* solution, that is to say, $v \equiv 0$ for each of the long Rossby waves that are generated at the eastern boundary as well as for the other parts of the solution. The reason can be understood by

inspecting Fig. 8.9: the long Rossby waves are piecewise constant in x, and therefore are Rossby waves in the limit $k = 0$, even though the waves do have some variation with x because of the jump discontinuities at the eastern boundary and at the front $x = X_E - t/(2n + 1)$. In the limit $k \to 0$, however, v, which is $O(k)u$, necessarily goes to 0. Thus, the spin-up of the equatorial ocean in response to a switched-on, x-independent meridional wind stress involves – excluding the very high frequency inertial oscillations and coastally-trapped waves – only u and ϕ in the interior of the ocean.

At first sight, this seems paradoxical since mass must normally be carried across the equator so that the latitudinal height gradient can balance the north-south wind stress. The resolution of this paradox is that the short Rossby waves of the western boundary layer do indeed have nonzero meridional velocities. The simplest illustration of the role of these short Rossby waves is when $G(y)$ is symmetric about the equator because there are then no Kelvin waves; since the net mass flux (latitudinal integral of u) is solely due to the Kelvin wave, the sole function of the western boundary layer for asymmetric north-south stress is to redistribute mass.

Figure 8.11 shows the response when the wind is from the south. The interior of the flow is dominated, at least for small times, by the Yoshida, x-independent current, which creates nonzero zonal velocity. Since the north-south stress is symmetric, the Yoshida flow is antisymmetric. The final Sverdrup solution is one in which $u = 0$, so the effect of the advancing front of Rossby waves, shown shaded in the figure, is to reduce the zonal velocity to 0 – not instantly, but in a step-by-step process as successively higher Rossby modes reach the point in question. Outside the western boundary layer, the total current (excluding inertial oscillations) is

$$(u, v, \phi) = (0, 0, \int G\, dy) - \sum_{n=N+1}^{\infty} a_n \bar{R}_n \tag{8.45}$$

$$2N + 1 \le t/(X_E - x) \tag{8.46}$$

The first term on the right in (8.45) is the final Sverdrup solution; the infinite series is the sum of all the higher order Rossby modes that have not yet reached the point in question. Note that $\bar{R}_n$ is a vector with non-zero components for both u and ϕ, so there is always a non-zero zonal velocity which becomes weaker and weaker as $N(x, t)$ increases with time at a fixed point x.

Behind the successive Rossby fronts, convergence occurs in the northern hemisphere and divergence in the southern hemisphere, consistent with the reduction in zonal velocity. The circulation is completed on the other side by the symmetric north-south current in the western boundary layer. The dominant flow in that boundary layer is along the wall; the net transport is northward. However, the Bessel function *oscillates* with x, and therefore there is always recirculation and some flow towards the south as shown in Fig. 8.11 also.

The spin-up problem with $G(y)$ antisymmetric about the equator is similar except the response is now *symmetric* about the equator, and a Kelvin wave is excited at the

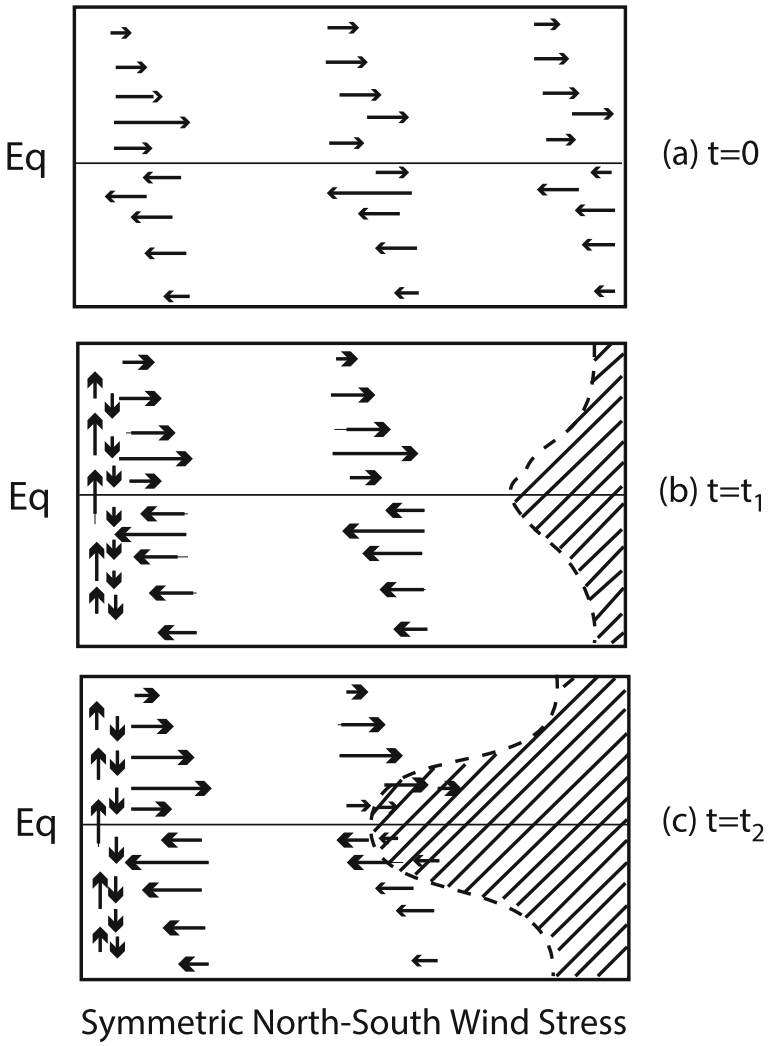

Fig. 8.11 Schematic of the response to a symmetric north-south wind stress at various times. Figure 5 of Cane and Sarachik [2]

western boundary to bring u to zero there even though no Kelvin mode is present in the unbounded Yoshida solution for any north-south stress. The mass transport is more complicated; Cane and Sarachik [5] show that a mass surplus, i.e., mean sea level higher than normal, develops at the eastern boundary while the Kelvin wave lowers sea level as it propagates away from the western boundary. Eventually, the Kelvin wave and Rossby fronts cross and the height begins to settle down towards its steady-state value, but because the modes all have different velocities and also

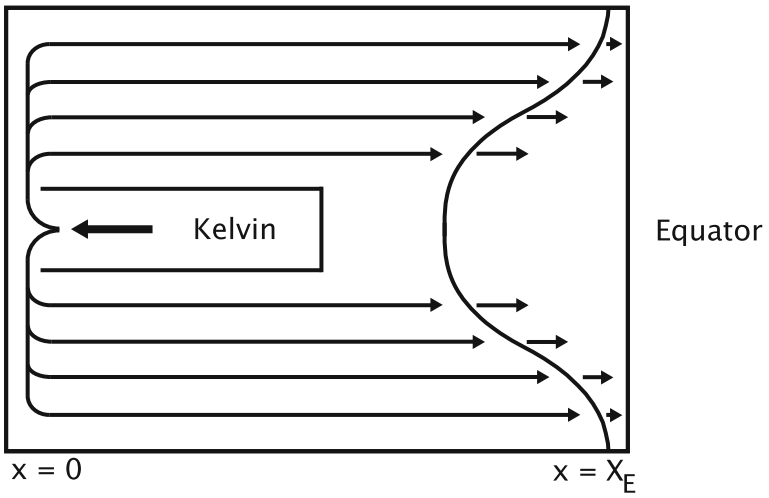

Fig. 8.12 Mass flux at a time $t < X_E$ when the wind is an antisymmetric north-south stress. Cane and Sarachik [2], Fig. 6

different latitudinal structures, some mass sloshes back and forth across the basin for rather a long time. Figure 8.12 shows the mass flux associated with the Kelvin wave and the Rossby front at a relatively early time.

The response for a zonal wind stress is more complex. In the first place, the curl of the wind stress involves only the y-derivative of F, so there can be non-zero currents in the Sverdrup solution, which is

$$\begin{aligned} u &= (x - X_E)\, F_{yy}; \qquad v = -F_y \\ \phi &= (x - X_E)\,[F - yF_y] + \phi_0 \end{aligned} \qquad \text{[Sverdrup flow; zonal wind stress]} \quad (8.47)$$

where $\phi_0 = 0$ if the forcing is antisymmetric about the equator, but otherwise may be nonzero.

Dynamically, the growing-in-time unbounded, Yoshida solution — u and ϕ — can be conceptualized as $k = 0$ Rossby waves (plus the Kelvin wave). The vector functions $\bar{R}_n(y)$ that will be used to write the solution to the general Yoshida problem in Chap. 9 do in fact give the u and ϕ components for Rossby waves in the limit $k \to 0$. In this respect, the Yoshida response to a zonal wind stress is similar to that for a meridional wind stress except that (i) the amplitude of the Rossby waves grows with time only for a zonal stress and (ii) there is a steady, but non-zero v component when the stress is zonal. (The steady v response to an x-independent zonal stress is simply the usual, off-equatorial Ekman transport: the Coriolis torque on the steady v balances the wind stress, and this is all that happens at high latitudes. However, this v field also gives the v component of the Rossby waves.)

The significance of this is that the eastern boundary does not act on these modes as a *reflecting barrier* since they carry energy away from it. Rather, the eastern boundary

represents a cut-off of the forcing region, and the eastern boundary response is the same as if the problem were done in a zonally unbounded ocean, but with the wind stress non zero only west of $x = X_E$. The eastern boundary response $\bar{u}^E$ is [ignoring Kelvin waves for the moment]

$$\bar{u}^E = \sum_{n=1}^{\infty} a_n \, H(\xi_n) \, \left\{ [t + (2n+1)(x - X_E)] \, \bar{R}_n \, + \, (0, \psi_n, 0) \right\} \tag{8.48}$$

where $n = t + (2n + 1)(x - X_E)$ and where $a_n = -r_n$, i.e., the coefficients of the boundary reaction are identical with those of the x-independent Yoshida solution except for sign. The result is that when $\bar{u}^E$ is added to the Yoshida solution $\bar{u}_I$, the factors of t cancel (ignoring the step-functions in (8.48)) to give a solution which is *independent* of time and depends on x as $(x - X_E)$. This sum is in fact none other than the steady Sverdrup solution (8.47).

This would seem to imply simple, instantaneous spin-up, but we have ignored two factors. One is the Kelvin wave, which complicates life considerably. When $F(y)$ is antisymmetric about the equator, however, the solution is free of Kelvin waves, which is why Cane and Sarachik [2] found it useful to consider this case separately from a symmetric zonal wind stress [2]. The second neglected factor is the appearance of step-functions in (8.48). At the eastern boundary, all the step functions are unity and spin-up *is* instantaneous. Farther to the west, only a finite number of Rossby modes have arrived, and the spin-up is therefore incomplete. However, as more and more modes arrive, the flow adjusts closer to the Sverdrup solution.

Since the Sverdrup solution has nonzero zonal currents for all time, it is necessary for the western boundary layer to redistribute mass for all time to insure that the western boundary condition is satisfied, and to preserve mass continuity. When the stress is meridional, however, the Sverdrup solution has $u = 0$, and so the western boundary layer of short Rossby waves decays with time.

When a Kelvin wave is directly excited by the zonal wind stress, the western boundary response includes a Kelvin wave to cancel it analogous to the Rossby modes in $\bar{u}^E$. Like these Rossby waves, the Kelvin wave grows linearly with time until it feels the effect of the other boundary (at $t = X_E$); the width of the basin therefore plays a role (for a symmetric stress) in how long the unbounded, secular growth of the Yoshida solution will continue, and thus is an essential parameter in determining the amplitude of the transient waves and of the Sverdrup currents. [Note that the amplitude of u at the western boundary is directly proportional to X_E] for the Sverdrup solution (8.47), but is independent of basin width (zero, in fact) in (8.45).]

An interesting special case is that in which $F(y) = \sqrt{1/2}\psi_0(y)$. The unbounded, Yoshida solution is

$$\bar{u}_I = t \, \bar{M}_K \, - \, (4/3) \, t \, \bar{R}_1 \, - \, (1/3) \, (0, \psi_1, 0) \tag{8.49}$$

and the final steady-state solution is

$$\bar{u} = \bar{u}_I + (x - t)\bar{M}_K + (4/3)\left\{[t + 3(x - X_E)]\bar{R}_1 + (0, \psi_1, 0) - X_E\,\bar{M}_K\right\} \tag{8.50}$$

The first Kelvin and Rossby terms in (8.50) arise directly from cutting off the forcing at the boundaries, and cancel the corresponding terms in (8.49). The Kelvin wave that is the rightmost term in (8.50) is the eastern boundary reflection of the sum of the two Rossby waves. It is also an illustration of how Kelvin waves complicate the picture: for a meridional stress or an antisymmetric zonal stress, no Kelvin waves are present, and the steady solution does not involve any reflected waves at all.

Of course, it takes a certain amount of time — $t = 4X_E$ — for all the modes present in (8.50) to be generated. This length of time is sufficient for the Rossby modes to propagate clear across the basin ($3X_E$) and for the reflected Kelvin wave to be present across the whole width of the ocean (another X_E), and is usually taken to be "the" spin-up time for the equatorial ocean. In reality, however, these modes do not settle down until a much larger time has passed. Figure 8.13 shows the amplitude of the Kelvin wave as a function of time for this forcing. The directly excited Kelvin wave reflects as Rossby waves of *all* mode numbers which later reflect as Kelvin waves. The overall conclusion is that, even without resorting to the quasi-normal modes of Gent, there can be sloshing motions in the basin long after the theoretical spin-up time of $4X_E$.

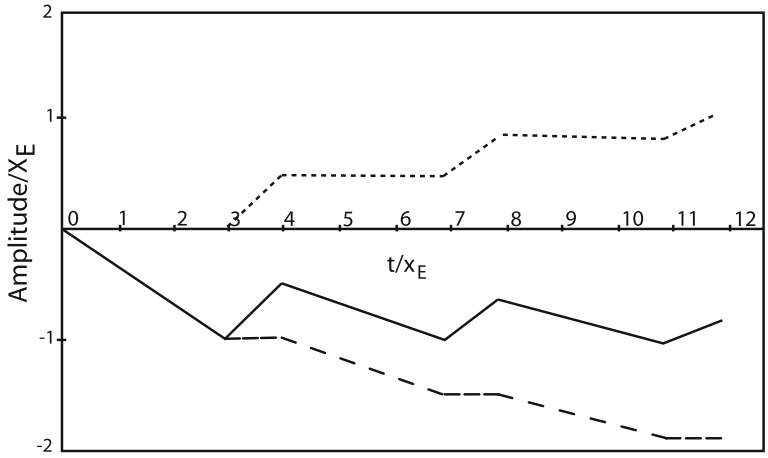

Fig. 8.13 Kelvin wave response at the western boundary, scaled by X_E, to the zonal wind stress $F(y) = 2^{1/2}\,\psi_0(y)$ [*solid line*]. The part of the Kelvin wave which is the unbounded response is *dotted*; the $n = 1$ Rossby component is dashed. From Cane and Sarachik [2], Fig. 10

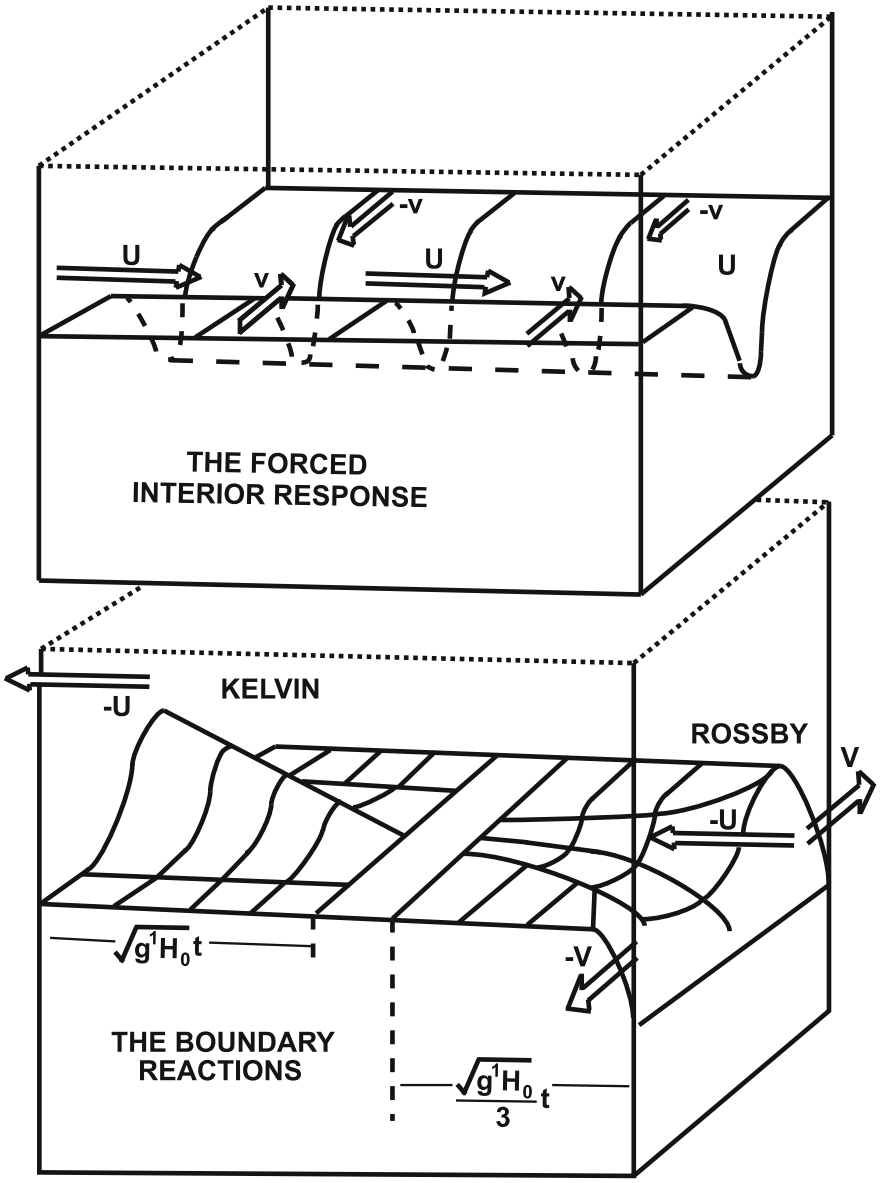

Fig. 8.14 Response to a uniform weakening of the tropical wind. *Upper graph* convergence, downwelling and the formation of an east-west jet in the interior of the ocean (away from boundaries). *Lower panel* There is downwelling and poleward flow along the entire eastern boundary associated with a packet of Rossby waves and also upwelling at the western boundary generated as a packet of equatorial Kelvin waves. [Coastal Kelvin waves, although excited, are assumed to have already propagated away from the equator and out of the domain shown by the time of the sketch.] From McCreary [8]

8.10 Equatorial Spin-Up: Summary

If there are no Kelvin waves [true for a symmetric meridional wind stress or an antisymmetric zonal wind stress], then spin-up is accomplished by Rossby waves emitted by the eastern boundary, much as on the midlatitude beta-plane. The spin-up is quick along the equator and slower and slower at higher and higher latitudes.

Kelvin waves complicate the picture considerably because they provide a means for wave energy to slosh back and forth from one side of the basin to the other – in the absence of Kelvin waves, the energy can only go west, there to be trapped in the thinning, decaying western boundary layer. Spin-up to a true steady state is therefore much longer when Kelvin waves are present than when they are not because one must wait until the "ringing" of the basin has died out through losses to dissipation and coastal Kelvin waves.

Figure 8.14 is a schematic of the analytic El Niño model of McCreary [8]. Although the changes in wind stress are turned-on over a month rather than instantaneously, this time scale is short enough so that the results are virtually identical with those for the idealized problems considered above. McCreary's detailed results give a qualitatively accurate model of El Niño despite all the assumptions and approximations inherent in the one-and-a-half layer, linear model.

References

1. Anderson DLT, Rowlands PB (1976) Role of inertia-gravity and planetary waves in response of a tropical ocean: to incidence of an equatorial Kelvin wave on a meridional boundary. J Mar Res 34(3):295–312
2. Cane MA, Sarachik ES (1977) Forced baroclinic ocean motions, II: the linear equatorial bounded case. J Mar Res 35(2):395–432
3. Anderson DLT, Rowlands PB (1976) Somali current response to southwest monsoon – importance of local and remote forcing. J Mar Res 34(3):395–417
4. Anderson DLT (1973) A low latitude spectral model using Chebyshev-parabolic cylinder functions. Report 7, GARP Programme on Numerical Experimentation W. G. N. E., World Meteorological Organization, Geneva
5. Cane MA, Sarachik ES (1976) Forced baroclinic ocean motions, I: the linear equatorial unbounded case. J Mar Res 34(4):629–665
6. Cane MA, Sarachik ES (1979) Forced baroclinic ocean motions, III: the linear equatorial basin case. J Mar Res 37(2):355–398
7. Cane MA, Sarachik ES (1981) The response of a linear baroclinic equatorial ocean to periodic forcing. J Mar Res 39:652–693
8. McCreary JP (1976) Eastern tropical ocean response to changing wind systems: with application to El Nino. J Phys Oc 6:623–645

Chapter 9
Yoshida Jet and Theories of the Undercurrent

Abstract This chapter is focused firstly on the response of the ocean to an x - independent wind stress that is turned on as a step function in time, first on the f-plane and then on the equatorial plane. In contrast to the midlatitudes where the steady-state flow is a geostrophically balanced jet *perpendicular* to the wind, Yoshida showed, through an infinite Hermite series with explicit, analytical coefficients, that a zonal wind stress would yield a zonal equatorial jet. The transient adjustment, which is radiation of gravity waves, is also explicitly given. In the second half of the chapter, four barotropic theories of the Equatorial Undercurrent are analyzed as a warmup to the chapter where more realistic continuously stratified modes are explored.

If all you have are observations, that's botany. If all you have is theory, that's philosophy.
— Michael S. Turner, quoted in *Scientific American*, October (1990), p. 117

9.1 Introduction

The earliest work on equatorially trapped dynamical phenomenon is the now-classic 1959 paper of Kozo Yoshida (1922–1978) who showed that the wind could drive a strong equatorial jet. The Equatorial Undercurrent, then called the Cromwell Current, was a recent discovery: a strong, shallow jet with velocities as high as 1.5 m/s — extraordinarily high for the sea — but with a very narrow latitudinal width of only a couple of hundred kilometers — and centered right at the equator.

One other example of a strong, intense jet was known to oceanographers of the day: the coastal jet analyzed by Jule Charney in 1955. There, the presence of a boundary made possible a very strong, boundary-trapped jet. Away from the coast, the Coriolis force tends to turn wind-driven motion so that the ocean quickly settles into a balance in which the current is roughly *perpendicular* to the wind stress so that the Coriolis force on the current balances the wind stress, halting the acceleration of the water. Yoshida realized that at the equator, this mechanism must fail, just as it does at a coast, because the Coriolis force goes to zero at the equator, permitting the wind stress to accelerate the current without limit.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_9

In this chapter, we will first discuss the midlatitude equivalent of Yoshida's problem to understand clearly what is and is not changed at the equator. Next, we will discuss Yoshida's equatorial solution. Finally, we will discuss modern theories of the Undercurrent and the results of numerical models for it.

9.2 Wind-Driven Circulation in an Unbounded Ocean: f-Plane

The model is the same as applied to model the coastal jet: the linearized shallow water wave equations on an f-plane. Because the model (and its solutions) are so simple, we shall not retreat to the nondimensional formulation used for Yoshida's problem in the next section, but shall instead allow quantities to assume their dimensional values. For conformity with the Yoshida jet, assume that:

1. flow is independent of x
2. wind stress only in the x-direction, and uniform in space and time
3. no coasts.

The dimensional symbols used are: $g =$ gravitational constant, $H =$ mean depth of ocean, $\phi =$ sea surface height relative to mean sea level, f is the (constant!) Coriolis parameter and $F\ (= \tau^x/(\rho\, H))$ represents the wind stress. The model equations are

$$u_t - fv = F \tag{9.1}$$

$$v_t + fu = -g\,\phi_y \tag{9.2}$$

$$\phi_t + Hv_y = 0 \tag{9.3}$$

Through straightforward algebra, one can reduce these three down to a single equation for v:

$$v_{yy} - \frac{f^2}{gH}\,v - \frac{1}{gH}\,v_{tt} = \frac{f}{gH}\,F \tag{9.4}$$

As true for any inhomogeneous differential equation, the general solution to (9.4) may be separated into (i) a particular solution, which for a steady wind is independent of time plus (ii) the general solution to the homogeneous form of (9.4), which is necessary to satisfy the initial conditions. For a y-independent wind, the particular solution is

$$v = -F/f \qquad \text{[steady solution; Ekman transport]} \tag{9.5}$$

This steady solution is commonly referred to as the "Ekman transport" because a well-known result of Ekman theory is that the vertically-averaged wind-driven ocean circulation is, away from boundaries, a steady circulation *perpendicular* to the wind and in geostrophic balance with it. In point of fact, (9.1)–(9.3) can be obtained by

vertically integrating the usual viscous Ekman model equations. Since the viscous terms are the z-derivatives of the stress, their integral is simply the stress at the surface of the sea. Consequently, it is quite legitimate to refer to (9.5) as the "Ekman transport" even though the viscosity does not appear *explicitly* in our model.

The time-dependent homogeneous solutions satisfy, when the flow is independent of y, the equation

$$v_{tt} + f^2 v = 0 \tag{9.6}$$

with the general solution

$$v = A \sin(ft) + b \cos(ft) \tag{9.7}$$

Elementary courses in dynamics analyze three simple balanced motions: pure geostrophy, pure cyclostrophic flow, and pure inertial motion, which is a balance between the pressure and centrifugal forces. Pure inertial motion consists of circular trajectories, always moving in an anticyclonic direction, with a period equal to $2\pi/f$. Equations (9.6) and (9.7) are a mathematical description of this pure inertial motion.

Thus, the general response of the ocean to a constant wind stress — away from boundaries and with f constant — is the steady Ekman transport plus transient inertial oscillations superimposed upon it. In a vertically integrated viscous model, the steady circulation is unchanged from the inviscid barotropic solutions of (9.1)–(9.3), but the inertial oscillations damp out quickly as shown by Fredholm a century ago [1]. The short time scale found by Fredholm — the flow is within a few percent of the steady solution within a period of π/f after an impulsive turn-on of the wind — justifies taking a time-independent wind stress; the ocean adjusts very quickly to changes in the wind so that v is always close to the steady-state value given by (9.5) where F is the instantaneous value of the wind stress.

When the wind is varying with y, both the steady and transient solutions will be forced to vary on more or less the same length scale, and (9.4) must be retained in full. Recall that the Rossby radius-of-deformation is

$$R \equiv \sqrt{gH}/f \qquad \text{["Rossby radius of deformation"]} \tag{9.8}$$

Equation (9.4) can be written as

$$v_{yy} - \frac{1}{R^2} v - \frac{1}{gH} v = \frac{f}{gh} F \tag{9.9}$$

Comparing the first and second terms, one see that v_{yy} will be negligible in comparison to $(-R^{-2}v)$ unless F and therefore v are varying on a length scale at least as small as the Rossby radius. However, the baroclinic Rossby radius is only about 100 km even for the first baroclinic mode. Since the large scale atmospheric weather patterns are varying only on much larger scales, it is quite reasonable to ignore the y-variations in (9.9) for both the steady and transient parts of the solution. A hurricane or other local wind system may produce wind-driven flows for which the y-derivatives

in (9.9) are important, but for the planetary scale flow, it is a good approximation to take the local-in-time-and-space wind stress and pretend it is constant in both time and space as done here.

The solution on the equatorial beta-plane bears many similarities to that on the f-plane. There are again inertial oscillations, but these propagate latitudinally since the equatorial beta-plane equations are not independent of y and spatially constant solutions are not possible. There is again a part of the motion which is steady, and directly proportional to the wind stress. However, in dramatic contrast to the steady, perpendicular-to-the-wind flow found here, the Yoshida problem contains a jet which is linearly accelerating — forever — in the x-direction. The equation beta-plane changes the solution rather a lot.

9.3 The Yoshida Jet

The model is the same as in the previous section except that the geometry is now that of the equatorial beta-plane. To simplify further, the x-wind stress is the constant, one; since the problem is linear, one can obtain the solution for a different constant wind stress F by multiplying all the solutions given below by F. The model equations are

$$u_t - yv = 1 \tag{9.10}$$

$$v_t + yu + \phi_y = 0 \tag{9.11}$$

$$\phi_t + v_y = 0 \tag{9.12}$$

The solutions take a simpler form when written in terms of the sum and difference variables, S and D, so, taking the sum and difference of (9.10) and (9.12) as before, we shall solve the set

$$S_t + \mathfrak{R}\, v = 1 \tag{9.13}$$

$$v_t + (1/2)\,(\mathfrak{L}S + \mathfrak{R}D) = 0 \tag{9.14}$$

$$D_t - \mathfrak{L}v = 0 \tag{9.15}$$

where $\mathfrak{L}$ and $\mathfrak{R}$ are the raising and lowering operators defined in Chap. 3.

An alternative is to reduce these three equations down to one for v alone:

$$v_{yy} - y^2 v - v_{tt} = y \tag{9.16}$$

We shall use both the three-equation and one-equation formulation as convenient.

To proceed with the triplet approach, the crucial next step is to introduce the Hermite expansion of the trivial function, $f(y) \equiv 1$:

$$1 = \sum_{n=0}^{\infty} I_{2n}\, \psi_{2n}(y) \tag{9.17}$$

where

$$I_{2n} = \frac{\sqrt{2}\,\pi^{1/4}\,\sqrt{(2n)!}}{2^n\, n!} \tag{9.18}$$

The equations of motion in spectral form are the single equation

$$S_{0,t} = I_0 \tag{9.19}$$

plus the infinite set of triplets of equations

$$S_{n+1,t} - \sqrt{2(n+1)}v_n = I_{n+1} \tag{9.20}$$

$$v_{n,t} + \sqrt{(n+1)/2}S_{n+1} - \sqrt{n/2}\, D_{n-1} = 0 \tag{9.21}$$

$$D_{n-1,t} + \sqrt{2n}v_n = -I_{n-1} \tag{9.22}$$

Because the constant one is symmetric about the equator, only even Hermite functions appear in its expansions. Consequently, (if the initial conditions are symmetric about the equator, too) we can dispense with the pair of equations that describe the forced Yanai wave and restrict n in (9.20)–(9.22) to be an *odd* integer.

As for the f-plane problem, the solution can be split into a particular solution and a homogeneous solution. The difference is that in the particular solution, now only v is steady; S_p and D_p are *growing linearly with time* and so therefore are u and ϕ. The necessity for this can be seen directly from (9.10), which reduces to

$$u_t = 1 \tag{9.23}$$

at the equator: u must increase linearly with time at $y = 0$ because no other term is available to balance the wind stress. The particular solution will therefore be denoted by subscript "sl", which reminds us that timewise, the components are either steady or linearly-growing.

It can be easily verified by direct substitution into (9.20)–(9.22) that the solution is

$$S_{sl} = t\left\{ I_0\psi_0(y) - \sum_{n=1,n\ \text{odd}}^{\infty} \frac{\sqrt{n}}{(2n+1)\sqrt{n+1}}\, I_{n-1}\, \psi_{n+1}(y) \right\} \tag{9.24}$$

$$v_{sl} = -\sqrt{2} \sum_{n=1,n\ \text{odd}}^{\infty} \frac{\sqrt{n}}{2n+1}\, I_{n-1}\, \psi_n(y) \tag{9.25}$$

$$D_{sl} = -t \left\{ \sum_{n=1,n \text{ odd}}^{\infty} \frac{1}{2n+1} I_{n-1} \psi_{n-1}(y) \right\} \tag{9.26}$$

while the corresponding homogeneous solutions are, assuming that the wind is turned on *impulsively* with initial conditions of a state of rest, i.e., $u = v = \phi = 0$ at $t = 0$,

$$S_{waves} = \sum_{n=1,n \text{ odd}}^{\infty} \frac{2n+2}{(2n+1)^{3/2}} \sin\left(\sqrt{2n+1}\, t\right) I_{n-1} \psi_{n+1}(y) \tag{9.27}$$

$$v_{waves} = \sum_{n=1,n \text{ odd}}^{\infty} \frac{\sqrt{2n}}{2n+1} \cos\left(\sqrt{2n+1}\, t\right) I_{n-1} \psi_{n}(y) \tag{9.28}$$

$$D_{waves} = \sum_{n=1,n \text{ odd}}^{\infty} \frac{2n}{(2n+1)^{3/2}} \sin\left(\sqrt{2n+1}\, t\right) I_{n-1} \psi_{n+1}(y) \tag{9.29}$$

where we have used the identity

$$I_{n+1} = \sqrt{\frac{n}{n+1}}\, I_{n-1} \tag{9.30}$$

which follows from the definition of I_n. The homogenous solutions are waves and so are labeled by subscript "waves".

The y-dependent parts of the particular solution are shown in Fig. 9.1. The zonal velocity is strongly concentrated around the equator, so the description of the Yoshida solution as a "jet" is very apt. By using the definition of the normalized Hermite polynomials and the explicit solution given above, one can show that

$$u_{sl} = t \left\{ 0.94 \exp(-[1/2]y^2) + \mathfrak{f}(y) \right\} \tag{9.31}$$

where it can be proved that

$$|\mathfrak{f}(y)| \leq 0.15, \qquad \forall\, y \tag{9.32}$$

Thus, to an excellent first approximation, the jet has the latitudinal structure of a Kelvin wave, a simple Gaussian decay away from the equator at exactly the same rate as for a Kelvin wave.

Nonetheless, it is also obvious from the Fig. 9.1 that the north-south current is very broad and decreases as $|y| \to \infty$ very slowly. If we accept for the moment that u decreases much more rapidly than v, we can return to the original shallow water equations to deduce the correct asymptotic behavior of all three fields.

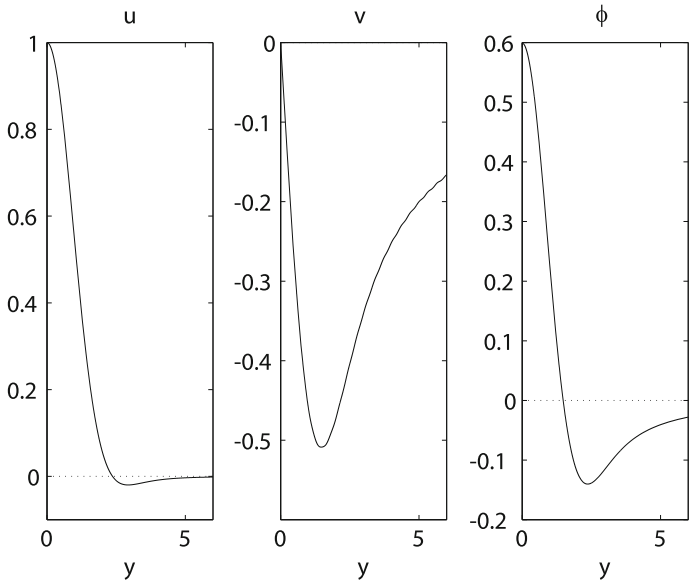

Fig. 9.1 The time-independent particular solution components, u_{sl}, v_{sl} and ϕ_{sl}. Because u_{sl} and ϕ_{sl} are symmetric with respect to $y = 0$ while v_{sl} is antisymmetric, the solutions are shown only for $y \geq 0$. The factors $u_{sl}(y)$ and $\phi_{sl}(y)$ are multiplied by t to obtain u and ϕ for the time-dependent solution

For large $|y|$, the time derivative of u in the x-momentum equation is negligible, reducing this ODE to simply $(-yv) = 1$ or in other words

$$v \sim -1/y, \qquad |y| \gg 1 \qquad \text{[“Ekman transport”]} \tag{9.33}$$

Recalling that $f = y$, we find that (9.33) is simply the Ekman transport: at high latitudes, $u \ll v$, and the wind again drives a current perpendicular to itself as on the midlatitude f-plane. (See Eq. (9.5).)

The height equation (9.12) is $\phi_t + v_y = 0$, which implies that

$$\phi \sim -t/y^2, \qquad |y| \gg 1 \tag{9.34}$$

The y-momentum equation, $yu + \phi_y = 0$, shows that

$$u \sim -2t/y^4, \qquad |y| \gg 1 \tag{9.35}$$

The increasing-with-time height and zonal velocity fields are needed for large y because the variation of the Coriolis parameter with y requires v to decrease with y even though the wind stress is independent of y. At a given northern latitude, more fluid is flowing towards the equator than is coming in from the north to replace it, so

this divergence requires a steady decrease in sea level as given by (9.34). Since this change in sea level varies with latitude, the jet can settle into geostrophic balance in the y-direction only if there is a weak zonal flow, also increasing with time, to balance it. Similar complications can be expected whenever the Ekman transport is forced to vary with latitude, whether the reason is due to changes in f or latitudinal variations in the wind. It is noteworthy, however, that the zonal flow and height both vary with y more slowly than v, and are much smaller than v for large $|y|$. Outside the region of the equatorial jet — for $|y| \geq 2$, say — the flow is again that of time-independent Ekman transport. Only around the equator does a strong zonal flow accelerate.

The free oscillations excited by the wind are shown in perspective plots in the $y - t$ plane in Fig. 9.7. The waves are inertial-gravity waves. The frequency of the n-th mode is $\omega = \pm\sqrt{2n+1}$, which is equal to the value of the Coriolis parameter at the turning point of the n-th Hermite function. (Mathematically, the turning points are the places where the Hermite function changes from oscillation to exponential decay, i.e., where the curvature of the function changes sign relative to the Hermite function itself. Physically, the Hermite functions describe waves bouncing back and forth between "walls" at the turning points.) If we examined the behavior of wave packets on the equatorial beta-plane, we would find that they would slow down and finally reflect back towards the equator off the turning latitudes. Thus, the free modes shown in Fig. 9.7 can still be legitimately interpreted as inertial oscillations: the frequency is not equal to *the* Coriolis parameter because the Coriolis parameter varies with latitude, but it *is* equal to the value of f in that neighborhood where the wave spends the largest amount of time. (In reaching this conclusion, we are visualizing the Hermite functions as the superposition of two travelling waves propagating in opposite directions to form a standing wave between the turning points.)

On the midlatitude f-plane, the inertial oscillations would make the sea rise and fall in place without latitudinal propagation, and a perspective plot similar to Fig. 9.7 would simply show ridges and troughs *parallel* to the y-axis. The equatorial beta-plane plots show *tilted* ridges and troughs, indicating propagation *towards* the equator. However, the tilt is not very pronounced: again, the free modes resemble pure inertial oscillations (Fig. 9.2).

Figure 9.3 shows contour plots of the *total* currents [particular solution plus inertial oscillations]. Moore and Philander [2] also consider (as we shall not) a wind stress blowing north-south. In this case, the particular solution is *time-independent*: Only the east-west component of the wind stress will drive an accelerating jet.

The contour plots provide some clarification of the free oscillations. At a given latitude y, the plots resemble those for inertial oscillations for $t < y$. For larger t, however, equatorial disturbances have propagated sufficiently far from the equator to superimpose themselves upon the inertial oscillations and mask their character. (Note that $y = \pm t$ gives the characteristics of Eq. 9.16.)

As on the midlatitude beta-plane, one would expect the inertial modes to be quickly damped by viscosity. At high latitudes, the flow would settle into the steady-state Ekman transport perpendicular to the wind stress; along the equator, friction or drag or nonlinear effects would eventually check the runaway acceleration of the zonal wind. Nonetheless, an intense zonal current will form. Continental boundaries will

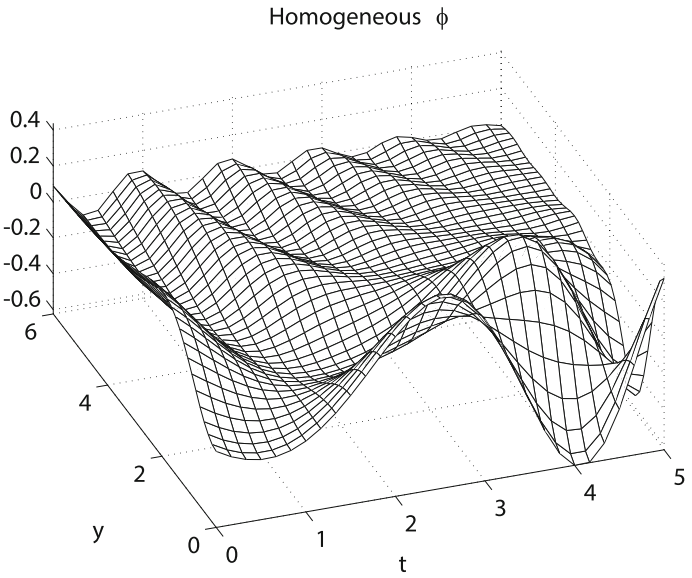

Fig. 9.2 The homogeneous part of the height, $\phi_{waves}(y, t)$, for the Yoshida jet. The zonal and latitudinal currents are qualitatively similar to ϕ and therefore not illustrated

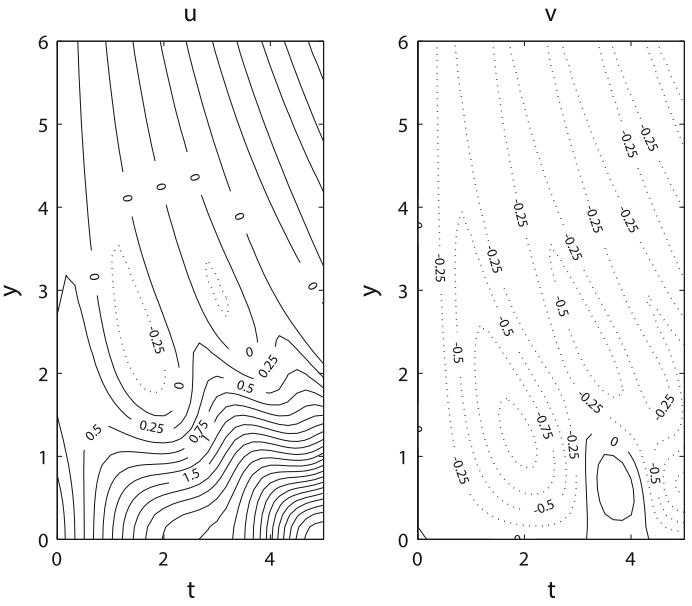

Fig. 9.3 The *total* velocity fields for the Yoshida jet, i.e., the sum of the homogeneous and particular solutions. The height, ϕ, is not shown because it is similar to u

presumably deaccelerate and turn the current, but only rather close to land. Over most of the width of the Pacific or Atlantic, neglecting the boundaries is a tolerable first approximation. Despite its simplicity, the Yoshida jet does explain how very strong currents like the Equatorial Undercurrent can form around the equator.

9.4 An Interlude: Solving Inhomogeneous Differential Equations at Low Latitudes

The success of Hermite series has obscured two important realities which are here explained before resuming our analysis of tropical jets: the Hermite series need sum acceleration methods and there are three other useful strategies for the Yoshida jet. Table 9.2 catalogs some of these additional representations. We review each in turn.

9.4.1 Forced Eigenoperators: Hermite Series

If $\mathscr{L}$ is a differential operator associated with a complete set of Sturm–Liouville eigenfunctions, then inhomogeneous differential equations whose operator is the eigenoperator plus the identity operator can be trivially solved by an eigenfunction series. The relationships

$$\mathscr{L}\phi_n = \lambda_n \phi_n, \qquad \& \qquad \mathfrak{B}(u) = 0, \tag{9.36}$$

where $\mathscr{L}$ is the differential operator (eigenoperator), λ_n and ϕ_n are the eigenvalue and eigenfunction and $\mathfrak{B}$ is the boundary condition operator, imply that

$$\mathfrak{L}u + qu = f(y) \tag{9.37}$$

where q is a constant is solved by

$$u(y) = \sum_n \frac{f_n}{\lambda_n + q}\ \psi_n(y) \tag{9.38}$$

$$f_n = \frac{< \phi_n(x),\, f(x) >}{< \phi_n(x),\, \phi_n(x) >} \tag{9.39}$$

An elementary example is

$$\mathscr{L} = \frac{d^2}{dx^2}, \qquad \mathfrak{B}(u) = \{u(0), u(\pi)\} \tag{9.40}$$

$$\phi_{n,xx} = \lambda_n \phi_n, \qquad \phi_n(0) = \phi_n(\pi) = 0 \tag{9.41}$$

$$\phi_n(x) = \sin(nx), \qquad \lambda_n = -n^2, \qquad n = 1, 2, \ldots \tag{9.42}$$

$$u_{xx} - u = f(x), \qquad u(0) = u(\pi) = 0 \tag{9.43}$$

$$f(x) = \sum_{n=1}^{\infty} f_n \sin(nx) \tag{9.44}$$

$$f_n = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(nx) dx \tag{9.45}$$

$$u(x) = \sum_{n=1}^{\infty} \frac{f_n}{-n^2 - 1} \sin(nx) \tag{9.46}$$

The Hermite function is simpler because the Hermite functions individually decay as $|y| \to \infty$ so no boundary conditions need be explicitly imposed. Note that $\psi_n(y)$ denotes a member of the orthonormal Hermite basis.

$$\mathfrak{H} v = v_{yy} - y^2 v \tag{9.47}$$

$$\mathfrak{H}\, \psi_n = -(2n+1)\psi_n \tag{9.48}$$

$$v_{yy} - y^2 v = f(y) \tag{9.49}$$

$$f(y) = \sum_{n=1}^{\infty} f_n \psi_n(y) \tag{9.50}$$

$$f_n = \int_{-\infty}^{\infty} dy\, \psi_n(y)\, f(y) \tag{9.51}$$

$$v(y) = -\sum_{n=1}^{\infty} \frac{f_n}{(2n+1)} \psi_n(y) \tag{9.52}$$

Yoshida's paper seems to have the first application of Hermite functions and series to tropical oceanography. It was a great advance both in ocean physics and in ocean mathematics.

Nonetheless, there are drawbacks to eigenfunction series. Eigenfunction series decay very slowly unless certain conditions are satisfied. (For the sine eigenfunctions, for example, an essential condition for rapid convergence is that the inhomogeneous term $f(y)$ is periodic.)

It can be shown directly from the parabolic cylinder differential equation (which the Hermite functions solve) that the $\psi_n(y)$ oscillate for $|y| \le \sqrt{2n+1}$ and decay exponentially fast — *Gaussian* fast — for larger y. This implies that a Hermite partial sum that includes $\psi_N(y)$ but nothing higher will have negligible amplitude for $|y| > \sqrt{2N+1}$. (The points $y = \pm\sqrt{2n+1}$ are the "turning points" for the n-th mode, the points where oscillation in latitude changes to exponential decay in y.) A necessary condition for fast convergence of a Hermite series is that $f(y)$ decay exponentially fast as $|y| \to \infty$.

9.4.2 Hutton–Euler Acceleration of Slowly Converging Hermite Series

The inhomogeneous term in the Yoshida jet differential equation, $f(y) = y$, does not decay with $|y|$. This implies that its Hermite coefficients cannot decrease with degree n as confirmed by column two of Table 9.1. Nevertheless, the Hermite series for $f(y) = y$ does approximate the linear function after a fashion as illustrated on the left in Fig. 9.4. The approximation is highly nonuniform in space; it is zero and therefore useless over most of the real axis but approximates $f(y) = y$ over a finite interval in y whose width increases steadily with increasing truncation.

Unfortunately the oscillations visible on the left in the figure, similar to the oscillations of Gibbs' Phenomenon in the Fourier series of a discontinuous function, mean that the approximation is horrible.

For $f(y) = y$, the corresponding Yoshida jet flow has very slow decay (or growth) as $|y| \to \infty$. This requires that the Hermite series converges very slowly. Theorems by Hille, Delves, Bain and Boyd, reviewed in [4], prove that for the Yoshida jet the normalized Hermite coefficients of $f(y) = 1$ decrease as $n^{-1/4}$, those of v_{sl} as $n^{-3/4}$, those of ϕ_p fall as $n^{-5/4}$ and those of u_{sl} as $n^{-9/4}$. That is to say, *inverse power* decay with $|y|$ rather than *exponential* decay with n implies that the coefficients of the corresponding Hermite function decrease *algebraically* with n, too. The unaccelerated Hermite series is almost useless.

Fortunately, Hermite series in oceanography are amenable to so-called "sum acceleration" or "sequence acceleration" methods. The partial sums of Hermite series alternately overshoot and undershoot their limit. It follows that the *average* of the partial sums truncated after N and $N - 1$ terms is a better approximation to the limit than either of the partial sums from whence it came. This averaging of two partial sums is a "Hutton acceleration". The averaging can be repeated; when all available partial sums are exhausted, the iterated Hutton acceleration is the "Euler acceleration".

Table 9.1 Normalized and unnormalized Hermite coefficients for $f(y) = y$ [columns two (normalized) and columns three (unnormalized)] and of the steady north-south current $v(y)$ of the Yoshida jet (columns four and five). $y = \sum a_n \psi_n(y) = \sum a_n^{unnorm} \exp(-y^2/2) H_n(y)$; $v = \sum b_n \psi_n(y) = \sum b_n^{unnorm} \exp(-y^2/2) H_n(y)$

n	a_n	a_n^{unnorm}	b_n	b_n^{unnorm}
1	2.66267	1.41422	−0.887557	−0.471403
3	3.26109	0.353550	−0.465870	−0.0505076
5	3.64601	0.0441942	−0.331456	−0.00 401765
7	3.93814	0.00368285	−0.262543	−0.000 245523
9	4.17703	0.000230176	−0.219844	− 0.0000121146
11	4.38091	0.0000115089	−0.190474	−5.00385E-7
13	4.55979	4.79536E-7	−0.168881	−1.77606E-8
15	4.71983	1.71263E-8	−0.152253	−5.52460E-10
17	4.86509	5.35199E-10	−0.139003	−1.5214E-11
19	4.99841	1.48665E-11	−0.128164	−3.8193E-13

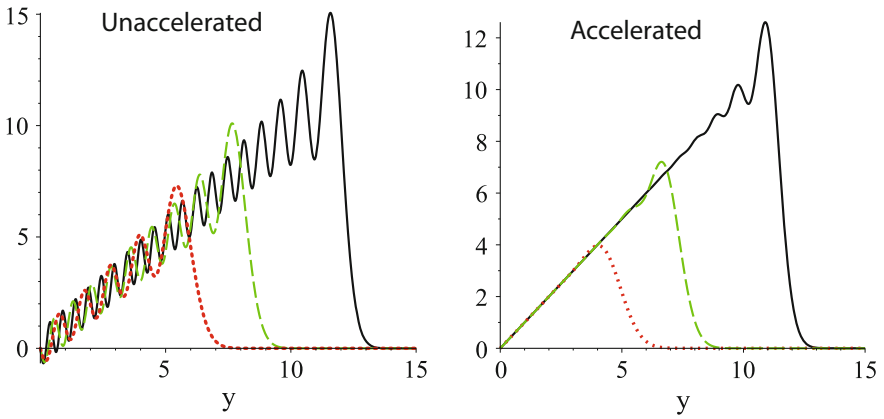

Fig. 9.4 *Left* partial sums of the linear function, $f(y) = y$, using 10, 18 and 38 odd degree Hermite functions [*red dots*, *green dashes* and *solid black*, respectively]. *Right* same except that the series was accelerated by multiplying the last 8 coefficients by (255/256), (247/256), (219/256), (163/256), (93/256), (37/256), (9/256) and (1/256), respectively

In his Ph.D. dissertation, Dennis Moore introduced these methods to equatorial oceanography. Although not explicitly described in his thesis, Moore applied four-fold Hutton acceleration. The procedure is very simple: The last four terms in the chosen truncation are multiplied by universal constants:

$$\begin{aligned} f_N(y) &\equiv \sum_{n=0}^{N} a_n \psi_n(y) \\ &\Downarrow\ \Downarrow\ \Downarrow \\ f_N(y) &\equiv \sum_{n=0}^{N-4} a_n \psi_n(y) + \frac{15}{16}\psi_{N-3}(y) + \frac{11}{16}\psi_{N-2}(y) + \frac{5}{16}\psi_{N-1}(y) + \frac{1}{16}\psi_N(y) \end{aligned} \tag{9.53}$$

A thorough review is presented by Boyd and Moore [5] and Appendix A here. Figure 9.4 is a visual proof of the effectiveness of Moore's series acceleration.

9.4.3 Regularized Forcing

An alternative strategy to prevent slow convergence in Hermite series solutions is to modify the wind stress so that the regularized inhomogeneous term $f(y)$ decays exponentially with y for large latitude. Cane and Sarachik often utilized this strategy by taking $f(y)$ to be a Gaussian $\exp(-y^2/W^2)$ or the first derivative of a Gaussian where the width W is large. An advantage is that the exact Hermite coefficients are known in explicit, analytic form (Appendix A). However, the forcing is modified at all latitudes, even close to the equator. This "regularization error" can be reduced by choosing large W, but one then needs a very large number of Hermite functions to obtain a accurate answer.

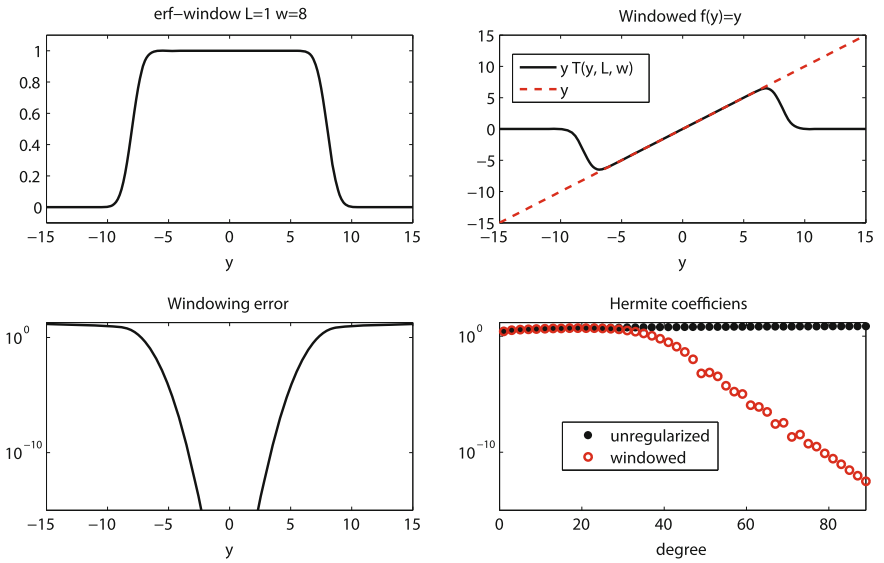

Fig. 9.5 *Upper left* a representative window. *Upper right* unmodified and windowed linear function, $f(y) = y$. *Lower left* regularization error. *Lower right* regularized and unmodified Hermite coefficients for $f(y) = y$

Another option is to multiply the desired $F(y)$ by a smoothed "top hat" or "box function" such as

$$T(y; L, w) = \frac{1}{2}\{\text{erf}(L(y + w)) - \text{erf}(L(y - w))\} \tag{9.54}$$

When L is large, this window approximates one on $y \in [-w, w]$ and zero otherwise as illustrated in the upper left of Fig. 9.5. The difference between the windowed and unregularized function is exponentially small, perhaps smaller than machine precision, over most of the interval $y \in [-w, w]$. Nevertheless, the Hermite coefficients of large degree decay exponentially fast as shown in the lower right.

9.4.4 Bessel Function Explicit Solution for the Yoshida Jet

The differential equation

$$v_{yy} - y^2 v = f(y) \tag{9.55}$$

can be transformed by the change of coordinate $\tilde{y} = (1/2)y^2$ into a differential equation whose homogeneous solutions are the $\sqrt{\tilde{y}}$ times a modified Bessel function of $\tilde{y}$. (Identity 9.1.52 on p. 362 of [6].) The transformed equation has the general solution

$$v = C_I \sqrt{y} \mathrm{I}_{1/4}((1/2)y^2) + C_K \sqrt{y} \mathrm{K}_{1/4}((1/2)y^2) + v^p(y) \tag{9.56}$$

$$v^p = \frac{1}{2}\sqrt{y}\left\{\frac{1}{2}\sqrt{y}\left(\mathrm{I}_{1/4}((1/2)y^2)\int\sqrt{z}\mathrm{K}_{1/4}((1/2)z^2)f\,(z)\,dz\right.\right.$$
$$\left.\left.-\mathrm{K}_{1/4}((1/2)y^2)\int\sqrt{z}\mathrm{I}_{1/4}((1/2)z^2)\,f\,(z)\,dz\right)\right\} \tag{9.57}$$

Because the K-Bessel function is singular at the origin, $C_K = 0$. The remaining constant C_I is determined by the condition $v(y) \to 0$ as $|y| \to \infty$; it can be evaluated by exploiting the known asymptotic approximations of the Bessel functions.

For the particular case of $f(y) = y$,

$$v = C_I\sqrt{y}\mathrm{I}_{1/4}\left([1/2]\,y^2\right) + v^p(y) \tag{9.58}$$
$$C_I = -1/4\sqrt{2\pi}\;\Gamma\,(3/4) = -.767916038651074 \tag{9.59}$$
$$v^p = 1/8\sqrt{2\pi}\,\Gamma\,(3/4)\,y^{5/2}\mathbf{L}_{1/4}\left([1/2]\,y^2\right)\times$$
$$\left(\mathrm{K}_{3/4}\left(1/2]\,y^2\right)\mathrm{I}_{1/4}\left([1/2]\,y^2\right) + \mathrm{K}_{1/4}\left(1/2\,y^2\right)\mathrm{I}_{-3/4}\left([1/2]\,y^2\right)\right) \tag{9.60}$$

where $\mathbf{L}_{1/4}(z)$ is the usual modified Struve function which solves $z^2u_{zz} + zu_z - (z^2 + (1/16))u = \left\{2^{3/4}/(\sqrt{\pi}\,\Gamma(3/4))\right\}z^{5/4}$.

Although these solutions appear complicated, routines to evaluate Bessel functions are included in Matlab, Maple and all Fortran and C++ libraries. The Struve function, which is sometimes omitted from numerical libraries, has power series and asymptotic approximations and the integral representation

$$\mathbf{L}_{1/4}\,(z) = \frac{2^{3/4}}{\Gamma(3/4)\sqrt{\pi}}\,z^{1/4}\int_0^{\pi/2}dt\,\sinh(z\cos(t))\,\sqrt{\sin(t)} \tag{9.61}$$

The trapezoidal rule has a geometric rate of convergence for this integral. However, these explicit Bessel–Struve solutions are not the only good non-Hermite option.

The function v found here is identical with v_{sl}, the part of the north-south velocity of the Yoshida flow which is independent of time

$$v_{sl} = v \tag{9.62}$$
$$u_{sl} = t\,(1 + yv_{sl}) \tag{9.63}$$
$$\phi_{sl} = t\left(-v_{y,sl}\right) \tag{9.64}$$

$$\phi_{sl}/t = -1/24\,y^{3/2}\left(-6\sqrt{2}\sqrt{\pi}\,\Gamma\,(3/4)\,\mathrm{I}\left(-3/4,\;1/2\,y^2\right)\right. \tag{9.65}$$
$$+3\sqrt{2}\sqrt{\pi}\,\Gamma\,(3/4)\,\mathbf{L}\left(1/4,\,1/2\,y^2\right)\mathrm{K}\left(3/4,\;1/2\,y^2\right)\mathrm{I}\left(1/4,\;1/2\,y^2\right)$$
$$+3\sqrt{2}\sqrt{\pi}\,\Gamma\,(3/4)\,\mathbf{L}\left(1/4,\,1/2\,y^2\right)\mathrm{K}\left(1/4,\;1/2\,y^2\right)\mathrm{I}\left(-3/4,\;1/2\,y^2\right)$$
$$+3\sqrt{2}\sqrt{\pi}\,\Gamma\,(3/4)\,\mathbf{L}\left(5/4,\,1/2\,y^2\right)y^2\mathrm{K}\left(3/4,\;1/2\,y^2\right)\mathrm{I}\left(1/4,\;1/2\,y^2\right)$$
$$+3\sqrt{2}\sqrt{\pi}\,\Gamma\,(3/4)\,\mathbf{L}\left(5/4,\,1/2\,y^2\right)y^2\mathrm{K}\left(1/4,\;1/2\,y^2\right)\mathrm{I}\left(-3/4,\;1/2\,y^2\right)$$
$$+4\,y^2\sqrt[4]{y^2}\mathrm{K}\left(3/4,\;1/2\,y^2\right)\mathrm{I}\left(1/4,\;1/2\,y^2\right)$$
$$\left.+4\,y^2\sqrt[4]{y^2}\mathrm{K}\left(1/4,\;1/2\,y^2\right)\mathrm{I}\left(-3/4,\;1/2\,y^2\right)\right)$$

Table 9.2 Properties of the Yoshida jet

Exact	$v = \mathfrak{Y} v_{hom}(y) + v_{particular}(y)$
$\mathfrak{Y}$	$-1/4\sqrt{2\pi}\ \Gamma(3/4) = -.767916038651074$
v_{hom} analytical	$v_{hom} = \sqrt{y}\, \mathrm{I}_{1/4}\left([1/2]\, y^2\right)$
$v_{particular}$ analytical	$v_{particular} = 1/8\sqrt{2\pi}\,\Gamma(3/4)\, y^{5/2}\mathbf{L}_{1/4}\left(1/2\, y^2\right) \times$ $\left(\mathrm{K}_{3/4}\left(1/2\right] y^2\right)\mathrm{I}_{1/4}\left([1/2]\, y^2\right) + \mathrm{K}_{1/4}\left(1/2\, y^2\right)\mathrm{I}_{-3/4}\left([1/2]\, y^2\right))$
Slope at the origin $v_y(0)$	$dv/dy(y=0) = -\Gamma(3/4)^2/\sqrt{2\pi} = -0.59907011736$
Large y series for $v(y)$	$v \sim -1/y - 2/y^5 - 60/y^9 - 5400/y^9 + \ldots$
Small y series for v_{hom}	$y + \frac{1}{20}y^5 + \frac{1}{1440}y^9 + \frac{1}{224640}y^{13} + \ldots$
Small y series for $v_{particular}$	$\frac{1}{6}y^3 + \frac{1}{252}y^7 + \frac{1}{27720}y^{11} + \frac{1}{5821200}y^{15} + \ldots$

Table 9.3 Yoshida Jet Summary

Initial conditions	$u = 0, \quad v = 0, \quad \phi = 0$
Shallow water eqs.	$u_t - yv = 1, \quad v_t + yu + \phi_y = 0, \quad \phi_t + v_y = 0$
Solution Form	$u = t u_{sl}(y) + u_{waves}(y,t), \quad v = v_{sl}(y) + v_{waves}(y,t), \quad \phi =$ $t\,\phi_{sl}(y) + \phi_{waves}(y,t)$
$u_{sl}(y) =$	$-2\sum_{m=0}^{\infty} \frac{1}{(4m+3)(4m-1)}\, I_m\, \psi_{2m}(y)$
$=$	$1 + y v_{sl}$
$v_{sl}(y) =$	$-2\sum_{m=0}^{\infty} \frac{\sqrt{m+1/2}}{4m+3}\, I_m\, \psi_{2m+1}(y)$
$=$	$1/8\sqrt{2\pi}\,\Gamma(3/4)\, y^{5/2}\mathbf{L}_{1/4}\left([1/2]\, y^2\right) \times$ $\left(\mathrm{K}_{3/4}\left(1/2\right] y^2\right)\mathrm{I}_{1/4}\left([1/2]\, y^2\right) + \mathrm{K}_{1/4}\left(1/2\, y^2\right)\mathrm{I}_{-3/4}\left([1/2]\, y^2\right))$
$\phi_{sl}(y) =$	$-\sum_{m=0}^{\infty} \frac{4m+1}{(4m+3)(4m-1)}\, I_m \psi_{2m}(y)$
$=$	$-(1/24)\, y^{3/2}\left(-6\sqrt{2}\sqrt{\pi}\,\Gamma(3/4)\,\mathrm{I}\left(-3/4,\ 1/2\, y^2\right)\right.$
	$+3\sqrt{2}\sqrt{\pi}\,\Gamma(3/4)\,\mathbf{L}\left(1/4,\ 1/2\, y^2\right)\mathrm{K}\left(3/4,\ 1/2\, y\right)\mathrm{I}\left(1/4,\ 1/2\, y^2\right)$
	$+3\sqrt{2}\sqrt{\pi}\,\Gamma(3/4)\,\mathbf{L}\left(1/4,\ 1/2\, y^2\right)\mathrm{K}\left(1/4,\ 1/2\, y^2\right)\mathrm{I}\left(-3/4,\ 1/2\, y^2\right)$
	$+3\sqrt{2}\sqrt{\pi}\,\Gamma(3/4)\,\mathbf{L}\left(5/4,\ 1/2\, y^2\right)\, y^2\mathrm{K}\left(3/4,\ 1/2\, y^2\right)\mathrm{I}\left(1/4,\ 1/2\, y^2\right)$
	$+3\sqrt{2}\sqrt{\pi}\,\Gamma(3/4)\,\mathbf{L}\left(5/4,\ 1/2\, y^2\right)\, y^2\mathrm{K}\left(1/4,\ 1/2\, y^2\right)\mathrm{I}\left(-3/4,\ 1/2\, y^2\right)$
	$+4\, y^2\sqrt[4]{y^2}\mathrm{K}\left(3/4,\ 1/2\, y^2\right)\mathrm{I}\left(1/4,\ 1/2\, y^2\right)$
	$\left.+4\, y^2\sqrt[4]{y^2}\mathrm{K}\left(1/4,\ 1/2\, y^2\right)\mathrm{I}\left(-3/4,\ 1/2\, y^2\right)\right)$
$u_{waves}(y,t) =$	$\sum_{m=0}^{\infty} I_m \psi_{2m}(y)\left\{\frac{2m+1}{(4m+3)^{3/2}}\ \sin\left(\sqrt{4m+3}\, t\right) + \frac{2m}{(4m-1)^{3/2}}\ \sin\left(\sqrt{4m-1}\, t\right)\right\}$
$v_{waves}(y,t) =$	$\sum_{m=0}^{\infty} I_m \psi_{2m+1}(y)\ \frac{2\sqrt{m+1/2}}{4m+3}\ \cos\left(\sqrt{4m+3}\, t\right)$
$\phi_{waves}(y,t) =$	$\sum_{m=0}^{\infty} I_m \psi_{2m}(y)\left\{-\frac{2m+1}{(4m+3)^{3/2}}\ \sin\left(\sqrt{4m+3}\, t\right) + \frac{2m}{(4m-1)^{3/2}}\ \sin\left(\sqrt{4m-1}\, t\right)\right\}$
$I_m =$	$\sqrt{2}\,\pi^{1/4}\,\frac{\sqrt{(2m)!}}{2^m\, m!}$ [coefficients of one, i.e., $1 = \sum_{m=0}^{\infty} I_m \psi_{2m}(y)$]

Tables 9.2 and 9.3 summarize the explicit Yoshida jet solution.

9.4.5 *Rational Approximations: Two-Point Padé Approximants and Rational Chebyshev Galerkin Methods*

Substituting a series of inverse powers of y into $v_{yy} - y^2 v = y$ gives

$$v \sim -\frac{1}{y} - \frac{2}{y^5} - \frac{60}{y^9} - \frac{5400}{y^9} - \frac{982800}{y^{17}} - \frac{300736800}{y^{21}} + \ldots \tag{9.66}$$

There are no undetermined constants and the expansion involves only every fourth power of y. The factorial growth of the coefficients shows that this series is asymptotic but divergent.

The homogeneous solution which is bounded at the origin has the power series

$$v_{hom} \approx C_I (y + \frac{1}{20} y^5 + \frac{1}{1440} y^9 + \frac{1}{224640} y^{13} + \ldots) \tag{9.67}$$

The particular solution has the parameter-free expansion

$$v_p = \frac{1}{6} y^3 + \frac{1}{252} y^7 + \frac{1}{27720} y^{11} + \frac{1}{5821200} y^{15} + \ldots \tag{9.68}$$

The constant C_I cannot be determined soley from small y behavior, but rather one must apply the large y asymptotic expansions to find that the solution is bounded as $|y| \to \infty$ only if

$$C_I = v_y(0) = -\frac{\Gamma(3/4)^2}{\sqrt{2\pi}} = -0.5990701170 \tag{9.69}$$

The ODE has coefficients and inhomogeneous terms which are free of singularities except at the origin (where $v(y)$ is nonsingular) and at infinity, so the series have infinite radii of convergence. In practice, roundoff error limits the usefulness of the power series. The various expa nsions and identities for the Yoshida $v(y)$ are summarized in Table 9.2.

It is possible to combine the small y and large y series into rational functions called "two-point Padé approximants". The approximation we shall denote $v_{[3/4]}$ was constructed so that its power series matches the first four terms of the power series of $v(y)$ and also that its inverse power series matches the first four terms of the asymptotic expansion for $v(y)$.

$$v_{[3/4]} = \left\{-0.00466626 y^7 - 0.157887 y^3 - 0.0286541 y^5 - 0.599070 y\right\} \\ / \left\{0.0046662 y^8 + 0.028654 y^6 + 0.541761 y^2 + 1.0 + 0.148554 y^4\right\}$$

Table 9.4 A Maple code to apply the Galerkin method with rational Chebyshev basis functions

```
restart; with(orthopoly); with(LinearAlgebra);
N:=3;  L:=3; G:=Matrix(N,N, fill=0):  a:=Vector(N, fill=0):  ff:=Vector(N, fill=0):
SB:= (n,L,y) -> (L/sqrt(L**2+y**2))*U(n-1,y/sqrt(L**2 + y**2) );
f:= y -> y;
for irow from 1 to N do
ff[irow]:=int(SB(2*irow,L,y)*f(y)/(L**2+y**2),y=-infinity..infinity);
  for jcol from 1 to N do
  R:= diff( SB(2*irow,L,y),y,y) - y*y*SB(2*irow,L,y);
G[irow,jcol]:=int(SB(2*irow,L,y)*R/(L**2 +y**2),y=-infinity..infinity);
od: od:
a:=LinearSolve(G,ff);
v:=simplify(a[k]* SB(2*k,L,y),k=1..N);
```

Alternatively, one can apply a standard Galerkin method using rational Chebyshev functions as the basis. The computer code is brief in Matlab or Fortran, and briefer still in a computer algebra system. The basis functions are $\mathrm{SB}_{2n}(y;L) \equiv (l/\sqrt{L^2+y^2})U_{2n-1}\left(y/\sqrt{L^2+y^2}\right)$ where $U_j(x)$ denotes the j-th degree Chebyshev polynomial of the second kind. The choice of basis and of the map parameter L and also why the SB_j series converges exponentially fast whereas the Hermite series converges only as $n^{-3/4}$ are explained in [7]. The basis functions are orthogonal with weight function $1/(L^2+y^2)$ when integrated over the entire real axis. Table 9.4 is the complete Maple code to generate rational Chebyshev solutions of any order. The most significant feature of the code is its brevity.

Truncating the $n = 10$ solution to three terms gives

$$v_{SB} = -y\frac{(440.8017 + 15.098y^2 + 1.1412y^4)}{(9+y^2)^3} \tag{9.70}$$

The left of Fig. 9.6 shows that both approximations are graphically indistinguishable from the true Yoshida jet solution. Both are much simpler than lengthy Hermite function series with Hutton or Euler acceleration and also simpler than Bessel and Struve functions of order one-fourth.

The right panel of Fig. 9.6 shows that, in contrast to truncated Hermite series, the rational function approximations are uniformly good approximations with maximum pointwise errors of only 0.0089 for the two-point Padé approximant and 0.0058 for the truncated SB series. Spectral approximations are notorious for their near-uniformity, so it not surprising that the SB approximation, despite lower degree, is noticeably more uniform than the Padé approximation, which is a composite of two very non-uniform local approximations.

Hermite function series are a beautiful way to solve many problems in equatorial dynamics. We have detailed alternatives in this section not to advocate a particular methodology, but simply to point out that there are options and alternatives. Computer algebra has made it almost childishly easy to convert traditional series expansion

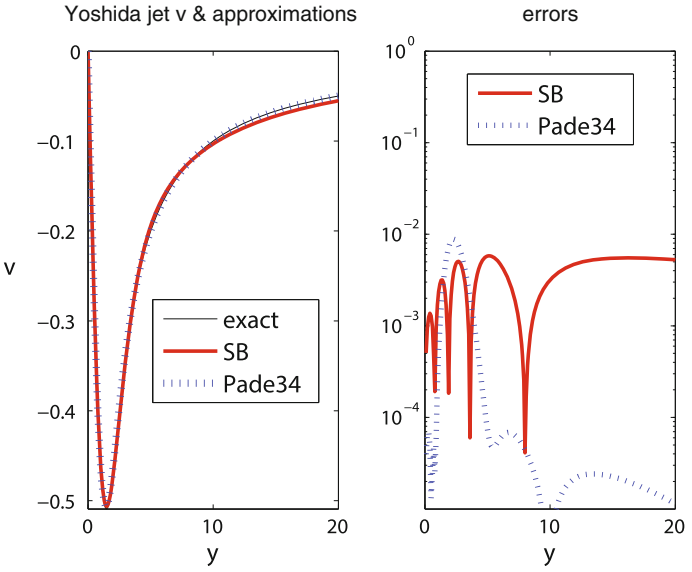

Fig. 9.6 The exact solution is compared with the pair of rational function approximations discussed in the text. There are three curves plotted but the errors, shown on the right, are so small that it is difficult to distinguish $v(y)$ from its approximations

methods — power series, inverse power series, Chebyshev, Fourier and rational Chebyshev spectral methods — into explicit analytic solutions [3, 8–10].

9.5 Unstratified Models of the Undercurrent

A rather confusing number of different theoretical explanations for the Equatorial Undercurrent have been offered over the past sixty years. The problem in creating a canonical theory is that many different physical mechanisms are known to be important. Provided certain essentials are included, one can then add in almost anything else — lateral viscosity, vertical viscosity, nonlinearity, thermal effects — and come up with a qualitatively reasonable theory of the Undercurrent. This relative success of many competing theories is due in part to the physical uncertainties in the free parameters of the model. One can vary the viscosity within the (large!) range allowed by measurements, and thus vary the width and strength of the Undercurrent until a good fit to observations is achieved. To add to the fun, the Undercurrent itself shows considerable seasonal and interannual variability and also varies in the zonal direction.

Nonetheless, considerable progress has been made. First, the "essentials" of the Undercurrent are now known. Second, the simple, analytic or quasi-analytic unstratified models illustrate the roles of different mechanisms in the Undercurrent even

though none of these models — because of the their neglect of other, equally important physics — is a *complete* model of the Undercurrent. Third, "honest" Undercurrent models, i.e., stratified, nonlinear numerical models, have given a good picture of how the different mechanisms come together. Because the stratified, nonlinear models are necessarily rather complicated, we shall confine our attention to simple generalizations of Yoshida's theory which like his theory use a single layer of homogeneous fluid.

The "essential" ingredient of the Undercurrent is a zonal boundary and its concomitant east-west pressure gradient. Yoshida's solution, which applies to an ocean without boundaries, gives only a flow in the direction of the wind stress, i.e., westward. This can be identified with the South Equatorial Current (SEC). Note that the SEC, despite its name, actually straddles the equator. The Undercurrent, however, flows *underneath* the SEC in a direction *opposite* the wind.

When the ocean flow is interrupted by a continent, the wind stress exerted by the easterly Trade Winds will pile up water on the western coast of the ocean, and thereby cause the sea surface to tilt from the east down towards the west.

9.5.1 Theory of Fofonoff and Montgomery (1955)

Theory One: Fofonoff and Montgomery [11]; also Veronis [12] and Cromwell [13].

Away from the equator, the hydrostatic pressure gradient created by the sea surface tilt would be balanced by a geostrophic flow *towards* the equator in both hemispheres. At the equator, however, the Coriolis force is zero, so there is nothing to balance the east-west pressure. The result is that fluid flows straight down the pressure gradient. Figure 9.7 illustrates this situation.

The equatorial thermocline tilts opposite the sea surface so that the pressure gradient *below* the thermocline is zero. It follows that if this simple idea is sensible, the Undercurrent should be confined above (or at least near) the thermocline rather than extending down into the deep ocean. In point of fact, the observed Undercurrent is surface-trapped and decays rapidly with depth below 200 m.

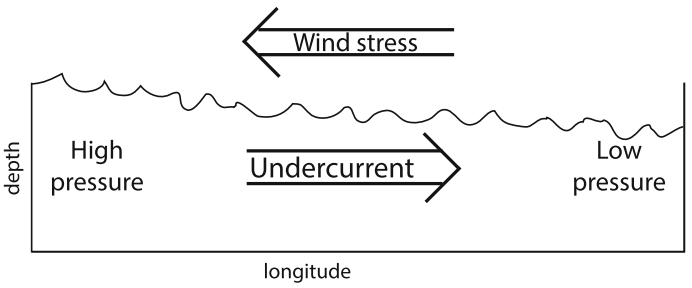

Fig. 9.7 Schematic of how the Equatorial Undercurrent is driven

The second fundamental idea is also contained in this same simple picture. The geostrophic flow that balances the zonal pressure gradient away from the equator is obviously advecting mass towards the equator. The Ekman transport, however, is directed *away* from the equator in both hemispheres (for a westward wind stress). At the surface, the Ekman stress must win but the flow below the Ekman layer must be towards the equator at some depths lest the tropical sea be denuded of water and turned into dry land. The picture is then one of equatorial upwelling with divergent flow away from the equator at the surface and convergent flow towards the equator beneath. Because the pressure gradient is small or zero below the thermocline and also because the thermocline is a stratification barrier to the upwelling created by Ekman suction at the surface, we do not expect the lower convergent flow to extend into the deep ocean, but rather to occupy a layer between the thermocline and the bottom of the Ekman layer.

Conservation of potential vorticity implies, however, that as particles are advected towards the equator, the decrease in f implies a gain in vorticity. A counterclockwise deflection of particles moving towards the equator requires an eastward deflection of the particles, just what is provided by the pressure gradient. Again, since the meridional flow is found only above the thermocline, this picture is consistent with the shallow depth of the observed Undercurrent.

In quantitative terms, if a particle starts at an initial latitude of y_0, conservation of vorticity shows that it will reach the equator with a velocity of

$$u = \beta\, y_0^2/2 \qquad \text{[dimensional]} \tag{9.71}$$

When $y_0 = 3^o$ N., $u = 100$ cm/s at the equator, which is about the observed magnitude of the Undercurrent.

This argument is only heuristic because, strictly speaking, it is the *potential* vorticity which is conserved, and predictions from vorticity conservation will obviously be modified by the convergence/divergence occurring in the equatorial region of upwelling. Still, it shows that a circulation which is highly localized about the equator can produce the necessary velocities. The east-west pressure gradient exists throughout the whole of the tropics (because the winds are from the east throughout the whole of the tropics), so it will certainly tend to drive a geostrophic flow towards the equator in the vicinity of ± 3 degrees of latitude. A number of theories have tried to go beyond these qualitative arguments to actually calculate an Undercurrent. We shall discuss a representative sample.

9.5.2 Model of Stommel (1960)

Theory Two: Stommel [14].

Stommel's model is classic Ekman layer theory in a single layer of incompressible fluid — constant coefficient viscous terms plus depth-independent pressure gradient terms — with three changes. First, the Coriolis parameter is allowed to vary with

y. Since standard Ekman theory — and Stommel's model — are solved point-by-point, allowing no y or x variations in the solution, the Coriolis parameter appears only parametrically in the Ekman equations anyway. Thus, looking at values of y for which f is very small requires no changes *per se* in Ekman theory at all.

The second change is more important: the depth of the fluid is taken to be *finite*. (In ordinary Ekman theory, the depth is taken to be infinite for the eminently sensible reason that away from the equator, the Ekman layer is no more than 100 m deep with the wind-driven currents decaying exponentially with depth.)

The third change is that the lower boundary condition to taken to be one of *no stress* rather than the seemingly more logical condition of *no slip*. The motive for this is that it is silly to suppose that the Ekman layer actually extends all the way to the bottom, 4000 m down. Interpreting the Ekman layer as the water *above* the thermocline; at the thermocline, a condition of no stress is more plausible than a rigid boundary. However, this is still a very ad hoc assumption: in a 1-1/2 layer model, there *is* a stress as the velocity jumps from its upper layer value to zero across the thermocline. Stommel's justification for choosing no stress is simply one of mathematical simplicity: "making the bottom a plane of no stress has the virtue of avoiding the necessity of calculating an additional functional layer at the bottom, thus simplifying the analysis".

With these assumptions, the model equations are

$$-fv = \nu u_{zz} - -(1/\rho)\, p_x \tag{9.72}$$

$$fu = \nu v_{zz} - -(1/\rho)\, p_y \tag{9.73}$$

where the pressure gradient terms are assumed to be constants and where the boundary conditions are

$$u_z = \tau, \qquad v_z = 0 \qquad z = 0 \tag{9.74}$$

$$u_z = 0, \qquad v_z = 0 \qquad z = -H \tag{9.75}$$

Now the usual infinite depth Ekman solutions have an e-folding scale of

$$D_e = \sqrt{2\nu/f} \tag{9.76}$$

As the equator is approached, $f \to 0$, so the depth $D_e \to \infty$, and the theory for an infinitely deep ocean is singular.

With an ocean of finite depth, however — regardless of the lower boundary condition — the singularity is *removed*. Instead, the Ekman layer simply fills the whole of the fluid, or more accurately, the whole of the fluid which feels the wind stress. Classical Ekman theory implicitly assumes that the eddy viscosity ν is constant at all depths. In reality, its magnitude drops off steeply below the mixed layer in the upper 100 m of the ocean so that it is only a thin layer that strongly feels the wind stress. Stommel's imposition of a no stress condition only a couple of hundred meters below the surface may be considered a crude way of allowing for this drastic decrease in

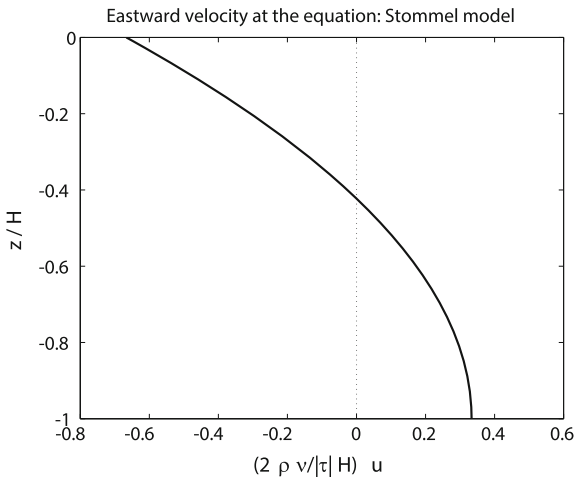

Fig. 9.8 $u(0, z)$ in Stommel's model, given in units of $|\tau|H/(2\rho\nu)$

viscosity while still remaining consistent with the constant coefficient mathematics of classic Ekman theory.

At the equator where f is zero, the x and y momentum equations become decoupled. The x momentum equation is the only one of interest since the wind stress is assumed to be purely zonal. This reduces to a direct balance between the zonal pressure gradient and the viscosity. Exactly this same balance occurs in the flow down a pipe in engineering fluid mechanics, and the solution is the same in either case, pipe flow or equatorial Ekman: the velocity is a parabolic function of the cross-stream coordinate. In the present case, we have only half a parabola because we have only one solid boundary, but the form is the same:

$$u = \tau \frac{H}{2\,\rho\,\nu} \left\{ \frac{(z+H)^2}{H} - \frac{1}{3} \right\} \tag{9.77}$$

as illustrated in Fig. 9.8.

Stommel's paper gives the finite-depth solution for general latitudes; this can be used to calculate the total (latitudinally-averaged) transport of the Undercurrent (though he does not do this himself).

9.5.3 Gill (1971) and Hidaka (1961)

Theory Three: Gill and Hidaka [15, 16].

Gill's model is very similar to Stommel's except for the addition of *latitudinal* viscosity. This is a significant complication because the presence of y derivatives implies that latitude can no longer be treated as a parameter, and the solution at one latitude depends on what is happening at all other latitudes. Otherwise, he includes

vertical viscosity, pressure, and Coriolis force, a wind stress acting only in the zonal direction, and a boundary condition of no stress on the bottom of an ocean of finite depth.

Gill divides the flow into a barotropic, depth-independent part (u^I, v^I) plus a depth-varying Ekman flow (u^∞, v^∞). He points out that by vertically-averaging the equations of motion, one finds that with a sidewall, the *vertically averaged* velocities in his model are *zero*; the hydrostatic pressure gradient produced by the tilting of the sea surface balances east-west wind stress. It follows that the vertically averaged transports of the barotropic and baroclinic parts of the flow must exactly balance.

Since Gill does not separate the flow into vertical modes, he uses a different nondimensionalization than that given in Chap. 2. Specifically, θ and λ below are latitude and longitude in radians, and his nondimensional quantities (unprimed) are related to the corresponding dimensional quantities (primed) by

$$z' = Hz, \qquad u' = \tau u/(2\Omega H), \qquad v' = \tau v/(2\Omega H) \tag{9.78}$$

$$w' = \tau w/(2\Omega a), \qquad \phi' = a\tau\phi/H \tag{9.79}$$

with the nondimensional parameters (Ekman numbers)

$$E_H = A/(2\Omega a^2), \qquad E_V = \nu/(2\Omega H^2) \tag{9.80}$$

where A is the latitudinal viscosity, ν is the vertical viscosity, Ω and a are the angular frequency and radius of the earth, and H the depth of the ocean. Note that Gill employs the rather odd convention of defining τ to be the *negative* of the wind stress. This is because a negative, i. e., westward, wind stress is normal near the equator. The model equations become

$$-\theta\, v + \phi_\lambda = E_H\, u_{\theta\theta} + E_V\, u_{zz} \tag{9.81}$$

$$\theta\, u + \phi_y = E_H\, v_{\theta\theta} + E_V\, v_{zz} \tag{9.82}$$

$$u_\lambda + v_\theta + w_z = 0 \tag{9.83}$$

subject to the boundary conditions

$$E_V\, u_z = -1, \qquad v_z = 0 \quad \text{at } z = 1 \quad \text{(top)} \tag{9.84}$$

$$u_z = 0, \qquad v_z = 0 \quad \text{at } z = 0 \quad \text{(bottom)} \tag{9.85}$$

For a layer of homogeneous fluid, there is only the hydrostatic pressure which isalways independent of depth. It follows that we can calculate the pressure gradient

terms above by vertically-averaging the equations of motion. However, as already pointed out earlier, the vertically-averaged velocities are 0, so the only allowed pressure gradient is that which balances the applied wind stress. Thus, we can replace the λ-derivative of ϕ by 0 and the θ-derivative by (-1) without approximation.

Away from the equator, the solution is rather obvious. (u^∞, v^∞) are the usual wind-driven Ekman spiral with latitude θ appearing only as a parameter. The barotropic part of the solution is

$$v^I = -1/\theta, \qquad u^I = 0 \qquad (\theta \text{ not small}) \tag{9.86}$$

This has the same form as the Ekman transport in our model of the Yoshida jet except that because of Gill's peculiar sign convention, the signs in (9.86) apply when the wind stress is westward whereas in the Yoshida jet problem, the $(-)$ sign for v was correct only for an eastward wind stress. The truth is that (9.86) gives a kind of anti-Ekman transport. It is simply the equatorward geostrophic current in balance with the zonal pressure gradient created by the tilted sea-surface. It has exactly the same form as the (non-geostrophic) vertically-averaged Ekman transport — but opposite sign — because the total vertically-averaged velocity must be zero.

At the equator, however, (9.86) and the Ekman spiral are both apparently singular. The depth independent solution satisfies the equations

$$-\theta\, v^I = E_H\, u^I_{\theta\theta} + 1, \qquad \theta\, u^I = E_H\, v_{\theta\theta} \tag{9.87}$$

A useful trick is to identify u^I and v^I as the real and imaginary parts of a complex velocity U. Next multiply the second part of (9.87) by i and add it to the first part to give a single differential equation for the complex velocity. This is the same trick that works in ordinary Ekman theory, and no wonder. The system (Eq. 9.87) has exactly the same form as that for conventional, wind-driven Ekman theory except they are not constant coefficient so the resulting differential equation is Airy's equation

$$E_H\, u_{\theta\theta} - i\,\theta\, U = -1 \tag{9.88}$$

Gill's paper gives both power series and asymptotic expansions for the solution of (9.88) plus a thorough discussion of how it scales as a function of the horizontal Ekman number. The Airy function solutions are graphed in Fig. 9.9. Figure 9.10 summarizes the different components of Gill's theory. There is strong equatorial upwelling and a weak but deep undercurrent under a shallower but much stronger flow towards the west.

A major shortcoming of Gill's model is that the mass transport in the model Undercurrent, which he shows is $-\pi\tau/\beta$ *independent* of the latitudinal or vertical viscosities, is smaller than that observed by a factor of 5. Gill is not too distressed; he notes that nonlinear effects can resolve the discrepancy. But this motivates the next theory: both observations and linear theory show that nonlinear effects are very important for the Undercurrent, so that a purely linear theory cannot hope to be quantitatively correct (except by accident!).

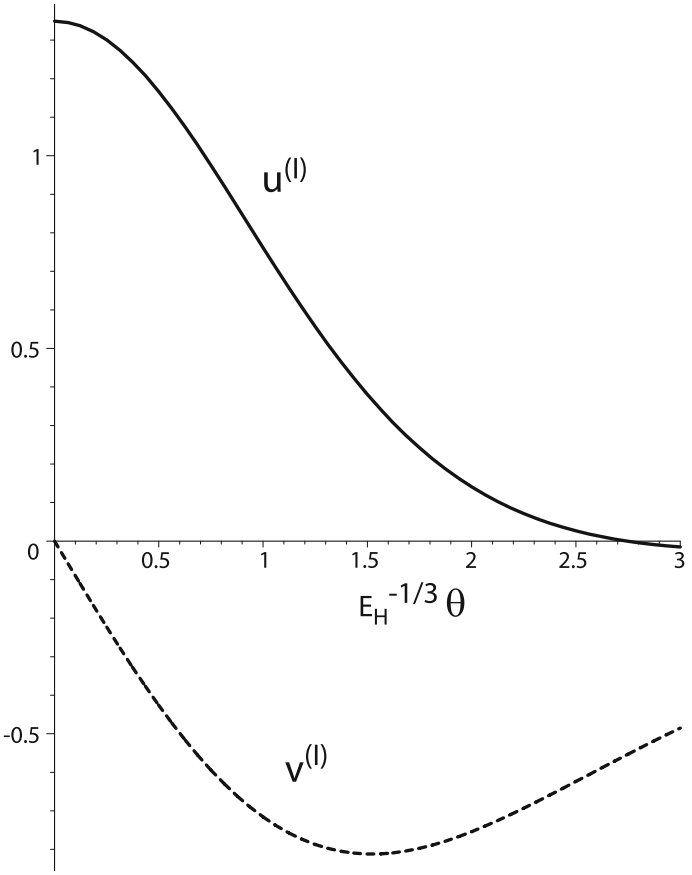

Fig. 9.9 u^I and v^I in Gill's model. The coordinate is $(E_H)^{-1/3}\,\theta$ where θ is latitude in radians

Theory Four: Charney, Charney and Spiegel [17, 18].

The basic equations and assumptions of Charney's model are identical to Gill's except that (i) the latitudinal friction is ignored (ii) the nonlinear terms are included and (iii) no slip lower boundary conditions are imposed although this is probably less significant than the other two changes. Since, as already pointed out above in the discussion of Gill's model, the only quantity which is allowed to vary with x is the pressure so that the pressure gradient can balance the wind stress via tilting of the sea surface, we can set all the zonal derivatives of the velocity components equal to 0 here and in all the Stommel and Gill models, too.

His model equations in dimensional form are

$$vu_y + wu_z - fv = -\phi_x + \nu\, u_{zz} \tag{9.89}$$

$$vv_y + wv_z + fu = -\phi_y + \nu\, v_{zz} \tag{9.90}$$

$$v_y + w_z = 0 \tag{9.91}$$

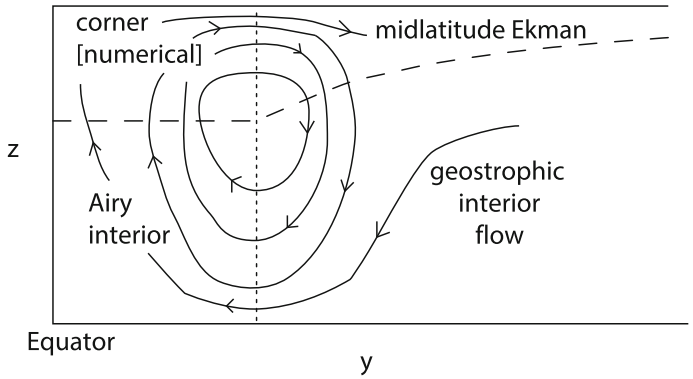

Fig. 9.10 The regions of Gill's model along with schematic streamlines of the meridional flow. The "Airy interior flow" is depicted in the previous figure; the analytic solutions are Airy functions. The "corner region" is described by coupled partial differential equations in both latitude and depth; numerical calculations by W.D. McKee appear as Figs. 2 and 3 of Gill [15]

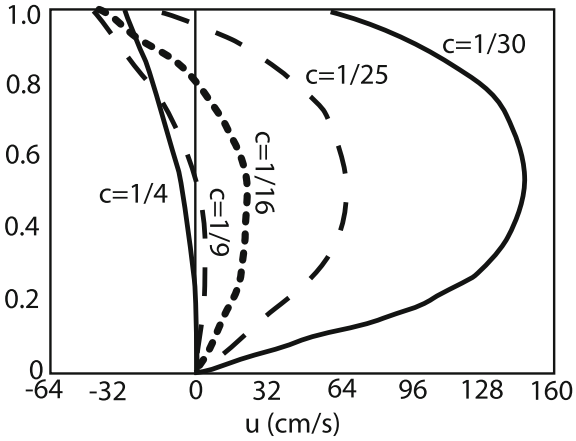

Fig. 9.11 East-west velocity in the model of Charney [17] as a function of scaled depth [unit depth = 150 m] for different values of the parameter $c = \nu/H^2\sqrt{P\beta}$

Naturally, Charney nondimensionalizes them, but it is not possible to find a single nondimensional scaling such that all the terms in the above equations are of $O(1)$ everywhere; with its rigid lower boundary, his model has both upper and lower Ekman layers with a geostrophic interior in between. Consequently, we will not repeat his scaling here, but simply note that the one important parameter of his model is the Ekman number defined by

$$c = \frac{\nu}{H^2\left\{\tau\beta/(\rho H)\right\}^{1/3}} \tag{9.92}$$

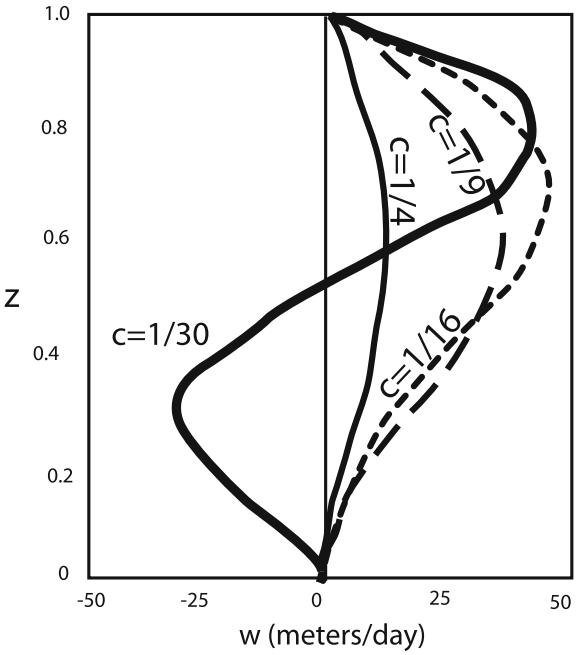

Fig. 9.12 Same as the previous figure, but showing the vertical velocity w as a function of scaled depth [unit depth = 150 m] for different values of the parameter $c = v/H^2\sqrt{P\beta}$. *Thick line* $c = \frac{1}{30}$, *thin solid line* $c = \frac{1}{4}$, *long dashes* $c = \frac{1}{9}$, *short dashes* $c = \frac{1}{16}$. Charney's [17] model

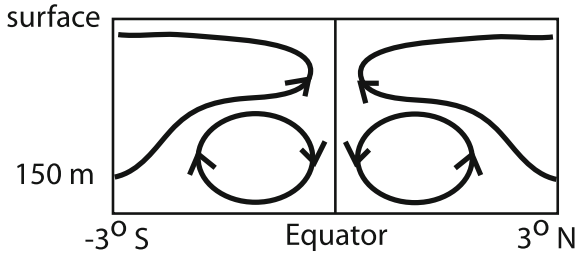

Fig. 9.13 Schematic of the meridional circulation in the model of Charney [17]

The steady-state was computed by reinserting the local time derivatives in (9.89) and then marching forward in time until the viscous terms had damped the transients.

For large c, Charney's model is strongly viscous and linear. As shown in Fig. 9.11, the flow is westward at all levels at the equator and there is no Undercurrent. The physical explanation is that the tendency of the pressure gradient to drive an eastward flow below the level of surface wind-driven westward flow is overwhelmed by the very rapid diffusion of westward momentum from above. To put it another way, the wind-driven Ekman flow extends all the way to the bottom when c is very large. (J. McCreary has shown that a linear Undercurrent is possible, but only when stable stratification is taken explicitly into account, as discussed later.)

When c is small, i.e., the flow is strongly nonlinear and the Ekman layers are trapped near the boundaries, one finds that the flow is *eastward at all depths* at the equator. Part of this effect is due to the equatorial upwelling: eastward momentum

is advected from the Undercurrent up into the Ekman layer. Robinson [19] showed, using perturbation theory, that the first correction to u is always eastward (for small nonlinearity), whatever the direction of the wind. For small nonlinearity, however, he showed that it was due to the divergence of the zonal momentum flux (Figs. 9.12 and 9.13).

Gill [20] gives a simple argument to explain this. In the vertically-integrated x-momentum equation, the pressure gradient cancels the wind stress; this is true of all the linear models discussed above, but it can be justified for nonlinear models as first done by Charney [17], who pointed out that the nonlinear terms are important only within two or three degrees of latitude of the equator. The wind stress varies little on the latitudinal scale of the Undercurrent because the Trade Winds extend roughly ten Undercurrent widths or more away from the equator. If the zonal pressure gradient varies rapidly with y, this implies that the x-dependent part of p — not merely the total pressure, but that part which is varying zonally — must be varying rapidly with y also. This would create a huge nonuniformity in x which contradicts a basic postulate of both theory and observations: that the Undercurrent varies slowly with x. Thus, the zonal pressure gradient and wind stress approximately balance each other even in the heart of the Undercurrent where the flow is strongly nonlinear.

Vertical integration of the equation of continuity with (i) assumption of no variations in x, which eliminates the u_x term and (ii) rigid upper and lower boundaries, which eliminate the boundary terms that result from the integration of the perfect z-derivative, w_z, gives the conclusion that the vertically-integrated north-south current is 0 — which is a conclusion common to all four of the theories discussed here. Thus, the vertically-integrated x-momentum equation reduces to (overbars denote a vertical average)

$$\bar{u}_t = -(\bar{u}\bar{v})_y - \nu\, u_z(z = \text{bottom})/H \tag{9.93}$$

The important point is that the divergence of the momentum flux is always positive. Looking at Fig. 9.14, which shows typical zonal and meridional flows for a nonlinear model, one sees that in the upper ocean where the zonal flow is westward (negative), the meridional flow is away from the equator ($v > 0$ in the northern hemisphere) because the wind-driven Ekman circulation is stripping mass away from the equator. Thus, $uv < 0$ in the upper ocean for $y > 0$. Near the bottom, meridional convergence must occur to supply mass to the upwelling at the equator, but this is also the region of the Undercurrent, so again $uv < 0$. It is possible for u and v to be of the same sign in the northern hemisphere in the middle of the layer, but both u and v are small there because they change sign somewhere in the middle, too. The conclusion is that the vertically-averaged momentum flux will be negative in the northern hemisphere and positive in the southern, and this will always give a divergence that is causing eastward accelerations in a zone around the equator (and driving westward currents at higher latitudes), regardless of the details of the model. Gill [20] notes that the flanking westward currents on either side of the Undercurrent (and like it, well below the surface) are often observed. What is more important is that nonlinearity greatly intensifies and narrows the Undercurrent, increasing its total mass transport as well.

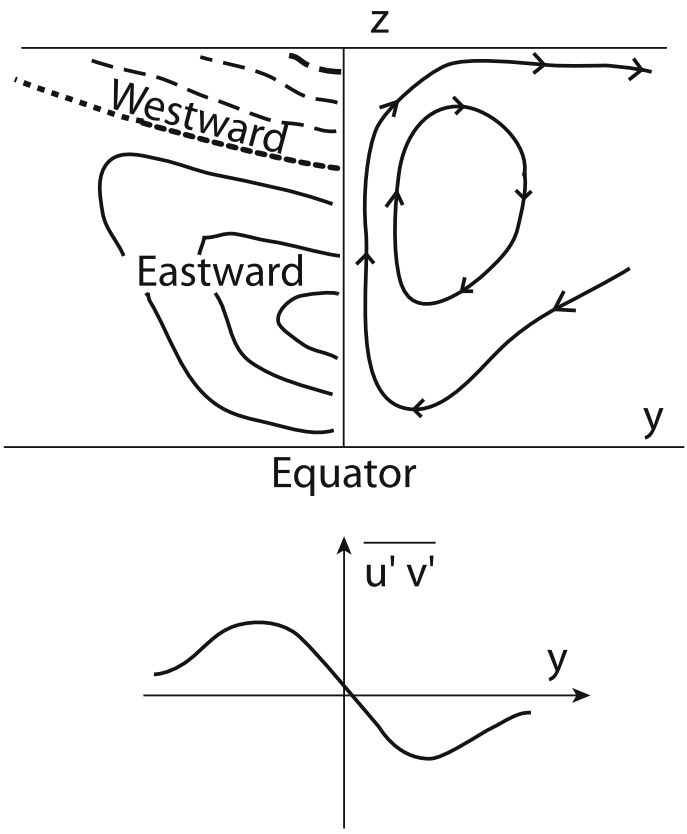

Fig. 9.14 *Upper left* isotachs of zonal velocity. Contours of westward flow are *dashed*; the zero eastward velocity contour is *dotted*. *Right* streamlines of the meridional circulation. *Bottom* latitudinal momentum flux. Theory of Gill [20]

When the winds reverse, both u and v reverse so that the sign of the momentum flux divergence is invariant. This implies that nonlinearity will weaken a westward Undercurrent, and this is important because any good theory for the Undercurrent must also explain why there isn't one in the Indian Ocean when the winds are from the west (Southwest Monsoon in summer).

In this vertically-averaged argument, the vertical and meridional advection of eastward momentum do not explicitly appear even though we have repeatedly referred to their role. These terms are implicit in the vertically-averaged momentum flux, however, because it is precisely the existence of the meridional circulation that creates the negative momentum flux at almost all (northern hemisphere) levels in the model.

Numerous other models could be mentioned, but we will return to discuss the Undercurrent in terms of stratified models later. Another limitation of our theory is the omission of zonal boundaries: the Yoshida jet is for an unbounded ocean, and the Undercurrent theories described above all assume that one is in the middle of the

ocean so that x-variations can be neglected. There must, alas, be boundary layers near the coasts where these x-variations are rapid and important. This, too, is something we must later return to.

One final complication we ignored here is that of the effects of cross-equatorial winds. Gill [20] describes a simple model due to W.D. McKee in which linear ("Rayleigh") friction is used to give a simple, but dissipative linear model. Its explicit solutions show — in accord with observations and the numerical nonlinear study of Charney and Spiegel [18] — that the Undercurrent tends to be moved off the equator upwind. For example, the jet axis is sometimes observed to be a degree south of the equator in the Indian Ocean when the winds are from the southeast.

In summary, there are three principal concepts in the Undercurrent. First, the wind stress created by the Trade Winds causes a sea surface tilt which hydrostatically creates an eastward pressure gradient; the Undercurrent then simply flows down this gradient below the westward flow in the wind-driven surface Ekman layer.

Second, the surface Ekman transport is *poleward*, so by mass continuity, the equator should be a region of strong upwelling, at least near the surface. (Note that some models predict small downwelling cells near the bottom, centered on the equator. These are too weak to alter the vertically-averaged momentum flux, however, but may help to extend the Undercurrent a bit below the thermocline, as observed.)

Third, nonlinear effects narrow an eastward Undercurrent and increase both the zonal velocity and the total mass transport. However, they weaken a (hypothetical!) westward Undercurrent, perhaps explaining why none is observed (Table 9.3).

References

1. Neumann G, Pierson WJ Jr (1966) Principles of physical oceanography, 2nd edn. Prentice-Hall, Englewood Cliffs
2. Moore DW, Philander SGH (1977) Modelling of the tropical oceanic circulation. In: Goldberg ED (ed) The Sea, no. 6 in The Sea, Wiley, New York, pp 319–361
3. Boyd JP (1993) Chebyshev and Legendre spectral methods in algebraic manipulation languages. J Symb Comput 16:377–399
4. Boyd JP (2001) Chebyshev and Fourier spectral methods. Dover, New York
5. Boyd JP, Moore DW (1986) Summability methods for Hermite functions. Dyn Atmos Oceans 10:51–62
6. Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
7. Boyd JP (1987b) Spectral methods using rational basis functions on an infinite interval. J Comput Phys 69:112–142
8. Boyd JP (2016) Correcting three errors in Kantorovich & Krylov's approximate methods of higher analysis: energizing perturbation series and Chebyshev and Legendre spectral algorithms with computer algebra. Amer Math Monthly 123(3):241–257
9. Boyd JP, Iacono R (2015) The Kidder equation: $u_{xx} + 2xu_x/\sqrt{1-\alpha u}$. Stud Appl Math 35(1):63–85
10. Boyd JP (1986) Solitons from sine waves: analytical and numerical methods for non-integrable solitary and cnoidal waves. Phys D 21:227–246
11. Fofonoff NP, Montgomery RB (1955) The equatorial undercurrent in the light of the vorticity equation. Tellus 7:518–521

12. Veronis G (1960) An approximate theoretical analysis of the equatorial undercurrent. Deep Sea Res 6(4):318–327
13. Cromwell T (1953) Circulation in a meridional plane in the central equatorial Pacific. J Marine Res 12(2):196–213
14. Stommel H (1960) Wind-drift near the equator. Deep Sea Res 6:298–302
15. Gill AE (1971) The equatorial current in a homogeneous ocean. Deep Sea Res 18(4):421–431
16. Hidaka K (1961) Equatorial upwelling and sinking in a zonal ocean with lateral mixing. Geophys J R Astr Soc 4:359–371
17. Charney JG (1960) Non-linear theory of wind-driven homogeneous layer near the equator. Deep Sea Res 6:303–310
18. Charney JG, Spiegel SL (1971) Structure of wind-driven equatorial currents in homogeneous oceans. J Phys Oceangr 1(3):149–160
19. Robinson AR (1966) An investigation into the wind as the cause of the Equatorial Undercurrent. J Mar Res 24:179–204
20. Gill AE (1975) Models of equatorial currents. In: Proceedings of a symposium on numerical models of ocean circulation. National Academy of Sciences

Chapter 10
Stratified Models of Mean Currents

Abstract The alternating surface jets and the Equatorial Undercurrent are important features that have long challenged both numerical modelers and the perturbation theorists. In this chapter, stratification and viscosity are added to the shallow water jets found by Yoshida. The models of McCreary and McPhaden are ancient and cannot fit observations as well as the ever-changing succession of ocean general circulation models. Yet they are well worth a careful analysis to understand the underlying mechanisms.

> If many remedies are prescribed for an illness, you may be certain that the illness has no cure.
>
> — Anton Chekhov, physician, playwright and short-story writer

10.1 Introduction

Figure 10.1 shows the major equatorial currents. The westward South Equatorial Current (SEC), the westward North Equatorial Current (NEC) and the eastward North Equatorial Countercurrent (NECC) which lies between them are all *surface* currents that are reasonably well predicted by the 1–1/2 layer model of Busalacchi, Takeuchi and O'Brien [1, 2]. Consequently, our attention in this chapter will be focused on the Equatorial Undercurrent (EUC) and its two satellites, the Subsurface Countercurrents (SSCC's) which are sometimes contiguous with the EUC (western Pacific) and sometimes separated from the EUC by bands of westward flow (eastern Pacific). For clarity, we shall largely concentrate on the Pacific currents because these are the most stable and the most studied, both observationally and theoretically, but it should be emphasized that the EUC is found in all three oceans.

In Chap. 9, we have already reviewed a number of attempts to explain the EUC using barotropic models. These pioneering calculations are still useful in delineating the major balance of forces in the EUC. The eastward pressure gradient, created by the piling up of water along the western coast of the ocean, provides the driving force for the current. The usual Ekman flow will strip surface water away from the

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_10

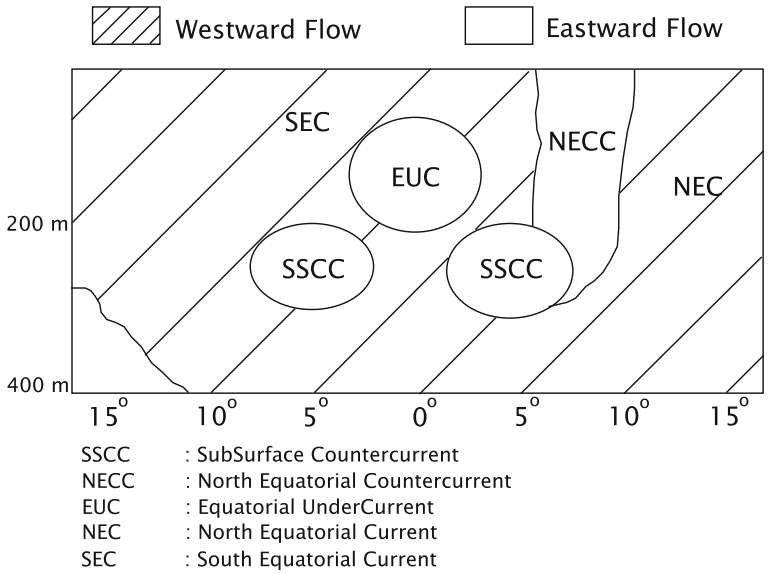

Fig. 10.1 Schematic of equatorial currents in a latitude-height cross-section (mid-ocean)

equator in both hemispheres, forcing a secondary circulation with upwelling along the equator itself. Finally, the barotropic and two-layer models show that nonlinear effects are very important to the Undercurrent. In particular, nonlinear advection tends to intensify and narrow an eastward Undercurrent, but to weaken a westward Undercurrent. This may explain why no westward Undercurrent is found in the Indian Ocean during the summer Southwest Monsoon, even though an eastward Undercurrent is present during the Northeast Monsoon when the winds resemble the normal trade winds.

The important role of nonlinearity is a modest limitation on the analytic theories discussed below: observations as well as the earlier theories imply that nonlinear numerical models are essential to completely capture the physics of the EUC. Nonetheless, McCreary [3] and McPhaden [4] approximate an EUC and the two SSCC's, respectively, with linear, analytical theories. The extension from the one baroclinic mode — the 1–1/2 layer models discussed in earlier chapters — to a sum over many baroclinic modes makes a vast difference in simulating these subsurface currents, even though most of the other artificialities of the Cane–Sarachik analytic articles are retained.

The east-west variations of the Undercurrent are an important test of a model. Fig. 10.2 is an $x - z$ section of isotherms across the Pacific. In the west, there is a thick, well-mixed surface layer; rapid temperature variations occur between 150 and 250 m down (notice that 15°, 20°, and 25 °C. isotherms are all scrunched up in this region) whereas the temperature changes by only a couple of degrees between this region and the surface. Eastward, the pycnocline becomes more and more shallow,

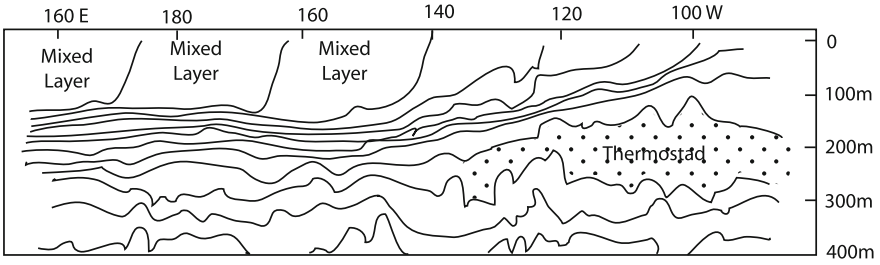

Fig. 10.2 Schematic of isotherms in an equatorial plane (longitude-depth cross-section) in the Pacific. The thermostad is marked by *dots*; the mixed layer, which lies above the closely-packed isotherms at all longitudes, is also labeled

and even the 15 °C. isotherm, buried more than 200 m down in the far west, nearly breaks the surface in the east.

Below the pycnocline in the east, however, is another thick layer, the "thermostad", which is the name applied to a layer of very uniform water of 34.9 parts-per-thousand salinity and 13 °C. temperature found between 150 and 300 m down with a latitudinal extent of about 5 ° N. to 5 ° S. This well-mixed sub-pycnocline is shaded in Fig. 10.2. A good baroclinic model must therefore explain two uniform layers: the thick, surface mixed layer in the west and the equally thick and equally uniform thermostad in the east.

In the past there was considerable debate about whether the Equatorial Undercurrent was or was not in geostrophic balance. The definition of geostrophic flow can be extended to the equator itself by noting that ϕ_y must vanish at the equator if the geostrophic zonal velocity is to remain finite. L'Hopital's Rule, which is equivalent to expanding both the numerator and denominator in Taylor expansions about $y = 0$, then gives

$$u(y = 0) \approx \phi_{yy}(y = 0) \tag{10.1}$$

The question still remains: Do the observations support geostrophy?

Figures 2 and 3 of Lukas and Firing [5] [not shown] give an emphatic answer: Yes! Tahiti and Hawaii lie very near the same meridian, but differ in latitude by about 40 degrees. The "Hawaii-Tahiti Shuttle" simply steamed back and forth along the equator between these two islands, collecting the most detailed and complete set of mid-ocean measurements of the $y - z$ cross-section of the mean currents obtained up to the mid-80's. Lukas and Firing show that the maximum absolute difference between the zonal velocity computed by geostrophy from the pressure field and the flow measured directly by a Profiling-Current-Meter (PCM) is less than 20% of the maximum current in the EUC. Since there are observational errors in both ϕ and u, this is an upper bound on the geostrophic error; the actual ageostrophic flow may be much weaker.

In the rest of this chapter, we shall concentrate on explaining these observational characteristics using the analytic theories of McCreary [3, 6] [for the Undercurrent] and McPhaden [4] [for the Subsurface Currents]. Both studies make somewhat artificial assumptions and neglect nonlinear terms so as to make it possible to separate variables. The solution is then a double sum over all baroclinic modes and over all latitudinal modes. First, however, we begin with a simple model in which Hermite functions are unnecessary.

10.2 Modal Decompositions for Linear, Stratified Flow

The mathematics problem is to solve a single partial equation in the three spatial coordinates in a domain that is either unbounded in both x and y, or is a strip that is unbounded in latitude but confined by two straight coastlines parallel to the y-axis. The flow is assumed to be steady-state. The first step is to reduce the three-dimensional boundary value problem to an infinite set of two-dimensional problems by expanding the solution in the baroclinic modes described in Chap. 3. As in Chap. 2, we use special forms for the viscous damping and diffusion so that separation of variable is possible.

The latitudinal dependence is usually represented as Hermite series. McCreary found that one hundred baroclinic modes and seventy Hermite functions resolved all his cases; this resolution was affordable even on the number-crunching hardware of forty years aago.

The equations of McCreary's [3] model for the Undercurrent are the same as Eqs. (2.29) and (2.30) with the time derivatives suppressed (because the forcing is assumed to be steady). Baroclinic mode by baroclinic mode, the equations for the x and y dependence of the coefficients of the m-baroclinic mode are

$$A\lambda_m u_m - f v_m + \phi_{m,x} = F_m \tag{10.2}$$

$$A\lambda_m v_m + f u_m + \phi_{m,y} = G_m \tag{10.3}$$

$$A\lambda_m^2 \phi_m + u_{m,x} + v_{m,y} = 0 \tag{10.4}$$

with the dependent equations

$$\rho_m = \phi_m (\rho_M / g) \tag{10.5}$$

$$w_m = A \lambda_m^2 \phi_m \tag{10.6}$$

Here, λ_m is the eigenvalue of the vertical structure differential equation, A is the viscosity coefficient, and ρ_M is the mean density of the ocean (a constant). The variables $\{u_m(x, y), v_m(x, y), \phi_m(x, y), w_m(x, y), \rho_m(x, y)\}$ are the coefficients of the vertical structure functions in the expansions of u, v, etc., in terms of the baroclinic

modes; note that w_m must be multiplied by the integral of $\zeta_m(z)$ and ρ_m by $\mathfrak{z}_{m,z}(z)$ while the other three are multiplied by the vertical eigenfunction $\mathfrak{z}_m(z)$ itself. The F_m and G_m are the projections of the wind stress forcing onto the m-th baroclinic mode, with the normalization $\mathfrak{z}_m(0) = 1$,

$$F_m = \tau^x \, \mathfrak{z}_m(0) \, / \int_{-D}^{0} \mathfrak{z}_m^2 \, dz \tag{10.7a}$$

$$G_m = \tau^y \, \mathfrak{z}_m(0) \, / \int_{-D}^{0} \mathfrak{z}_m^2 \, dz \tag{10.7b}$$

where τ^x and τ^y are the x and y components of the wind stress vector and $z = -D$ is the bottom of the ocean. Although this formulation can handle quite general wind stress distributions, we specialize to a purely zonal wind stress and further take the stress to be the product of a function of x times a function of y, i.e.,

$$F_m = \tau_m \, X(x) \, Y(y) \tag{10.8}$$

where τ_m is a constant.

One complication introduced by stratification is that the parameter-eliminating nondimensionalization employed for the "1–1/2 layer" model of earlier chapters is *different* for each different baroclinic mode. One can go ahead anyway, but then the wind stress parameters, viscosity and frequency (for a time-dependent problem) are all different from mode to mode, which is rather confusing. Most workers have therefore imitated meteorological practice and employed a nondimensionalization which is the same for all baroclinic modes. This implies that A, τ, and the frequency σ (if there is one) are the same for each mode, but it also implies that the eigenvalue of the vertical structure equation will appear in many places in our equations.

The normal choice in meteorology is to use the radius of the earth a as the horizontal length scale, take the latitudinal coordinate to be latitude (measured in radians), and use $1/(2\,\Omega)$ as the time scale where Ω is the angular frequency of the earth's rotation. With this choice, $\lambda_m = \varepsilon$, Lamb's parameter for that particular mode. For oceanographic purposes, it might be more convenient (since the low latitudinal modes are confined within an equatorial zone that is far smaller than the radius of the earth a) to nondimensionalize all the modes using the Lamb's parameter-dependent scaling appropriate to the first baroclinic mode. In this case, λ_m in the formulas below would be interpreted as the ratio of Lamb's parameter for the n-th mode to that for the first baroclinic mode. The choice matters not a whit as long as one is consistent. However, either way, it will be necessary to carry around factors of $\beta(\equiv df/dy)$ and λ_m throughout the rest of our analysis.

Let y denote the latitudinal variable of the chosen mode-independent nondimensionalization. We can then define the nondimensional argument of the Hermite functions,

$$\xi \equiv \mathfrak{a}_0 \, y \tag{10.9}$$

where

$$\mathfrak{a}_0 = \sqrt{\beta}\,\lambda^{1/4} \tag{10.10}$$

where the subscript "m" has been omitted in the last two equations for notational simplicity.

The Hermite functions $\psi_n(\xi)$ satisfy the differential equation (in y!)

$$\psi_{n,yy} - \lambda\,\beta^2\,y^2\,\psi_n = -\mathfrak{a}_n^2\,\psi_n \tag{10.11}$$

where

$$\mathfrak{a}_n^2 \equiv \mathfrak{a}_0^2\,(2n+1) \tag{10.12}$$

The latitudinal width varies as $O(1/\sqrt{m})$ where m is the vertical mode number. Figure 10.3 illustrates a typical Hermite function for each of the first fifty vertical modes with the width assumed to narrow as predicted by the estimate. Mode fifty has a latitudinal width which is an order of magnitude narrower than the first baroclinic mode.

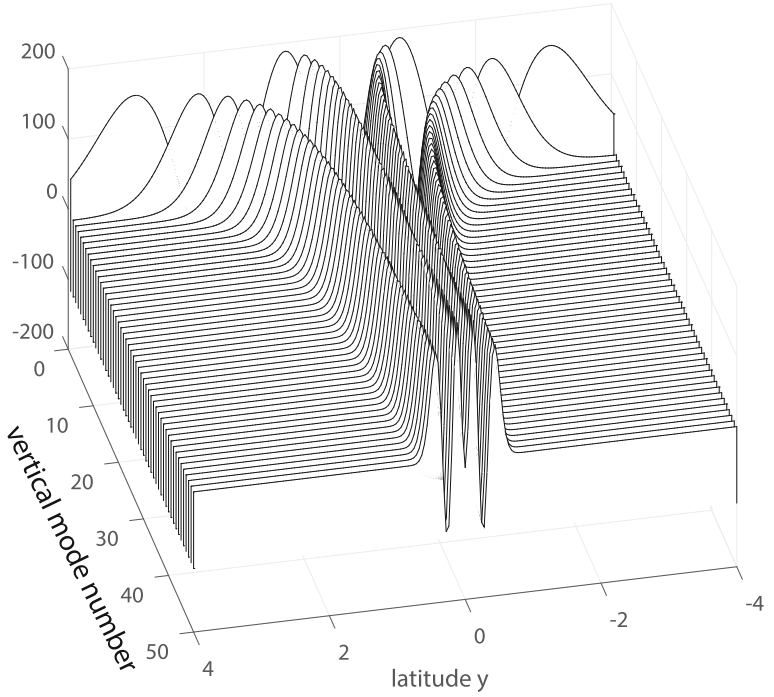

Fig. 10.3 The latitudinal width of the Hermite modes of a stably stratified ocean is strongly sensitive to the vertical wavenumber with the scalefactor shrinking approximately as $s \sim 1/\sqrt{m}$ where m is the vertical mode number. (The precise curve of width versus m depends on the assumed vertical stratification profile.) The waterfall plot shows $\psi_6(\sqrt{m}\,y)$; each slice graphs a different vertical eigenmode m

Algebraic reduction of the system of three shallow water wave equations to a single equation for v alone yields:

$$\boxed{v_{xx} + v_{yy} - i(\beta/\omega)v_x + [\lambda(\omega^2 - \beta^2 y^2]v = \lambda\beta y F + i(1/\omega)F_{yx}} \tag{10.13}$$

where the complex frequency ω is defined as

$$\boxed{\omega = -i\lambda A} \tag{10.14}$$

Note that this convention — time dependence proportional to $\exp(+i\omega t)$ — is opposite to the frequency convention assumed elsewhere in these notes.

Expand both sides of (10.13) in terms of Hermite functions, i.e.

$$v(x, y) = \sum_{n=1}^{\infty} v_n(x)\,\psi_n(\xi); \qquad F(x, y) = \sum_{n=1}^{\infty} F_n(x)\,\psi_n(\xi) \tag{10.15}$$

WARNING: Remember that the scale factor for ξ is different for each baroclinic mode; this implies that we must repeatedly calculate the expansion coefficients for each mode even when $F(x, y)$ is of separable form. The expansion coefficients $v_n(x)$ satisfy the ordinary differential equations

$$v_{n,xx} - i(\beta/\omega)v_x - [\mathfrak{a}_n^2 - \lambda\,\omega^2]v = \tau\mathfrak{a}_0\left\{\lambda^{1/2}[\xi Y]_n\,X(x) + i(1/\omega)\,[Y_\xi]_n\,X_x(x)\right\} \tag{10.16}$$

where $X(x)$ and $Y(y)$ are the factors of the zonal wind stress for the m-th baroclinic mode: $F(x, y) = \tau_m X(x)Y(y)$ where we have dropped the subscript on the constant τ in (10.16).

The sums in (10.15) start with $n = 1$; the reduction to a single equation for v omits the Kelvin wave entirely and introduces a spurious branch for the $n = 0$ Yanai, so we must return to these waves later. Parenthetical note: if we had assumed a non-separable wind stress, the only complication would be that the function $X(x)$ in (10.16) would be different for each latitudinal mode number n.

For simplicity, we shall omit the subscript "m" denoting the baroclinic mode number in the rest of this chapter, except where it might cause confusion.

Up to now, the approximations are only those so carefully delineated in Chap. 2. All the approximations made in the rest of the chapter are alterations of (10.2)–(10.4).

10.3 Different Balances of Forces

McCreary [6], in the absence of any accepted terminology for describing the different balances of forces in a baroclinic mode, introduced some terminology of his own which is summarized in Table 10.1. He omitted a label for the system (10.17)–(10.19).

Table 10.1 Definitions of balances in baroclinic modes. A is the viscosity coefficient and λ_m is the eigenvalue of the m-th baroclinic mode. All definitions except the first are from McCreary [6]

BJERKNES
$-fv_m + \phi_{m,x} = F_m$
$fu_m + \phi_{m,y} = G_m$
$A\lambda_m^2 \phi_m + u_{m,x} + v_{m,y} = 0$
EKMAN BALANCE:
(PSEUDO-EKMAN omits the terms in braces)
$[A\lambda_m u_m] - fv_m = F_m$
$\{A\lambda_m v_m\} + fu_m = G_m$
$A\lambda_m^2 \phi_m + u_{m,x} + v_{m,y} = 0$
SVERDRUP BALANCE: A vertical mode in which the dominant balance is
$-fv_m + \phi_{m,x} = F_m$
$fu_m + \phi_{m,y} = G_m$
$u_{m,x} + v_{m,y} = 0$
YOSHIDA BALANCE: (independent of x)
(PSEUDO-YOSHIDA omits terms in braces)
$A\lambda_m u_m - fv_m = F_m$
$\{A\lambda_m v_m\} + fu_m + \phi_{m,y} = G_m$
$A\lambda_m^2 \phi_m + v_{m,y} = 0$

We shall call this the "Bjerknes" approximation because J. Bjerknes [7] [see also the review by Hendershott [8]] made a similar approximation to describe ultra-long Rossby waves.

10.3.1 Bjerknes Balance

It is useful to study the approximation of neglecting the terms proportional to λ_m in the momentum equations only:

$$-f v_m + \phi_{m,x} = F_m \tag{10.17}$$

$$fu_m + \phi_{m,y} = G_m \tag{10.18}$$

$$i\omega\lambda_m \phi_m + u_{m,x} + v_{m,y} = 0 \tag{10.19}$$

This approximation is more drastic than the "long wave" or "latitudinal geostrophy" approximation employed in earlier chapters because the 'long wave" approximation demanded geostrophic balance in one horizontal coordinate only. The Bjerknes approximation is legitimate only when both the x and y length scales are large in comparison to the Rossby radius of deformation:

$$L_x >> 1/\mathfrak{a}; \qquad L_y >> 1/\mathfrak{a} \tag{10.20}$$

$$1/\mathfrak{a} \equiv 1/(f\,\lambda^{1/2}) = \text{Rossby radius} \tag{10.21}$$

As might be expected, the Bjerknes approximation fails in many situations where latitudinal geostrophy is highly accurate, but because of its simplicity, it is still useful to analyze its consequences.

This Bjerknes Balance system can be reduced to the single equation, dropping the baroclinic mode number m for notational simplicity,

$$v_x - i\,\frac{\omega\,\lambda\,f^2}{\beta}\,v = i\,\frac{\omega\,\lambda\,f}{\beta}\,F - \frac{1}{\beta}\,F_{yx} \tag{10.22}$$

or using $\omega = -i\,A\lambda$,

$$\boxed{v_x - \frac{A\,\lambda^2\,f^2}{\beta}\,v = \frac{A\,\lambda^2\,f}{\beta}\,F - \frac{1}{\beta}\,F_{yx}} \tag{10.23}$$

The general solution is

$$\mathfrak{Q} = -\,i\,\frac{-\omega\lambda f^2}{\beta} \tag{10.24}$$

$$= -\,\frac{A\,\lambda^2 f^2}{\beta} \tag{10.25}$$

$$v_x - \mathfrak{Q}v = \mathfrak{F} \tag{10.26}$$

$$v = -\beth,\exp(\mathfrak{Q}x) + \exp(\mathfrak{Q}x)\int_{-\infty}^{x} dx'\exp(-\mathfrak{Q}x')\,\mathfrak{F}(x') \tag{10.27}$$

where Ξ is a constant determined from the boundary condition. Since we have made the long wave approximations, all homogeneous solutions have westward phase and group velocities and the correct boundary condition is $v(\infty) = 0$.

An enlightening special case is a top-hat wind stress, constant everywhere between $|x| < L$ and zero elsewhere

$$F(x,y) = \begin{cases} 0, & |x| > L \\ 1, & x \in [-L, L] \end{cases} \tag{10.28}$$

$$v = \exp(\mathfrak{Q}x)\left\{\int_{-\infty}^{x} dx'\exp(-\mathfrak{Q}x')\,\mathfrak{F}(x') - \int_{-\infty}^{\infty} dx'\exp(-\mathfrak{Q}x')\,\mathfrak{F}(x')\right\}$$

$$\beth = (\exp(\mathfrak{Q}L) - \exp(-\mathfrak{Q}L))/\mathfrak{Q} \tag{10.29}$$

$$v = \exp(\mathfrak{Q}x)\begin{cases} -\beth, & x < -L \\ -\beth + (\exp(-\mathfrak{Q}\mathfrak{x}) - \exp(\mathfrak{Q}L))/\mathfrak{Q}, & |x| < L \\ 0, & x > L \end{cases} \tag{10.30}$$

10.4 Forced Baroclinic Flow in the "Bjerknes" Approximation

Bjerknes' approximation describes only ultra-long Rossby waves which have a westward group velocity. Consequently, we only need to impose a single boundary condition: the solution eastward of the region of the wind stress must be zero. This is automatically guaranteed by setting

$$v(x = \infty, y) = 0 \qquad \forall y \tag{10.31}$$

Since $\mathfrak{Q}$ is proportional to f, and f increases with latitude, it follows that we can make the exponential in (10.29)–(10.30) vary with x as fast as we wish simply by moving to a sufficiently high latitude. Conversely, the solution will break down for sufficiently large $|x|$ because the y-dependence of f implies that the solution is varying in y at a rate that varies directly with x. Thus, Bjerknes' approximation is never uniform over all space. This limitation is discussed in greater detail in Schopf et al. [9]; these authors use Bjerknes' approximation — which is not dispersive with respect to either zonal or latitudinal wavenumber — as a starting point for an analysis of ray paths on the equatorial beta-plane.

When ω is pure imaginary, the Rossby waves simply decay exponentially westward of the patch of wind stress.

Figure 10.4 illustrates weak, strong and intermediate damping for *steady* flow ($\sigma = 0$) with the wind stress in the form of a "tophat", i.e., constant over some finite patch of water. Since there is no propagation or decay in y, there is no coupling from

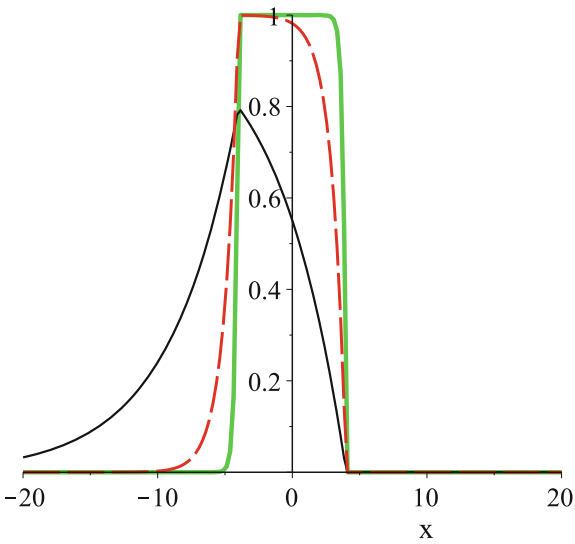

Fig. 10.4 Limiting cases of forced "Bjerknes" flow. *Thin black curve*: Weak damping ($\lambda A \to 0$) [Sverdrup flow]. *Thick green curve*: Strong Damping ($\lambda A \to \infty$) [Pseudo-Ekman flow]. *Red dashed curve*: Intermediate Damping. The wind stress is (-1) on $x \in [-4, 4]$. $\mathfrak{Q} = 5$ for strong damping, 1 for the intermediate case, and 1/5 for weak damping. Note that $\mathfrak{Q}v$ is plotted so that the maxima for all three cases coincide

one latitude to another and the problem is essentially one-dimensional, so it suffices to graph the solution as a function only of x.

When the damping is weak (small $\mathfrak{Q}$), the e-folding scale of the decaying Rossby waves is large in comparison to the width of the wind stress patch. This limit is

$$v \approx -(1/\beta) \int_{\infty}^{x} F_{yx}\, dx' = F_y/\beta \qquad \text{[Sverdrup flow]} \tag{10.32}$$

Figure 10.4 shows the opposite limit of strong damping as well. The Rossby waves decay rapidly before crossing the wind patch. Although we can derive the results in this limit from the integrals in (10.29)–(10.30), too, it is simpler and more illuminating to return to (10.17)–(10.19). Since ω appears in (10.19) as a multiplier of the ϕ term, it follows that $\phi << u, v$ in this limit. Then the pressure gradient terms drop out of the horizontal momentum equations, leaving what McCreary calls a "pseudo-Ekman" balance:

$$v \approx -F/f \tag{10.33}$$

In a *vertically-averaged* model, we would omit the "pseudo": Eq. (10.33) is always true for the vertically-averaged meridional velocity in the classic Ekman problem. However, we have actually solved the shallow water wave equations for a *single* baroclinic mode; these equations are not vertically-averaged. As McCreary points out, assuming that (10.33) is true on a mode-by-mode basis — instead of simply for the *sum* of the modes after averaging — is a fairly drastic approximation. It implies that the whole mixed layer moves as a slab at right angles to the wind stress, and the Ekman spiral is suppressed.

In the real ocean, weak damping — Sverdrup balance — is a good assumption for the lowest few modes. However, since viscosity is scale dependent, the strength of the vertical viscosity increases rapidly with the baroclinic mode number, so the high order modes must inevitably be strongly damped, implying a pseudo-Ekman balance.

10.4.1 Other Balances

In the rest of this section, we will pursue models that make less drastic assumptions. However, the methodology will remain quite similar. Again the low order baroclinic modes satisfy one type of balance while the high order modes are governed by a quite different sort of dynamics.

McCreary [3, 6] used a wind profile "representative of steady trades in the Pacific" by picking a zonal wind stress in the form

$$F(x, y) = \tau\, X(x)\, Y(y) \tag{10.34}$$

where $\tau = -0.5\,\text{dyn/cm}^2$ and

$$X(x) \equiv \begin{cases} \cos(\pi x/\Delta x) & |x| \le 0.5\,\Delta x \\ 0 & |x| > 0.5\,\Delta x \end{cases} \tag{10.35}$$

$$Y(y) \equiv \left(1 + y^2/(\Delta y)^2\right) \exp(-y^2/(\Delta y)^2) \tag{10.36}$$

where $\Delta x = 5000\,\text{km}$ and $\Delta y = 1000\,\text{km}$. For small y, this wind stress is almost y-independent like many of the idealized profiles employed before. The wind stress is confined to a patch that extends over half the longitudinal width of the ocean (Fig. 10.5). Leetmaa, McCreary and Moore [10] show that this shape is surprisingly close to the observed mean wind stress in the Pacific and the maximum is about $0.5\,\text{dyn/cm}^2$. This localized wind stress is different from the x-independent stresses of earlier models, but still drives an Undercurrent that extends across the whole ocean.

In order to separate variables, McCreary was forced to assume (unrealistically) that the viscosity ν and diffusion κ are both equal to A/N^2 where $N^2(z)$ is the Brunt-Vaisala frequency and A is a constant. If there is a surface mixed layer (i.e., a layer of uniform density), the viscosity coefficient is infinite there. ν dips to a minimum ν_{min} just below the thermocline — the thermocline is always accompanied by a large positive spike in N^2 — and then gradually increases downward into the deep ocean where N^2 is slowly decreasing. McCreary somewhat arbitrarily adjusted A so that

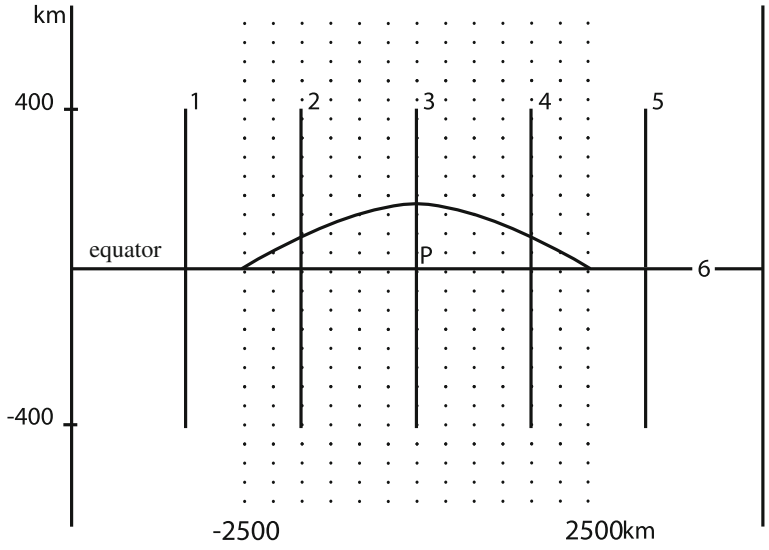

Fig. 10.5 *Top* Schematic of the model of McCreary [3], showing the location of Sects. 10.1–10.6, the ocean boundaries, etc. The shading marks the extent of the wind; the thin line is the longitudinal profile of the wind stress. The wind is almost independent of latitude. The wind stress is everywhere westward (or zero) with a maximum of $(1/2)\,\text{dyn/cm}^2$ in the center of the basin. Redrawn after McCreary [6]. (Note that although the ocean extends from 140° E. to 80° W., 140 degrees *in toto*, only 70 degrees of longitude are shown here)

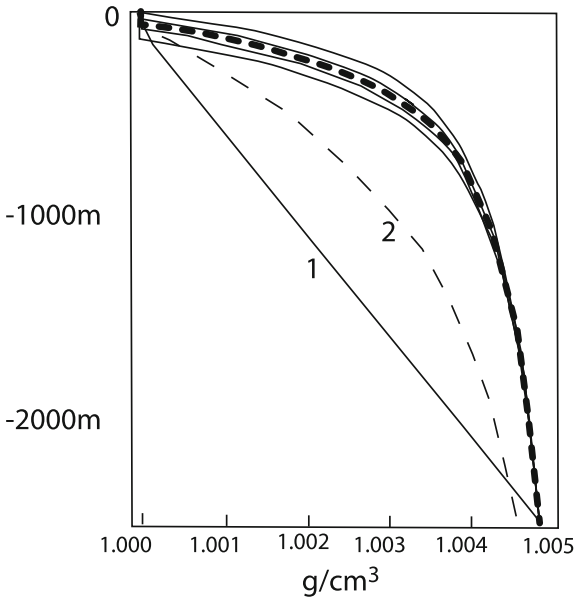

Fig. 10.6 Density profiles in g/cm^3 versus depth in meters for McCreary's model. The *curve* labeled "1" has [unrealistically] a linear gradient at all depths below the mixed layer. The *dashed curve* "2" has a mixed layer in the upper 75 m of the sea where the density $\rho = \rho_0$, a constant; at greater depths, $\rho = \rho_0 + (1/200)\{1 - \exp(-(z + 75m)/1000m)\}$. The five uppermost profiles are more realistic and are employed in most of his calculations. All have $\rho = \rho_0$ in a mixed layer of thickness H and at greater depths $\rho = \rho_0 + (3/1000)\{1 - \exp(-(z + 75m)/200m)\} + (1/500)\{1 - \exp(-(z + 75m)/1000m)\}$ where the mixed layer depth $H = 0, 50, 75, 100$ m and 150 m. The *thick dotted curve*, the middle of this group of five, is McCreary's benchmark. His analytical formulas are cataloged as a reminder that density profiles are always empirical, that is, are freehand sketches disguised as formulas. Redrawn after Fig. 2 of McCreary [6]

$\nu_{min} = 0.55$ cm^2/s for each of the density profiles illustrated in Fig. 10.6. The thicker curve is representative of the Pacific Ocean and is used for the results graphed in the rest of his article.

Despite the approximations, the model Undercurrent is a surprisingly faithful qualitative representative of what is observed. There is an *eastward* equatorial jet with the core just below the main thermocline perhaps 120 m down. [The thermocline in the model, the thick curve in Fig. 10.6, is 75 m under the surface.] There is westward flow, a model South Equatorial Current, at the surface. The jet is strongly confined about the equator with a latitudinal e-folding width of only about 200 km. Like the real thing, the Undercurrent is only O (100 m) thick. The maximum strength of the Undercurrent is about 150 cm/s slightly to the east of the wind stress maximum. The Tahiti-Hawaii shuttle observed a peak Undercurrent of only 90 cm/s, but 153 W. is somewhat west of the wind maximum, so the jet maximum in the model is certainly in the right ballpark. As is also observed, there is westward flow at all depths below the model Undercurrent.

As noted in Sect. 10.1, one of the more useful results of the NORPAX shuttle was to show conclusively that the EUC is geostrophic. The model Undercurrent (and all the other zonal currents) are in geostrophic balance (of the usual $u = -(1/f)\phi_y$ sort) even a slight distance from the equator, and they are consistent with the more general geostrophic relation (10.1) at the equator itself.

In the interior sections (2–4), the meridional circulation is consistent with simple Ekman transport in the sense that there is strong upwelling along the equator and poleward flow near the surface. Because of the strong stratification, however, the return flow towards the equator does not occur in a deep layer, but rather is strongest only 100 m down. The largest north-south velocities — just below the thermocline in the return flow, and *not* in the surface current — are O(10 cm/s). The upwelling just above the core of the Undercurrent is as large as 0.025 cm/s. These maxima are achieved in Sect. 10.3 where the wind is also a maxima; the flows in Sects. 10.2 and 10.4 are similar but noticeably weaker.

There is also a weak downwelling underneath the Undercurrent. Since each meridional cell has a mirror image in the other hemisphere, the overall pattern is four-celled; two in each hemisphere stacked on top of each other.

Outside the region of wind stress, cross-Sects. 10.1 and 10.6, all three velocity components are relatively weak, but the overall pattern persists. In particular, there is an Undercurrent at longitudes with a westward current at the surface, and there is also a four-celled circulation pattern. However, the usual upwelling along the equator is replaced by a weak down-welling in cross-Sect. 10.1 near the western boundary although the pattern of meridional velocity is qualitatively the same as under the wind stress. East of the wind, there is still upwelling at the equator, but the meridional velocity pattern is reversed in sign at all depths.

These changes in meridional circulation are to be expected. The equation of continuity, $w_z = -(u_x + v_y)$, tells us that upwelling and downwelling require horizontal convergence or divergence at some level in the water column. Within the region of wind stress, this convergence occurs because Ekman transport in the surface boundary layer is stripping water away from the equator. Outside the area of the wind patch, convergence still occurs because the zonal currents must deaccelerate to 0 at both the eastern and western boundaries, so u_x will drive meridional cells. Maximum divergence occurs near the eastern edge of the wind patch and maximum convergence near the western edge even though the stress is small near the edges.

The density field is consistent with thermal current balance for the zonal jets, but it also has a number of highly unrealistic features (McCreary [3] provides no graphs). As discussed in Chap. 2, there is necessarily a density inversion at the base of the mixed layer wherever there is downwelling. (The reason is that the perturbation density $\rho = 0$ at the surface and throughout the mixed layer. A downwelling particle therefore has its density jump from 0 to some finite value. One can account for the density jump that occurs when a particle upwells into the mixed layer because the mixing coefficients are ∞ in the mixed layer. There is no mechanism to cause a density jump when a particle's density changes below the thermocline in a region of rather modest diffusion, so the model becomes physically unreasonable.) As his contour plots make painfully clear, there will be one region of static instability off

the equator in the central ocean and everywhere in the western part of the sea. There is another region of instability below the core of the Undercurrent in the central and eastern ocean.

McCreary notes that these regions of static instability coincide with the two layers of strong mixing in the real ocean: the thick surface mixed layer in the west and thermostad in the east. His precise phrase is "It is interesting that …"; he does not make any stronger claim because the model really is screwing up the density field in these regions, and agreement with reality is likely to be more coincidence than cause.

The upper left panel of Fig. 10.7 compares the zonal velocity at points in mid-ocean, right at the equator — for three different cases with wildly different viscosity coefficients. The vertical profile of ν is the same for all three runs, but the magnitude is roughly 10 times stronger than the benchmark case for the curve marked "High" and about 8 times weaker for the graph labeled "Low". As the viscosity weakens, the Undercurrent becomes stronger, closer to the surface, and thinner, but the qualitative shape of the surface westward flow, a thin eastward jet, and slow westward flow in the deep ocean – remain unaltered. The maximum speed of the Undercurrent and its depth and width and so on all vary more slowly than $1/\nu^{1/2}$. The e-folding scale in latitude (not graphed) varies by only a factor of 2 as ν changes by a factor of 75, so it is even less sensitive than the depth profile of the jet.

This slow variation with the diffusion is very comforting because eddy viscosity coefficients and McCreary's artificial way of including them in the model are correct only in order-of-magnitude. On the other hand, the thinness of the jet labeled "Low" shows that the viscosity *is* a parameter of the problem. There are many geophysical phenomena which have well-defined limits as $\nu \to 0$. The Equatorial Undercurrent in McCreary's model is not one of these since it seems to be tending towards a δ-function at the thermocline in the limit of vanishing damping. The viscosity does seem to be essential to the EUC's physics. (Alas!)

The upper right panel of Fig. 10.7 compares the zonal jets at the same points for three different mean density profiles. The dark curve is the benchmark; the graphs labeled "1" and "2" correspond to the two profiles with the same numbers shown in Fig. 10.6. The mixed layer is the same for all three profiles, but the density differs drastically below the thermocline. Profile "1" has a constant Brunt-Vaisala frequency at all depths below the thermocline in contrast to the benchmark and the real ocean for which N^2 decreases rapidly with depth. Nonetheless, we obtain Undercurrents of similar shape and magnitude. Although not visible from the graph, there is no sign that the latitudinal width of the current changes at all.

This is *very* comforting. Since the errors in measurement in density are tiny in comparison to those in the diffusion – the difference between the exact profile and the thick curve used for the benchmark is surely small in comparison to the difference between the benchmark and "1" — we conclude that variations and errors in mean density *below* will have little effect on modeling the Undercurrent. Note that since the viscosity coefficient is inversely proportional to the Brunt- Vaisala, the insensitivity to density in the deep ocean simultaneously implies insensitivity to values of the diffusion in the deep sea. It is only changes in ν in the upper ocean that matter.

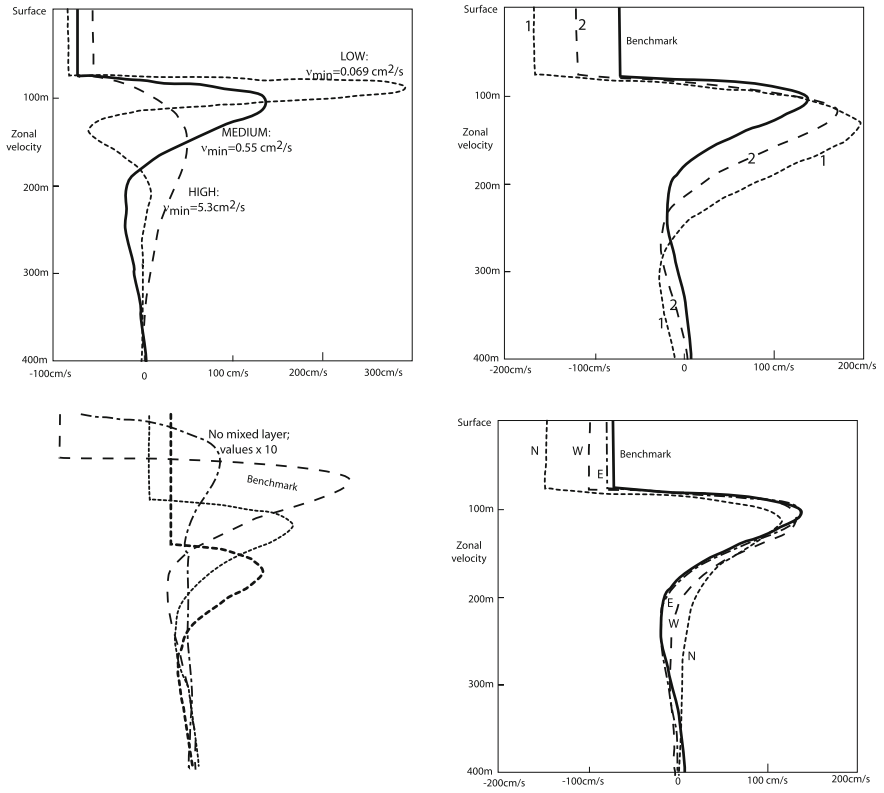

Fig. 10.7 *Upper left* Zonal current at mid-ocean: Effects of varying viscosity. After Fig. 5 of McCreary [3]. *Upper right* Zonal current at mid-ocean: Effects of varying $\rho_B(z)$ *below* the thermocline. The *thick solid curve* is the benchmark. After Fig. 6 of McCreary [3]. *Lower left* Zonal current at mid-ocean: Effects of different mixed layer depths. *Thick curve* benchmark (mixed layer depth = 75 m). The curve for no mixed layer has a surface flow so strong it woudl be off-scale were it not reduced by a factor of 10 as indicated by the label. Figure 7 of McCreary [3]. *Lower right* Zonal current at mid-ocean: Effects of varying boundaries. *Thick curve* benchmark with both eastern and western boundaries. "E": eastern boundary only. "W": western boundary only. "N": *no* boundaries. Figure 8 of McCreary [3]

Unfortunately, we have not yet explored the consequences of variations in the *depth* of the *thermocline*. He compares five runs that are identical except for using a different density profile corresponding to one of the unlabelled graphs of Fig. 10.6. This shows that the depth of the mixed layer is an important parameter. With a shallow thermocline, the Undercurrent becomes stronger and more shallow, but its thickness in depth and its latitudinal width are unchanged. The highly sensitive part of the flow is the surface flow: it increases by a factor of 7 as the thermocline depth decreases from 150 to 50 m. When there is no mixed layer, the westward surface velocity increases to over 14 m/s! Clearly, a mixed layer of at least 50–75 m in depth is necessary to obtain a reasonable surface flow.

Since all the one-layer theories of the Undercurrent depend entirely upon a westward boundary where water was piled up by the wind to create an eastward pressure gradient extending some hundreds of meters deep, it is rather shocking that we obtain pretty much the same Undercurrent *without boundaries* as with. For a baroclinic ocean, the crucial mechanism for driving an Undercurrent is the fact that the wind stress is a patch of finite size.

To be sure, the boundaries are important. They drastically reduce the strength of the westward surface flow, and the westward flow underneath the EUC — the "Equatorial Intermediate Current" as it is usually called — disappears completely in the absence of coasts. This is rather ironic; the westward surface flow was the current that was always easily explained as being directly forced by the wind, but the coasts are far more important in determining its magnitude than they are for the Undercurrent.

Looking carefully at the curves, we can assess the relative importance of the two coasts out in mid-ocean. At all levels, the difference between the solution with an eastern boundary ("E") is two or three times larger than the difference between the benchmark and the flow with only a western boundary. More irony: although it is the western boundary where water is piled up by the wind stress, the eastern boundary is a good deal more important throughout the whole depth of the ocean. The moral seems to be that fairy tales spun of one layer models are no substitutes for summing many baroclinic modes.

We must note one qualification: The preceding figures are all samples in mid-ocean. Boundaries are obviously much more important at the edges of the ocean. However, this does not alter the fundamental conclusion that coastal effects are not responsible for creating the Equatorial Undercurrent.

10.5 The Sensitivity of the Undercurrent to Parameters

As McCreary [3] notes, a description of the balance of forces for the model as a whole is not very illuminating; a better approach is to look at the balance of forces within a given baroclinic mode. We already know that viscosity will be important for the high order modes, but not for the lowest few. We have already seen above that viscosity plays an essential role in the dynamics of the Undercurrent; we cannot obtain an Undercurrent which is as thin as observed without including many high order modes [3].

One possible approach would be to locate the baroclinic mode which is "on the fence" between an inviscid dynamics versus a viscous balance of forces; for McCreary's benchmark, this is $n = 8$. One could then make graphs of the sum of the first 8 modes and then a separate graph of the sum from $n = 9$ to ∞. This in and of itself would not be too helpful, however, because we already know that both partial sums are important. McCreary therefore adopts an alternative approach of taking infinite sums of modes which satisfy one of the lower three balances listed in Table 10.1.

Figure 9 of McCreary [3] shows a pure baroclinic Ekman flow. McCreary defines this to be a flow in which the pressure gradients have been omitted for all *baroclinic* modes. (In classical Ekman theory, of course, the pressure forces are retained, but the approximation is made that these pressure terms are *independent* of depth, i.e., barotropic. However, the barotropic mode is omitted from all the sums presented in this book because it is never equatorially trapped. Consequently, McCreary's definition actually is consistent with classical Ekman theory.)

At first glance, Fig. 9 of his article *looks* okay aside from the complete absence of the Undercurrent (which was, after all, what we were trying to model!), but the latitude scale is *kilometers* rather than meters, and the contour interval is not meters per second, but *kilometers/sec*. The flow in the center of the jet exceeds earth escape velocity!

We can understand why omitting the pressure gradients leads to such hilariously wrong results by writing down the x and y momentum equations for "baroclinic Ekman flow":

$$A\lambda u - fv = F \tag{10.37}$$

$$A\lambda v + fu = 0 \tag{10.38}$$

Solving (10.38) for v and substituting into (10.37) gives

$$\left\{A\lambda + f^2/(A\lambda)\right\} u = F \tag{10.39}$$

or equivalently

$$u = A\lambda F/\left\{A\lambda + f^2/(A\lambda)\right\} \tag{10.40}$$

Now with realistic values of the viscosity, the term $A^2\lambda^2$ is extremely small in comparison to f^2 at all latitudes except within a few hundred meters of the equator; thus, the denominator of the factor in square brackets is approximately $1/[y^2 + \varepsilon^2]$ where $\varepsilon << 1$. Now a representation of the Dirac delta-function is

$$\delta(y) = (1/\pi) \lim_{\varepsilon \to 0} \left\{\varepsilon/[y^2 + \varepsilon^2]\right\} \tag{10.41}$$

so it follows that u as given by (10.41) is proportional to $(1/\varepsilon)$ times this representation of the δ-function for small but finite ε. In the limit of vanishing viscosity, the amplitude of the zonal flow goes to infinity while the width of the jet goes to 0.

McCreary gives a full discussion of Ekman flow, but since our purpose is scientific understanding rather than comedy, it seems appropriate to move on to more realistic cases. The Ekman flow shows quite clearly, however, that baroclinic pressure gradients are indeed essential to modelling the Undercurrent, just as in the simple one-layer models discussed in Chap. 9. The reason that it is possible to obtain an Undercurrent even in the absence of boundaries is that a finite wind patch will create the necessary pressure field.

The next simplified flow, that of Yoshida balance in which the x-derivatives are omitted, reiterates this point about the primacy of the zonal pressure gradient. The Yoshida model retains the pressure gradient in y, but McCreary's Fig. 9 shows that the Undercurrent is still completely absent from a sum of modes that each have no zonal pressure force. The *latitudinal width* of the currents has been greatly improved; the y-scale is now kilometers. However, the velocity scale is still kilometers per second: the flow velocity still exceeds the escape velocity of the earth!

The reason for these absurdly large zonal velocities is the omission of the zonal pressure gradient. Without it, the zonal wind stress can be balanced only by the Coriolis torque, $(-fv)$, and by diffusion. However, the Coriolis force is zero at the equator, so the zonal wind accelerates to a magnitude limited only by the strength of the viscosity: $u(y=0) = F/(A\lambda)$, which is an enormous value for the lowest few modes.

Nonetheless, the Yoshida balance is important because one can show that for sufficiently large values of the vertical mode number n, the x-derivatives do indeed drop out, and the Yoshida balance is a good approximation to the high order baroclinic modes.

It would have been nice to have found a simpler way to show that strong damping makes the x-derivatives important, but this is difficult because in this limit, the zonal wavenumber becomes larger at the same time we are trying to show that the x-derivatives become small.

Nonetheless, note that it is plausible the x-derivatives are important because the dynamics in the presence of strong damping is local. The general solution for forced baroclinic modes is "non-local" in the sense that the response depends upon *integrals* of the wind stress. Thus, one can have a flow even where there is no wind stress because the propagation of Rossby and Kelvin waves can carry the effects of the stress far from the patch of wind. When the mode is strongly damped, however, the waves decay so rapidly that the flow cannot be influenced by distant events, but only by the local stress. The Yoshida flow, which is proportional to $F(x, y)$ itself rather to integrals over x, certainly displays this expected local behavior.

The opposite limit of vanishing damping gives the Sverdrup balance listed in Table 10.1. The corresponding solutions are

$$v = -(1/\beta)F_y \qquad \text{[Sverdrup flow]} \tag{10.42a}$$

$$u = (1/\beta)\int_{\infty}^{x} F_{yy}dx' + \left\{(1/2)\lambda^{1/2}\,\tau\psi_0(\xi)\,Y(y=0)\int_{-\infty}^{\infty} X(x')dx'\right\} \tag{10.42b}$$

$$\phi = \int_{\infty}^{x}\left\{F - yF_y\right\}dx' + \left\{(1/2)\tau\psi_0(\xi)\,Y(y=0)\int_{-\infty}^{\infty} X(x')dx'\right\} \tag{10.42c}$$

The terms in the curly braces are present only for an unbounded ocean. When there is a coast, the expressions in braces are omitted and we have the equatorial Sverdrup solutions (8.39) and (8.42) of Chap. 8. In the absence of land, imposition of a radiation

condition does not exclude the constant of integration (independent of x) that was set to zero in the earlier chapter because we imposed $u(x = X_E) = 0$. However, initial value experiments with a finite wind McCreary [3] show that, via the radiation of Kelvin and Rossby waves, the ocean spins up to Sverdrup flow (10.42a). The spin-up is not uniform in time; distant regions never come to a steady state in finite time since all the flow is being driven by a finite patch of wind, and it takes time for waves to carry the effects of the wind to remote areas where there is no wind.

McCreary's Fig. 11a contrasts the Sverdrup flow for a bounded versus and unbounded ocean. The coastal boundary condition is extremely important *for this mode* [$m = 1$] even though it has a relatively minor effect on the solution as a whole as in the previous section.

No $y - z$ cross-section for a pure Sverdrup flow is shown because it is too boring to need one. With boundaries, the sum over vertical modes can be done explicitly — note that without the terms in braces, the Sverdrup flow is independent of baroclinic mode number m — and one finds that the currents are confined entirely to the surface mixed layer. If the wind has no curl, there are not even currents: baroclinic pressure gradients everywhere balance the stress (again, provided there are coastal boundaries). In an unbounded ocean, the extra terms make it possible to obtain an Undercurrent – in the form of a δ-function just under the thermocline.

The conclusion of looking at these limits is that neither Sverdrup nor Yoshida flow will do: the observed Undercurrent must be a sum of a few modes that are only weakly damped in combination with an infinite sum of higher order modes that are as viscous as if the sea were molasses.

McCreary's Fig. 11b illustrates this transition from Sverdrup to Yoshida flow. The $m = 8$ mode is a hybrid of the two; for McCreary's viscosity, the e-folding scale for the Kelvin wave is exactly equal to the width of the wind patch.

One noteworthy characteristic is the strong equatorial trapping of the $m = 15$ mode. The equatorial Rossby radius, $1/\mathfrak{a}_0$, is 380 km for the $m = 1$ mode, far larger than the observed width of the Undercurrent. For the $m = 15$ mode, the same scale is only 134 km (with McCreary's stratification profile). The narrowness of the Undercurrent is therefore a direct consequence of the narrowness of the jets in the higher baroclinic modes — and the fact that they, too, and not just the lowest mode or modes are important to the Undercurrent.

10.6 Observations of Subsurface Countercurrents (Tsuchiya Jets)

The Subsurface Countercurrents (SSCC's) occur on the flanks of the thermostad, the 150 m-thick layer of water of very uniform temperature (about 13 C.) and salinity (34.9 parts per thousand), which is found in the western Pacific between 150 and 300 m underneath the surface in the band between 5 N. and S. (McPhaden [4]). These jets are often called the "Tsuchiya Jets" in the literature because of Tsuchiya [11].

Their speed is O(20 cm/s) and the transport is 0(5–10 sverdrups). Compared to the 0(1 m/s) flow and 20–50 sverdrup transports in the EUC, the SSCC's are rather weak. [A sverdrup is one million cubic meters/second.] Nonetheless, the SSCC's are stable and well- defined flows and currents similar to them have also been identified in the Atlantic. (They have not been found in the Indian Ocean, but because the dynamics there is dominated by the monsoon, the Indian Ocean is very wierd in comparison to the Atlantic and Pacific.) More important, the SSCC's are subsurface features that can only be explained with baroclinic models, so they furnish a useful test of our understanding of the effect of density stratification on low-latitude dynamics.

In the central and western Pacific, the SSCC's are driven deeper by the downward tilt of the thermocline. This is a reminder that the "Subsurface" adjective in their name is meant quite literally; they are always below the thermocline. In the west, the SSCC's are closer to the equator and contiguous or nearly contiguous with the EUC. In the east, the SSCC's are farther from the equator and are usually separated from the EUC by thin bands of westward flow (extensions of the "Equatorial Intermediate Current", which is the westward flow underneath the EUC) as shown schematically in Fig. 10.1. In the Tahiti-Hawaii Shuttle data, however, (which is mid-ocean, circa 153° W.), the EUC and the SSCC's blend into one another.

10.7 Alternate Methods for Vertical Structure with Viscosity

McPhaden's [4] model of the SSCC's is an analytical sum-over-vertical-modes similar to McCreary's; a fuller discussion is in McPhaden [12]. There are, however, a number of important differences: (i) No boundaries (ii) Constant N^2, ν, and κ [Brunt-Vaisala frequency, diffusion and thermal conduction coefficients] and (iii) Newtonian cooling instead of vertical diffusion. The reason for the first restriction, no boundaries, is that McPhaden's solutions have the form

$$v(x, y, z) = \int_{-\infty}^{\infty} dk \, \exp(ikx) \sum_{m=0}^{\infty} v_n(k, y, z) \tag{10.43}$$

where

$$v_n(k, y, z) = \sum_{j=1}^{6} a_{n,j}(k) \, \mathfrak{z}_{n,j}(z; k) \, \psi_m(\xi) \tag{10.44}$$

where the six vertical structure functions $\mathfrak{z}_j$ are the linearly independent solutions to the sixth order ODE (10.45) below and similarly for the other variables (plus a contribution to u and ϕ from the Kelvin wave). The only way a series like (10.44) can vanish for all y is if the coefficient of each Hermite function is 0. Since the integral over k cannot normally be done in closed form, the condition that each $u_m(x, z) = 0$ at $x = 0$ or $x = X_E$ leads, even if we could somehow ignore the coupling between

different latitudinal modes, to a set of linear equations in $6N_{quad}$ unknowns where N_{quad} is the number of quadrature points used to perform the numerical integration over k. Unfortunately, we cannot ignore the coupling; we must solve a banded system in which all the different modes are coupled, but now each row has $O(24\,N_{quad})$ unknowns instead of just 4.

Worse still, we probably do not have enough degrees of freedom in the vertical to justify setting up such a system, never mind solving it. If, for example, we excited only a single wavenumber k, or if the $v_n(z)$ were somehow independent of k (they are not!), then we would have to choose the 6 numbers a_{mj} so as to make $u_m(0, z)$ vanish for *all* z. This is impossible except in degenerate cases. McCreary's model evades this problem by expanding $u(x, y, z)$ in a complete, infinite set of vertical modes first.

The second restriction, to constant Brunt-Vaisala frequency and constant damping coefficients, is necessary so that the single equation for v – sixth order in z and second order in x and y – will be constant coefficient in z, which is essential for applying the method of separation-of-variables. The vertical structure functions are *not* those of McCreary, which are the same as for an inviscid ocean of the same density structure, but rather are the six linearly independent solutions of

$$\mathfrak{z}_{zzzzzz} + E_1\,[(2m+1)/L + k]\,\mathfrak{z}_{zz} + ikE_2\,\mathfrak{z} = 0 \tag{10.45}$$

where E_1, E_2, and L are constants defined in terms of N^2, K, A, and β. For simplicity, McPhaden imposes the lower boundary conditions as $z = -\infty$ [another difference from McCreary, who used a finite lower boundary] by discarding the three solutions that grow with depth and keeping only the three, one quasi-geostrophic mode and two viscous-dominated modes, which decay with depth.

The third restriction, which is to replace the usual thermal diffusion by a Newtonian cooling, i.e.,

$$(KT_z)_z \quad \rightarrow \quad -(K/D^2)\,T \tag{10.46}$$

where D is a constant with units of depth, is also necessary for separating variables.

In mid-ocean, far from boundaries, McPhaden's model works fairly well, at least for the EUC and SSCC's. It is, however, inherently more limited than McCreary's because we cannot impose coastal boundaries, not even straight ones. Further, it is more difficult to relate the structure of McPhaden's model to the 1–1/2 layer concepts we have been using throughout this work because McCreary's vertical modes are the same as for an inviscid ocean, and McPhaden's are not. For this reason, we shall not derive McPhaden's model in detail.

It must, however, receive at least mention in passing. First, it is successful for some currents and second, it shows that there are alternatives to McCreary's approach for including viscosity in a separable, analytical modes.

10.8 McPhaden's Model of the EUC and SSCC's: Results

McPhaden [4] forced his model using a steady wind stress with a Gaussian shape in latitude and a sinusoidal shape in x with a wavelength of 20,000 km. As noted in the previous section, the model does not (and cannot) have zonal boundaries. The results at the longitude of the maximum zonal wind stress are plotted in Figs. 5 and 10 of McPhaden [4]. Note that because the forcing is symmetric about the equator, u and ϕ must also be symmetric while v and the vorticity ζ are antisymmetric about $y = 0$. Consequently, McPhaden has only plotted the solutions for the northern hemisphere.

The upper left-hand panel shows success! There is an Equatorial Undercurrent (EUC) of about 120 cm/s eastward sandwiched between surface westerly flow and an Equatorial Intermediate Current (EIC) of about 5 cm/s westward. The band of eastward flow is not a single jet of circular cross-section but rather extends off the equator. There is no band of westward flow separating the SSCC's from the EUC as is found in the western Pacific, but that is all right: the graphs are intended to represent mid-ocean where the EUC and the SSCC's blend smoothly into one another. The region of deep eastward flow on the right of (a) is thus a credible model SSCC. It should be noted that a similar tongue of eastward flow can also be found in McCreary's results.

A major flaw of the model, however, is that the SEC, the surface westward flow, is ridiculously large. McPhaden notes that this is a consequence of using a constant stratification and viscosity coefficient. McCreary's model does not fail in this way because he is able to include a surface mixed layer; the strong vertical viscosity of the upper 75 m reduces the SEC to a reasonable magnitude. [12] [unpublished thesis] carried out some experiments with a strong near-surface viscosity (by modifying his model) and found that it did indeed reduce surface currents without affecting the flow in the deeper ocean.

As in McCreary's model, the zonal pressure gradient is essential to the eastward currents; again as in McCreary's model, it is perfectly possible to obtain such a gradient without coastal boundaries provided that the wind patch is finite (or oscillating in x, as here). The model agrees well with the NORPAX Shuttle data where both are negative, but the reversal below the core of the EUC is about 4 times weaker in the model than in the data. Oops!

As expected, the velocities reveal the expected Ekman divergence at the surface, but this is a strong flow confined to the upper 20 m. At the level of the EUC, there is a convergent flow towards the equator which is driven by the zonal pressure gradient off the equator; this reverses at the depth of the SSCC because the pressure gradient reverses, but v is too weak because ϕ_x is.

Why does the pressure field imply convergent flow at the EUC level? His figures show clearly that the *pressure* gradient is very *broad* in latitude whereas the eastward zonal flow is narrow. (The pressure is broad because the wind stress driving it is.) It follows that at a latitude of 3 degrees, for example, the zonal velocity term in the x-momentum equation is very small, but ϕ_x is not. The only term available to balance the pressure is the Coriolis torque on the north-south flow, so the result is

equatorward drift where the pressure gradient is negative and poleward, divergent flow where the pressure gradient is positive.

McPhaden helpfully provides balance-of-forces graphs. It is in fact the normal state of affairs in the ocean for Coriolis force to balance pressure gradient; only at the equator do we get a non-geostrophic zonal flow driven directly by a zonal wind stress.

The zonal flow in his model is in almost perfect geostrophic balance at all levels; at 480 m, the difference between fu and ϕ_y is less than the thickness of the curves. The zonal momentum equation is more complex. At high latitudes (where high means anything larger than 2 degrees) there is a balance between the pressure gradient and the Coriolis torque as already explained. Nearer the equator, the Coriolis force is small and the balance is between the pressure gradient and diffusion of momentum, which reiterates the importance (unfortunately) of the momentum damping for steady zonal flows.

McPhaden offers the vorticity balance to explain the SSCC's. Within two degrees of the equator, he shows that there is a boundary layer in which downward diffusion of cyclonic vorticity balances the poleward advection of planetary vorticity (βv). The divergence term is negligible in this boundary layer because (fw_z) is forced to vanish at the equator itself. Outside this boundary layer, the dominant balance of forces changes: the diffusion is negligible and the advection of planetary vorticity is balanced by vortex stretching.

This high latitude convergence tends to intensify the vertical temperature gradient. Within two degrees of the equator, however, there is a compensating mass divergence with a maximum at about ± 1.5 degrees of latitude. This only reduces the vertical temperature a little, producing a thermostad-like area which is too weak and too narrow ($|y| \leq 2$ degrees versus the observed $|y| \leq 5$ degrees). Nonetheless, thermal current balance requires eastward flow on the flanks of this thermostad, i.e., the SSCC. A contour plot of the vorticity shows that the latitudinal shear of the EUC is the source of the cyclonic vorticity that diffuses downward to drive the divergence.

Because of the symmetry with respect to the equator, this vorticity is zero at the equator itself. Consequently, the strongest divergence and the resulting eastward flow occurs not at the equator but a couple of degrees away from it. Once the SSCC is established, its off-equatorial location implies that the source of vorticity for the deeper flow is considerably further from the equator than for the EUC, so the deep SSCC is displaced further away from the equator than the part of the SSCC which is just below the Undercurrent. Observations illustrated in his article show quite clearly that the SSCC is displaced further and further from the equator as we go deeper. The same trend is very much in evidence in the corresponding mid-ocean zonal velocity contours for McCreary's model.

McPhaden's earlier [12] paper also used an alternative diffusion model in which he obtained separability by adding two z-derivatives to both the momentum and thermal damping, giving a biharmonic (fourth order viscosity) for the momentum equations and an ordinary (second order) diffusion for the heat equation. The result was qualitatively similar to McCreary's [3, 6] model and to McPhaden's [4] diffusion/Newtonian cooling calculation. We can thus end this section on the positive

note that linear calculations give an Undercurrent and contiguous SSCC's for a wide range of conditions. Coastal boundaries are not very important (in mid-ocean!), so McCreary's assumption of straight, north-south running coasts is not an important defect except near land. McCreary and McPhaden were both forced to assume rather artificial forms for the damping in order to separate variables, but since three different artificial forms all give the EUC/SSCC pattern, this is probably not too harmful either.

10.9 A Critique of Linear Models of the Continuously-Stratified, Wind-Driven Ocean

Despite their qualitative success, the McCreary/McPhaden models do have some serious flaws. First, they fail near the coasts; sloping boundaries (McCreary) or any boundaries at all (McPhaden) can only be included at the cost of separability. Nonetheless, it is perfectly possible to numerically extend linear models to realistic coastlines.

The second flaw is that both McCreary and McPhaden are obliged to force their dissipation into a straitjacket. The former has wildly unrealistic density inversions below the thermocline; the latter must employ a quite unrealistic constant Brunt-Vaisala frequency and depth-independent damping. One could choose depth-variable damping and mean density independently, but this generates a non-separable BVP in y and z. Schneider and Lindzen [13] have done this for axisymmetric atmospheric flows, but their method has not been applied to the ocean in a boundary-free, one-zonal-wavenumber formulation like McPhaden's.

Why has no one made these obvious linear extensions? Figure 10.8, which shows results of the three-dimensional numerical model of Philander and Pacanowski [14], is part of the reason: the third major flaw of the analytical models is that they neglect nonlinear effects. It is clear from the figure that the nonlinear terms are numerically

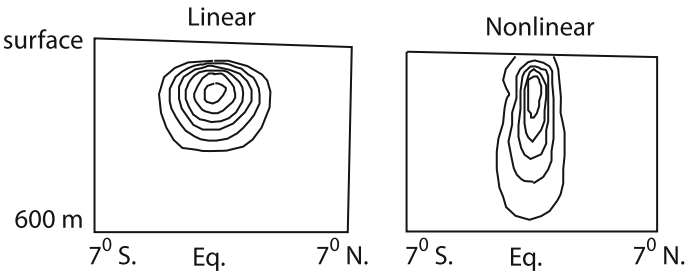

Fig. 10.8 Depth-latitude section at the east-west center of u for a numerical model employing a basin 5000 km wide. The contour interval is 10 cm/s. Only a small part of the computational domain is illustrated. *Left* linear case, instantaneous flow at day 300. *Right* nonlinear integration, average of days 300–350. Sketched based on computations of Philander and Pacanowski [14]

important; the intensity and shape of the nonlinear EUC are strikingly different from those of its linear counterpart. The jet core is much narrower in latitude, but taller in depth. In addition [not shown in our sketch, but clear in the plots of Philander and Pacanowski which show the entire computational domain], the nonlinear SSCC's are separated from the EUC by bands of westward flow as is observed in the eastern Pacific. The linear numerical calculation and the models of both McCreary and McPhaden completely miss these regions of westward flow.

In consequence, the attitude seems to have been that if one had to lose the conceptual power of independent, separable modes, then one might as well move on to general, time-dependent, nonlinear numerical models instead of building equally complicated but much more limited linear boundary value models.

References

1. Busalacchi AJ, OBrien JJ, (1981) Interannual variability of the equatorial Pacific in the 1960s. J Geophys Res-Oceans Atmos 86(NC11):901–907
2. Busalacchi AJ, Takeuchi K, OBrien JJ, (1983) Interannual variability of the equatorial Pacific - revisited. J Geophys Res-Oceans Atmos 88(NC12):7551–7562
3. McCreary JP (1981) A linear stratified ocean model of the equatorial undercurrent. Phil Trans R Soc Lond A 298(1444):603–635
4. McPhaden MJ, Semtner AJ (1984) On the dynamics of equatorial subsurface countercurrents. J Phys Oceanogr 14(7):1216–1225
5. Lukas R, Firing E (1984) The geostrophic balance of the Pacific Equatorial Undercurrent. Deep-Sea Res 31(4):61–66
6. McCreary, Jr JP (1980) Modelling wind-driven ocean circulation. Technical report, Hawaii Institute of Geophysics/JIMAR, Honolulu, School of Ocean and Earth Science and Technology, 64 pp
7. Bjerknes J (1937) Die Theorie der Aussertropischen Zyklonen building. Meteorologische Zeitschrift 54:462–466
8. Hendershott MC (1981) Long waves and ocean tides. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge
9. Schopf PS, Anderson DLT, Smith R (1981) Beta dispersion of low-frequency Rossby waves. Dyn Atmos Oceans 5(3):187–214
10. Leetmaa A, McCreary JP Jr, Moore DW (1981) Equatorial currents: observations and theory. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge, pp 184–234
11. Tsuchiya M (1975) Subsurface countercurrents in eastern equatorial Pacific Ocean. J Mar Res 33(Supplement S):145–175
12. McPhaden MJ (1981) Continuously stratified models of the steady-state equatorial ocean. J Phys Oceangr 11(3):337–354
13. Schneider EK, Lindzen RS (1977) Axially symmetric steady-state modells of basic state for instability and climate studies. 1. linearized calculations. J Atmos Sci 34(2):263–279
14. Philander SGH, Pacanowski RC (1981) Response of equatorial oceans to periodic forcing. J Geophys Res-Oceans 86(NC3):1903–1916
15. McCreary JP (1976) Eastern tropical ocean response to changing wind systems: with application to El Nino. J Phys Oc 6:623–645

Chapter 11
Waves and Beams in the Continuously Stratified Ocean

Abstract The equatorial ocean features deep jets that oscillate rapidly with depth. One possible mechanism for creating these jets is the propagation of wind-driven waves deep into the interior. However, the jump in density at the main thermocline will reflect waves. Deep penetrating waves, on the other hand, will bounce off the bottom. Using ideas borrowed from quantum scattering, vertical beams are analyzed through a mix of theory and number-crunching.

"Ex umbris et imaginibus in veritatem" "From shadows and appearances into truth"
— Tombstone of Blessed John Henry Cardinal Newman

11.1 Introduction

11.1.1 Equatorial Beams: A Theoretical Inevitability

A classic laboratory experiment to illustrate the generation of gravity waves is to oscillate a disk-and-rod in a tank of density-stratified water. The layering is produced by making a dozen bottles of salt water of varying salinity. Each is added to the tank slowly, a steady drip rather than a pour, so that each new, lighter layer does not disturb the denser layers lying placidly below. The resulted stairstep density profile is smoothed by patience, that is, by allowing sufficient time for diffusion of salt to smooth the discontinuous jumps in density. Jiggled up and down at a controllable frequency by an electric motor, the oscillating disk-and-rod are a periodic forcing that generates internal gravity waves. The forcing is resonant, and therefore generates a large amplitude response, when the frequency of the oscillator matches the intrinsic normal mode free oscillation frequency of gravity waves.

The result is that the response is dominated by vertically-and-horizontally-propagating beams shown schematically in Fig. 11.1.

The angle of the beams is controlled by the frequency of the forcing relative to the Brunt–Vaisala frequency. When the frequency is greater than the Brunt-Vaisla frequency, the beams disappear and the response is localized in the neighborhood of

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_11

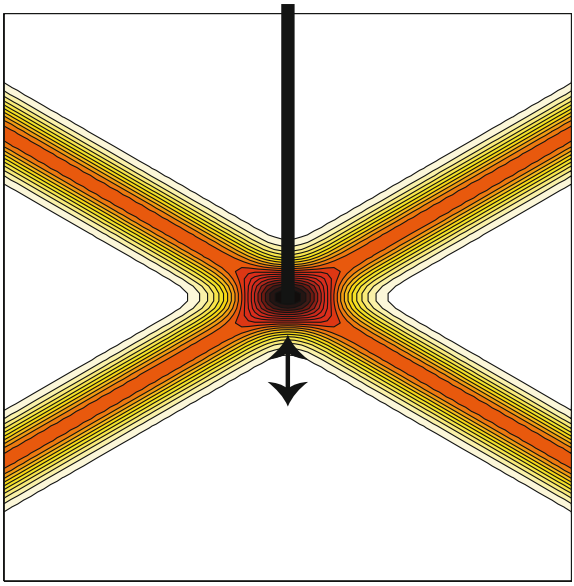

Fig. 11.1 In this side view of a tank of a stratified fluid, a localized forcing, here a disk on the end of an oscillating rod, will generate beams of internal gravity waves. The double-headed *arrow* shows the *up*-and-*down* oscillations of the rod

the disk, decaying exponentially wih distance away from the forcing. Mowbray and Rarity [1] combined a detailed theoretical analysis with careful laboratory experiments that generated half a dozen beautiful photographs of beams. Turner's book reprints some of their photographs as Fig. 2.7 of Plate II and makes some approximations to give a very readable analysis of beams [2].

When the frequency of the gravity waves is $N\cos(\theta)$ where N is the Brunt–Vaisala frequency, the particles osciilate along slanting planes which make an angle θ to the horizontal. The physics is that of a frictionless particle accelerating down a ramp ("inclined plane" in freshman textbooks) which makes an angle θ where $\theta = 0$ is a horizontal ramp and $\theta = \pi/2$ is a vertical ramp. The acceleration is reduced to the gravitational constant g times $\sin(\theta)$ because when the gravitational force is resolved into components parallel and perpendicular to the ramp, only the component parallel to the ramp, $g\ \sin(\theta)$, accelerates the particles. Similarly the frequency of gravity waves can be much smaller than the Brunt–Vaisala frequency if the particles oscillate in nearly horizontal planes.

Seasonal and other periodic wind stress variations will repeat Mowbray and Rarity's lab experiment on a planetary scale. When the forcing is at the surface instead of the deeps, the upward-radiating beams are absent and the "X-beam" is replaced by a pair of beams with the shape of an upside-down letter "v" or an uppercase Greek "Lambda". Mowbray and Rarity and also Turner neglected Coriolis forces. It therefore behooves us to understand beams in the equatorial ocean.

The principal observational focus of this discussion will be the vertically oscillating deep jets or features discovered by Luyten and Swallow [3]. (It is now thought that inertial instability of mean jets is the primary generator of these motions.)

However, theory (and the Mowbray and Rarity experiments) show that the tropical ocean ought to be beamy, and the study of beams is therefore a Theoretical Inevitability.

11.1.2 Slinky Physics and Impedance Mismatch, or How Water Can Be as Reflective as Silvered Glass

A spring in the shape of a helical coil of wire is a popular children's toy known in America as a "Slinky". For a high school audience, Slinkies are splendid illustrations of wave physics. The helical springs support both longitudinal and transverse waves. A shake of the wrist suffices to launch a wave pulse. The wave packets travel slowly enough so as to be easily observed by eye, but not slowly so as inducee sleep. My high school yearbook shows photographs of myself and other honor physics students deploying very long slinkies in the hallways.

Our Slinkies were very long because some were actually two coils of different stiffness joined end to end. The purpose of these double Slinkies was to illustrate that the notion of a subsurface ocean mirror is not crazy.

"Mechanical impedance" is an engineer's term for how much a structure resists motion when a time-periodic force is applied. The impedance of a light spring is low, meaning a small child can easily excite a big wave. A spring made of thick wire, all else the same as for the light spring attached to it, has a large impedance. The boundary between the thin and thick wire is a place of impedance mismatch, a point where the wave propagation properties change.

A wave that is tall on thin wire cannot be tall on thick wire because much more energy is required to make a wave of a given height on the stiff spring than on the thin. Yet the wire cannot tear and energy and momentm must be consered.

The limit of an infinitely stiff spring is easiest to analyze because a very stiff spring does not move at all, and the joining point becomes a rigid boundary condition, a homogeneous Dirichlet boundary condition. The boundary does not move because the incoming wave *reflects* as an outgoing wave of *equal amplitude* but *opposite sign*. The two cancel identically at the interface for all time.

In the opposite limit of a perfect impedance match across the interface, that is, no change in properties of the wave-transmitting material from one region to another, the incoming wave propagates across the interface without change.

The most realistic situation is a finite change of impedance. As shown in the schematic, Fig. 11.2, the incoming wave is partly transmitted and partly reflected.

This wave physics is not restricted to Slinkies. Who has not looked out a window from the interior and seen his image reflected from the glass, superimposed on the exterior view? The air-to-glass transition is also a jump in impedance.

Indeed, rigid boundaries are always unphysical because no boundary is perfectly clamped. Fortunately, the fraction of transmitted wave may be so tiny that the mathematical accuracy of the rigid boundary condition approximation is very high. So-

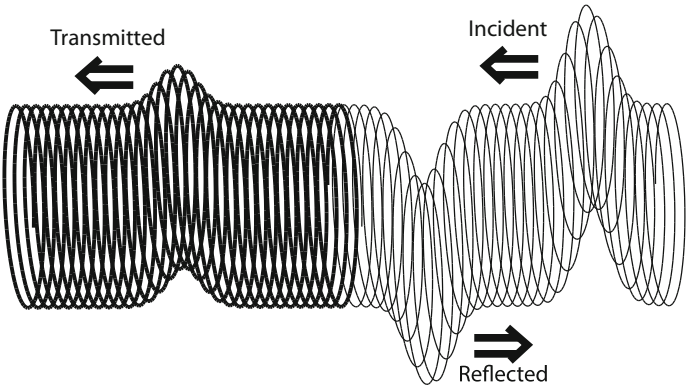

Fig. 11.2 A popular educational aid to illustrate partial reflection and partial transmission consists of two long springs tied together; one of the pair is much stiffer than the other because it is made of much thicker wire. Here, the stiff spring is on the *left*. Shaking the *right* end of the pair of springs ("Slinkies" to American children) generates either transverse or longitudinal waves that propagate *leftward*. (Only transverse waves are shown here.) The "impedance mismatch" between the two springs implies that at the interface where the two springs are tied together, part of the incident wave is reflected as a *right*-moving wave of the opposite sign, and part is transmitted to propagate *leftwards* in the stiff, heavy spring

called "penalty" boundary conditions, which treat an island as a part of the ocean with a very high viscosity, are often very useful for numerical modeling.

Thus, if wave-relevant properties of the medium such as the static stability vary drastically, ocean water may be highly reflective without turning into concrete or milk chocolate.

11.1.3 Shallow Barriers to Downward Beams

Sharp variations in the coefficients of a wave equation can reflect some or most of the energy of an incident wave packet. This implies that the very rapid density variations in the thermocline could stop wind-excited waves as surely as light bouncing off polished silver. If this is in fact what happens, then studies of vertical beams in the abyss are physically meaningless and surface forcing cannot explain the deep internal jets. Gent and Luyten argued that this in fact is true: that the thermocline is a mirror and that vertical beams are a figment of the imagination [4].

The same question was important in meteorology. Early work in atmospheric dynamics and numerical weather forecasting made heavy use of rigid lid boundary

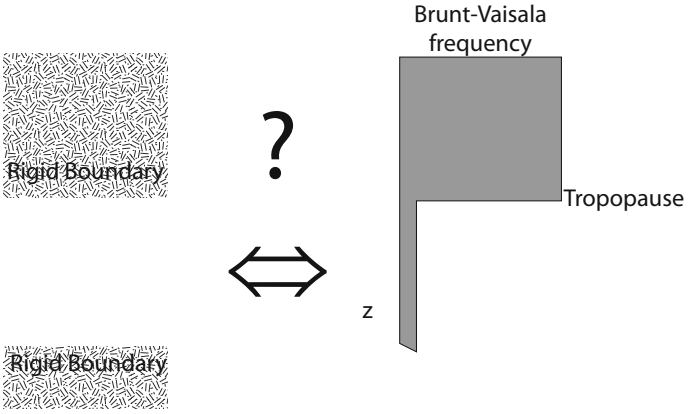

Fig. 11.3 The Eady model is an idealized troposphere, 10–15 km thick [*left*]. The static stability, proportional to the Brunt–Vaisala frequency, jumps at the interface (*right*). The question mark is over the equivalence of the two rather different scenarios: Does the tropopause behave as a mirror, too?

conditions. It is obviously absurd, outside a science fiction novel, to imagine that the tropopause is a rigid layer of glass or plastic.[1]

Nevertheless Erik Eady posed his famous model, the simplest for baroclinic instability, in a troposphere bounded by rigid upper and lower boundaries. Early numerical models, limited to a handful of vertical levels, did the same. The true motive for the rigid upper lid was desperation: the growth rate of Eady modes falls exponentially fast as the rigid roof is (mathematically) moved upwards to infinity; a two-layer model cannot implement a sponge layer nor a radiation condition for a multichromatic spectrum of upward-propagating waves (Fig. 11.3).

Lindzen, Batten, and Kim (1968) showed, however, that the tropopause is as transparent as glass to most waves [5].

11.1.4 Equatorial Methodology

The linear, analytical model of this chapter is the same as that employed in Chap. 9; the major difference is simply that we will allow the complex frequency ω to have a real part, representing a periodic forcing, instead of merely the imaginary part that is due to viscosity and thermal diffusion. The three major technical issues are (i) defining "beam", (ii) the ability of surface-forced waves to propagate through the

[1] In Isaac Asimov's *Foundation* series, the galactic capital, Trantor, has a population of over a trillion and is completely roofed-in. Even the highly imaginative Dr. Asimov would not have placed the roof 10–15 km above ground level!.

spike of huge Brunt–Vaisala frequency at the thermocline and (iii) the structure of narrow beams of waves propagating zonally and vertically through the ocean.

The statement that a localized surface forcing will excite beams just like the Mowbray–Rarity laboratory experiment is a little lie. The experiment generated only Coriolis-free gravity waves whereas the large-scale forcing in the tropical ocean will excite a great diversity of waves. Strong dispersion will rapidly destroy much of the coherence of the response. In contrast, the nondispersive Kelvin and weakly dispersive long Rossby waves will retain their coherence.

Thus, when we talk of vertically propagating beams, we are implicitly focusing on a wave packet in a particular latitudinal mode. The general linear solution will be a superposition of many such mobile wave packets.

A related issue is an error analysis that up to now has heavily relied upon vertical normal mode expansions. A coherent, vertically propagating wave packet can always be decomposed into a series of eigenfunctions of the vertical structure equation, but such expansions will converge slowly for beams and it is difficult to extract much information or insight from the normal mode expansion. We must keep these complications in mind as we follow the rest of the discussion of vertically propagating beams. Kelvin beams will do a "star-turn" in what follows.

Sharp variations in the coefficients of a wave equation can reflect some or most of the energy of an incident wave packet as surely as armor plate. If this is in fact what happens, then it implies that (i) the vertical beam studies of McCreary [6] and Rothstein, Moore and McCreary [7] are physically meaningless and (ii) it is impossible for surface forcing to explain the deep internal jets. However, the reflectivity of the thermocline is sensitive to both the shape and strength of the thermocline and to the vertical wavelength of the waves. Our first order of business will therefore be to compute the reflection coefficient of a model thermocline. In the rest of the chapter, we will then discuss ray paths and vertically propagating beams.

Although the discussion will focus almost entirely on reflection due to stratification, shear layers can also reflect (Grimshaw) [8].

11.2 Alternate Form of the Vertical Structure Equation

The three fields u, v, and ϕ are all proportional to a vertical structure function $\mathfrak{z}(z)$ that is an eigensolution of

$$[\mathfrak{z}_z/N^2(z)]_z + \lambda\,\mathfrak{z} = 0 \tag{11.1a}$$

subject to

$$\mathfrak{z}_z(0) = \mathfrak{z}_z(-D) = 0 \tag{11.1b}$$

Unfortunately, the presence of the first derivative in (11.1a) is rather awkward; it is much easier to calculate and interpret reflection coefficients for an equation without a first derivative. Fortunately, the function $G(z)$ which gives the vertical structure for the vertical velocity w satisfies the simpler equation

$$G_{zz} + \lambda N^2(z)\, G = 0 \tag{11.2a}$$

with

$$G(0) = G(-D) = 0 \tag{11.2b}$$

[Note that as in earlier chapters, we omit the subscript m denoting the baroclinic mode number wherever it is possible to do so without causing confusion. Strictly speaking, λ, $\mathfrak{z}(z)$, and $G(z)$ should all have subscript "m".]

The proof of (11.1b) begins with the integral equation (2.23)

$$\mathfrak{z}_z + \lambda N^2(z) \int_{-D}^{0} \mathfrak{z}(z')\, dz' = 0 \tag{11.3}$$

which is just the vertical integral of (11.1a). Since the vertical structure of w is given by the *integral* of $\mathfrak{z}(z)$ as shown in (2.20), (11.3) becomes

$$\mathfrak{z}_z + \lambda N^2(z)\, w = 0 \tag{11.4}$$

Equation (11.4) immediately justifies the boundary conditions (11.2b). If we now differentiate (11.1a) with respect to z and then use (11.4), we obtain (11.2a).

11.3 The Thermocline as a Mirror

Gent and Luyten (1985) argued that the sharp increase of density in the thermocline layer, which creates a large spike in a graph of $N^2(z)$, is an almost perfect reflector of wave energy except for very short waves which are destroyed by viscosity [4]. Their model has the virtue of great simplicity, but it has some limitations, too. It seems likely that enough energy does leak into the deep ocean to justify the discussion of vertical beams later in the chapter. Nonetheless, the thermocline *does* reflect at least some of the energy of the surface-excited waves, and it is important that we try to get a feel for how this partial barrier alters the wave spectrum.

This issue of reflection-by-changes-in-static-stability first became important in meteorology. Observations clearly show that most of what we think of as "weather" stops at the tropopause, the thin layer where the static stability rapidly increases from the modest value of the troposphere to the significantly larger stability of the stratosphere. Most early weather prediction models therefore stopped at the tropopause, imposing conditions mathematically equivalent to a rigid lid (usually $dp/dt = 0$) either there or at $p = 0$ (for a model using pressure as the vertical coordinate). The confinement-to-the-troposphere of the weather seemed to argue the happy coincidence that the rigid upper boundary condition which was analytically or numerically necessary was also conistent with both observations and wave theory.

Lindzen, Batten, and Kim (1968) showed, however, that the tropopause is as transparent as glass to most waves [5]. To be sure, there is very little *mass* exchange

between the troposphere and stratosphere, but small-amplitude waves transmit energy without transporting mass. Years later, this same controversy was argued for the ocean thermocline.

Gent and Luyten solve (11.2a). This is identical with the Schroedinger equation if we identify $(\lambda N^2(z))$ with $(E - V(z)$ where E is the energy (the eigenvalue) and $V(z)$ is the potential energy, as a quantum mechanics scattering problem. They consider several idealized models of the Brunt–Vaisala frequency, but for our purposes, it will suffice to solve

$$G_{\underline{zz}} + \left\{E - V(\underline{z})\right\} G = 0 \tag{11.5}$$

where

$$E \equiv M_1^2; \qquad V(\underline{z}) \equiv -M_2^2 \operatorname{sech}^2(\underline{z}) \tag{11.6}$$

with

$$\underline{z} \equiv \sigma z; \qquad \lambda N^2(z) = \{E - V\}\, \sigma^2 \tag{11.7}$$

where E, σ, M_1, and M_2 are all constants.

The reason that it is possible to scale out the σ dependence and compute the reflection/transmission coefficients for (11.5) in closed form is that the quantum mechanics scattering problem ignores finite boundaries. Instead, one solves the problem on the interval $\underline{z} \in [-\infty, \infty]$. As $|\underline{z}| \to \infty$, the differential equation reduces to

$$G_{\underline{zz}} + E\, G = 0 \tag{11.8}$$

which, since $E = M_1^2$ is a constant, has the asymptotic solution

$$G = \alpha \exp(i M_1 \underline{z}) + \beta \exp(-i M_1 \underline{z}), \qquad \underline{z} \to \infty \tag{11.9}$$

$$G = \gamma \exp(i M_1 \underline{z}) + \delta \exp(-i M_1 \underline{z}), \qquad \underline{z} \to -\infty \tag{11.10}$$

where the α, β, γ, and δ are constants. Since our interest is how the spike in the thermocline both reflects and transmits downward propagating waves, we demand a downward propagating wave of unit amplitude ($\alpha = 1$) and impose a radiation condition below the thermocline, $\delta = 0$, so that the only waves in the deep ocean are those transmitted downward from the thermocline. Then $|\beta|^2$ is the reflection coefficient and $|\gamma|^2$ is the transmission coefficient (Fig. 11.4).

Numerically, we choose boundaries at $\underline{z} = \pm \Xi$ where Ξ is sufficiently large so that the $\operatorname{sech}^2(\underline{z})$ term is negligible away and impose the conditions

$$G_{\underline{z}} + i\, M_1\, G = \frac{\exp(i M_1\, \Xi)}{(i M_1)} \qquad \underline{z} = \Xi \tag{11.11a}$$

$$G_{\underline{z}} - i\, M_1\, G = 0 \qquad \text{at } \underline{z} = -\Xi \tag{11.11b}$$

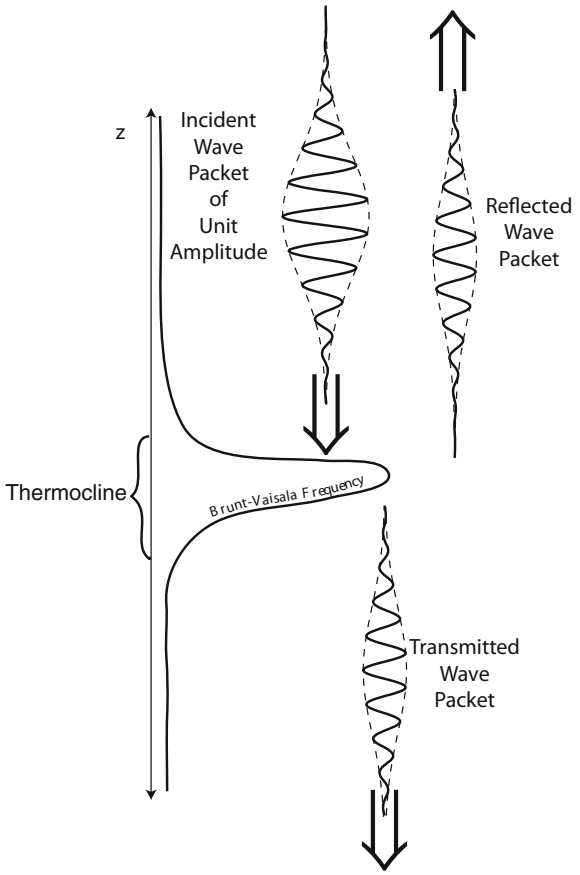

Fig. 11.4 Schematic of the Gent–Luyten Scattering problem. The Brunt–Vaisala frequency N^2 has a big spike at the thermocline where the density increases very rapidly with depth. Note that in the analogous quantum mechanics problem, the positive spike in N^2 is equivalent to scattering off a potential well

The operator on the L. H. S. of (11.11a) annihilates the reflected wave, thereby allowing β to be arbitrary (until we have solved (11.5)), but it forces $\alpha = 1$. The operator in the radiation condition in the deep ocean, (11.11a, b), annihilates the downward-propagating wave, leaving γ arbitrary, but it imposes $\delta = 0$ so that there are no upward-propagating waves beneath the thermocline. Using second order finite differences on (11.5) and one-sided differences on (11.11a) gives a tridiagonal matrix problem which can be solved by band-Gaussian elimination in $O(5N)$ multiplications where N is the number of grid points (Lindzen and Kuo) [9]. A slower but spectrally accurate algorithm is described in Boyd [10].

For the sech^2 potential, it is possible to compute the reflection coefficient in closed form as

$$R = \frac{1 + \cos\left(2\pi\sqrt{M_2^2 + 1/4}\right)}{\cosh(2\pi M_1) + \cos\left(2\pi\sqrt{M_2^2 + 1/4}\right)} \tag{11.12}$$

Because of the cosine factors, R is a rather complicated oscillatory function of both M_1, the strength of the Brunt–Vaisala frequency in the deep ocean, and M_2, which measures the intensity of the spike in N^2 at the thermocline.

This same scattering problem – with the sech-squared profile and the analytic reflection coefficient — was solved by Mied and Dugan [11, 12] for nonrotating gravity waves. They give some excellent perspective plots of the reflection coefficient as a function of both parameters. Another good reference from a quantum mechanics viewpoint is Morse and Feshbach (p. 1657) [13]; the problem was first solved by a physicist named Epstein in 1930 for the scattering of electromagnetic waves. Analytic calculations of R are also possible when the profile is generalized so that the Brunt–Vaisala frequency asymptotes to *different* constants above and below the thermocline.

Fortunately, one simplification of this messy two-parameter reflection coefficient is that when the scale of the wave becomes small in comparison to the width of the Brunt–Vaisala frequency peak, (11.5) can be solved by the WKB approximation. This predicts that the reflection coefficient is *zero* to all orders in the relevant small parameter, which is the ratio of the slow variation of the "potential" to the fast variation of the wave phase (equal to the local wavelength divided by 2π.) (Hyperasymptotic methods show that the tiny reflection is proportional to the exponential of the reciprocal of the small parameter [14–16].) This short wavelength limit is mathematically equivalent to $M_1 >> 1$; (11.12) then shows that we can make a stronger statement. Because of the hyperbolic cosine factor in the denominator of (11.12), the reflection coefficient decreases exponentially as M_1 increases. A more refined hyperasymptotic WKB treatment, which requires locating "turning points" in the complex $\underline{z}$-plane, shows that this is in fact a general result: for any smooth potential, the reflection decreases exponentially fast as the wavelength decreases.

Unfortunately, for small M_1, the WKB approximation cannot be justified. How small must M_1 be before the "potential well" generates a significant amount of scattering? The key physical parameter is the ratio of the *thickness* of the "potential well" to the asymptotic *wavelength*. If we (arbitrarily) define the thickness of the $\text{sech}^2(\underline{z})$ to be the distance between the two depths where the potential has decreased to 1/2 of its maximum value, which is 1.76, then noting that the vertical wavelength is $2\pi/M_1$, we can define the thickness/wavelength parameter

$$q = 0.28\, M_1 \equiv \text{thickness/wavelength} \tag{11.13}$$

Although R itself is a function of both M_1 and M_2, we can put an upper bound on R which is a function *only* of M_1, i.e., of the *vertical wavelength* outside the thermocline zone. Replacing the cosine functions in (11.12) by ± 1, whichever leads to the largest value of R, gives the upper bound:

$$R \le 2/\{\cosh(2\pi M_1) - 1\} \qquad \forall\, M_1,\ M_2 \tag{11.14}$$

The bound is graphed in Fig. 11.5.

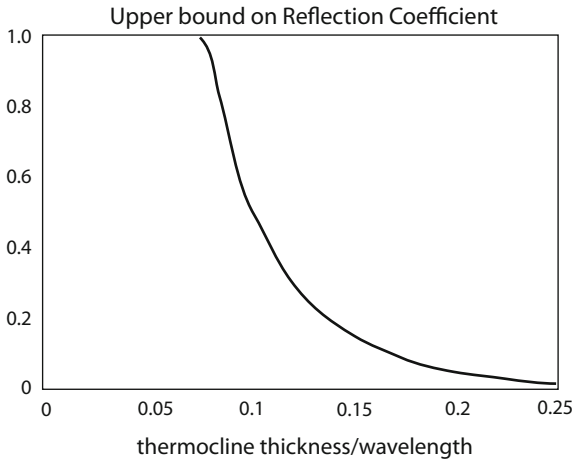

Fig. 11.5 Upper bound on the reflection coefficient R from a sech-squared potential. The reflection coefficient bound is analytically given by $R \leq 2/\{\cosh(2\pi M_1) - 1\}$

We see that even when the thickness of the potential is only 1/10 that of the asymptotic vertical wavelength, no more than half of the incident energy can be reflected. With the parameters Gent and Luyten (1985) employ (an asymptotic Brunt–Vaisala freqency squared of $10^{-5}\mathrm{s}^{-2}$, a thermocline half-maximum-to-half-maximum thickness of 88 m), we conclude that all waves with wavelengths shorter than 800 m — equivalent depths greater than 1.8 cm — will pass freely through the thermocline.

For longer wavelengths, reflection *may* occur, but the reflection coefficient is oscillatory. However, both M_1 and M_2 are implicitly proportional to the equivalent depth so that they go to zero together as the vertical wavelength increases to give

$$R \approx 1.0, \qquad M_1, M_2 << 1 \tag{11.15}$$

Figure 11.6 gives the reflection coefficient as calculated by Gent and Luyten as a function of equivalent depth. Table 11.1 shows the equivalent depths for the first ten baroclinic modes of the ocean along with bound on the reflection coefficient as given by (11.14). Where the bound is larger than 1 (which is impossible), we have replaced the limit given by (11.14) by 1, that is to say no bound at all.

Collectively, the figure and table show that the lowest few modes have such large asymptotic wavelengths that they can be strongly reflected by the thermocline — or not, depending on the precise vertical wavelength and strength of the spike of the thermocline. However, the equivalent depths for near-perfect transmission are in the range of $m = 4$ and higher; the first three modes will be strongly reflected.

We have gone through the sech^2 model at some length to show that the reflection is characteristic of the wave scattering generated by a narrow region of rapid change, even if that change is described by a very smooth analytic function. Figure 11.7 shows that if we replace the idealized profile by an observed profile, prepared by K. O'Neill from Indian Ocean data, we obtain strong reflection even for asymptotic wavelengths as short as 500 m (= equivalent depth of 1 cm = baroclinic mode number

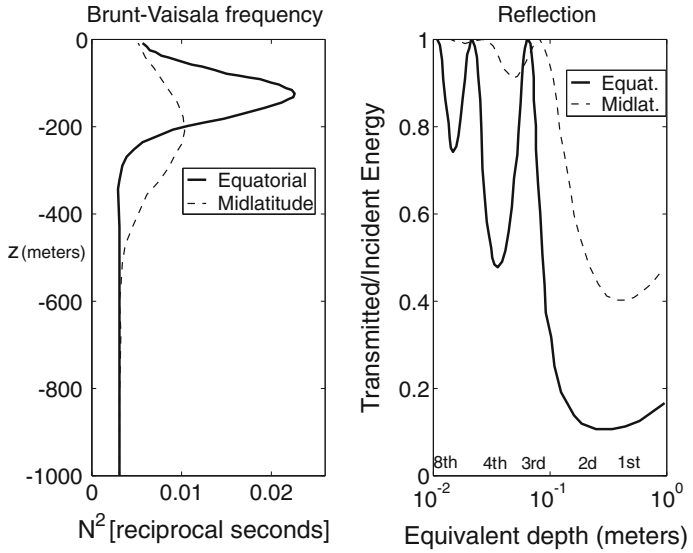

Fig. 11.6 *Left* two profiles of Brunt–Vaisala frequency; the thick solid curve is typical of the tropical ocean while the *thin dashed curve* is representative of the middle latitudes. *Right* plot of the ratio of the transmitted energy to the incident energy as a function of equivalent depth for the equatorial profile (*solid*) and middle latitude $N^2(z)$ (*dashed*). The labels "1st", "2nd", etc., mark the equivalent depths of the lowest few baroclinic modes. Redrawn after Gent and Luyten

Table 11.1 Bounds on reflection coefficients. The equivalent depths for the first four baroclinic modes for the Atlantic are taken from Moore and Philander [17]. The remaining 6 were estimated by assuming that the equivalent depths decrease as l/m^2. The asymptotic vertical wavelengths were computed using $N^2 = 10^{-5}/\mathrm{s}^2$ (far from the thermocline) to facilitate comparison with Gent and Luyten (1985). The bound R_{max} is defined by (11.14). L_v is the vertical wavelength. The parameter q, which is the ratio of thickness of the thermocline to the vertical wavelength, is defined by (11.13)

m	h (cm)	R_{max}	L_V (m)	q
1	60	1.	4900	0.018
2	20	1.	2800	0.031
3	8	1.	1800	0.049
4	4	1.	1250	0.07
5	2.5	0.75	1000	0.088
6	1.7	0.43	820	0.11
7	1.3	0.29	720	0.12
8	1.0	0.19	630	0.14
9	0.8	0.13	560	0.16
10	0.64	0.08	500	0.18

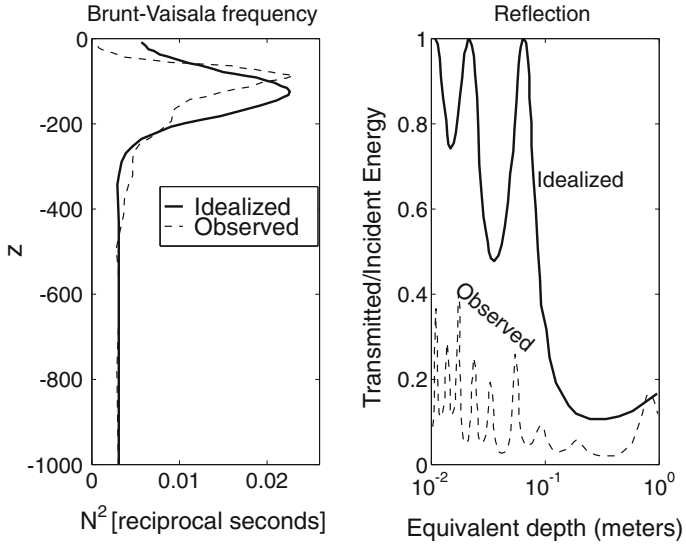

Fig. 11.7 *Left* observed N^2 profile from the Indian Ocean (*dashed*) and the idealized equatorial profile used in the calculation of the previous graph (*solid*). *Right* Transmission coefficient versus the equaivalent depth for the observed Indian Ocean Brunt–Vaisala frequency (*dashed*) and for the equatorial profile of the previous figure (*solid*). After Gent and Luyten

8) because of scattering from the little wiggles and other small scale features with a scale of a few meters superimposed on the main peak with a width of 50–60 m. For comparison, the sech^2 profile and the corresponding transmission coefficient are also shown in Fig. 11.7 as dashed curves. Although the two Brunt–Vaisala frequencies look rather similar to the eye, the transmission coefficients bear not the slightest resemblance to one another. This clearly shows the paramount role of small-scale features in reflecting downward-propagating waves.

Gent and Luyten [4] conclude: "For realistic buoyancy profiles, very little, about 10%, of the surface energy flux reaches the deep equatorial oceans". This has several implications.

First, they assert "the observed, very low frequency, small vertical scale deep jets cannot be explained by linear wave theory either as caused by surface forcing or by instability in the main thermocline". An alternative theory for the Luyten–Swallow features, inertial instability, is discussed in the Chap. 13.

Second, WKB analysis, which is widely used by observationalists to interpret wave data, must be employed with great caution. Standard WKB predicts that a "potential well", i.e., variations in N^2 that never change the sign of N^2, can never cause *any* reflection whatsoever. (As noted above, it is possible to compute the tiny reflection coefficient by applying WKB on paths in the complex plane, but this shows that the reflection coefficient is exponentially small in the thickness/wavelength ratio; in other words, the reflection coefficient always decays exponentially fast as the

wavelength decreases). It follows that for those equivalent depths or vertical wavelengths where in fact strong reflection is occurring, the WKB approximation is horribly wrong. In particular, it overestimates the vertical energy flux into the ocean.

In addition, it also will give very poor fits to wave dispersion curves. The WKB analysis predicts only downward phase propagation; the numerical solution shows that the mode may be more aptly described as a vertical standing wave. The tiny residual phase propagation may badly underestimate the phase velocity of the two component waves, one propagating downward and one upward, which superpose to form the standing wave.

Thus, the Gent–Luyten paper will blow up several major lines of inquiry, if true. But is it right?

11.4 The Mirror-Thermocline Concept: A Critique

The Gent–Luyten model is very simple and certainly raises some important issues, but it does have a number of defects. The first is that it calculates reflection/transmission coefficients on the unbounded interval $z \in [-\infty, \infty]$. The real ocean, however, is bounded, and the boundary is extremely close to the thin layer which is causing the scattering.

As explained in Sect. 10.3, both the numerical and analytical calculations of scattering coefficients assume that the asymptotic solution, far from the "potential well" or thermocline, is a sinusoidal plane wave. However, this is only the *mathematical* posing of the problem. *Physically*, we must imagine that the asymptotic solution is a wave *packet*, that is to say, a plane wave modulated by an amplitude factor $A(z, t)$ that varies *slowly* with position and is exponentially small for sufficiently large distances away from the center of the wave packet:

$$G(z, t) = A(z, t)\, \exp(i N \lambda^{1/2} z - i \omega t) \tag{11.16}$$

The shape of $A(z, t)$ depends on the initial conditions and is arbitrary except for the requirements that (i) $A(z, t)$ varies slowly compared to the "carrier wave" $\exp(i N \lambda^{1/2} z - i \omega t)$ and (ii) is exponentially small for large $|z - c_{group} t|$ so that the wave packet is really a packet (see further discussion in Chap. 17 for envelope solitary waves.)

If $A(z, t)$ varies sufficiently slowly with z, then all reflection/transmission of the packet is determined entirely by the complex exponential in (11.16), and the exact shape of $A(z, t)$ is then totally irrelevant. To put it another way, if we compute the Fourier transform of (11.16), the effect of the "envelope" factor $A(z, t)$ is to smear out the spectrum, which for a simple plane wave is a δ-function spike at $k = N \lambda^{1/2}$, into a peak of finite width and height. However, as $A(z, t)$ becomes broader and broader, the peak in the Fourier spectrum becomes narrower and narrower. If we calculate the transmission/reflection coefficients as outlined in the previous section for each of the components of the peak in vertical wavenumber, we find that if the peak of N^2 is

sufficiently narrow, then R(vertical wavelength) is approximately the same for all of them. However, we can make the Fourier spectrum as narrow as we want by taking the "envelope" function $A(z, t)$ to be sufficiently broad in z. We can then speak of a *single* reflection coefficient for the wave packet as a whole and ignore the envelope function in calculating that reflection coefficient by the method explained above.

This may all seem a rather long-winded digression. However, since no experimental or observational configuration is really infinitely long, the substitution of the wave packet for an infinitely long plane wave is absolutely essential for the conceptual reasonableness of the scattering problem. A wave packet can reflect, but a plane wave extends to infinity in both directions and simply *is*.

However, the thermocline is only about 75 m below the surface of the equatorial sea and the shortest wave in Table 11.1 is 500 m in wavelength. It is not possible to meaningfully insert a wavepacket of such long waves into such a short space between the sea surface and the thermocline. Particle physicists do not have this problem; even a cyclotron a few centimeters in diameter has a scale enormous in comparison to the size of an atomic or nucleonic potential well. Wind-driven waves in the equatorial ocean, however, do not constitute a properly-posed unbounded scattering problem. We really need to consider the effects of the upper boundary.

The first effect of the boundary is that the thermocline variations in N^2 extend all the way to the surface. There is no interval between the surface and thermocline where one can approximate $G(z)$ as a plane wave. The whole concept of a reflection coefficient is rather poorly defined without such a layer where the wave tends to a simple, analytic form.

Even if we ignore this difficulty, we must consider the effects of re-reflections. Since the upper boundary condition is $w = 0$, i.e., a rigid lid, the sea surface is a mirror, a perfect reflector. Consequently, any wave which is reflected by the thermocline will need travel upward no more than a hundred meters before bouncing off the sea surface. If, for example, we pretend that the scattering problem is well- defined and that we calculate a transmission coefficient of only 25%, the remaining 75% of the incident energy will quickly reflect off the top of the ocean. Then 25% of that, or 18.75% of the initial energy, will penetrate the thermocline on the rebound, so to speak. An additional 14% of the initial energy will enter the deep ocean after two reflections from the thermocline and two from the surface. Viscosity will ultimately limit the effectiveness of this repeated reflection process, of course, but after two reflections, more than 57 % of the surface-forced energy has reached the deep sea in spite of a reflection coefficient of only 25 %.

It is worrying enough that the reflection coefficient is not well defined and that it is a significant underestimate of the energy that actually penetrates the deep, but there is still more left out: the forcing that excites the wave in the first place. Implicitly, the Gent–Luyten theory supposes the energy is input directly at the surface so that one need only solve an unforced, homogeneous differential equation with an inhomogeneous upper boundary condition. In reality, the turbulent stresses carry the momentum transferred from the wind throughout the whole depth of the mixed layer. The earliest study of wind-driven equatorial motions, Lighthill (1969), in fact distributed the wind stress as a *body force* acting on the whole of the mixed layer

[18]. This is obviously crude, even though we don't understand the dynamics of the turbulent transfer well enough to do much better, but Lighthill's artifice serves as a metaphor for what is certainly reality: that the forcing extends deep into the mixed layer, and that one should not be solving a homogeneous wave equation until one is quite far down.

The "equatorial beamers", discussed in later sections, are naturally unhappy with the Gent–Luyten paper because it would put them out of business. Moore and collaborators have recalculated the reflection coefficients when the forcing is a body force extending below the surface into the region of the spike in N^2. An extended forcing layer magnifies the reflection coefficients by a factor of 2–5 times compared to a forcing limited to $z = 0$. It is obviously a lot easier for forced waves to penetrate the barrier of the thermocline when the forcing itself, like a Trojan horse, extends halfway into that barrier. The model of Rothstein et al., which we shall discuss below, allows for variable N^2, and they do obtain interesting surface-driven motion in the deep sea. Their model does include both viscosity and the proper boundary conditions, so it has more credibility than the idealized quantum mechanics scattering problem.

Another issue is that it is difficult to justify the statement "deep jets cannot be explained ... by instability in the main thermocline" in the abstract of their paper when they have not in fact solved any sort of a stability problem.

Finally, there is the issue of choosing a proper model thermocline. Gent and Luyten note that a smooth profile like their idealized sech^2 is probably a proper choice for annual forcing while the observed profile analyzed by O'Neill is appropriate for calculating the reflection of higher frequency waves. The reason is simply that wiggles and bumps in the N^2 profile that have a short scale in space also probably have a short persistence in time, too. This clouds the whole issue considerably.

A number of authors have in fact considered the propagation of waves (usually non-rotating gravity waves, for simplicity) through a Brunt–Vaisala frequency profile that consists of a constant plus a fluctuating random component. Examples are McGorman and Mysak (1973), and Tang and Mysak [19, 20] and Mysak's review [21]. Their conclusions, obtained through two rather different methods, would take us too far afield to derive, but the punchline is that the mean wave field decays exponentially as it propagates through the thermocline because of repeated scattering by the small-amplitude, random inhomogeneities in the density. It would be very useful to repeat this sort of calculation for forced planetary waves in the equatorial ocean if one could obtain data on the spectrum of the Brunt–Vaisala fluctuations and allow for the finite boundary effects and the distributed forcing.

The overall conclusion is that the Gent–Luyten paper has asked an important question which equatorial oceanography must continue to address.

11.5 The Zonal Wavenumber Condition for Strong Excitation of a Mode

The explicit mode-by-mode solution for McCreary's model is given in Chap. 10, and will not be repeated here. It is important, however, to recall that all the terms are proportional to integrals of the form

$$I \equiv \int_L^x X(x)\, \exp(-ikx)\, dx \tag{11.17}$$

where $X(x)$ gives the x-dependence of the wind-stress. Strictly speaking, k, L, and I should all have double subscripts to denote the baroclinic and latitudinal mode number, but we shall omit them in this section for the sake of notational simplicity. It is difficult to make general statements about (11.17), but the lower limit L is always $\pm\infty$. A sense of the magnitude of I can be obtained by taking the other limit to be infinite also, which converts (11.17) into the Fourier transform $\underline{X}(k)$ of the wind stress.

A well-known general property of Fourier transforms is the "Heisenberg Uncertainty Principle". This states that if x is the characteristic length scale of $X(x)$ — in our case, it will be the width of the wind patch, assuming that the wind stress is smooth — and Δk is defined to be the width of the Fourier transform $\underline{X}(k)$, i.e., $\underline{X}(k)$ decays rapidly for $|k| \geq \Delta k/2$, then the two widths are related via

$$\Delta x\, \Delta k \sim 2\pi \qquad \text{[Heisenberg Uncertainty Principle]} \tag{11.18}$$

A good discussion of this principle in quantum mechanics (where the wavenumber k is interpreted as the momentum) is given in Merzbacher [22].

This principle implies that high zonal wavenumbers will *not* be excited by a wind patch of width $\Delta x = 5,000$km. In words,

$$|k| = \sqrt{k_r^2 + k_i^2} \leq 2\pi/\Delta x \tag{11.19}$$

is a necessary condition for a mode to strongly respond to the wind.

McCreary provides a graphical interpretation of this condition. His lower axis is k divided by the baroclinic mode dependent scale factor $\mathfrak{a}_0$ so that the thin slanting curves, which show the dispersion relation for various latitudinal modes including both Rossby and Kelvin waves, are correct for all vertical modes. The possibility of nondimensionalizing each baroclinic mode so as to eliminate λ (or E) as an explicit parameter implies that the *shape* of each dispersion curve is independent of the baroclinic mode number n; one curve describes all Kelvin waves provided that we appropriately rescale k and σ for each mode.

The one complication that we cannot evade is that if we drive the motion by forcing at a fixed dimensional period — a year, for example, then the corresponding nondimensional values of σ will be different for each baroclinic mode.

The Kelvin waves of the first 20 baroclinic modes lie between the crosses, but only the first four latitudinal mode 1 Rossby waves and only the lowest one or two baroclinic mode Rossby waves of latitudinal modes 3 and 5 can strongly couple to the wind. For the latitudinal mode number 5 Rossby mode, it is rather silly to talk about "vertical beams" because it is a good approximation to represent the whole of that part of the solution which involves this Rossby wave as being in the first baroclinic mode alone. Consequently, for the high latitudinal mode number Rossby waves, one should talk about "vertical modes" rather than vertical beams. The behavior of these latitudinal modes in a baroclinic model will differ little from that of the corresponding wave in the 1–1/2 layer, which forced all the energy to be in the first baroclinic mode.

For the Kelvin waves, however, it is a rather different story. The 1–1/2 layer model is a much less satisfactory representation of Kelvin waves of annual period because so many waves of this species are strongly excited by the wind, but completely filtered from our earlier model. It is difficult to visualize exactly how so many contributions from so many different modes add up if we stick with the discrete series representation. It is useful to pretend that the distribution of vertical wavenumber is *continuous* and this is the approach taken in later sections; this makes it possible to identify Kelvin wave beams propagating down into the deep ocean.

The topic of equatorial beams is a relatively narrow one in the sense that we will be concerned almost entirely with beams of Kelvin waves and $n = 1$ mode Rossby waves. The higher Rossby waves have very small amplitude in the higher baroclinic modes. The same is true of gravity waves. McCreary's figure shows that beams are also confined to low frequencies. For a forcing with a monthly period, only the gravest Kelvin mode will be strongly forced, and no Rossby waves at all. There is also a spectrum of mixed Rossby-gravity waves of intermediate m which can be strongly excited, but these all have $m > 10$, and thus would be fairly strongly damped by viscosity, and only that part of the zonal wind stress which is antisymmetric about the equator (or the symmetric part of the meridional stress) can force them. Consequently, we shall only be interested in exploring equatorial beams for periods on the order of a year or longer.

Note that we have been implicitly assuming that the wind stress — at least the annual component of the wind stress — is a smooth function so that Δx is the width of the wind stress. If the wind patch is highly oscillatory or has a lot of small scale structure, then Δx in the Uncertainty Principle would have to be taken to be the scale of the variability of the stress, even if the total width of the stress were much larger. Such variations would excite a much broader range of wavenumbers than indicated in McCreary's graph. However, it is also true that low frequency motions generally have much less small-scale spatial structure than high frequency motions. In any event, we lack the data to improve much upon the cosine wind patch.

11.6 Kelvin Beams: Background

The two pioneering articles on equatorial beams are McCreary [6] and Rothstein, Moore, and McCreary [7]. Although the McCreary paper was published first, we shall begin with Rothstein et al. because their model is simpler, and addresses fundamental issues rather than details. Their paper concentrates on Kelvin waves only whereas McCreary also discusses beams of Rossby modes. Their work is further restricted in that they do not include side boundaries, but instead use a model ocean 50,000 km wide (sic!). McCreary describes the bottom and side reflections of beams.

The Rothstein et al. model is almost identical to McCreary's model of the Undercurrent except that (i) the forcing frequency is taken to be ($\sigma = 2\pi/$(one year) instead of 0 (ii) there are no coastal boundaries and (iii) the wind stress is applied as a body force over a mixed layer 50 m thick instead of being applied as a surface boundary condition on $\partial u/\partial z$ and $\partial v/\partial z$. The first change does not alter the mathematics at all; it merely adds a real part to the complex frequency ω that we carried throughout our calculations in Chap. 10 anyway. The alteration in the stress implies that

$$\tau_m = \frac{(\tau_0/H)\ \int_{-H}^{0} \mathfrak{z}_m(z)\, dz}{\int_{-D}^{0} \mathfrak{z}_m(z) dz} \tag{11.20}$$

where H is the depth of the mixed layer. The corresponding coupling constant in McCreary [23] and Chap. 2 replaced the numerator integral (and $1/H$) by $\mathfrak{z}_m(0)$. This avoids Gibbs' phenomenon as discussed earlier; since the stress is now an inhomogeneous term in the Navier-Stokes equations, it is possible to take $u_z = v_z = 0$. In either case, the wind stress provides inhomogeneous terms in the set of shallow water wave equations for each mode, so none of our equations are changed except (11.20) itself.

Let c_{gx} and c_{gz} denote the zonal and vertical group velocities. Standard ray theory then implies that θ_e, the angle that the ray paths make to the horizontal is given by

$$\tan(\theta_e) \ = \ c_{gz}/c_{gx} \tag{11.21}$$

The dispersion relation for Kelvin waves is

$$\sigma = k/\lambda^{1/2} \qquad = k\, N\, /\, \tilde{m} \tag{11.22}$$

where the vertical wavenumber m is given by

$$\tilde{m} = N\, \lambda^{1/2} \tag{11.23}$$

Equation (11.23) is exact for constant Brunt–Vaisala frequency N, but it is also true in the sense of a WKB approximation to the vertical structure Eq. (11.2a) even when

$N^2(z)$ is varying slowly with depth. The zonal and vertical group velocities are given by the k and m derivatives of G. Performing the differentiations gives

$$\tan(\theta_e) \;=\; -k/\tilde{m} = -\,\sigma/N \qquad \text{[beam angle]} \tag{11.24}$$

for waves with *downward* group velocity and *upward* phase velocity. Like ordinary, non-rotating gravity waves, Kelvin waves have vertical group and phase velocities which are of opposite signs. Because Kelvin waves are nondispersive with respect to zonal wavenumber (but not with respect to $\tilde{m}$!), the *zonal* group and phase velocities are equal.

The lines of constant phase in the $x - z$ plane have a slope given by for any waves, regardless of their dispersion relation, by

$$\tan(\theta_p) = -k/\tilde{m} \tag{11.25}$$

Comparing (11.25) with the first form in (11.24), we see that $\theta_e = \theta_p$ for all k and $\tilde{m}$: the lines of *constant phase* are *parallel* to the *beams*. Put another way, this implies that we will not see variations *along* the beam, but only *perpendicular* to it.

This happy coincidence implies that the beams will show up clearly no matter what quantity we choose to graph. The contours of all quantities will run parallel to the beam; large values mark the core of the beam.

It is important to observe that the conclusion $\theta_e = \theta_p$ is true for Rossby waves as well. Since Rossby beams exist only for very low frequencies, we can legitimately make the "long wave" (alias "latitudinal geostrophy", alias "low frequency" approximation). This gives a dispersion relation identical to that for the Kelvin wave except that a numerical factor of $+1$ is replaced by $-1/(2m+1)$. Since this numerical factor does not change the dependence of the frequency on k and m, everything else follows just as for the Kelvin wave. Only the sign and slope of the beam and phase lines are altered:

$$\tan(\theta_e) = \tan(\theta_p) = (2m+1)\,\sigma/N \qquad \begin{array}{c}\text{[Rossby waves of latitudinal mode } m\text{;}\\ \text{downward group velocity]}\end{array} \tag{11.26}$$

Note that since Rossby waves carry energy westward as well as downward, and since the R. H. S. of (11.26) is very small, the actual slopes are approximately given by $\theta_e \approx \pi - |\tan(\theta_e)|$ where the $\tan(\theta_e)$ is used as a shorthand for the R. H. S. of (11.26).

The slopes are extremely shallow: with $N^2 = 0.005/\text{s}^2$, which is about right for the deep ocean, the Kelvin beams sink only 400 m in a distance of 10,000 km, the width of the Pacific Ocean. This is the reason for the artificial 50,000 km ocean of Rothstein et al. (1985): without coastal boundaries to reflect the beams, the Kelvin waves can reach the 2000 m level, just halfway to the bottom, only by traveling a zonal distance greater than the circumference of the earth! Rossby waves travel at steeper angles, but the $m = 1$ Rossby beam still sinks only 1200 m in a zonal journey of 10,000 km. We conclude that reflections from coasts are extremely important, but for the moment will postpone a discussion of their role.

11.7 Equatorial Kelvin Beams: Results

Figure 11.8 shows the Kelvin beam for the benchmark case: the mixed layer depth (which is always also the depth of the "body force" exerted by the wind on the water) is 50 m, the Brunt–Vaisala frequency is a constant 0.005/s, and there is no viscosity. Because the beam slope is so small, the plot is 25,000 km wide in x — double the width of the real Pacific. The real part of the complex-valued zonal energy flux is graphed. Figure 11.9 illustrates the contributions of each mode to the vertically integrated zonal energy flux.

The beam is clear and well-defined, but its latitudinal extent L is wider than that of the wind patch, Δx. Figure 11.10 is a schematic which helps to explain why. Since the wind stress is applied as a body force throughout the whole mixed layer, all of the shaded rectangle in the diagram serves as a source of waves. In addition, a ray which is emitted *up* from the lower right hand-corner of the box, as denoted by the arrow, will reflect off the sea-surface and generate a downward-propagating ray that marks the eastern limit of the beam. We must always remember that while the zonal group velocity of a Kelvin wave is always eastward, a forcing source within the ocean will generate Kelvin waves with *vertical* group velocities of both signs. The diagram thus predicts

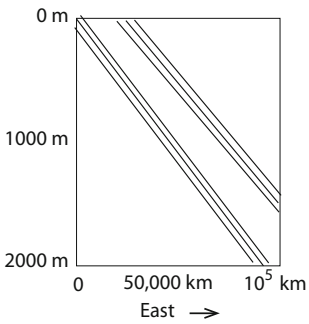

Fig. 11.8 Kelvin beam along the equator. The drawing is a schematic, but the slope is about *right* for an ocean of constant $N^2 = 0.000025\text{s}^{-2}$. Contours of $\Re(u\phi^*)$

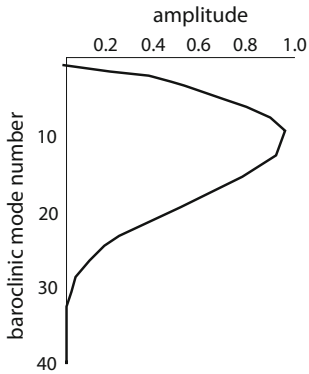

Fig. 11.9 Amplitude of baroclinic modes for Kelvin beams in an ocean of constant $N^2 = 0.000025\,\text{s}^{-2}$

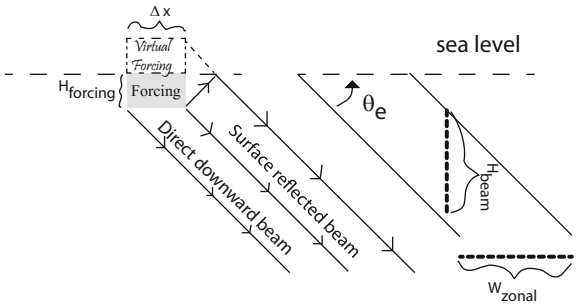

Fig. 11.10 Schematic of the geometry of a beam. H is the depth of the mixed layer; since the forcing is a body force over the entire depth of the mixed layer, this is also the depth of the forcing reigon. Δx is the width of the wind patch. The forcing region, a rectangle $\Delta x \times H$, is shaded. w_{zonal} is the longitudinal width of the beam, which is wider than the wind patch for reasons explained in the text. H_{beam} is the vertical thickness of the beam. θ_e is the angle of the beam with respect to the horizontal

$$w_{zonal} = \Delta x + 2H \arctan(\theta_e) \approx \Delta x + 2HN/\sigma \tag{11.27}$$

$$H_{beam} = w_{zonal} \tan(\theta_e) \approx \Delta\sigma/N + 2H \tag{11.28}$$

where w_{zonal} is the east-west wind of the beam. Thus, the beam width is sensitive to the width of the mixed layer as well as to that of the wind patch. The apparent narrowness of the beam on the unrealistically large scale of Fig. 11.8 hides the fact that beam is more than 5,000 km across, wider than the whole of the continental United States.

Figure 11.9 illustrates the amplitudes (in terms of zonal energy flux) for the vertical modes. The spectrum is dominated by intermediate mode numbers with a maximum around $m = 10$ and an exponential fall-off for higher m with almost no contributions from $m > 35$. The reason why intermediate m are important is that for an annual period, σ/N is so small that even with a broad wind patch, the vertical extent of the beam T is only a couple of hundred meters. This is small in comparison to the depth of the ocean, so many vertical modes are necessary to represent the narrow beam. It is not obvious, however, why the $m = 1$ mode has a very small amplitude in comparison to $m = 10$ as is true of most Fourier expansions, even for narrow functions. This is what the graphs show, however.

Increasing the depth of the mixed layer broadens the beam (i.e., increases both w_{zonal} and H_{beam}) and the peak in the spectrum shifts to smaller m. Halving the mixed layer depth to 25 m broadens the spectrum and also shifts the peak to larger m (around $m = 12$). The effects are not dramatic, however; w_{zonal} and H_{beam} and the beam spectrum are all primarily determined by the width of the wind patch Δx with a weaker sensitivity on H.

The viscosity is a far more significant parameter. Figure 11.11 is a summary of many cases analyzed in detail by Rothsstein, Moore and McCreary, to which we

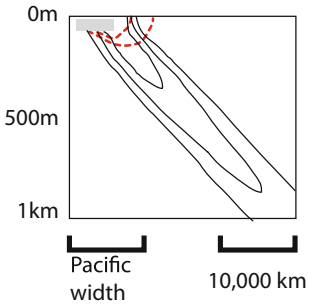

Fig. 11.11 Schematic of the effects of strong viscosity on Kelvin beams. *Black contours* small or zero viscosity. *Red dashed* same isolines with large viscosity

leave the details. Note that strongly viscous beams are localized near the surface. Dissipation, as well as impedance matching, can confine equatorial waves to a shallow surface layer. Philander and Pacanowski [24] used large viscosity and provide further illlustrations of surface-localized viscous beams.

Note also the shallowness of the beam. Although the domain is three times wider than the Pacific Ocean, the graph shows only the upper quarter of the ocean. It is likely real beams bounce off continental sidewalls before striking the bottom except in shallow parts of the oceans. This shallow angle of descent is a strike against simple beam theories. More realistic theories must incorporate sidewall reflections, more accurate models of wind stress forcing, and so on. Dynamic oceanographers must tolerate a messiness, an untidiness, that is not present in the crystallinity of general relativity or particle physics, which are very hard, but untroubled by the fractal unpredictability and microstructure of the ocean edges.

References

1. Mowbray DE, Rarity BSH (1967) A theoretical and experimental investigation of the phase configration of internal waves of small aamplitude in a density stratified liquid. J Fluid Mech 28:1–16
2. Turner JS (1973) Buoyancy effects in rotating fluids. Cambridge University Press, Cambridge
3. Luyten JR, Swallow JC (1976) Equatorial undercurrents. Deep Sea Res 23:999–1001
4. Gent PR, Luyten JR (1985) How much energy propagates vertically in the equatorial oceans. J Phys Ocean 15(7):997–1007
5. Lindzen RS, Batten ES, Kim JW (1968) Oscillations in atmospheres with tops. Mon Weather Rev 96(3):133–140
6. McCreary JP (1984) Equatorial beams. J Marine Res 42(2):395–430
7. Rothstein LM, Moore DW, McCreary JP (1985) Interior reflections of a periodically forced equatorial [Kelvin] wave. J Phys Ocean 15(7):985–996
8. Grimshaw R (1976) Reflection of internal gravity-waves from a shear-layer. Quart J Mech Appl Math 29(Nov):511–525
9. Lindzen RS, Kuo HL (1969) A reliable method for the numerical integration of a large class of ordinary and partial differential equations. Mon Weather Rev 97:732–734
10. Boyd JP (1990) A Chebyshev/radiation function pseudospectral method for wave scattering. Comput Phys 4:83–85

11. Mied RP, Dugan JP (1974) Internal gravity wave reflection by a layered density anomaly. J Phys Ocean 4(3):493–498
12. Mied RP, Dugan JP (1976) Internal wave reflection from a sinusoidally corugated surface. J Fluid Mech 76:259–272
13. Morse PM, Feshbach H (2005) Methods of theoretical physics, 2nd edn. Feshbach Publishing, Wellesley, Massachusetts, 2000 pp, (in two volumes), reissue of book published in 1953
14. Boyd JP (1998b) weakly nonlocal solitary waves and beyond- all-orders asymptotics: generalized solitons and hyperasymptotic perturbation theory, mathematics and its applications, vol 442. Kluwer, Amsterdam
15. Boyd JP (1999) The devil's invention: Asymptotics, superasymptotics and hyperasymptotics. Acta Appl 56(1):1–98
16. Boyd JP (2005a) Hyperasymptotics and the linear boundary layer problem: why asymptotic series diverge. SIAM Rev 47(3):553–575
17. Moore DW, Philander SGH (1977) Modelling of the tropical oceanic circulation. In: Goldberg ED (ed) The Sea, no. 6 in The Sea, Wiley, New York, p 319–361
18. Lighthill MJ (1969) Dynamic response of the Indian Ocean to onset of the S. W. monsoon. Phil Trans 265:45–92
19. McGorman RE, Mysak LA (1973) Internal waves in a randomly stratified fluid. Geophys Fluid Dyn 4(3):243–266
20. Tang CL, Mysak LA (1976) Note on internal waves in a randomly stratified fluid. J Phys Ocean 6(2):243–245
21. Mysak LA (1978) Wave-propagation in random media, with oceanic applications. Revs Geophys 16(2):233–261
22. Merzbacher E (1970) Quantum Mechanics, 2nd edn. Wiley, New York, p 400
23. McCreary, Jr JP (1980) Modelling wind-driven ocean circulation. Tech. rep., Hawaii Institute of Geophysics/JIMAR, Honolulu, School of Ocean and Earth Science and Technology, 64 pp
24. Philander SGH, Pacanowski RC (1981) Response of equatorial oceans to periodic forcing. J Geophys Res-Oceans 86(NC3):1903–1916

Chapter 12
Stable Linearized Waves in a Shear Flow

Abstract Instability is not the only trick of waves interacting with mean flows. Equatorial waves may be refracted by the currents and absorbed at critical surfaces where the phase speed matches the mean current. Atmospheric sciences pioneered such studies in the 60s; here, similar perturbative and arithmurgical methods are deployed in the ocean. Linearized about a mean flow that varies with both latitude and depth, the dynamics can be reduced to a single second order partial differential equation of mixed elliptic-hyperbolic type. This can be solved numerically by finite differences accompanied by sparse direct elimination methods. (Iterative methods are unreliable because the equation is of mixed elliptic-hyperbolic type.) The numerical computations are credible if care is taken near the critical surfaces where the mean flow matches the phase speed of the waves. The same wave-in-bivariate-mean-shear problems can also be attacked analytically by multiple scales perturbation theory. The horrendous algebra usual to this approach is enormously reduced by invocation of wave-mean flow conservations laws.

> Because philosophy arises from awe, a philosopher is bound in his way to be a lover of myths and poetic fables. Poets and philosophers alike being big with wonder.
>
> St. Thomas Acquinas (1225-1274)

12.1 Introduction

The tropical ocean is riddled with jets: the Equatorial Undercurrent, the South Equatorial Current, the North Equatorial Counter-Current and so on. To a first approximation, these currents are "mean" in the sense of being independent of both time and longitude. To understand the various instabilities of these currents and also to understand how the currents distort neutrally-stable waves, the essential first step is to solve the equations of motion as linearized about a mean flow.

These linearized studies of tropical waves in the ocean and atmosphere fall into five main categories:

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_12

1. Multiple scales perturbation theory for flows with vertical shear
2. Numerical studies of neutral waves in two-dimensional shear
3. Instability of latitudinal shear: analytical and numerical studies
4. Instability of flows varying in both latitude and depth: numerical only

A crucial point is that much of "shear effect physics" for Rossby waves carries over directly from the middle latitudes to the tropics. The same is true for gravity waves; however, because of their higher phase speeds, mean current effects on gravity waves are much more limited than for Rossby waves. The Kelvin wave will dominate our discussion because this mode has no midlatitude counterpart.

Linearized inviscid wave equations in fluids are usually *singular* whenever the phase speed matches the zonally-averaged mean flow. This motivates the following:

Definition 12.1 (*Critical Latitude/Level/Surface*) A spatial surface where the zonal mean longitudinal velocity $U(y, z)$ and the phase speed match,

$$U(y, z) = c \tag{12.1}$$

is a "critical surface". If the mean current is a function $U(y)$ of latitude only, the critical surface is usually labeled the "critical latitude"; similarly, if U is a function only of height or depth, $U(z_c) = c$ defines the "critical level" z_c.

If the amplitude of the wave is sufficiently small compared to the viscosity or unsteadiness of the wave (Dickinson [1]), then the critical latitude is *linear* and a Rossby wave impinging on such a critical latitude is *completely absorbed*. The wave amplitude rises steeply as the critical latitude approached, so if damping and transience are weak, the critical latitude may be dominated by *nonlinearity*; the wave is then completely *reflected* by a nonlinear critical latitude, level or surface.

This wave physics developed for Rossby waves on the midlatitude beta-plane carries over to the equatorial beta-plane without change.

In this chapter, we shall begin by discussing the eigenvalue problem when the mean currents are idealized as functions of latitude only.

12.2 $U(y)$: Pure Latitudinal Shear

Our model is the 1-1/2-layer model, that is, the nonlinear shallow water wave equations in the equatorial beta-plane approximation. Denoting the zonal wavenumber by k and the phase speed by c as before, the equations are linearized about a zonal mean east-west current $U(y)$ and a mean depth field $1 + \Phi(y)$:

$$ik(U - c)u + (U_y - y)v + ik\phi = 0 \tag{12.2}$$

$$ik(U - c)v + yu + \phi_y = 0 \tag{12.3}$$

$$ik(U-c)\phi + iku + v_y + \aleph\left\{ik\Phi(y))u + \Phi(y)v_y\right\} - \beth\, yU\, v = 0 \qquad (12.4)$$

where the mean flow and height field are in geostrophic balance: $\Phi_y = -yU(y)$ and where subscript y denotes differentiation with respect to latitude. In the literature, this system has actually been applied in three flavors. The full shallow water model with all its conservation laws is obtained by setting the flags $\aleph = \beth = 1$ (Ripa and Marinone [2]). A similar set of equations can be derived by assuming U is a function of latitude only and then separating variables in the linearized equations for a continuously stratified atmosphere: the "separation-of-variables" set used by Boyd [3, 4] and Boyd and Christidis [5–7]. Other authors such as Bennett and Young [8] and Zhang and Webster [9] have solved the intermediate system that depends only on $U(y)$ and not also on $\Phi(y)$ by setting $\aleph = 0$.

To solve these equations by a shooting method, it is convenient to reduce them to two equations in two unknowns by solving the x-momentum equation for u in terms of the other two variables. The equation for ϕ_y is derived from the y-momentum equation while that for v_y is just the equation of continuity with u replaced by its expression in terms of v and ϕ. Let us introduce

$$w = k(c - U(y)) \qquad (12.5)$$

Also replace U_y by $\Gamma(y)$. One obtains

$$v_y = \left\{k\,\frac{y-\Gamma}{w(y)} + \beth\,\frac{yU}{1+\aleph\Phi}\right\} v + \left\{i\,\frac{w(y)}{1+\aleph\Phi} - ik\,\frac{k}{w(y)}\right\}\phi \qquad (12.6)$$

$$\phi_y = \left\{iw(y) - i\,\frac{y(y-\Gamma)}{w(y)}\right\} v - y\,\frac{k}{w(y)}\,\phi \qquad (12.7)$$

In turn, this system can be reduced to a single equation in either ϕ or v. To obtain the equation for v, one can solve (12.6) for ϕ and substitute the result into (12.7) to obtain a second order equation in v only. Unfortunately, the resulting differential equation for nonzero $\aleph$ and $\beth$ and the similar equation in ϕ only are horribly complicated. This explains why simplified models with either or both of $\aleph$ and $\beth$ equal to zero have been more widely used than the full shallow water equations. Because of their complexity, we shall omit these one-unknown equations, but see Boyd [3] for the $\aleph = \beth = 0$ case.

When the mean shear is weak — this always *excludes critical latitudes* — regular perturbation series, that is, power series in some measure of the shear strength, have been usefully applied by Boyd [3, 4], McPhaden and Knox [10] and Greatbatch [11].

Dickinson [12] analyzed linear critical latitudes for midlatitude waves and obtained solutions in shear flows in terms of hypergeometric functions and Whittaker functions. Obtaining explicit solutions in terms of simpler transcendentals has proved impossible.

WKB methods with asymptotic matching to local solutions that are Bessel functions of order one in the neighborhood of the critical latitude were very successful for midlatitude Rossby waves (Dickinson and Boyd [12, 13]). Unfortunately, efforts to extend this to the Kelvin wave have so far failed.

12.3 Neutral Waves in Flow Varying with Both Latitude and Height: Numerical Studies

When the mean zonal wind is allowed to be a function of both latitude and height, it is still possible to reduce the system of linearized equations to a single partial differential equation for the eddy geopotential Φ'. There are two complications.

The first is that the ensuing second order partial differential equation has excruciatingly complicated coefficients, imposing serious perils of coding and/or algebra-in-the-derivation errors. This complexity has motivated various authors to make multiple *ad hoc* approximations. For example, Holton [14] solved

$$\frac{y^2-w^2}{S}\left\{\Phi'_{zz}-\Phi-\sigma\Phi'_z-\frac{\sigma}{2}\Phi'\right\}+(y^2-w^2)\left\{\frac{(1-y^2)}{y^2-w^2}\Phi'_y\right\}_y$$
$$+\left\{\frac{k}{w}\frac{y^2+w^2}{y^2-w^2}-\frac{k^2}{1-y^2}\right\}\Phi'=-i\frac{y^2-w^2}{w}\{(Q/S)_z-(Q/S)\} \quad (12.8)$$

where Φ' is the wave geopotential and $S(z)$ is the static stability, $Q(y,z)$ is convective heating and $\hat{\omega}(y,z)=\omega-kU(y,z)$. Even though he assume that the mean wind has the restrictive form $U(y,z)=\tilde{U}(z)(1-y^2)$ and furthermore neglects U_y compared to y, he still obtains a very messy two-dimensional PDE which is of mixed elliptic-hyperbolic type.

Similarly, McPhaden, Proehl and Rothstein [15] apply the long wave (meridional geostrophy) approximation. The resulting PDE f or the pressure is still so ghastly that they write it only in symbolic form as $Ap_{yy}+Bp_{yz}+Cp_{zz}+Dp_y+Ep_z+F=0$ and then display the disheartening equations for the coefficients, some so lengthy as to be split over multiple lines.

Schoeberl and Geller [16, 17] derived a quasi-geostrophic wave equation simpler than those of Holton and of McPhaden, Proehl and Rothstein. Even so, they were forced to publish a corrigendum [18] after an error was discovered in their waves-in-shear code.

The numerical solution of the nonseparable two-dimensional wave equation is straightforward. Centered, second order finite differences yield a block tridiagonal matrix problem. If the grid is a tensor product of M points in latitude and N in depth, the $MN\times MN$ matrix has the structure of an N-dimensional tridiagonal matrix — all elements H_{ij} are zero except for the diagonal and the first superdiagonal and first subdiagonals, H_{jj}, $H_{j,j+1}$ and $H_{j+1,j}$—except that the tridiagonal elements are $M\times M$ matrices. The usual fast recurrences for solving a tridiagonal system still

apply with division interpreted as solving an M-dimensional matrix problem. The cost is roughly $O((2/3)NM^3)$ operations nd $O(NM^2)$ memory storage. Lindzen and Kuo [19] give a good description. Boyd [20] observes that since "fill-in" occurs, there is no penalty for using a *high order* discretization in *one* coordinate.

A modern reader, at least with a good background in numerical analysis, is likely to snort with disgust and throw the book across the room at this point. Direct methods like block Gaussian elimination are very old-fashioned and an order-of-magnitude slower compared to modern iterative methods. Indeed, multigrid yields a solution for the Poisson equation on an $M \times N$ grid is just $O(MN)$ operations where the proportionality constant is $O(10^2)$. Unfortunately, the two-dimensional waves-in-shear PDE is generally of *mixed elliptic-hyperbolic type*. As explained in [21], this wrecks most iteration schemes.

Today, it is not necessary to code a block-tridiagonal solver. Sparse matrix solvers are available in most software libraries.

It is common to impose a radiation boundary condition: waves propagate freely upward to be dissipated off the grid in the thermosphere [atmospheric models] or downward into the abyssal ocean to be dissipated before completing a round trip to the surface [ocean]. Matsuno [22] smoothly tapered his mean wind to a separable form, solved the one-dimensional eigenproblem, projected his finite difference solution onto the discrete eigenfunctions, and imposed the analytical radiation condition of upward propagation only. Similar tricks are now in widespread use as the "perfectly matched layer" strategy.

It is much simpler and equally effective to impose a "sponge layer". If the physical ocean is of depth $z = -H$, one can extend the model to depth $z = -(H + L)$ and add a rapidly increasing artificial dissipation in the unphysical regime. As long as the scale of the wave, which is roughly the wavelength W divided by 2π, is small compared to the spatial scale of variations in the damping, the spurious reflection induced by the dissipation will be exponentially small in the ratio of scales, and the sponge layer will almost completely absorb the wave.

Durran, who devotes the entirety of his 44-page Chap. 9 to "nonreflecting boundary conditions", is a comprehensive discussion [23].

The remaining numerical complication is also both an essential part of the physics and a numerical obstacle: the wave equation is all its various forms is *singular* at the critical surface where $U(y, z) = 0$. Finite differences and spectral methods and so on explicitly assume that the solution is smooth, but the flow is logarithmically singular at the critical surface. When the mean shear is a function of latitude only, Dickinson showed that the singularity is of the form $y \log(y)$.

When $U(y)$ allows reduction to an ordinary differential equation, it is easy to calculate the phase speed c by detouring the path of integration into the complex plane [6, 24–26]. However, this not helpful for boundary value problems. A more general strategy is add a little bit of damping to smear out the singularity into a thin internal high-gradient layer of finite thickness and then apply a change of coordinate that concentrates grid points in the vicinity of the near-singularity (Boyd [27]).

Modifying the momentum equations with a simple undifferentiated term such that

$$u_t = \ldots \quad \Rightarrow \quad u_t = \ldots - \alpha_{damping} u, \tag{12.9}$$

and similarly for the other equations of motion, is called a "Rayleigh friction" in the momentum equations and "Newtonian cooling" in the temperature equation. This is preferable to adding viscous terms because viscosity *raises the order* of the PDE whereas Rayleigh friction and Newtonian cooling do not.

Some examples of equatorial waves with critical latitudes in a flow $U(y)$ are given in [27]. For Rossby waves, eddy fluxes are towards the critical latitude from the side where $U(y) - c$ is positive; there is no flux on the other side, and the wave decays exponentially with distance from the critical surface on the "flux-free" side. The absorption of energy and momentum at the critical latitude is a major form of wave-mean flow interaction. Rossby waves generically provide eastward acceleration; Kelvin waves produce positive mean accelerations and these propositions are true for both equatorial waves and global modes. However, the acceleration need not be everywhere one-signed [27].

12.4 Vertical Shear and the Method of Multiple Scales

> [Of Lindzen's treatment of equatorial waves in vertically-varying shear by multiple scales perturbation theory] "The most complicated damned thing I've ever seen."
>
> Robert E. Dickinson, private communication (1974)

Lindzen [28] and Holton [14, 29, 30] etc., performed numerical studies of equatorial waves in flow that was varying with both latitude and height. However, Lindzen made an important observation. The vertical wavelength of three-dimensional solutions to Laplace's tidal equations is inversely proportional to ε. This implies that $\varepsilon \to \infty$ not only implies equatorial trapping and a narrowing of latitudinal scale. The equatorial beta-plane limit is also the *short vertical wavelength* limit.

Lindzen's numerical study confirmed his belief that the scale[1] of the vertically-propagating atmospheric waves of interest was only $O(1\,\text{km})$, an order of magnitude smaller than the vertical scale of variation of the winds. Comparison of his perturbative results (Lindzen [31, 32]) with his numerical computations indeed showed close agreement.

The analysis is complicated for two reasons. First, direct imposition of the usual multiple scales nonsecularity condition leads to terrifying amounts of algebra. This chirugery can be greatly reduced by applying conservation laws (Boyd [33] and Andrews and McIntyre [34].)

The second difficulty is the variation of the latitudinal scale with height, and this cannot be fixed because it is fundamental to the physics. Without mean shear, the vertical wavelength is specified as a function of frequency by the explicit dispersion

[1] The wave scale is not the wavelength, but rather the wavelength divided by 2π; $\cos(kz)$ varies on a length scale of $1/k$ since each differentiation extracts a factor of k, i.e., $u_{zz} = -k^2 u$, but the wavelength is $2\pi/k$.

relation for equatorial waves. The obvious strategy is to insert the local Doppler-shifted frequency into the dispersion relation, thereby defining a local vertical wavenumber which varies slowly with height on the same scale as the mean zonal wind. One can then apply the well-known WKB method to the vertical structure equation. This is what Lindzen did.

However, Lamb's parameter ε controls both the vertical wavelength and the latitudinal scale of the wave. The perturbative approximation must therefore expand or contract latitudinally with height at the same time that the local vertical wavenumber is changing with height, too.

The "slow" vertical scale is defined as

$$\tau \equiv \mathfrak{s}\, z \tag{12.10}$$

where $\mathfrak{s}$ is the strength of the vertical shear. The "fast" variable is defined by the indefinite integral of the local vertical wavenumber

$$\hat{z} = \int^{z} \tilde{f}(z')\, dz', \qquad \tilde{f}(\tau) \equiv \sqrt{N^2\, \varepsilon(\tau) - 1/4} \tag{12.11}$$

where N^2 is the usual Brunt-Vaisala frequency. All this is identical with applying the WKB method [35] to the usual (atmospheric) vertical structure equation.

The wave variables are treated as functions of both $\hat{z}$ and τ as if they were independent so that

$$\frac{\partial}{\partial z} = \mathfrak{s}\frac{\partial}{\partial \tau} + \tilde{f}(\tau)\, \frac{\partial}{\partial \hat{z}} \tag{12.12}$$

In contrast, the mean zonal wind and vertical stratification are only functions of the slow variable τ. Defining the Doppler-shifted frequency by

$$\hat{\omega}(y, \tau) \equiv k(U(y, \tau) - c) \tag{12.13}$$

it follows that

$$\frac{d\hat{\omega}}{dz} = \mathfrak{s}\frac{d\omega}{d\tau} \tag{12.14}$$

The slowly-height-varying latitudinal coordinate is

$$\xi \equiv \varepsilon^{1/4}\theta \tag{12.15}$$

where θ is latitude. Figure 12.1 illustrates the simultaneous contraction of both latitudinal width and vertical wavelength as the Doppler-shifted frequency varies. To

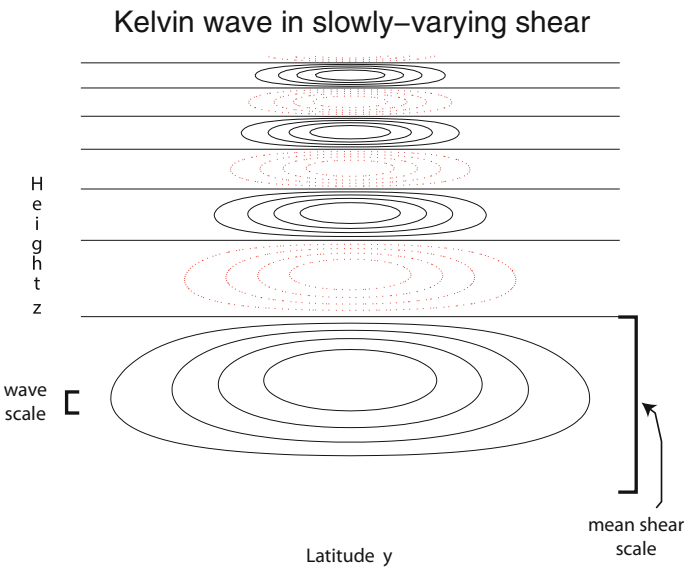

Fig. 12.1 Schematic isolines of a field such as the wave's zonal velocity for an equatorial wave in slowly-varying vertical shear. Note that the latitudinal scale contracts when the vertical wavelength contracts. For simplicity, the latitudinal structure is that of a Kelvin wave, but the same qualitative picture applies to any mode. If the sign of the shear is reversed, the vertical wavelength will grow with height instead of the contraction with increasing z illustrated here

lowest order, ε is calculated from the usual equatorial wave dispersion relation, given analytically earlier for waves linearized about a resting atmosphere, merely by replacing ω by the Doppler-shifted frequency:

$$\varepsilon = f(k, \omega) \;\rightarrow\; \varepsilon = f(k, \hat{\omega}(z)) \tag{12.16}$$

For conceptual purposes, the variation of the mean wind and therefore the variation of $\hat{\omega}$ are assumed to be slow and thus z should be replaced by τ in applying the method of multiple scales. For numerical purposes, that is, for graphing the multiple scales solution, z is applied without modification or approximation.

The effects of variations of the mean flow in latitude cannot be treated by multiple scales perturbation theory, but regular perturbation series in powers of the strength of the horizontal shear have been developed by several authors [3, 4, 10, 11, 27, 36].

As explained in Sect. 8 of [3], it is very important to note that the multiple scales solution (at lowest order) is almost identical to the wave solution *without vertical shear*. The changes induced by vertical shear are merely a *parametric* dependence on the slow variable τ and the replacement of z by the WKB fast variable $\hat{z}$. Thus, the north-south velocity is

$$v \sim A(\tau) P(\xi) \exp(i\hat{z}) \; \exp(ikx + \omega t) \tag{12.17}$$

where $P(\xi)$ is a Hermite function ψ_n and the amplitude $A(\tau)$ varies only with the slow height variable. Latitudinal shear will modify $P(\xi)$ but this shape change can be calculated at a given height by using the mean zonal wind at that height. The only dependence of the lowest order multiple scales solution which is not determined by the no-vertical-shear theory is the dependence of the amplitude A on the slow height variable τ.

Oh, dear. This is where Lindzen [31, 32] descends into the Infernal Regions of Very Lengthy Algebra. Lindzen himself suggested a remedy after his student Boyd [33] generalized the Eliassen-Palm and Charney-Drazin wave-mean flow interaction theorems [33]: apply these conservation laws [4]. Andrews and McIntyre independently proved similar theorems and worked out a conservation-law-simplified theory for equatorial waves in shear in the same year [34, 37]. When dissipation is neglected, the conservation law is nondivergence of wave flux:

$$\frac{d\overline{\phi' w'}}{dz} = 0 \tag{12.18}$$

where $\overline{\phi' w'}$ is the usual vertical energy flux and where we have modified the original (2.13) of [4] to a form appropriate for oceanography.

Dissipation can be included, but there is, unless the damping is very strong, a near-cancellation in the wave fluxes that forced Andrews and McIntyre to proceed to first order in $\mathfrak{s}$, the vertical shear strength. Boyd [4] showed that this is unnecessary if the wave-driven acceleration of the mean zonal flow is rewritten to analytically remove the cancellations as in his (8.10).

Unfortunately, the Equatorial Undercurrent has a small vertical scale which is roughly the same as the wave scale. However, asymptotic methods are successful (often) when the relevant perturbation parameter [here, the ratio of wave and vertical shear length scales] is $O(1)$ as illustrated by many examples in Bender and Orszag [35] and Miller [38].

The only application of multiple scales to equatorial waves in the ocean is Ponte [39].

Long and Chang [40] is a rare oceanographic article that cites Lindzen's work. Multiple scales perturbation theory is indeed applied by Long and Chang, but to the quite different problem of a nonlinear Kelvin pulse propagating along a shoaling thermocline.

Constantine and Johnson have kept the perturbative tradition in the ocean alive, barely, [41, 42]. Madja and his "Courant School" have devastated whole forests applying advanced mathematics to the tropical atmosphere and air-sea interaction. It is ironic that algebra-intensive methods have faded even as the capacity for computer algebra has grown exponentially both in computer cycles and in the sophistication and in the variety of the software tools and library routines to exploit them.

Table 12.1 Shear Effects on Stable Equatorial Waves: Selected References "$U(y)$" denotes use of the 1-1/2-layer model for oceanic applications. Perturbation theory has not yet been applied to two-dimensional mean flows, $U(y, z)$.

Reference	Remark
Philander [43]	$U(y)$;
Philander [44]	$U(y)$;
Philander [45]	$U(y)$; all wave species; numerical
McPhaden and Knox [10]	1-1/2-layer model; numerical eigensolutions for Kelvin and gravity waves; perturbation series for Kelvin wave concentrates on low wavenumber
McPhaden, Proehl and Rothstein [15]	$U(y, z)$, Kelvin only; long wave (meridional geostrophy) approximation
Constantin and Johnson [41], Johnson [42]	nonlinear wave-mean flow interaction with EUC

Selected references are cataloged in Table 12.1.

References

1. Dickinson RE (1970) Development of a Rossby wave critical level. J Atmos Sci 27(4):627–633
2. Ripa P, Marinone SG (1983) The effect of zonal currents on equatorial waves. In: Nihoul JCJ (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 291–318
3. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part I: theory and methods. J Atmos Sci 35:2236–2258
4. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part II: applications to the atmosphere. J Atmos Sci 35:2259–2267
5. Boyd JP, Christidis ZD (1982) Low wavenumber instability on the equatorial beta- plane. Geophys Res Lett 9:769–772
6. Boyd JP, Christidis ZD (1983) Instability on the equatorial beta-plane. In: Nihoul J (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 339–351
7. Boyd JP, Christidis ZD (1987) The continuous spectrum of equatorial Rossby waves in a shear flow. Dyn Atmos Oceass 11:139–151
8. Bennett JR, Young JA (1971) The influence of latitudinal wind shear upon large-scale wave propagation into the tropics. Mon Weather Rev 99(3):202–214
9. Zhang CD, Webster PJ (1989) Effects of zonal flows on equatorially trapped waves. J Atmos Sci 46(24):3632–3652
10. McPhaden MJ, Knox RA (1979) Equatorial Kelvin and inertia-gravity waves. J Phys Oceanogr 9(2):263–277
11. Greatbatch RJ (1985) Kelvin wave fronts, Rossby solitary waves and nonlinear spinup of the equatorial oceans. J Geophys Res 90:9097–9107
12. Dickinson RE (1968) Planetary Rossby waves propagating vertically through weak westerly wind wave guides. J Atmos Sci 25:984–1002
13. Boyd JP (1981) A Sturm-Liouville eigenproblem with an interior pole. J Math Phys 22:1575–1590
14. Holton JR (1971) A diagnostic model for equatorial wave disturbances: the role of vertical shear of the mean zonal wind. J Atmos Sci 28:55–64
15. McPhaden MJ, Proehl JA, Rothstein LM (1986) The interaction of equatorial Kelvin waves with realistically shear currents. J Phys Oceanogr 16(9):1499–1515
16. Schoeberl MR, Geller MA (1976) The structure of stationary planetary waves in relation to the polar night jet intensity. Geophys Res Lett 3:177–180

17. Schoeberl MR, Geller MA (1977) A calculation of the structure of stationary planetary waves in winter. J Atmos Sci 34:1235–1255
18. Schoeberl MR, Geller MA, Avery SK (1979) The structure of stationary planetary waves in winter: a correction. J Atmos Sci 36:365–369
19. Lindzen RS, Kuo HL (1969) A reliable method for the numerical integration of a large class of ordinary and partial differential equations. Mon Weather Rev 97:732–734
20. Boyd JP (1978) Spectral and pseudospectral methods for eigenvalue and nonseparable boundary value problems. Mon Weather Rev 106:1192–1203
21. Boyd JP (2001) Chebyshev and Fourier Spectral Methods. Dover, New York, p 680
22. Matsuno T (1970) Vertical propagation of stationary planetary waves in winter northern hemisphere. J Atmos Sci 27(6):871–883
23. Durran DR (2010) Numerical methods for fluid dynamics with applications to geophysics, texts in applied mathematics, vol 32, 2nd edn. Springer, New York
24. Boyd JP (1985) Complex coordinate methods for hydrodynamic instabilities and Sturm-Liouville problems with an interior singularity. J Comput Phys 57:454–471
25. Gill AW, Sneddon GE (1995) Complex mapped matrix methods in hydrodynamic stability problems. J Comput Phys 122:13–24
26. Gill AW, Sneddon GE (1996) Pseudospectral methods and composite complex maps for near-boundary critical latitudes. J Comput Phys 129(1):1–7
27. Boyd JP (1982) The effects of meridional shear on planetary waves, part II: critical latitudes. J Atmos Sci 39:770–790
28. Lindzen RS (1970) Internal equatorial planetary-scale waves in shear flow. J Atmos Sci 27:394–407
29. Holton JR (1970) The influence of mean wind shear on the propagation of Kelvin waves. Tellus 22:186–193
30. Holton JR (1972) Waves in the equatorial stratosphere generated by tropospheric heat sources. J Atmos Sci 29:368–375
31. Lindzen RS (1971) Equatorial planetary waves in shear: part I. J Atmos Sci 28(4):609–622
32. Lindzen RS (1972) Equatorial planetary waves in shear: part II. J Atmos Sci 29(8):1452–1463
33. Boyd JP (1976) The noninteraction of waves with the zonally averaged flow on a spherical earth and the interrelationships of eddy fluxes of heat, energy, and momentum. J Atmos Sci 33:2285–2291
34. Andrews DA, McIntyre ME (1976) Planetary waves in horizontal and vertical shear: asymptotic theory for equatorial waves in weak shear. J Atmos Sci 33:2049–2053
35. Bender CM, Orszag SA (1978) Advanced mathematical methods for scientists and engineers. McGraw-Hill, New York, p 594
36. Boyd JP (1982) The effects of meridional shear on planetary waves, part I: nonsingular profiles. J Atmos Sci 39:756–769
37. Andrews DA, McIntyre ME (1976) Planetary waves in horizontal and vertical shear: the generalized Elissen-Palm relation and the mean zonal acceleration. J Atmos Sci 33:2031–2048
38. Miller PD (2006) Applied asymptotic analysis, graduate studies in mathematics, vol 75. American Mathematical Society, Providence
39. Ponte RM (1988) Equatorial Kelvin waves embedded in mean flow, with applications to the deep jets. J Atmos Sci 93(C11):13,941–13,946
40. Long B, Chang P (1990) Propagation of an equatorial Kelvin wave in a varying thermocline. J Phys Oceanogr 20:1826–1841
41. Constantin A, Johnson RS (2015) The dynamics of waves interacting with the Equatorial Undercurrent. Geophys Astrophys Fluid Dyn 109(4):311–358
42. Johnson RS (2015) An ocean undercurrent, a thermocline, a free surface, with waves: a problem in classical fluid mechanics. J Nonlinear Math Phys 22(4):475–493
43. Philander SGH (1976) Instabilities of zonal equatorial currents. J Geophys Res 81(21):3725–3735
44. Philander SGH (1978) Instabilities of zonal equatorial currents, 2. J Geophys Res 83:3679–3682
45. Philander SGH (1979) Upwelling in the Gulf of Guinea. J Marine Res 37:23–33

Chapter 13
Inertial Instability, Pancakes and Deep Internal Jets

Abstract The necessary conditions for inertial instability, alias symmetric instability and centrifugal instability, are very easily satisfied near the equator if $dU/dy(0) \neq 0$, that is, if there is nonzero latitudinal shear at the equator. Inertial instability is the primary suspect in the formation of ("pancake") alternating jets in the tropical stratosphere and of deep internal jets in the equatorial ocean. Particle models, Taylor-Couette flow, and instability criteria are reviewed. The "gamma-plane" approximation is analytically soluble. Its prediction of a dynamical equator shifted from the geographical equator is confirmed by more realistic models. The large zonal scales and short vertical length scales of theory are amply confirmed in both pancake instabilities and deep internal jets.

> There are well documented cases of structures near the equatorial stratopause, with a vertical wavelength of about 10 km, bearing a strong resemblance to unstable modes of the linear stability theory (e.g. Hitchman et al. [1], Hayashi, Shiotani and Gille [2], Smith and Riese [3]).
>
> There seem to be several large-amplitude events of inertial instability per year.
>
> Stephen D. Griffiths, introduction [4]
>
> Equatorial deep jets are a ubiquitous feature of the circulation in all three oceans, since their original discovery by Luyten and Swallow [5] in the western Indian Ocean.
>
> p. 347 of Hua, Moore and Le Gentil [6]

13.1 Introduction: Stratospheric Pancakes and Equatorial Deep Jets

Inertial instability has a short vertical wavelength and a long or zero zonal wavelength. It thus generates layers of jets that alternate in direction. The Nimbus 7 satellite, launched in 1978, carried among other instruments the cryogenically-cooled Limb Infrared Monitor of the Stratosphere (LIMS). Before the last of the coolant boiled away, long before the satellite itself decayed into a flaming fireball in 1994, LIMS detected "vertically stacked temperature extrema of alternating sign" in the equatorial lower mesosphere, to quote the abstract of Hitchman, Leovy, Gille and Bailey [1].

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_13

Dubbed "pancakes", these alternating layers of hot and cold are coherent with winds of alternating sign which could not be directly seen by the radiometer. The Pancakes were seen in periods of strong cross-equatorial shear. "Their occurrence is confined to regions of very weak or negative inertial stability [as measured by a diagnostic (13.1)] and their meridional to vertical aspect ratio, meridional structure and zonal spectrum are consistent with disturbances predicted by inertial instability theory [1]."

In the deep ocean, similar vertically stacked structures of alternating sign were discovered by Luyten and Swallow [5] and dubbed "deep internal jets" [5]. Hua, Moore and Le Gentil [6], amplifying their quote above, [6] note (p. 347) that "Equatorial deep jets are a ubiquitous feature of the circulation in all three oceans."

Inertial instability, sometimes misleadingly called "symmetric instability", can occur at all latitudes, but is especially common and important at low latitudes. The reason is that the fundamental criterion for the instability is

$$f\left(f - \frac{dU}{dy}\right) < 0 \tag{13.1}$$

at some latitude where $U(y)$ is the mean zonally-averaged current. Since the Coriolis parameter f is zero at the equator and small in the tropics, it is particularly easy for the mean shear to overcome f at low latitudes and trigger the instability.

In addition, inertial instability is the Great Symmetrizer. Like other instabilities, it tends to correct whatever condition created it. The instability criterion shows that the mean flow can return to stability everywhere only if the shear is *zero* at the *equator*. However, if there is no shear at the equator, then the mean flow $U(y)$ — and to stay in geostrophic balance, the mean height field $\Phi(y)$ also — must be *symmetric* with respect to the equator, at least at low latitudes.[1]

One is reminded of Sherlock's remark about the "curious incident of the dog in the night". When Watson replies, "but the dog did nothing in the night", Holmes responds, "that was the curious incident". In a similar way, inertial instability may play a powerful role in the ocean and atmosphere even when absent.

To understand inertial instability, we shall describe a hierarchy of models of increasing complexity: first particle and parcel arguments, then Boyd's "equatorial gamma-plane" approximation, and finally vertical structure, numerical models and observations.

13.2 Particle Argument

13.2.1 Linear Inertial Instability

Elementary meteorology and physical oceanography classes discuss three elementary pairwise balances of forces in steady horizontal flow: geostrophic, cyclostrophic and

[1] The first three terms in the Taylor series for $U(y)$ must be symmetric, but higher order terms may violate the symmetry without breaking down geostrophy or the stability of $U(y)$ at the equator.

inertial. The last neglects pressure forces as well as all nonlinear terms. Textbooks such as Holton and Hakim [12] show that in pure inertial flow a blob of fluid will orbit anticyclonically with a frequency equal to the local value of the Coriolis parameter.

Inertial instability adds a single ingredient: linearization about a spatially-varying mean current. For expository simplicity, we assume that the mean flow is zonal and the shear is purely latitudinal, but one can rotate the horizontal coordinate system for a mean flow with arbitrary orientation. Only the *local* value of shear and *local* value of the Coriolis parameter are relevant because the neglect of pressure gives equations which are independent of the horizontal coordinates:

$$u_t - (f - U_y)v = 0; \qquad v_t + fu = 0 \tag{13.2}$$

(A particle sliding frictionlessly on a rotating sphere, as opposed to a blob of fluid, is well-described by these equations because the particle dynamics is intrinsically pressure-free.)

Eliminating v gives

$$u_{tt} + f(f - U_y)u = 0 \tag{13.3}$$

In pure, shear-free inertial motion, a parcel of fluid loops round and round in anticyclonic circles. If the blob is given a small eastward velocity in the northern hemisphere, then Coriolis force creates a southward force, always perpendicular to the velocity, that turns the velocity without altering its magnitude.

When $\mathrm{sgn}(f - U_y) = \mathrm{sgn}(f)$, the motion is periodic in time; the shear merely increases or decreases the frequency. When

$$f(f - U_y) < 0, \qquad \text{[Necessary Particle Condition for Inertial Instability]} \tag{13.4}$$

the velocity grows exponentially with time as $\exp(\sigma t)$ where the growth rate is

$$\sigma = \sqrt{-f(f - U_y)} \tag{13.5}$$

$$u(t) = A\exp(\sigma t) + B\exp(-\sigma t), \qquad v = -(f/\sigma)A\exp(\sigma t) + (f/\sigma)\,B\exp(-\sigma t). \tag{13.6}$$

13.3 Centrifugal Instability: Rayleigh's Parcel Argument

We can pass from particle to parcel by applying a century old analysis of Lord Rayleigh's to the Taylor-Couette apparatus. This is an annular shell of liquid confined between two concentric cylinders which can be spun independently by electric motors. If one cylinder is fixed while the other spins at a constant speed, the flow will settle into a steady-state that solves the diffusion equation, a roughly linear variation of tangential velocity with radius. The flow is in cyclostrophic balance with inward pressure exactly counteracting the outward centrifugal force (Fig. 13.1).

If a blob of fluid is displaced radially, the pressure will almost instantly adjust to the change in ambient pressure by radiating sound waves. A radial push, however, will not change the tangential velocity. If the inner wall is stationary and the outer cylinder is rotating, a parcel originating near the inner cylinder will have small angular momentum. Displaced farther from the center of the apparatus, the parcel will feel the same strong inward pressure force that balances the strong centrifugal force acting on the fast-moving fluid near the outer wall. However, the blob displaced outward towards the spinning outer cylinder does not have large tangential velocity to generate a strong centrifugal force. The strong ambient pressure overcomes the outward centrifugal force to push the blob back towards its original radius. A high-angular momentum blob displaced inward will be restored to its original radius by the strong centrifugal force acting on the blob to push it outward. Either way, the blobs are still moving when they cross their original radii. The overshoot creates waves known as "inertial oscillations".

In contrast, when the *inner* cylinder is moving, a high velocity blob, pushed outward, moves into a low pressure force environment. Strong centrifugal force accelerates the blob outward. The flow is unstable because perturbations are amplified and accelerated. Similarly, slow-moving fluid initially near the unmoving outer cylinder will be hurled inward by the strong pressure force near the inner cylinder.

In summary, a flow with absolute angular momentum increasing radially outward is stable, but mean angular momentum decreasing outward is unstable. A careful quantitative discussion of Rayleigh's argument, with further embellishments by von Karman, is given in Yih's textbook, pp 450–459 [7].

For large-scale atmospheric and oceanic flows, Coriolis force is more important than centrifugal, but the concepts are little changed.

With a bit of handwaving, we can make some inferences about the ensuing inertially-unstable normal modes.

First, the Taylor-Couette apparatus is axisymmetric. The observed "Taylor roll vortices" are indeed independent of the angular coordinate θ in the usual (r, θ, z) cylindrical coordinate system. This argument is only semi-good because so-called "spontaneous symmetry breaking" is possible in many physical systems. Indeed, at high rotation rates, the axisymmetric Taylor roll vortices bifurcate into wavy vortices. Still, spontaneous symmetry breaking needs additional physics to break the symmetry – here unstable Kelvin waves and gravity wave physics will do the honors — so a strong tendency to axisymmetry is characteristic of inertial instability.

The vertical scale of inertially-unstable roll vortices is a vexing question. However, a large vertical scale makes no sense because the fundamental mechanism is about what happens in the horizontal. Indeed, one might expect that the unstable waves would have a very short vertical scale limited only by diffusion because this would minimize the ratio of useless up-and-down flow relative to the radial (latitudinal) motion. This is more or less what is observed in normal mode calculations with some caveats noted below.[2]

[2]The solution to the diffusion equation, $u_t = \nu u_{xx}$ with homogeneous Dirichlet conditions $u(0) = u(\pi) = 0$ is $u(x,t) = \sum_{n=1}^{\infty} a_n(t=0)\sin(nx)\exp(-\nu n^2 t)$. This explicitly shows that diffusion

13.4 Equatorial Gamma-Plane Approximation

Boyd [8] showed that, after linearization about a mean zonal current on the equatorial beta-plane, the atmospheric dynamics reduces to a single equation for v alone

$$v_{yy} - \frac{2\Gamma(y)k}{\omega}\left(\frac{\varepsilon}{\varepsilon - k^2/\omega^2}\right)v_y + \left\{\frac{k}{\omega}\left(\beta - \Gamma_y\right) - k^2 - \frac{2\Gamma k}{\omega}\left(\frac{\varepsilon}{\varepsilon - k^2/\omega^2}\right)\frac{k}{\omega}(\beta y - \Gamma)\right.$$
$$\left. -\varepsilon\left[\beta y(\beta y - \Gamma) - \omega^2\right]\right\} v = 0$$

This is a mess; the analogous equation derived from the nonlinear shallow water equations includes many additional terms all from the mean, latitudinally-varying height $\Phi(y)$ which is omitted from the atmospheric analog, and is even messier.

Fortunately, Boyd [8] showed by means of a perturbation series in the parameter $k\omega$ that this ODE can be simplified. His simplified equation comes in three flavors. His simplified equation for Rossby waves, his (4.12), will not be discussed here. The second version, cubic in the frequency or phase speed, which mixes the Kelvin and $n = 0$ gravity modes together, is analyzed later in this chapter. In this section, we will concentrate on the version appropriate for gravity modes.

Definition 13.1 (*Equatorial Gamma-Plane Approximation*) Treat the Doppler -shifted frequency $\hat{\omega}(y)$ as a constant except that $dU/dy = \Gamma$ is treated not equal to zero but instead as a constant.[3] The ODE for the n-th-south wave current is

$$v_{yy} - \varepsilon\left\{\beta y(\beta y - \Gamma) - \omega^2\right\} v = 0 \tag{13.7}$$

Although quadratic shear was analyzed in [8, 9], it is helpful to concentrate on a linear shear, $U(y) = \Gamma y$ where Γ is a constant.

This mimicks the way the usual midlatitude beta-plane treats the Coriolis parameter — as a constant except where differentiated, and then defining $df/dy = \beta$ where the derivative β is treated as a constant,

This gamma-plane approximation is very convenient because the eigenfunctions and eigenvalues can be found *analytically*.

Theorem 13.1 (Generalized Parabolic Cylinder Equation) *The eigenfunctions of*

$$v_{yy} + \left\{A + By - Cy^2\right\} v = 0 \tag{13.8}$$

are given by, introducing

(Footnote 2 continued)
and viscosity are *highly scale-selective* in that shorter spatial scales are damped more rapidly than lower wavenumbers. Thus, viscosity and/or diffusion will kill the growth of very short vertical wavelengths while allowing larger vertical wavelengths to grow exponentially.

[3] The reader is warned that other, quite different meanings have been applied to "gamma-plane" by later authors.

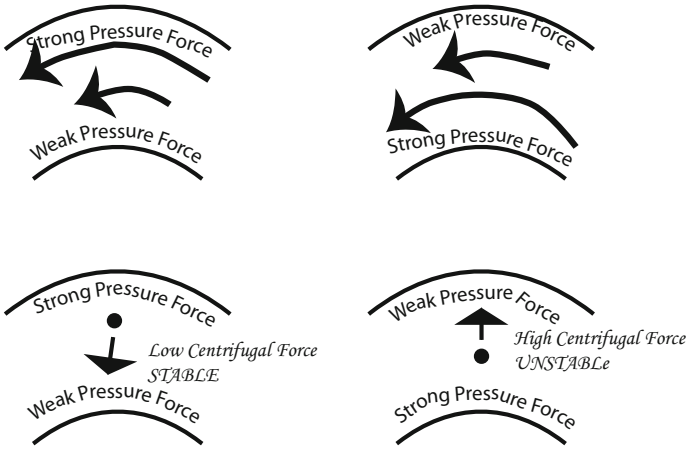

Fig. 13.1 Schematic of centrifugal instability in the Taylor-Couette apparatus of nested cylinders. Only part of the two concentric cylinders is shown. *Left two panels* the stable case in which the outer cylinder rotates at a constant speed Ω while the inner cylinder is stationary; the basic state angular momentum increases with radius in polar coordinates (r, θ) centered on the rotation axis common to both cylinders. A parcel of low angular momentum, pushed radially outward, is restored to its original radius near the stationary inner cylinder by the strong inward pressure force near the outer cylinder. *Right* unstable case in which the inner cylinder rotates at a constant speed Ω while the outer cylinder is stationary; the basic state angular momentum decreases with radius in polar coordinates. A blob pushed outward has higher angular momentum than its surrounding ambient fluid, and the strong centrifugal force, opposed only by a weak inward pressure force, accelerates the blob farther from its original radius

$$\xi = \sqrt[4]{C}\left(y - \frac{B}{2C}\right) \tag{13.9}$$

$$v = \psi_n(\xi), \qquad n = 0, 1, 2, \ldots \tag{13.10}$$

with the eigencondition

$$2n + 1 = A/\sqrt{C} + B^2/(4C^{3/2}) \tag{13.11}$$

Appendix of Boyd [8].

Here

$$C = \varepsilon\beta^2 \tag{13.12}$$

$$B = \varepsilon\beta\Gamma \tag{13.13}$$

$$A = \varepsilon\omega^2 \tag{13.14}$$

$$\xi = \sqrt[4]{\varepsilon}\sqrt{\beta}\left(y - \frac{1}{2\beta}\Gamma\right) \tag{13.15}$$

$$\omega^2 = (2n+1)\frac{\beta}{\sqrt{\varepsilon}} - \frac{\Gamma^2}{4}. \tag{13.16}$$

13.5 Dynamical Equator

Figure 13.2 shows how shear shifts the peak of an equatorial mode, here the north-south velocity of the $n = 0$ Yanai wave, from the geographical equator to a shear-dependent "dynamical equator". The Hermite functions are centered a the point where the scaled coordinate ξ is zero:

$$y_{dyn.eq} = \frac{1}{2\beta}\Gamma \tag{13.17}$$

Meteorologists have been familiar with the notion of a "dynamical equator". In idealized models of the Hadley circulation, the updraft is centered on the equator. The convergence of moist, warm air into this upwelling leads to a band of clouds that rings the globe, the Inter-Tropical Convergence Zone (ITCZ). However, the center is not at the equator, but shifts with the seasons with a mean position several degrees north of the geographical equator.

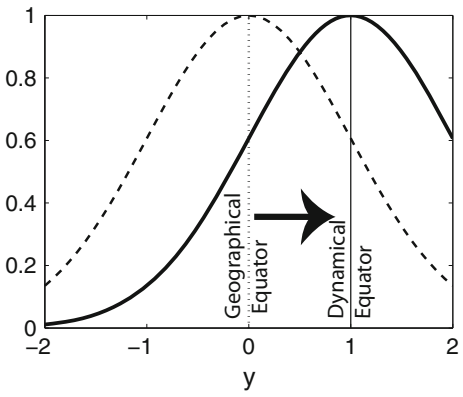

Fig. 13.2 In the gamma-plane approximation, the lowest order solutions are still described by Hermite functions, but shear alters the width of the functions. Shear also translates the symmetry point of each Hermite function to $y = y_{dny.eq} = \Gamma/(2\beta)$ [*solid curve*]

13.6 Gamma-Plane Instability

One possible mode of instability is that when the frequency is specified, and specified as a real constant, the dispersion relation will yield complex-valued eigenvalues for one or more modes. Recall that Lamb's parameter ε is, when inserted into the vertical structure equation, the square of the nondimensional vertical wavenumber. A complex-valued eigenvalue implies the vertical wavenumber is complex-valued, too, and therefore the mode does not merely oscillate with height or depth, but rather decays as it propagates, a "damped oscillation", or grows exponentially with height while oscillating. However in the gamma-plane approximation

$$\varepsilon = \frac{(2n+1)^2 \beta^2}{\left\{\omega^2 + \Gamma^2/4\right\}^2} \tag{13.18}$$

It is immediately obvious that the right-hand-side, with every parameter squared, is positive definite. Fixed-frequency, complex-valued Lamb's parameter inertial instability is impossible (Fig. 13.3).

One might suppose that this would prohibit growing-in-time instability, too, but rearranging this same gamma-plane dispersion relation gives

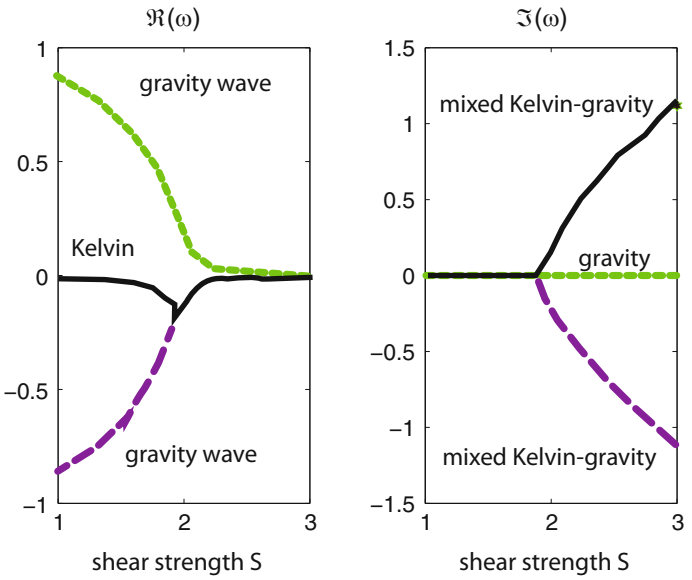

Fig. 13.3 When the gamma-plane is generalized to include one additional term so that the dispersion relation becomes a cubic polynomial in the frequency, sufficiently strong shear causes the Kelvin mode and a gravity wave mode to merge, forming an unstable complex-conjugate pair of modes for larger shear strength. The nondimensional zonal wavenumber $k = 1/100$, but cross-sections for other small to moderate k are similar

$$\omega = \sqrt{(2n+1)\frac{\beta}{\sqrt{\varepsilon}} - \frac{\Gamma^2}{4}} \tag{13.19}$$

This predicts temporal instability whenever

$$\Gamma^2 > 4\sqrt{(2n+1)}\frac{\sqrt{\beta}}{\varepsilon^{1/4}} \tag{13.20}$$

Unfortunately, $\sqrt{\varepsilon}$ is proportional to the vertical wavenumber. No matter how weak the mean latitudinal shear Γ, sufficiently short vertical wavelengths, that is, baroclinic modes of sufficiently high baroclinic mode number, are *always unstable*.

In reality, vertical viscosity and diffusion, whose effects grow quadratically with vertical mode number, will reshape the growth rate so that the inertial-instability-with-viscosity growth rates peak at a small but finite vertical scale. Indeed, both mesospheric "pancakes" and oceanic "deep internal alternating jets" are vertically oscillating with short vertical wavelength.

13.7 Mixed Kelvin-Inertial Instability

The Kelvin wave is unstable due to a critical latitude instability with a growth rate that is an exponential function of the reciprocal of the square of the strength of the shear. This is discussed in the next chapter. For strong shear, the complicated relationship between the Kelvin wave and the lowest gravity modes becomes more intricate still.

In the absence of shear, the mode which matches the equatorial beta-plane in the limit that Lamb's parameter $\varepsilon \to \infty$ becomes the lowest symmetric inertial-gravity wave as $\varepsilon \to 0$. Longuet-Higgins gave a clear analysis of this smooth transition in [10]. Equatorial oceanographers have rightly resisted the temptation to relabel this wave as the "mixed Kelvin-gravity" wave; ε is always huge and the small-ε limit is irrelevant for baroclinic modes in the sea.

When the shear is strong, even on the equatorial beta-plane, the Kelvin wave will merge with a lowest latitudinal mode gravity wave to form a complex conjugate pair of modes, one unstable, for larger shear. The gamma-plane needs to be generalized slightly to the "uniform gamma-plane", which identical with the gamma-plane latitudinal velocity equation except for one additional term (boxed):

$$v_{yy} - \varepsilon\left\{\boxed{\tfrac{k}{\omega}} + \beta y(\beta y - \Gamma) - \omega^2\right\} v = 0 \tag{13.21}$$

For small zonal wavenumber k or high frequency ω, this term is negligible. However, increasing shear greatly increases the frequency of the Kelvin wave until it merges with an $n = 0$ gravity wave with instability for larger shear. A fuller analysis is given in Boyd and Christidis [11] where it is shown that by careful tracking of modes as the shear strength varies that the mode which merges with the gravity wave is indeed the mode which is the Kelvin wave for zero shear.

13.8 Summary

The history of inertial instability is rich with irony. Deep alternating jets were first discovered in the ocean by Luyten and Swallow in [5], but the first steps toward a theory were made by three atmospheric scientists, Boyd, Dunkerton and Stevens, a few years later before there was any convincing evidence of inertial instability in meteorological observations. The discovery of "pancake instabilities" circa 1987 vindicated the atmospheric dynamicists, but another decade would pass before the magisterial theory-numerical-experiment article of Hua, Gentile and Moore made a convincing case that the ocean equatorial jets were the offspring of inertial instability.

Although the paper of Hua, Gentile and Moore and also the recent atmospheric articles of Griffin are full of insights, and Plougonven and Zeitlin have followed the instability into the nonlinear regime, the heart of inertial instability is captured in the simple equatorial gamma-plane model of Boyd, Dunkerton and Stevens. Latitudinal shear is the star; vertical shear, though strong, is only a modifier. The instability eats at the shear that spawned it and in so doing is a strong force for local symmetrization of the flow with respect to the equator. The concept of a "dynamical equator", shifted by cross-equatorial shear away from the geographical equator, was introduced by Boyd in [8, 9]. It is prominent in both the numerical simulations of Hua, Moore and Le Gentil and later authors. In their abstract, Hua, Moore and Le Gentil write,"a meridional shear of the basic state leads to a vertical stacking of equatorially-trapped zonal flows of alternate signs, with a new centre of symmetry located at the dynamical equator."

The short vertical scale of inertial instability is an essential part of the dynamics in both the deep equatorial ocean and in the tropical middle atmosphere. A good way to provoke a riot among tropical dynamicists is to strongly assert a single cause for what exactly determines the vertical scale and avoids the "ultraviolet catastrophe" of the arbitrarily short vertical wavelengths predicted by the simplest theories of inertial instability.

The alternative name for this phenomenon, "symmetric instability", is not entirely appropriate for the tropics because the most unstable waves are not usually of zero wave number, but merely ultralong in the east-west direction. Nevertheless,the slow variation in longitude, as opposed to the very fast oscillations in depth, is an important theoretical simplifier.

The vanishing of the Coriolis parameter allows inertial instability to flourish in equatorial regions in a manner impossible at higher latitudes where inertial instability is found only in violent frontal zones.

References

1. Hitchman M, Leovy CB, Gille JC, Bailey PB (1987) Quasi-stationary zonally asymmetric circulations in the equatorial lower mesosphere. J Atmos Sci 44:2219–2236
2. Hayashi H, Shiotani M, Gille JC (1998) Vertically stacked temperature disturbances near the equatorial stratopause as seen in cryogenic limb array etalon spectrometer data. J Geophys Res 103:19,469–19,483
3. Smith AK, Riese M (1999) Cryogenic infrared spectrometers and telescopes for the atmosphere CRISTA observations of tracer transport by inertially unstable circulations. J Geophys Res 104:19,171–19,182
4. Griffiths SD (2003) Nonlinear vertical scale selection in equatorial inertial instability. J Atmos Sci 60(7):977–990
5. Luyten JR, Swallow JC (1976) Equatorial undercurrents. Deep Sea Res 23:999–1001
6. Hua BL, Moore DW, LeGentil S (1997) Inertial nonlinear equilibration of equatorial flows. J Fluid Mech 331:345–371
7. Yih CS (1968) Fluid mechanics. West River Press, Ann Arbor
8. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part I: theory and methods. J Atmos Sci 35:2236–2258
9. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part II: applications to the atmosphere. J Atmos Sci 35:2259–2267
10. Longuet-Higgins MS (1968) The eigenfunctions of Laplace's tidal equations over a sphere. Philos Trans R Soc Lond A 262:511–607
11. Boyd JP, Christidis ZD (1982) Low wavenumber instability on the equatorial beta- plane. Geophys Res Lett 9:769–772
12. Holton JR, Hakim GJ (2012) An Introduction to Dynamic Meteorology, vol 88. Academic Press, New York

Chapter 14
Kelvin Wave Instability: Critical Latitudes and Exponentially Small Effects

Abstract The Kelvin wave has a critical latitude instability in a linear shear, $U(y) = \Gamma y$. There is no neutral curve, but instead with viscosity neglected there is instability for all $|\Gamma|$, however small. A power series in Γ is useless because $\Im(c) \approx \frac{0.14}{\Gamma^3} \exp\left(-\frac{1}{\Gamma^2}\right)$ which goes to zero faster than any finite power of Γ as $\Gamma \to \infty$. To capture "beyond-all-orders" effects, Hermité-Padé approximants, exponential asymptotics and numerical methods that loop off the real axis are applied to a hierarchy of models to illuminate this peculiar instability of this most important mode.

> One remarkable fact of applied mathematics is the ubiquitous appearance of divergent series, hypocritically renamed asymptotic expansions. Isn't it a scandal that we teach convergent series to our sophomores and do not tell them that few, if any, of the series they meet will converge? The challenge of explaining what an asymptotic expansion is ranks among the outstanding taboo problems of mathematics.
>
> Gian-Carlo Rota, p. 222 of his book *Indiscrete Thoughts* [1]

Kelvin wave instability is conceptually challenging because the growth rate is exponentially small in the reciprocal of the small parameter. Because $\exp(-1/\varepsilon^2)$ goes to zero faster than any finite power of $\varepsilon(= 1/\Gamma)$, the instability is invisible to a power series in the small parameter. A million terms of the power series, nay, even a billion terms, will yield only a real phase speed. Such power-series-invisible effects are said to be "beyond-all-orders" [in powers of ε]. A subfield of applied mathematics dubbed variously "hyperasymptotics" or "exponential asymptotics" has developed to interpret and calculate such "beyond-all-orders effects" as catalogued in the review article [3] and books [4, 5]. Unfortunately, Rota's quote is still true twenty years later even though [6] shows that only undergraduate mathematics is needed to understand it.

Kelvin wave instability is a tale of two regimes. When the shear strength is large, the Kelvin wave becomes mixed up with inertial gravity waves. This part of the story already has been told as part of the discussion of inertial instability. Here, we focus on the "beyond-all-orders" instability.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_14

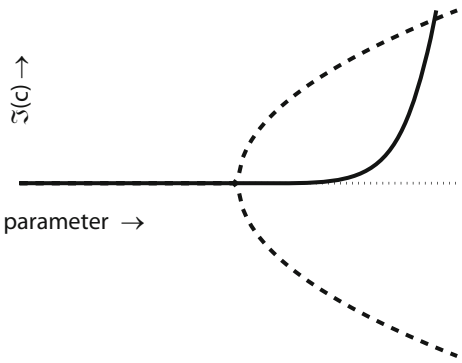

Fig. 14.1 Imaginary part of the phase speed for a conventional instability [dashed]. Both a growing mode and a decaying mode bifurcate at a certain value of the parameter. Imaginary part of the phase speed for the Kelvin wave instability [solid]

When the shear is weak, the Kelvin wave may have a critical latitude far from the equator where its amplitude is exponentially small. A critical latitude is a point where the mean current $U(y)$ matches the phase speed c. It has long been known that critical latitudes play a critical role in instability [7, 8]. (The imaginary part of the phase speed shifts the critical latitude off the real y-axis, but does not diminish the importance of the critical latitude.) Because the Kelvin wave decays exponentially fast away from the equator, the growth rate for a critical latitude instability will be *exponentially small* in the distance of the critical latitude from the equator. This requires us to dive into the depths of hyperasymptotics.

Figure 14.1 shows, as the dotted curves, the usual transition to instability. At some finite value of a parameter Γ, two neutrally stable modes merge to become a complex-conjugate pair for larger S. The phase speed for both modes has a square root singularity at the critical point, $\Gamma = \Gamma_c$. Because the first derivative of a square root is a reciprocal square root which is unbounded as the singularity is approached, $dc/d\Gamma(\Gamma_c) = \infty$.

In contrast, the Kelvin instability has no transition point. In the absence of viscosity or other dissipation, the wave is *unstable* for *all shear strengths*, however small. Note that because the growth rate is exponentially small in the reciprocal of the shear strength, growth is significant only for large and moderate shears. For small shears, even a tiny amount of dissipation (physical or computational) will suppress the instability.

14.1 Proxies and the Optical Theorem

The most direct path to a good theory would be to imitate the program successfully carried out for Rossby waves with critical latitudes by Dickinson, Boyd and Lindzen and Rosenthal [9–11] and for singular wave equations in plasma physics and many

other fields as skillfully cataloged by Adam [8]. Over most of the domain, a WKB approximation is satisfactory. The crucial final step to a globally-valid analytic solution is to apply the method of matched asymptotic expansions to join the WKB solution to a local solution around the critical latitude in terms of Bessel functions. Unfortunately, this game plan has proved too difficult for the Kelvin wave. However, much has been learned from simple proxies and numerical studies.

Boyd and Natarov investigated a proxy, a Sturm-Liouville Eigenproblem of the Fourth Kind [10], the "Hermite-with-Pole" equation

$$u_{yy} + \left\{ \frac{1}{y - 1/\varepsilon} - y^2 \right\} u = \lambda u \tag{14.1}$$

where λ is the eigenvalue. This is very similar to the *Rossby* wave eigenproblem studied by Dickinson and Boyd. However, it is not a Kelvin wave model except in the sense that it is an eigenproblem on the equatorial beta-plane with a critical latitude singularity. Nevertheless, much has been learned about this exemplar, and perhaps we shall eventually know as much about the Kelvin wave.

Boyd and Natarov proved the following theorem [12]. By analogue with similar theorems in particle physics, they dubbed it the "Optical Theorem". Note that their theorem is considerably more general than stated here, and is based on a theorem proved in [10].

The theorem employs a Rayleigh friction δ. This is necessary in the theorem because generically, solutions to geophysical wave equations have logarithmic singularities at the critical latitude. For neutrally-stable waves, as in the theorem below, the singularity is on the real axis. Adding a small amount of dissipation, such as the "Rayleigh friction" used in the theorem, is necessary to resolve the ambiguity of the branch cut — should it be in the upper y-plane or the lower half-plane? This subtlety is explained by Dickinson [9, 13] and Boyd [10]. Friction is necessary only when the singularity is otherwise on the real axis as in the Hermite-with-pole equation; when the phase speed has an imaginary part, the critical latitude is shifted off the real axis and the correct choice of branch is any branch cut which does not cut the real axis.

Theorem 14.1 (Optical Theorem) *Let the differential equation be*

$$u_{yy} + \left\{ \frac{1}{y - 1/\varepsilon - i\delta} - y^2 \right\} u = \lambda u, \qquad y \in [-\infty, \infty] \tag{14.2}$$

where $\delta > 0$ is the coefficient of Rayleigh friction. Then in the limit $\delta \to 0_+$,

$$\Im(\lambda) = -\pi \frac{|u(y_c)|^2}{\int_{-\infty}^{\infty} |u(y)|^2 \, dy} \tag{14.3}$$

where the critical latitude is $y_c = -1/\varepsilon$.

Proved by Boyd and Natarov [12].

For the lowest mode, $u(y) \sim \exp(-[1/2]y^2)$ and $\Re(\lambda) \sim 1$, just like the Kelvin wave. A WKB approximation combined with matched asymptotic expansions and the optical theorem give

$$\Im(\lambda) \sim \sqrt{\pi} \exp\left(-\frac{1}{\varepsilon^2}\right) \{1 - 2\varepsilon \log(\varepsilon) + \varepsilon \log(2) + \gamma \Gamma + \ldots\} \tag{14.4}$$

Thus, the exponential smallness for this eigenproblem can be captured by explicit analytical approximations.

One remarkable and unexpected fact is that the imaginary part of the eigenvalue is proportional to the *square* of the eigenfunction at the critical latitude, not to the eigenfunction itself. Is the same true of the Kelvin wave?

14.2 Six Ways to Calculate Kelvin Instability

An analytical answer is not known, but it is straightforward to perform numerical experiments by solving

$$ik(U - c)u + (U_y - y)v + ik\phi = 0 \tag{14.5}$$

$$ik(U - c)v + yu + \phi_y = 0 \tag{14.6}$$

$$ik(U - c)\phi + iku + v_y + \aleph \left\{ik\Phi(y)u + \Phi(y)v_y\right\} - \beth\, yU\, v = 0 \tag{14.7}$$

As noted earlier, meteorologists obtain an equation set of this form by the method of separation of variables in a continuously stratified atmosphere, and then the factors $\aleph$ and $\beth$ are zero. Oceanographers usually rationalize their depth-independent models from the 1-1/2-layer model. When linearized about a state of rest, the two approaches give the same linear equations. When linearized about a flow with mean currents, however, the equivalency breaks down. Much of the earlier work on equatorial instability has solved the "atmospheric" equations, that is, the trio above with $\aleph = \beth = 0$ as in the articles by Boyd and Christidis [2, 14]. Others have solved these equations with $\aleph = \beth = 1$. These differences must be kept in mind in comparing different studies.

Six ways to compute $\Im(c)$ are listed in Table 14.1. Four are useless, but two work well. We discuss each in turn.

Table 14.1 Six Ways to Calculate the Imaginary Part of an Eigenvalue That Is Exponentially Small in $1/\Gamma$ where $\Gamma \ll 1$ Is a Small Parameter

Numerical integration on real axis	Fails; singularity exponentially near real axis
WKB with matched asymptotic expansions	Fails; too complicated
Γ power series ["Rayleigh-Schrodinger"]	Fails; always real-valued
Padé Approximants	Fails; always real-valued
Hermite-Padé approximants	Success! Hermite-with-Pole
" "	Success! Gauss-Stieltjes integral [15]
Spectral method in $z \equiv y - i$	Success! [14, 16, 17]

Conventional numerical methods solve the differential equation on a computational interval which is the real axis. These can be dismissed almost immediately. When $|\Im(c)|$ is very tiny, the critical latitude is almost on the real axis, a computational land-mine.

As noted earlier, WKB-with-matched-asymptotic-expansions has been very successful in midlatitude applications, but on the equatorial beta-plane the technical difficulties have so far been insurmountable.

The other four methods will be discussed in more detail in the following sections.

14.2.1 Power Series for the Eigenvalue

Boyd, Greatbatch, McPhaden and Knox and others have applied power series in the strength of the shear [18–21]. Boyd employed computer algebra to go to second order; thirty-six years later, modern computer algebra (CAS) allows one to go to much higher order. The critical latitude does not create any problems in computing terms of the series and the expansion is demonstrably asymptotic to the solution although such expansions are usually divergent. Unfortunately, even an expansion to the billionth-order would give no direct information about $\Im(c(\Gamma))$ for the Kelvin wave!

Observe that the power series for the exponential of a reciprocal of the small parameter is trivial and useless:

$$\left.\begin{array}{c}\exp(-q/\Gamma)\\ \exp(-q'/\Gamma^2)\end{array}\right\} \sim 0 + 0\Gamma + 0/\Gamma^2 + \cdots \tag{14.8}$$

where $q, q' > 0$ are constants. The reason is that

$$\lim_{\Gamma\to 0} \frac{d^k}{d\Gamma^k} \exp(-q/\Gamma) = 0, \qquad k = 0, 1, 2, \ldots \tag{14.9}$$

References [22, 23]. Since these derivatives are the coefficients of the Taylor series of $\exp(-q/\Gamma)$ about $\Gamma = 0$, (14.8) follows.

Exponentials-of-a-reciprocal are *invisible* to power series. The destabilizing, exponentially tiny $\Im(c)$ is said to lie "beyond-all-orders" in Γ. When beyond-all-orders effects appear, the power series not only miss the exponentially small terms but are usually divergent as well.

Padé approximants can cure the divergence [24], but Padé approximants derived from power series with real coefficients are real, too, for real Γ and thus also fail utterly to yield beyond-all-orders imaginary parts.

14.2.2 Hermite-Padé Approximants

Hermite observed in the 19th century that to approximate a function $f(x)$ one can define an approximation $f_{[K/L/M]}(z)$ to be the solution of the quadratic equation

$$P_K(z) f_{[K/L/M]}(z)^2 + R_L(z) f_{[K/L/M]}(z) + Q_M(z) = 0 \tag{14.10}$$

where the polynomials P_K, R_L, and Q_M of degrees K, L, and M are chosen so that the power series of $f_{[K/L/M]}(z)$ matches that of the "target" $f(z)$ through the first $K + L + M + 1$ terms. Such generalizations of ordinary Padé approximations are called "Hermite-Padé" approximants [15, 25, 26]. The solution to a quadratic equation can be complex even if its coefficients are real.

Boyd and Natarov showed that quadratic Hermite-Padé approximants are successful for the imaginary part of the eigenvalue of the Hermite-with-pole equation and also for the exponentially small imaginary part of the Gauss-Stieltjes integral,

$$\sigma(\varepsilon) \equiv \lim_{\delta \to 0} \int_0^\infty \exp\left(-t^2\right) \left\{ \frac{1}{t - [1 + i\delta]/\varepsilon} + \frac{1}{t + [1 + i\delta]/\varepsilon} \right\} dt \tag{14.11}$$

14.2.3 Numerical Methods

Exponentially small quantities are best computed by applying spectral methods, which have an exponential rate of convergence.

When the growth rate is very small so that the critical latitude is very near the real axis, disaster will strike any orthodox numerical method. Boyd showed, with later improvements by Gill and Sneddon [14, 16, 17, 27, 28], that the problem could be removed by using an integration in the complex plane that detours around

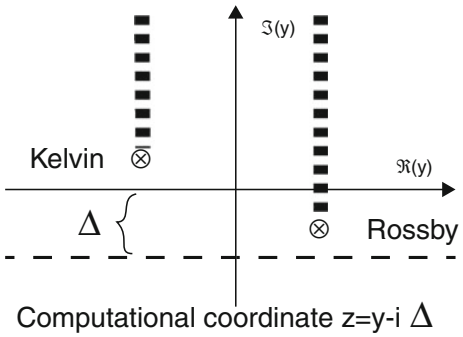

Fig. 14.2 Schematic of a numerical method which is applied in the computational coordinate $z \equiv y - i\,\Delta$ where $\Delta > 0$ is a real constant. The continuity of the numerical solution along the dashed computational path for real z implicitly forces the branch cuts (thick cross-hatched lines) to pass to infinity in the upper half-plane. The Rossby cut then crosses the real axis, forcing an unphysical jump for real y

the singularity in the lower half-plane. For Kelvin wave instability and mean shear profiles which generate only a single critical latitude, it suffices to displace the path downward by a constant as illustrated in Fig. 14.2.

The spectral series are in rational Chebyshev functions, $\mathrm{TB}(z;\, L)$, where $L > 0$ is a user-choosable constant, the "map parameter", and

$$z = y - i\,\Delta \tag{14.12}$$

where $\Delta > 0$ is a user-choosable constant. Table 14.2 and the references show that for sufficiently large truncation N of the spectral series, results are independent of N and the other numerical parameters. The computations for Fig. 14.4 and Table 14.2 were performed using a rational Chebyshev ("TB functions") pseudospectral method [28–30].

Each term in a TB series is bounded by its Fourier coefficient. Fig. 14.3 shows the exponential convergence of the coefficients of the rational Chebyshev series. Even though the smallest coefficient computed and illustrated for this truncation of $N = 125$ is only smaller than the imaginary part of the eigenvalue by about a factor of 250, this truncation gives $\Im(c)$ correct to the first ten decimal digits.

Table 14.2 Complex phase speeds for the Kelvin mode in linear shear $U(y) = \Gamma y$. Numerical parameters are also listed to demonstrate convergence to values independent of L, N, Δ and "Digits". (The last- named is the number of decimal digits of floating point precision, the Maple environmental parameter "Digits".) Correct digits are in boldface.

Γ	N	L	Digits		$\Re(c)$	$\Im(c)$
0.125	200	4	48	−1	**1.0077817**396586	**0.43616**3059753E-31
0.125	160	4	64	−1	**1.007781740**	**0.43616**29713E-31
0.15	200	4	64	−1	**1.011185987838818480755038636821**359	**0.2336**320418E-22
0.15	140	5	64	−1	**1.011185987838818480755038636821**439	**0.2336**17E-22
0.2	150	3	48	−1	**1.019795801358241207964480**24418165	**1.530896590**5197081E-14
0.2	125	4	48	−3/2	**1.019795801358241207964480**50	**1.53089659056**1E-14
0.3	120	4	48	−1	**1.043935871150942875754**3601	**0.55062692108551**9047284-7
0.3	100	3	48	−3/2	**1.04393587115094287575477**	**0.55062692108551**46E-7
0.3	80	4	16	−1	**1.043935871**	**0.550626921**3e-7
0.3	60	2	16	−1	**1.043935871**	**0.5506**18e-7
0.4	120	4	32	−1	**1.07646412260889623597798338265 75**	**0.000016874867741525008**595
0.4	100	3	32	−3/2	**1.0764641226088962359**78233818	**0.000016874 86774 15250 0** 76
0.5	120	3	32	−1	**1.1157527678438701943**4116480	**0.0003181293258021223 7**858
0.5	100	4	32	−2	**1.115752767843870194**323	**0.0003181293258021223 7** 28

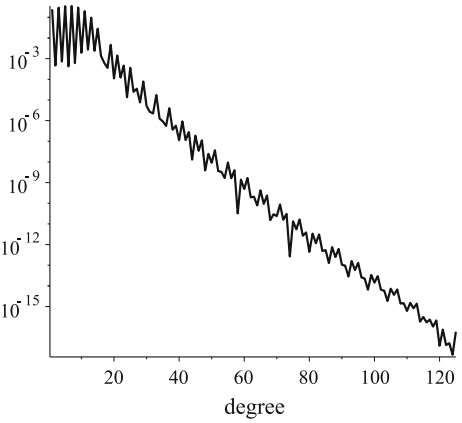

Fig. 14.3 Rational Chebyshev coefficients for the Kelvin wave for the shear strength $\Gamma = 1/5$. $\Delta = -3/2$, the Chebyshev map parameter $L = 4$ and computations were performed in 48 decimal digit precision in Maple. The eigenvalue was calculated as **1.01979580135824120796448050** + i **1.5308965905**61E-14 where all the digits in bold face are correct

14.3 Instability for the Equatorial Kelvin Wave in the Small Wavenumber Limit

Boyd and Christidis observed that one can eliminate the zonal wavenumber k as a parameter, greatly simplifying the numerical problem, when the zonal wavenumber is *small*. The reason is that when k is very small compared to one, the long wave approximation of assuming meridional geostrophy is accurate. (It is exact for the Kelvin wave in the absence of shear.) The y-momentum equation becomes

$$yu = -\phi_y \tag{14.13}$$

With $\aleph = \beth = 0$, the zonal momentum equation and height equation become

$$-ikU(y)\phi_y + ikc\phi_y + (U_y - y)yv + iky\phi = 0 \tag{14.14}$$

$$ikyU(y)\phi - ikyc\phi - ik\phi_y + yv_y = 0 \tag{14.15}$$

Now introduce

$$\tilde{v} \equiv v/k \tag{14.16}$$

The two equations now have k as a common factor of every term

$$-ikU(y)\phi_y + (U_y - y)yk\tilde{v} + iky\phi = -ikc\phi_y \tag{14.17}$$

$$ikyU(y)\phi - ik\phi_y + ky\tilde{v}_y = ikyc\phi \tag{14.18}$$

Dividing by k, we obtain an eigenproblem which is *independent* of k; the only remaining parameter is the shear:

$$-iU(y)\phi_y + (U_y - y)y\tilde{v} + iy\phi = -ic\phi_y \tag{14.19}$$

$$iyU(y)\phi - i\phi_y + y\tilde{v}_y = iyc\phi \tag{14.20}$$

14.3.1 Beyond-All-Orders Rossby Wave Instability

Is there a Rossby wave instability with growth rates exponentially small in the shear strength? Tung [31] shows that the answer is No! There is a neutrally-stable *continuous spectrum*.

When a numerical strategy — spectral, finite difference or finite volume or whatever — discretizes a differential equation into an algebraic eigenproblem, the *matrix* will necessarily only have discrete modes. For modes whose phase speed lies between the minimum and maximum of $U(y)$, and thus (unless unstable) have critical latitude singularities on the real axis, the matrix eigensolver obediently spits out a finite number of discrete eigenvalues in the range $c \in [\min(U), \max(U)]$ as well as discrete modes that are nonsingular because their velocity is outside this range of the mean flow. Varying N, the number of degrees of freedom, reveals that these discrete-but-continuous modes are not converged and indeed never converge as $N \to \infty$.

Kasahara [32] plots these spurious modes in his Fig. 5 anyway, but states in the caption, "The dots in the continuous spectrum have no physical significance, but they are plotted merely to show what Galerkin's method gives when one tries to represent singular, continuum modes in terms of a finite set of Hough functions." [32, 33].

As explained in [14], the complex integration path method of Boyd and Christidis seems to offer the ability to calculate discrete Rossby modes, but this is an illusion. Because the sign of the phase speed of Rossby waves is opposite that of Kelvin waves, the branch point for these discrete Rossby modes is in a different half-plane. The spectral method calculates solutions that are smooth on the numerical integration path. This implies that the branch cut for the Rossby waves must cross the real axis, implying that these modes have unphysical jumps on the real axis.

The Rossby spectrum is indeed continuous for $c \in [\min(U), \max(U)]$. There are no discrete modes except those whose phase speeds c lie outside the range of values of the mean wind.[1]

[1] For example, the zonal wavenumber one, lowest symmetric latitudinal free oscillation known as the "five day wave" is readily observed because its westward phase speed of 90 m/s greatly exceeds the magnitude of tropospheric mean winds [34].

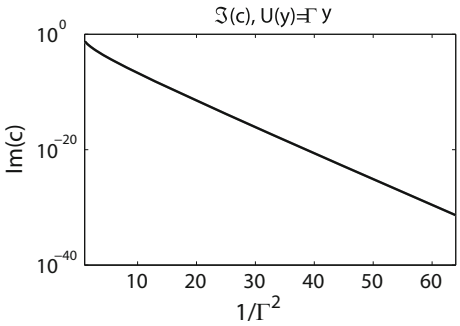

Fig. 14.4 The curve is actually the superposition of two curves which are visually indistinguishable. One connects the imaginary of the Kelvin phase as computed by a pseudospectral rational Chebyshev method for the mean current $U(y) = \Gamma y$ in the "long wave" approximation of Boyd and Christidis [2]. The other is the empirical fit $\Im(c) \approx 0.14 \exp(-1/\Gamma^2)/\Gamma^3$

14.3.2 Beyond-All-Orders Kelvin Wave Instability in Weak Shear in the Long Wave Approximation

Figure 14.4 shows the imaginary part of the phase speed as calculated by a very accurate pseudospectral method. Under the black curve is a red dashed curve which fits the numerical results so well that the red curve is invisible. The fitted curve is

$$\Im(c) \approx \frac{0.14}{\Gamma^3} \exp\left(-\frac{1}{\Gamma^2}\right) \tag{14.21}$$

The $0.14/\Gamma^3$ is empirical; a future WKB/matched asymptotics theory would presumably establish this or something like it on a firm deductive basis. The exponential factor is consistent with the behavior of the eigenvalue of the Hermite-with-pole equation: the growth rate is proportional to the square of the amplitude of the Kelvin mode at the critical latitude.

For further graphs and discussion, see Boyd and Christidis [2, 14].

14.4 Kelvin Instability in Shear: The General Case

Although primarily focused on critical latitude effects on (neutrally-stable) midlatitude Rossby waves, Boyd [35] does contain one section on equatorial waves in a mean flow that varies with latitude only. Fig. 13 of [35] shows a Kelvin wave in moderate shear, calculated in spherical coordinates; the critical latitude is only five degrees from the equator, but the imaginary part of the phase speed is very small.

Two features are noteworthy. Rossby waves, at least the low order modes, are one-sided; there is no amplitude for $U(y) - c < 0$ except an exponentially decaying tail with zero latitudinal fluxes (Boyd [10], gives further details) [10, 35]. In contrast, the Kelvin mode has a large secondary lobe on the poleward side of the critical latitude.

Kelvin waves transfer positive momentum to the mean flow when absorbed near a critical latitude. Rossby waves have opposite phase speed and acceleration. The net acceleration of the mean flow is negative equatorward of the critical latitude, but there is a negative acceleration spike just poleward of the critical latitude. Lindzen observed that the mode behaves like a fusion of Kelvin and Rossby modes (private communication, 1975). There is no easy way to quantify such insights. However, resonances between pairs of contra-rotating modes have become the primary theoretical paradigm for barotropic and baroclinic instability. Perhaps a reader will devise a similar theory for Kelvin/Rossby resonance instability.

References

1. Rota G (2008) Indiscrete thoughts, 2nd edn. Birhauser, Boston, p 280
2. Boyd JP, Christidis ZD (1982) Low wavenumber instability on the equatorial beta- plane. Geophys Res Lett 9:769–772
3. Boyd JP (1999) The devil's invention: asymptotics, superasymptotics and hyperasymptotics. Acta Appl 56(1):1–98
4. Boyd JP (1998) Weakly nonlocal solitary waves and beyond- all-orders asymptotics: generalized solitons and hyperasymptotic perturbation theory, mathematics and its applications, vol 442. Kluwer, Amsterdam, p 608
5. Segur H, Tanveer S, Levine H (eds) (1991) Asymptotics beyond all orders. Plenum, New York, p 389
6. Boyd JP (2005) Hyperasymptotics and the linear boundary layer problem: why asymptotic series diverge. SIAM Rev 47(3):553–575
7. Lindzen RS (1988) Instability of plane parallel shear flow (toward a mechanistic picture of how it works). Pure Appl Geophys 126(1):103–121
8. Adam JA (1986) Critical layer singularities and complex eigenvalues in some differential equations of mathematical physics. Phys Rep 142(5):263–356
9. Dickinson RE (1968) Planetary Rossby waves propagating vertically through weak westerly wind wave guides. J Atmos Sci 25:984–1002
10. Boyd JP (1981) A Sturm-Liouville eigenproblem with an interior pole. J Math Phys 22:1575–1590
11. Lindzen RS, Rosenthal AJ (1981) A WKB asymptotic analysis of baroclinic instability. J Atmos Sci 38:619–629
12. Boyd JP, Natarov A (1998) A Sturm-Liouville eigenproblem of the fourth kind: a critical latitude with equatorial trapping. Stud Appl Math 101:433–455
13. Dickinson RE (1970) Development of a Rossby wave critical level. J Atmos Sci 27(4):627–633
14. Boyd JP, Christidis ZD (1983) Instability on the equatorial beta-plane. In: Nihoul J (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 339–351
15. Natarov A, Boyd JP (2002) Shafer (Hermite-Padé) approximants for functions with exponentially small imaginary part with application to equatorial waves with critical latitude. Appl Math Comput 125:109–117
16. Gill AW, Sneddon GE (1995) Complex mapped matrix methods in hydrodynamic stability problems. J Comput Phys 122:13–24

17. Gill AW, Sneddon GE (1996) Pseudospectral methods and composite complex maps for near-boundary critical latitudes. J Comput Phys 129(1):1–7
18. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part I: theory and methods. J Atmos Sci 35:2236–2258
19. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, part II: applications to the atmosphere. J Atmos Sci 35:2259–2267
20. Greatbatch RJ (1985) Kelvin wave fronts, Rossby solitary waves and nonlinear spinup of the equatorial oceans. J Geophys Res 90:9097–9107
21. McPhaden MJ, Knox RA (1979) Equatorial Kelvin and inertia-gravity waves. J Phys Oceanogr 9(2):263–277
22. Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
23. Olver FWJ, Lozier DW, Boisvert RF, Clark CW (eds) (2010) NIST handbook of mathematical functions. Cambridge University Press, New York
24. Bender CM, Orszag SA (1978) Advanced mathematical methods for scientists and engineers. McGraw-Hill, New York, p 594
25. Baker GA Jr, Graves-Morris P (1996) Padé Approximants, part II, extensions and applications, encyclopedia of mathematics, vol 14, 2nd edn. Cambridge University Press, Cambridge, p 762
26. Boyd JP (2002) Chebyshev polynomial expansions for simultaneous approximation of two branches of a function with application to the one-dimensional Bratu equation. Appl Math Comput 143:189–200
27. Boyd JP (1985) Complex coordinate methods for hydrodynamic instabilities and Sturm-Liouville problems with an interior singularity. J Comput Phys 57:454–471
28. Boyd JP (2001) Chebyshev and fourier spectral methods. Dover, New York, p 680
29. Boyd JP (1987) Spectral methods using rational basis functions on an infinite interval. J Comput Phys 69:112–142
30. Boyd JP (1987) Orthogonal rational functions on a semi- infinite interval. J Comput Phys 70:63–88
31. Tung K (1983) Initial-value problems for Rossby waves in a linear flow with critical level. J Fluid Mech 133:443–469
32. Kasahara A (1980) Effect of zonal flows on the free oscillations of a barotropic atmosphere. J Atmos Sci 37(5):917–828
33. Kasahara A (1981) Corrigendum: effect of zonal flows on the free oscillations of a barotropic atmosphere. J Atmos Sci 37:3284–3285
34. Madden R (2007) Large-scale, free Rossby waves in the atmosphere — an update. Tellus A 59(5):571–590 (correnction in the same journal, vol. 60 (2008), pg. 394)
35. Boyd JP (1982) The effects of meridional shear on planetary waves, part II: critical latitudes. J Atmos Sci 39:770–790

Chapter 15
Nonmodal Instability

Abstract The classic theory of Orr, Farrell, Trefethen, Boyd and others for nonmodal growth and decay in non-equatorial flow is reviewed. Long-established arguments show that nonmodal perturbations must always decay as $t \to \infty$. However, plane Couette flow is always unstable in the laboratory in a broad range of Reynolds numbers where all normal modes are stable. Orr pointed out in 1907 that nonmodal disturbances may exhibit a transient growth that may push their amplitudes to such large values that secondary instabilities are triggered, leading to self-sustained turbulence. Farrell greatly generalized Orr's theory to baroclinic flows, with and without sidewalls. Trefethen and collaborators invented toy models that rigorously demonstrate Orr's conjecture, and also related transient growth to nonnormality of discretization matrices and from thence to similar properties in the differential operators they represent. How do these scenarios play out in the tropics? Boyd and Christidis have made a tentative beginning at answering this question, but this chapter is less a review of equatorial dynamics than a research agenda.

> It accordingly appears that, in this simple case, although the disturbance, if sufficiently small, must ultimately decrease indefinitely, yet, before doing so, it may be very much increased ... [Indeed] the ratio of increase may be made as great as we like.
>
> William McFadden Orr (1866–1934) (1907, p. 32 [1]).

15.1 Introduction

Linearized stability theory has traditionally focused almost exclusively on *discrete, exponentially-growing normal modes*. However, the general initial-value solution to linearized equations with mean shear flow has *three* parts:

1. Discrete neutral modes
2. Unstable normal modes with exponential growth [and a complex conjugate mode which decays exponentially]
3. A continuous spectrum of waves.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_15

The discrete neutral modes are stable and will be not be discussed here. However, the continuous spectrum or "continuum modes" can trigger instability.

The discrete normal modes are collectively a sum. The continuous spectrum cannot be described as being proportional to $\exp(ik[x - ct])$ except as an integration over the phase speed c with limits ranging from the minimum value of the mean current $U(y)$ to the maximum value of $U(y)$. All the continuum modes are singular at the critical latitude where the phase speed c equals $U(y)$. Tung [2] and Kasahara [3] discuss the theory and numerics of the continuous spectrum.

Sir William Thomson [27], five years before he became the first Lord Kelvin, showed that as $t \to \infty$, the *continuous spectrum always decays*. Consequently, it was long assumed that for instability theory, only *discrete modes* mattered.

However, William McFadden Orr [1] showed that there could be a *transient growth* stage before the final decay begins. He deduced, correctly, that sometimes the transient would become so large that the linearization would fail, and nonlinearity could lead to self-sustaining turbulence.

In this chapter, we give an introduction to his ideas. Modern thinking is that the "Orr Mechanism" is fundamental not only to turbulence initiated by the continuous spectrum, but plays an essential role in some types of exponentially-growing normal mode instability [4]. We shall end by describing early work on nonmodal instability on the equatorial beta-plane by Boyd and Christidis. However, the role of the continuous spectrum in instability at low latitudes is still but imperfectly understood.

15.2 Couette and Poiseuille Flow and Subcritical Bifurcation

One motive for the continued interest in Orr's work is the failure of normal mode instability for two widely studied mean flows illustrated schematically in Fig. 15.1.

Definition 15.1 (*Plane Couette flow*) A flow problem in which the mean flow varies linearly between two infinite parallel plates. These plates are in steady, relative motion, driving the flow.

Definition 15.2 (*Plane Poiseuille flow*) The steady flow between two parallel plates is Poiseuille flow. (Cylindrical Poiseuille flow is the flow in an infinitely long straight pipe). The flow is driven by a steady pressure gradient. The velocity profile is a parabola — maximum at the midpoint between the boundaries and zero on all boundaries.

Linearized stability theory predicts that for *infinitesimal amplitude perturbations*:

- plane Poiseuille stable for all $Re < 5772$
  Experimentally, unstable for $Re > 1000$[1]

[1] $Re > 2000$ in really careful experiments.

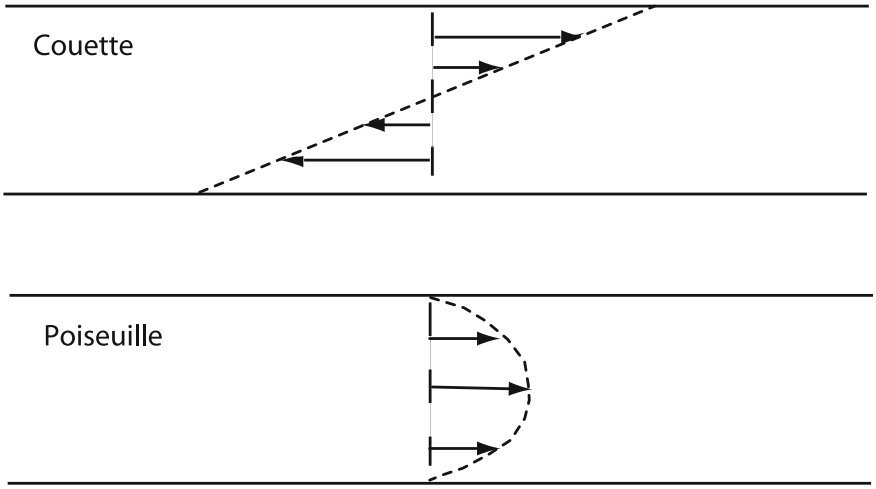

Fig. 15.1 Velocity profiles for a pair of much-studied two-dimensional flows. *Top* Couette flow, which is a steady flow which varies linearly with the cross-channel direction, driven by moving walls. *Bottom* Poiseuille flow, which is a steady flow whose velocity is a parabola, zero at both rigid walls

- plane Couette stable for *all* Re
  Experimentally, unstable for $Re > 350$

These two flows are said to exhibit "subcritical bifurcations" in the sense that instability due to *finite amplitude* perturbations arises for parameter values where all *infinitesimal amplitude* perturbations are stable (Chapman [5], Schmid [6]). In contrast, baroclinic instability is a supercritical bifurcation. Small finite amplitude perturbations grow only when the vertical shear is larger than the critical value for infinitesimal perturbations.

Subcritical bifurcations are a great challenge to experimentalists. Fitzgerald [7] gives a good review of the work of Hof, Treacher and Mullin, who built a pipe almost sixteen meters long, 785 times the diameter of the pipe, to minimize endpoint effects and maximize experimental control. They defined the amplitude of the perturbation as the perturbing injected flux divided by the steady flux down the pipe. A relative perturbation of only 0.007 was sufficient to trigger instability at a Reynolds number of 2000; the threshold varied inversely with amplitude. Earlier experimenters had a terrible time getting even close to the critical point because even very tiny perturbations can trigger turbulence. The dismal failure of linear normal mode instability theory is often labeled the "Sommerfeld Paradox".

15.3 The Fundamental Orr Solution

Orr offered an explanation for subritical bifurcation through a (non-equatorial) explicit solution. The model is the inviscid barotropic vorticity equation on an *infinite*

domain on the *midlatitude* f-plane [Cartesian coordinates with constant Coriolis parameter f] linearized about a zonal mean shear flow that varies with latitude only, $U(y)$. For simplicity, $\beta = 0$, but Tung [2] and Boyd [8] have shown that the beta-effect is secondary. The essential physics is that vorticity ζ is conserved following the motion:

Theorem 15.1 (Orr Solution in an Unbounded Domain) *1. The general solution to*

$$\frac{\partial \zeta}{\partial t} + U(y)\frac{\partial \zeta}{\partial x} = 0, \qquad \zeta = \nabla^2 \psi \tag{15.1}$$

is

$$\zeta(x, y, t) = \zeta(x - U(y)t, y, t) \tag{15.2}$$

2. If the mean current is written as $U(y) = S\tilde{U}(y)$ where $S = \max|u(y)|$, the maximum mean current in the domain, the solution is a function of St, and not S or t separately, so that the role of S is merely to alter the timescale without changing anything else. This theorem is true for general initial conditions.

3. Linear Couette Flow is the special case

$$U(y) \equiv S\, y, \qquad S \text{ is a constant} \tag{15.3}$$

A useful special solution is the "Advected Cosine":

$$\begin{aligned} \zeta \equiv \Lambda(x, y, t, k, S, \phi) &= \cos(kx - kS[t - \phi]y) \\ &= \{\cos(kx)\cos(kS[t + \phi]y) + \sin(kx)\sin(kS[t - \phi]y)\} \end{aligned} \tag{15.4}$$

where ϕ is an arbitrary constant phase factor. The corresponding streamfunction is

$$\psi = -\frac{\zeta}{k^2 + k^2S^2[t - \phi]^2} \tag{15.5}$$

As $t \to \infty$, $\psi \sim O(1/t^2)$ while the zonal velocity $u = -\psi_y$ decays as $O(1/t)$.

This gives a two-parameter family of solutions where the zonal wavenumber k and the phase constant ϕ are effectively the only parameters.

Proof The first and third propositions follow by substituting the assumed solution into the equation and canceling common factors until it is clear that the function is indeed a solution. The second proposition is proved by making the change of coordinate $\tau \equiv St$, which transforms the problem to $\zeta_\tau + y\zeta_x = 0$. This is independent of the shear strength S. ■

15.4 Interpretation: The "Venetian Blind Effect"

The rotation of isovorts in the "Advected Cosine" resembles the opening and closing of the window treatment known as "venetian blinds". When the slats are parallel, light can stream into the room, but the room can be darkened by rotating the slats until they touch one another. The isovorts at different times for the "Advected Cosine" solution very much resemble a cross-section of the slats in venetian blinds as illustrated in Fig. 15.2.

If the phase constant ϕ is chosen so that at $t = 0$, the contours of constant vorticity are tilted *against* the shear, that is, run northwest-to-southeast when dU/dy is positive, then the zonal velocity and streamfunction will amplify until the isovorts are rotated to run due north-south. For larger times, the isovorts continue to rotate clockwise (as viewed from above) when $S > 0$ and the streamfunction and velocity decay.

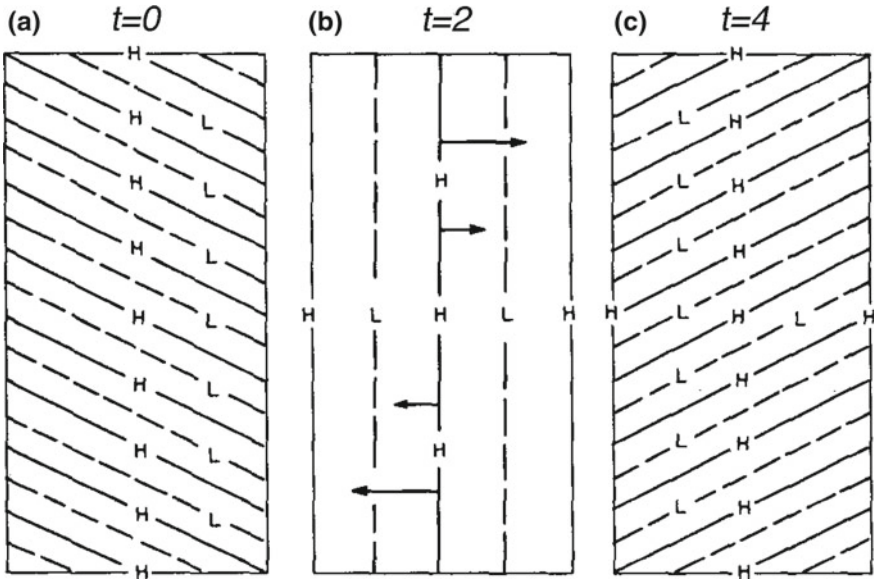

Fig. 15.2 Isovorts (contours of constant vorticity) for a plane wave solution to the barotropic vorticity equation, linearized about a mean shear flow. The *arrows* in the *middle panel* show the mean, advecting velocity. The isovorts rotate about ($x = 0$, $y = 0$) [where the mean flow is zero and particles do not move]. When the isovorts are tilted against the mean shear (*left panel*), the streamfunction and velocities AMPLIFY until a maximum amplitude is reached when the isovorts are parallel to the y-axis (*middle panel*). The streamfunction and velocities DECAY when the tilt is WITH the SHEAR. The phase ϕ is chosen so that at $t = 0$, the tilt is against the shear. ($S = 1$, $\beta = 0$) From Boyd [8]

The vorticity is conserved and therefore its magnitude never changes. However, the Poisson equation connecting ψ and ζ implies that for a sinusoidal vorticity, the streamfunction is ζ/κ^2 where $\kappa = \sqrt{k^2 + m^2}$ is the total wavenumber. The zonal wavenumber k is invariant under advection by a zonal mean wind. However, the latitudinal wavenumber m changes with time. When the isovorts rotate parallel to a meridian, the meridional wavenumber m reaches a minimum of zero, and the streamfunction is a maximum. For very large time, the continued clockwise rotation of the isovorts (for $S > 0$) causes the north-south wavenumber m to grow without bound, so κ grows without bound and the streamfunctions decays proportional to $1/\kappa^2$.

15.5 Refinements to the Orr Solution

Boyd [9], Tung [2] and Yamagata [10] generalized Orr's solution to include the beta-effect. This produces a westward movement at a rate which varies with time because the north-south wavelength varies with time. However, the beta-effect does not modify the growth and decay of the waves.

Farrell [11] showed that adding sidewalls requires thin irrotational boundary layers to enforce the boundary conditions. However, again growth and decay are unaltered.

Farrell [11–13] and Butler and Farrell [14] showed that the Orr mechanism works similarly when the shear is vertical instead of horizontal, bounded only by a lower boundary in a semi-infinite atmosphere.

There is nothing sacred about trigonometric functions as shown by the following.

Theorem 15.2 (Streamfunction Rotation) *Let $f(z)$ denote a univariate function and let $g(z)$ be a bounded solution to $g_{zz} = -f$. Let (X, Y) denote a new, rotated coordinate such that the Y axis is the result of rotating the y axis by an angle Θ in a clockwise direction. If the angle Θ is chosen so that $\tan(\Theta) = (t + \phi)$, then the contours of the sheared function will all parallel the rotated Y-axis. Then*

$$\zeta = f(x - y[t + \phi]) = f\left(\frac{1}{\cos(\Theta)} X\right) \tag{15.6}$$

where ϕ is an arbitrary constant phase and the corresponding solution for the streamfunction is

$$\Delta\psi = -\zeta, \qquad \psi = \frac{1}{1 + [t + \phi]^2}\, g\left(\frac{1}{\cos(\Theta)} X\right) \tag{15.7}$$

Proof From any elementary physics or computer graphics book

$$\begin{vmatrix} X \\ Y \end{vmatrix} = \begin{vmatrix} \cos(\Theta) & -\sin(\Theta) \\ \sin(\Theta) & \cos(\Theta) \end{vmatrix} \begin{vmatrix} x \\ y \end{vmatrix} \quad \Leftrightarrow \quad \begin{vmatrix} x \\ y \end{vmatrix} = \begin{vmatrix} \cos(\Theta) & \sin(\Theta) \\ -\sin(\Theta) & \cos(\Theta) \end{vmatrix} \begin{vmatrix} X \\ Y \end{vmatrix} \tag{15.8}$$

The line implicitly defined by $x - yt = 0$ has a slope equal to $1/t$. If we draw a triangle with unit hypotenuse, one vertex at the origin, and two sides that lie on the old and new y axes, the vertex on the X axis is $(x = \sin(\Theta), y = \cos(\Theta))$; the line from the origin to this point has a slope of $1/\tan(\Theta)$. This requires that $\tan(\Theta) = t$. Then

$$\begin{aligned} z &\equiv x - yt && (15.9)\\ &= \cos(\Theta)X + \sin(\Theta)Y - [-\sin(\Theta)X + \cos(\Theta)Y]t && (15.10)\\ &= [\cos(\Theta) + \sin(\Theta)t]X + [\sin(\Theta) - \cos(\Theta)t]Y && (15.11)\\ &= [\cos(\Theta) + \sin(\Theta)\tan(\Theta)]X + [\sin(\Theta) - \cos(\Theta)\tan(\Theta)]Y && (15.12)\\ &= \left[\frac{1}{\cos(\Theta)}\cos^2(\Theta) + \sin^2(\Theta)\frac{1}{\cos(\Theta)}\right]X && (15.13)\\ &= \frac{1}{\cos(\Theta)}X && (15.14) \end{aligned}$$

It is explained in classes in computer graphics that a shearing transformation is not merely a *rotation* but rather a combination of this with a *deformation*. Because the function $f(z)$ is univariate, the deformation is a simple dilation [stretching] by the factor of $1/\cos(\theta)$. The form of the Laplace operator is left unaltered by rotation: For any smooth function $u(x, y)$,

$$\partial_{xx} + \partial_{yy} = \partial_{XX} + \partial_{YY} \tag{15.15}$$

Nevertheless, the shearing transformation amplifies the streamfunction by the square of the dilation factor. This factor can be expressed in terms of t via a trigonometric identity:

$$\begin{aligned} \frac{1}{\cos^2(\Theta)} &= 1 + \tan^2(\Theta) && (15.16)\\ &= 1 + [t + \phi]^2 && (15.17) \end{aligned}$$

■

Thus, a single ridge of vorticity described by $f = \exp(-z^2)$ will experience the same transient growth as the "Advected Cosine" if aligned against the shear at the same angle.

15.6 The "Checkerboard" and Bessel Solution

The Advected Cosine solution shows clearly that temporary growth is possible even in a flow, such as linear Couette flow, which has no unstable normal modes. It is surely true that if a wave disturbance grows sufficiently large that linearization about the mean flow is no longer trustworthy, then linear theory cannot disprove the possibility that the linearly stable flow will achieve self-sustaining turbulence.

However, this is a long way from proving that nonmodal instability is significant. Boyd [8] pointed out that an initial perturbation tilted against the flow is rather artificial; a more realistic initial perturbation is a standing wave in latitude, that is, a superposition of waves tilted both against the shear and with the shear.

15.6.1 The "Checkerboard" Solution

He was thus motivated to consider the initial condition

$$\zeta(x, y, 0) = \cos(kx)\ \cos(my) \tag{15.18}$$

The exact (linearized) solution is, with $\Lambda(x, y, t, k, S, \phi) \equiv \cos(kx - kS[t - \phi])$

$$\zeta(x, y, t) = \frac{1}{2}\{\Lambda(x, y, t, k, S, m/(kS)) + \Lambda(x, y, t, k, S, -m/(kS))\} \tag{15.19}$$

This is illustrated in Fig. 15.3.

Theorem 15.3 (Growth of the Checkerboard)

1.

$$\frac{\max|\psi(t = m/(kS))|}{\max|\psi(t = 0)|} \approx \frac{1}{2}\left(1 + \frac{m^2}{k^2}\right) \tag{15.20}$$

2. Large growth requires

$$m >> k \tag{15.21}$$

(Boyd [8])

Shepherd [15] and Boyd and Christidis [16] showed that if the initial condition was a broad distribution of wavenumbers, growth would be *small*. Growth requires a *latitudinal scale* much *smaller* than the *zonal scale*.

Shepherd used Bessel functions to examine what would happen if the initial disturbance was a superposition of many wavenumbers, restricted to certain sectors of the $k - m$ plane.

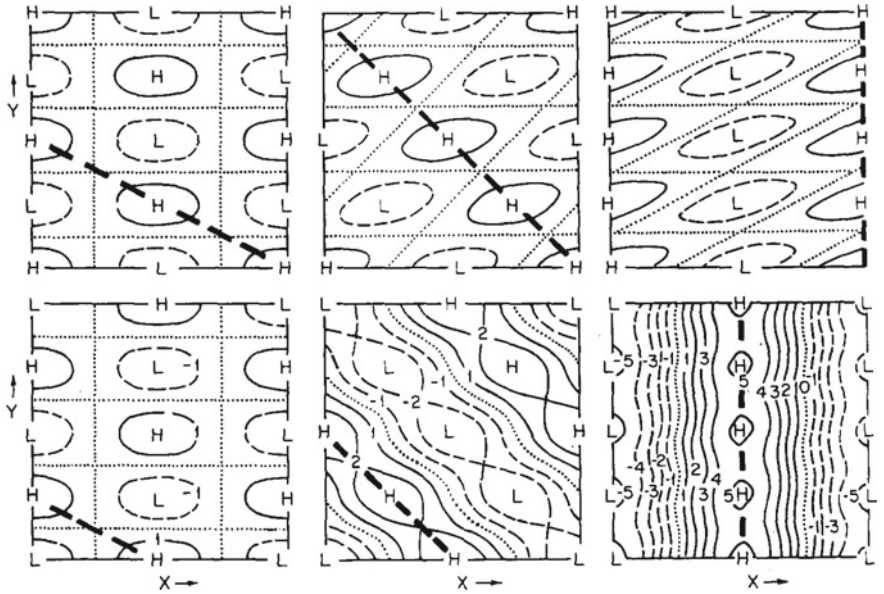

Fig. 15.3 The vorticity (*top three panels*) and streamfunction (*lower three panels*) for a standing wave initial condition ("Checkerboard Solution") at $t = 0$, $t = 1$ and $t = 2$ with zero beta effect, $S = 1$, and $m/k = 2$. The heavy *dashed lines* connect the same set of high vorticity or streamfunction centers throughout the evolution of the flow. The advection causes the heavy dotted lines to rotate clockwise until, at $t = 2$ [*right panel*], these lines parallel the y axis to produce a state reminiscent of the middle ($t = 0$) panel of Fig. 1. The number of streamfunction contours increases from *left-to-right* in the *lower panels* because the streamfunction amplifies by approximately a factor of 5. From Boyd [8]

In spite of the mathematical rigour and plausibility of the critiques of Boyd, Shepherd and Christidis, Farrell's revival of Orr's ideas has made many converts to the notion that nonmodal instability is the reason that plane Couette and Poiseuille flows are experimentally turbulent for Reynolds numbers where normal mode theory predicts stability. Boberg and Brosa [17], Reddy and Henningson [18], and Gustavsson [19], who independently arrived at similar ideas, helped to propagate these concepts in engineering fluid mechanics.

There are several developments that have strengthened this argument:

1. The Dandelion Strategy
2. Transient Growth of Three-dimensional Perturbations
3. Rigorous Analysis of Nonlinear ODE Systems and the Importance of Nonnormal Matrices

15.7 The Dandelion Strategy

One thing that literature would be greatly the better for
Would be a more restricted employment by the authors of simile and metaphor.…
What does it mean when we are told
That the Assyrian came down like a wolf on the fold?
In the first place, George Gordon Byron had enough experience
To know that it probably wasn't just one Assyrian, it was a lot of Assyrians.

Ogden Nash in his poem *Very Like a Whale*

A flow is perturbed not by a single perturbation, but by a lot of perturbations in succession. It is unlikely that a single perturbation will be strongly tilted against the shear. It is obviously much more likely that such a wave will be found in an ensemble of a hundred thousand perturbations than just one. What seems to happen is that sooner or later, a perturbation that can grow by orders of magnitude occurs, and then — kaboom!

In a sense, fluid mechanics is as profligate of perturbations as nature is with seeds. A dandelion scatters its seeds on the wind, and almost all perish. But one seed that grows another dandelion is enough to sustain the species. One perturbation out of thousands is enough to trigger self-sustaining turbulence.

15.8 Three-Dimensional Transients

Squire's Theorem (1933) asserts that for *exponentially growing normal modes* in a non-rotating inviscid, *two-dimensional* flow perturbations *always grow faster* than *three-dimensional* perturbations. This focused attention for decades on two-dimensional eigenmodes called "Tollmien-Schlicting" modes, which have turned out to be remarkably illusive in laboratory experiments. However, Squire's Theorem does *not* apply to the continuous spectrum.

Indeed, a common feature of turbulence is streamwise vortices, that is, vorticity whose vorticity vector is parallel to the mean flow as illustrated in Fig. 15.4. These are extremely effective in generating streamwise streaks of anomalous velocity which are also ubiquitous in experimental turbulence. Butler and Farrell [14] showed that *three-dimensional* perturbations to plane Couette flow could grow by over a factor of a *thousand*. Chapman [5] provides both an excellent review and an improved theory for the linear development of streamwise velocity streaks, [5].

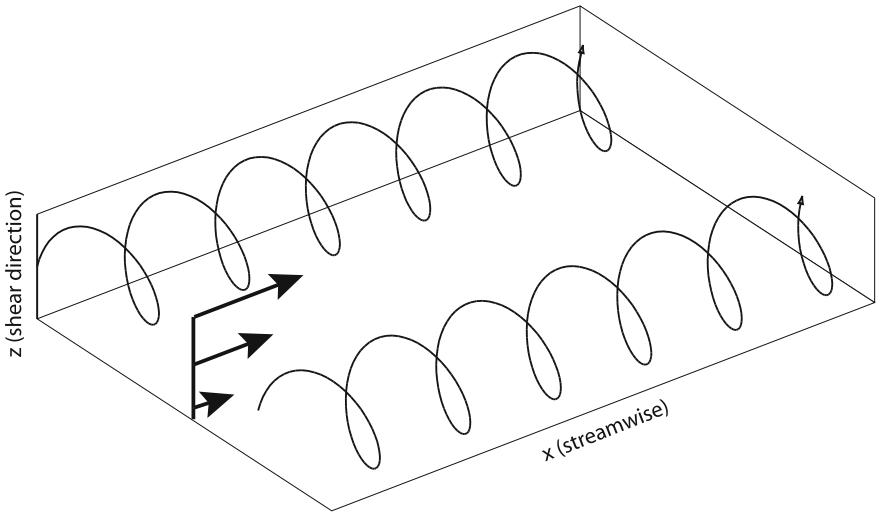

Fig. 15.4 Schematic of streamwise vortices (helical trajectories) in a vertically sheared flow in the x-direction, indicated by the thick lines and arrows

15.9 ODE Models and Nonnormal Matrices

The six page article by Trefethen, Trefethen, Reddy and Driscoll [20] is a very good review that also invented a coupled pair of nonlinear ODEs that evolve to self-sustaining chaos ("turbulence") from a tiny perturbation. Although linearized dynamics shows that the perturbation must always decay as $t \to \infty$ in the linearized approximation, the nonlinear dynamics is that the perturbation is amplified in a transient stage until secondary instabilities initiate the unending chaos.

Trefethen and collaborators noted that nonmodal instability can be greatly generalized (and divorced from an exclusively fluid mechanics context). Observe that linearizing about a time-independent basic state and discretizing the spatial dependence will always yield a linear system of coupled ODEs in time for the evolution of the perturbation $\boldsymbol{u}$

$$\frac{d\boldsymbol{u}}{dt} = \boldsymbol{A}\,\boldsymbol{u} \tag{15.22}$$

where $\boldsymbol{A}$ is a square matrix whose elements are independent of time. This is true regardless of whether the original nonlinear problem has any connection to fluid mechanics or not; it is true whether the spatial discretization is finite difference, spectral or finite volume. The linearized system is known in dynamical systems theory as the "tangent model" [to the nonlinear system] and its general solution is the "tangent propagator" or "tangent propagator".

The behavior of this abstract, linear instability problem depends tremendously on whether the matrix $\boldsymbol{A}$ has the property of being *normal*. Whenever $\boldsymbol{A}$ is normal,

nonmodal instability is impossible. The formal definition of this crucial matrix property is the following.

Definition 15.3 (*Normal/Nonnormal*) An $N \times N$ matrix is "normal" if it has N orthogonal eigenvectors, equivalent to diagonalizability. A matrix that isn't normal is "nonnormal".

Normality is important because of the following.

Theorem 15.4 (Uncoupling and Energy Norm) *1. If the matrix* $\boldsymbol{A}$ *is normal, then by making a change-of-basis for* N*-dimensional space to the eigenvector basis,*

$$\boldsymbol{u} = \boldsymbol{Q}\,\boldsymbol{a} \qquad \leftrightarrow \qquad \boldsymbol{a} = \boldsymbol{Q}^*\,\boldsymbol{u} \tag{15.23}$$

the system of ODEs is transformed to the uncoupled *system*

$$\frac{d\boldsymbol{a}}{dt} = \boldsymbol{\Lambda}\,\boldsymbol{a} \tag{15.24}$$

or equivalently in components

$$\frac{da_j}{dt} = \lambda_j a_j \tag{15.25}$$

with the exact solution

$$a_j(t) = a_j(0)\ \exp(\lambda_j t) \tag{15.26}$$

where $a_j(0) = Q_j^* \cdot u_0$ *and where* $u(t=0) = u_0$.

2. The "energy norm" of the system $||\boldsymbol{u}||^2$ *is*

$$||\boldsymbol{u}||^2 = \sum_{j=1}^{N} [a_j(0)]^2\ \exp(2\lambda_j t) \tag{15.27}$$

This implies the following.

Theorem 15.5 (Stability of a System with a Normal Matrix) *Let* $\boldsymbol{A}$ *be a normal matrix with elements which are independent of time. Assume the conditions and conclusions of the previous theorem apply. Represent the solution in the eigenvector basis as* $\boldsymbol{u} = \sum_{j=1}^{N} a_j(t)\,\boldsymbol{Q}_j$ *with the* $\boldsymbol{Q}_j$ *as the eigenvectors of* $\boldsymbol{A}$*. Then if*

$$\Re(\lambda_j) < 0 \qquad \textit{for all } j \tag{15.28}$$

it follows that

$$\frac{d||\boldsymbol{u}||^2}{dt} < 0 \tag{15.29}$$

and the amplitudes of $\boldsymbol{u}$ *in the eigenvector basis,* $a_j(t)$*, decay monotonically with time proportionally to* $\exp(-|\Re(\lambda_j)|)$*, that is,*

$$\frac{d|a_j|}{dt} < 0 \tag{15.30}$$

Thus, when the matrix of a constant coefficient linear system of ODEs is normal, stability theory is very boring and straightforward, and completely controlled by the eigenvalues.

Some types of matrices are always normal.

Theorem 15.6 (Matrix Categories That Are Normal and Diagonalizable) *The following categories of matrices are always unitarily diagonalizable and normal:*

1. *real symmetric*
2. *Hermitian*
3. *skew-symmetric*
4. *skew-Hermitian*
5. *unitary*
6. *circulant*
7. *any of the first four plus a multiple of the identity matrix*

Proof pg. 187 of [26].

Unfortunately, the matrices that arise in hydrodynamic stability are often very, very nonnormal, allowing spectacular transient instability. Chapter 7 of the book by Trefethen and Embree [21] is a good description of this general theory of nonmodal/nonnormal dynamics as applied specifically to fluid mechanics.

15.10 Nonmodal Instability in the Tropics

Boyd and Christidis [16] is the only study of transient growth on the equatorial beta-plane. They considered only the barotropically-stable linear shear $U(y) = Sy$ where S is a constant and U is the mean zonal current. They found the following notions from earlier, midlatitude investigations of Yamagata [10], Tung [2], Farrell [11] and Boyd [8] are still true on the equatorial beta-plane:

1. The Rossby wave spectrum in a linear shear is *continuous*, not discrete.
2. Lines of constant phase rotate clockwise as viewed from above for $S > 0$.
3. The latitudinal scale tends to zero as $t \to \infty$.
4. There is little latitudinal spreading of the initial disturbance.
5. A latitudinally-concentrated wavepacket will move north while growing and will move south while decaying ($S > 0$); the southward motion will eventually cease, leaving the packet "stalled out" at a particular latitude.

The following conclusions are modified or qualified:

1. The wave energy decays algebraically with time as $1/t^2$ as $t \to \infty$.

2. A sinusoidal wave tilted against the shear (phase lines running northwest-southeast for $S > 0$) will amplify for a finite time before entering the asymptotic decay stage.

The first part of their study was to solve the shallow water equations, linearized about a linear shear flow, for initial conditions that were a Rossby wave of zonal wavenumber k and latitudinal mode number n as given by the usual formulas for the eigenmodes of an ocean with no mean flow. The rationale is that gravity modes are not drastically modified by the shear, so the no-mean-flow Rossby initial condition would still excite little gravity wave activity. Higher mode Rossby waves are better and better described, as the mode number n increases, by the *midlatitude* beta-plane over most of the latitudinal range where they have significant amplitude; this was confirmed by numerical experiments. Therefore the interesting question is: How do the *lowest* couple of modes behave?

The answer is that equatorial Rossby waves behave very much like waves on the midlatitude beta-plane except for the two differences noted above. The disturbance moves south of the equator (if $S > 0$), narrows in latitudinal scale, and tilts more and more with the shear as illustrated in Fig. 15.5. The first alteration is that the energy decays with time for large time at a rate intermediate between $1/t$ and $1/t^2$. The slope of decay is monotonically increasing, so there is a suggestion that for *very* large times, the expected $1/t^2$ decay occurs, but even for a nondimensional time of 200, which is much longer than the trans-Pacific crossing time for the $n = 1$ and $n = 2$ Rossby waves, the observed decay is enigmatically slower. The second unexpected finding is that there is no sign of transient growth.

These puzzling differences motivated Boyd and Christidis to retreat to the midlatitude beta-plane where the linearized streamfunction equation is

$$\psi_{xxt} + \psi_{yyt} + U(y)\left(\psi_{xxx} + \psi_{yyx}\right) + \left(\beta - U_{yy}\right)\psi_x = 0 \tag{15.31}$$

The initial condition was

$$\psi(x, y, t = 0) = \exp(ikx)\ \exp(-(1/2)y^2)\ H_n(y) \tag{15.32}$$

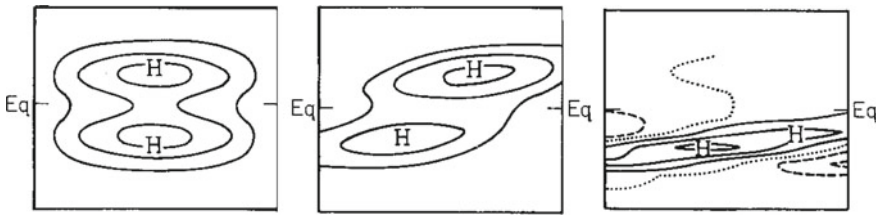

Fig. 15.5 Contours of the height field for $t = 0$, 30 and 90 [*left to right*] when the shallow water equations, linearized about $U(y) = Sy$, are solved for an initial condition which is the $n = 1$ Rossby mode (of an ocean with no mean motion) and $k = 0.1$. Note that the east-west scale is greatly compressed

For a linear shear, $U(y) = Sy$, it is convenient to introduce the parameters $\tau = kSt$, $B = \beta/(kS)$. The phase function is

$$\Phi \equiv \arctan\{(\tau + m)/k\} - \arctan(m/k) \tag{15.33}$$

The streamfunction then has the exact solution

$$\psi(x, y, t) = \frac{1}{\sqrt{2\pi}} \exp(ikx - iy\tau) \int_{-\infty}^{\infty} dm \mathscr{A}(m) \frac{\exp(-imy + iB\Phi)(k^2 + m^2)}{k^2 + (m + \tau)^2} \tag{15.34}$$

where the Fourier transform of the initial condition is

$$\mathscr{A}(m) = i^n \exp(-[1/2]m^2) H_n(m) \tag{15.35}$$

(The Hermite functions are (to within a factor of i) their own transform.)

The factor $iB\Phi$ encapsulates the usual phase behavior of Rossby waves; this produces the southward movement of the disturbance evident in Fig. 15.5 and is well understood. This is true even though the Fourier transform $\mathscr{A}$ is broad in m, not narrowly peaked about some wavenumber m_0 as usually assumed in ray-tracing (Yamagata [10]).

The critical factor is

$$\widetilde{\Lambda}(m, k, \tau) \equiv \frac{k^2 + m^2}{k^2 + (m + \tau)^2} \tag{15.36}$$

When $k << 1$, $\Lambda(m, k, \tau)$ grows to a maximum of m^2/k^2. Therefore, consistent with the Orr solution, large growth is possible for *individual* latitudinal wavenumbers m. However, here the initial condition is a *broad* spectrum in m. Boyd and Christidis show that because the amplitude is exponentially small for the fast-growing large m waves, the *integrated* spectrum decays *monotonically* with time.

We have already pointed out the limitations of this analysis: random perturbations may have spectra more or less regular than that of a Hermite function, and eventually one such perturbation may amplify enough to violate the linearization and create self-sustaining finite amplitude turbulence.

15.11 Summary

An alternative viewpoint for explaining subcritical bifurcation and resolving the "Sommerfeld Paradox" has gained adherents. Laboratory experiments long ago showed the existence of coherent structures, persistent, long-lived and organized, within the frothing fury of turbulence. Waleffe, Cvitanovic, Viswanath, Li and Lin and many others have numerically calculated these coherent structures [22–25]. Although unstable, these coherent structures organize the flow and also provide a

route for transition to turbulence from very small initial perturbations. The coherent structures approach is not incompatible with linear nonmodal instability, but linear theory is always focused on the infancy of turbulence whereas the unstable coherent structures concept tells stories about the adolescence and maturity of turbulence.

There has been no follow-up to the work of Boyd and Christidis on nonmodal instability in the tropics. There has been no work at all on calculating unstable coherent structures nor in exploring the role they might play in equatorial dynamics.

References

1. Orr WM (1907) Stability or instability of the steady motion of a perfect liquid. Proc Roy Ir Acad 27(1):9–69
2. Tung K (1983) Initial-value problems for Rossby waves in a linear flow with critical level. J Fluid Mech 133:443–469
3. Kasahara A (1980) Effect of zonal flows on the free oscillations of a barotropic atmosphere. J Atmos Sci 37(5):917–828
4. Lindzen RS (1988) Instability of plane parallel shear flow (toward a mechanistic picture of how it works). Pure Appl Geophys 126(1):103–121
5. Chapman SJ (2002) Subcritical transition in channel flows. J Fluid Mech 451:35–97
6. Schmid PJ (2007) Nonmodal stability theory. Ann Rev Fluid Mech 39:129–162
7. Fitzgerald R (2004) New experiments set the scale for the onset of turbulence in pipe flow. Phys Today 57(2):21–21
8. Boyd JP (1983) The continuous spectrum of linear Couette flow with the beta effect. J Atmos Sci 40:2304–2308
9. Boyd JP (1983) Second harmonic resonance for equatorial waves. J Phys Oceanogr 13:459–466
10. Yamagata T (1976) On trajectories of Rossby wave-packets released in a lateral shear. J Oceangr Soc Jpn 32:162–168
11. Farrell BF (1982) The initial growth of disturbances in a baroclinic flow. J Atmos Sci 39:1663–1686
12. Farrell BF (1984) Modal and nonmodal baroclinic waves. J Atmos Sci 41:668–673
13. Farrell BF (1985) Transient growth of damped baroclinic waves. J Atmos Sci 42:2718–2727
14. Butler KM, Farrell BF (1992) Three-dimensional optimal perturbations in viscous shear flow. Phys Fluids A 4(8):1637–1650
15. Shepherd TG (1985) Time development of small disturbances to plane Couette flow. J Atmos Sci 42:1868–1871
16. Boyd JP, Christidis ZD (1987) The continuous spectrum of equatorial Rossby waves in a shear flow. Dyn Atmos Oceass 11:139–151
17. Broberg L, Brosa U (1988) Onset of turbulence in a pipe. Z Naturforschung Physik, physikalische Chemie, Kosmophysik 43:697–726
18. Reddy SC, Henningson DS (1993) Energy growth in viscous channel flows. J Fluid Mech 252:209–238
19. Gustavsson LH (1991) Energy growth of three-dimensional disturbances in plane Poiseuille flow. J Fluid Mech 224:241–260
20. Trefethen LN, Trefethen AE, Reddy SC, Driscoll TA (1993) Hydrodynamic stability without eigenvalues. Science 261:578–584
21. Trefethen LN, Embree M (2005) Spectra and pseudospectra: the behavior of nonnormal matrices and operators. Princeton University Press, Princeton

22. Viswanath D, Cvitanovic P (2009) Stable manifolds and the transition to turbulence in pipe flow. J Fluid Mech 627:215–223
23. Li YC, Lin Z (2011) A resolution of the Sommerfeld Paradox. SIAM J Math Anal 43(4):1923–1954
24. Waleffe F (2003) Homotopy of exact coherent structures in plane shear flows. Phys Fluids 15(6):1517–1534
25. Waleffe F (2009) Experiments with semi-lagrangian wavelets. In: Turbulence and interactions: keynote lectures of TI 2006 conference. Notes on numerical fluid mechanics and multidisciplinary design, vol 105. Springer, Berlin, pp 139–158
26. Trefethen LN, Bau III D (1997) Numerical Linear Algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
27. Thomson W (1887) Stability of fluid-motion – Rectilinear motion of viscous fluid between two parallel planes. Phil Mag 24:188–196

Chapter 16
Nonlinear Equatorial Waves

Abstract Nonlinearity spawns a plethora of phenomena. In weakly nonlinear theory, each wave has the Hermite function latitudinal structure of infinitesimal amplitude waves and likewise the vertical structure of a baroclinic eigenfunction. East-west, equatorial waves are free both to propagate and also to evolve; the zonal-and-time amplitude function, $A(x,t)$, satisfies one of several generic PDEs or PDE systems. Each one-space-dimensional PDE is derived by the method of multiple scales or other singular perturbation procedure. In the absence of mean currents, the Kelvin mode is nondispersive, steepens and "breaks" when the slope becomes vertical; $A^{noshear}_{Kelvin}(x,t)$ evolves in obedience to the physics of the inviscid Burgers' equation. Mean currents induce weak dispersion and the Kelvin wave then forms roundish solitary waves with a longitude-and-time structure described by the Korteweg-deVries (KdV) equation. Rossby waves are weakly dispersive, and only weakly dispersive, with or without mean currents, and the latitudinally symmetric modes are governed by the KdV equation, too. The $n = 1$ Rossby soliton is stable even for large amplitude, a pair of contra-rotating anticyclones that resembles middle latitude vortex pairs ("modons"; "Lamb-Chaplygin dipoles"), even to a volume of recirculating fluid, trapped within the coherent structure, if the amplitude is sufficiently large. Higher mode Rossby solitons are "weakly nonlocal" in the sense that they radiatively decay by emitting zonally-sinusoidal waves of smaller latitudinal mode number whose amplitude is an exponential function of the reciprocal of the amplitude of the soliton. This exponentially small radiation lies "beyond-all-orders" in the amplitude parameter and therefore must be calculated by "exponential asymptotics". Lastly, selected numerical studies of idealized nonlinear wave dynamics, not necessarily small amplitude, are reviewed.

> It is a small irony to dub nonlinear phenomena nonsomething when they constitute by far the most common class of things in the universe. . .. This is like referring to the class of animals that are not elephants as non-elephants.
>
> Arthur Fisher and Stanislaw Ulam

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_16

16.1 Introduction

It has been said that the great idea of the nineteenth century was that everything was linear; the great idea of the twentieth century is that everything is nonlinear. The great oceanographer Henry Stommel once observed that perhaps there should be special prayers not only for the mariners who go to sea but also for those who study the sea as well.

Equatorial oceanography was born around 1959 with the publication of Yoshida's work on equatorially-trapped zonal jets, described by infinite series of Hermite functions with explicit coefficients. The first half century of the study of equatorially-trapped dynamics has had the same reliance on linear thinking as the physics and fluid mechanics of the nineteenth century. It is not unjust to say that linear thinking is the infancy and youth of all physical sciences. But oceanography cannot live in its cradle forever.

Nonlinear theory for the tropical oceans has employed two families of tools: numerical simulations of idealized scenarios and reductive singular perturbation theory, which is "reductive" in the sense that the output is a model of three-dimensional reality that is a partial differential equation in one space dimension plus time such as the Korteweg-deVries (KdV) equation. Most of this work has employed the nonlinear shallow water wave equations, also known as the one-and-a-half layer model, on the equatorial beta-plane.

$$\begin{aligned} u_t - y\, v + \phi_x &= -u\, u_x - v\, u_y \\ v_t + y\, u + \phi_y &= -u\, v_x - v\, v_y \\ \phi_t + u_x + v_y &= -u\, \phi_x - u_x\, \phi - v\, \phi_y - v_y\, \phi \end{aligned} \tag{16.1}$$

where u and v are the eastward and northward currents and ϕ is the height (or pressure) field. This is also called the "one-and-a-half-layer" model because it describes a two-layer fluid in the hydrostatic approximation when the lower layer is infinitely deep. The height ϕ is then the thickness of the upper layer or equivalently, the depth of the interface between the two layers. These equations have already been nondimensionalized.

Marshall and Boyd [1], Ripa [2] and Greatbatch [3] have extended this perturbative framework to a continuously stratified fluid. We shall describe their analysis in the next section. We shall continue with the much larger body of perturbation theory applied to the one-and-a-layer model just described. In later sections, we shall return to describe numerical simulations.

As cataloged in Table 16.1, each type of linear wave dispersion determines a corresponding category of interesting nonlinear behavior. Reductive perturbation theory is a tool for classification as well as computation.

Table 16.1 Dispersive categories and nonlinear species

Categories	Nondispersive	Weakly dispersive	Strongly dispersive
Definitions	Nonlinear $\gg\gg$ dispersion c independent of k	Nonlinearity $\approx$ dispersion $c(k) = c(0) + O(k^2)$ $k \ll 1$	Nonlinearity $\ll$ dispersion $c(2k) - c(k) \sim O(1)$
Nonlinear behavior	Frontogenesis and wave-breaking	Bell solitons	Envelope solitons
Evolution equation	One-dimensional advection (ODA)	Korteweg-deVries (KdV)	Nonlinear Schroedinger (NLS)
Wave species	Kelvin (no shear) ultralong Rossby	Kelvin-in-shear long Rossby	Yanai, gravity short Rossby

16.2 Weakly Nonlinear Multiple Scale Perturbation Theory

16.2.1 Reduction from Three Space Dimensions to One

Singular perturbation theory is a very general tool that can describe many nonlinear wave phenomena. The nonlinearity is assumed to be sufficiently weak so that nonlinear effects can be calculated by a power series in a parameter ε which is proportional to the amplitude of the wave. A crucial point is that weakly nonlinear theory is deeply rooted in *linear* theory.

Equatorial waves are trapped in both latitude and depth. This implies that when the wave amplitude is small, nothing very interesting happens in those two coordinates. However, the waves freely propagate east and west. Weak nonlinearity will only reshape the longitudinal structure of the waves slowly, but the accumulation of small changes can lead to large alterations over time. The Kelvin wave, for example, steepens in the absence of dispersion-inducing mean currents. When the nonlinearity is $O(\varepsilon)$ where $\varepsilon \ll 1$, the changes on a $O(1)$ time scale will be small. On a long time scale, however, the Kelvin wave steepens and steepens until it breaks, that is, develops a vertical slope $u_x = \infty$. Astronomers call such slow, accumulating changes "secular effects", and the term has become universal in perturbation theory.

"Reductive perturbation theory" collapses three-space-dimensional reality into nonlinear partial differential equations in just one space dimension plus time. As shown schematically in Fig. 16.1, each weakly nonlinear solution is the product of factors in depth and latitude which to *lowest order* are the *same* as in *linear* theory multiplied by a factor of longitude and time which satisfies one of these simplified "reduced" *nonlinear* partial differential equations.

It is usually necessary to approximate other complications such as dispersion by a perturbative expansion also.

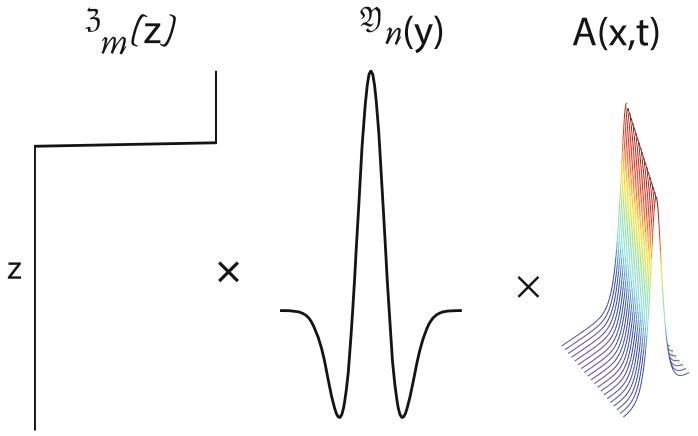

Fig. 16.1 In weakly nonlinear theory, the lowest order is the product of three functions. One factor is a vertical mode [*left*, represented by the baroclinic mode of a 1-1/2-layer model], which is a function only of depth. Another multiplier is $\mathfrak{Y}(y)$ [a Hermite function or the sum of two Hermite functions] and finally multiplied by a function of longitude x and time t. The z and y factors are the solutions to *linear* Sturm–Liouville eigenproblems. The third, longitude-and-time factor, $A(x, t)$, is the solution to a *nonlinear* evolutionary reduced PDE such as the KdV, MKdV, or NLS equations

In the absence of resonance, modes that are distinct in linear theory, and therefore do not interact, do not interact at lowest order in weakly nonlinear theory either. Resonance, however, may produce $O(1)$ interactions between different modes on an $O(1)$ time scale, so resonant triads, second harmonic resonance, long wave resonance, Landau poles and sideband instability will all be treated in due course.

The reduced one-dimensional PDEs form a fairly small set. These same equations have been derived in many different branches of physics because each is *generic* for a certain category of behavior as will be developed equation by equation in what follows.

Although these equations are one-space–dimensional, they collectively capture a very wide range of nonlinear behavior. Equatorial wave dynamics seems to follow the principle that Everything Not Forbidden is Observed; though not all listed phenomena have been observationally confirmed, theory predicts:

1. Kelvin frontogenesis and breaking
2. Rossby solitary waves of KdV type
3. Rossby solitary waves of MKdV type and also mixed cubic-and-quadratic nonlinearity
4. Rossby radiating solitons [“weakly nonlocal solitons”] (latitudinal mode number $n > 2$);
5. Kelvin round solitons
6. envelope solitary waves
7. sideband instability
8. resonant triad instability

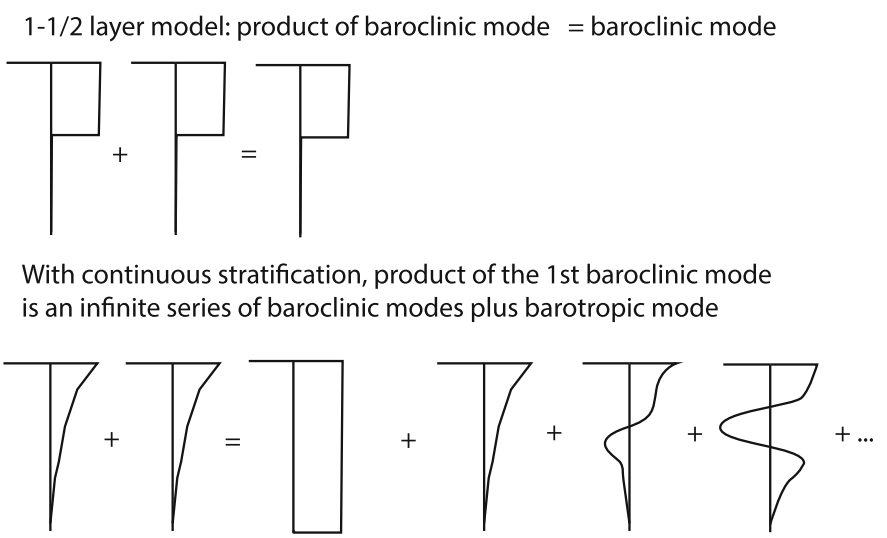

Fig. 16.2 Projections of nonlinear terms onto vertical modes. Each mode is shown schematically as little graphs with depth as the vertical axis. The *vertical line* in the middle of each graph is the axis of eigenfunction values. The higher modes [*lower right*] have vertical nodes where the eigenfunction is zero at each depth where its curve intersects the axis. Top: the single baroclinic mode of the nonlinear shallow water wave equations, also known as the 1-1/2-layer model, is a step function. The product of two step functions as in a nonlinear term is the same stepfunction, so nonlinearity in the baroclinic mode remains entirely confined within that vertical mode. In a continuously stratified mode, the product of two baroclinic modes is an infinite series including the barotropic mode [independent of depth, the first mode to the right of the equals sign in the lower schematic] plus a summation over all the denumerable infinity of baroclinic modes

9. second harmonic resonance
10. long wave/short wave resonance
11. double triad resonance
12. corner waves

We will define and elaborate each of these phenomena in the rest of the chapter.

16.2.2 Three Dimensions and Baroclinic Modes

Almost all theory has been confined to the 1-1/2 layer model in which the tropical ocean is idealized as a two-layer model in which the lower layer is infinitely deep and therefore motionless. There is only a single baroclinic mode; its vertical structure is a step function, nonzero only above the thermocline. There is no coupling to the barotropic mode because in the nonlinear terms, the product of a step function with another step function is again a step function. Thus the nonlinearity projects entirely onto the baroclinic mode.

Only Marshall and Boyd [1] and Ripa [4] have reached beyond the 1-1/2 layer model with perturbation theory, but some useful conclusions have been reached even so. The linear, fully separable model of Chap. 2 is most easily extended to weak nonlinearity by employing a Galerkin method in which, just as in linear theory, all fields are expanded in the eigenfunctions of the vertical structure equation and a Galerkin discretization is applied. The modes are not step functions. The product of the first baroclinic mode with itself projects onto the depth-independent barotropic mode plus an infinite series of baroclinic modes as shown schematically in Fig. 16.2. This has two effects that are filtered by the 1-1/2 layer model:

1. Nonlinearity is *weakened* relative to the 1-1/2-layer model by a factor of one minus the fraction of the square of the first baroclinic mode that projects onto that mode, a projection that would be 100% in the nonlinear shallow water model.
2. The barotropic mode is not equatorially trapped, so nonlinear projection onto this mode is a mechanism for equatorially-trapped motion in baroclinic modes to leak energy to extratropical latitudes.

Figure 1 of [5] is a good illustration of radiative decay of baroclinic equatorial Rossby solitons, theoretically predicted by Marshall and Boyd [1], through leakage into the barotropic mode. The solitary wave remains coherent for a very long time in spite of the leakage.

Dispersion plays a key role in the story. When infinitesimal amplitude waves of different wavelengths $2\pi/k$ (where k is the wavenumber) travel at different phase speeds $c(k)$, a soliton-like isolated peak, the superposition of many different waves $\exp(ik(x - c(k)t))$, will spread out into small peaks and troughs ("disperse") like runners in a marathon (Fig. 16.3). Nonlinearity often acts as an anti-dispersion, steepening the wave until, if dispersion is absent, the wave "breaks" (that is, develops a near-infinite slope or becomes multivalued) (Fig. 16.4).

16.3 Solitary and Cnoidal Waves

Solitary waves are among the most interesting phenomena of nonlinear dynamics since they have no linear counterparts [6, 7].

Definition 16.1 (*solitary wave/soliton*) A "solitary wave" or "soliton" is a localized, *finite amplitude* disturbance in which *dispersion* and *nonlinearity* balance to create a steadily translating wave of permanent form.

As amplified below, theory predicts that solitons of various wave types should be easily generated, and numerical models bear this out, at least for Rossby and Kelvin solitary waves of Korteweg-deVries type. The strength of the linear dispersion determines the form of the solitary waves – or even whether they exist – so the next section will present an overview of the role of dispersion. We will then take up each of the three major cases in turn: no dispersion [frontogenesis and breaking],

weak dispersion [bell solitons and KdV dynamics], and strong dispersion [envelope solitons and NLS dynamics].

After solitary waves, we will turn to resonant triad interactions and turbulence. There has been some useful preliminary work by Loesch and collaborators, Ripa, and a discussion by Salmon, and many articles by Majda and his collaborators, but the collective interactions of equatorial modes are complicated and poorly understood.

Later, we shall explain why "solitary wave" is the worst appellation in fluid mechanics.

Definition 16.2 (*cnoidal wave*) A "cnoidal" wave is a periodic generalization of a solitary wave. It, too, is a *finite amplitude* disturbance in which *dispersion* and *nonlinearity* balance to create a steadily translating wave, but the requirement of "localization" is replaced by the weaker constraint of spatial periodicity. A soliton is a special case of a cnoidal wave: it is the limit that the spatial period goes to infinity for fixed amplitude or the limit that the amplitude goes to infinity (and the width of the peak to zero) for fixed period.

Purists like to make a distinction between "solitary wave" and "soliton" by restricting the latter to solitary waves which collide elastically, that is, do not lose energy either to each other or to daughter waves during collisions. We shall use these terms as synonyms, however. It is now known that "solitons" in this narrow sense exist only for certain one-dimensional partial differential equations that approximate two- and three-dimensional systems. The solitary waves of the *full* equations, as opposed to those of the models, *always* seem to *lose energy* in *collisions* (at least in fluid mechanics), so the distinction between "solitons" and "solitary waves" is of only academic interest.

Purists also restrict "cnoidal" to those periodic waves which can be expressed in terms of the elliptic "cn" function. Modern research has found many periodic wave disturbances which are generalizations of solitary waves, and which tend to a train of evenly spaced solitary waves in the limit that the solitons are tall and narrow in comparison to the spacing between them. Most of these spatially periodic waves cannot be expressed in terms of elliptic functions, but we shall call these "cnoidal" waves, too.

A much more detailed treatment of solitons and cnoidal waves will be given in the later sections on the phenomenology of the time-and-one-space-dimensional PDEs that fall out of reductive singular perturbation theory.

16.4 Dispersion and Waves

The main tool of our theoretical analysis will be small amplitude perturbation theory. This method would seem to be quite restrictive. We shall see in our discussion of modons, however, that the expansion is quite accurate even for rather large amplitude solitary waves. This is a common property of singular (as opposed to regular)

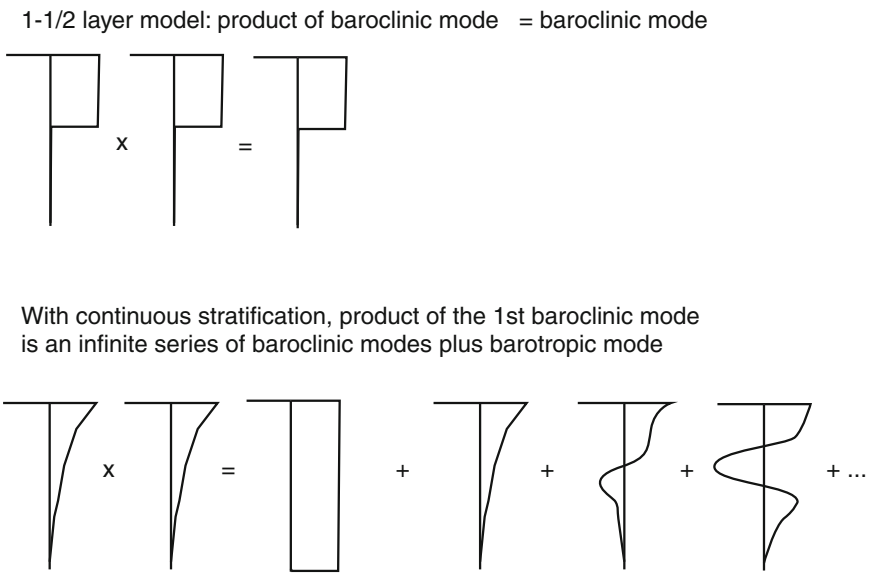

Fig. 16.3 Dispersion in a marathon race. *Top* runners at the startline. *Bottom* near the finish a couple of hours later.When runners, like waves of different wavenumbers, travel each at his/her own pace, what once was concentrated disperses

perturbation expansions: the first non-trivial approximation is usefu1 even when the "small" parameter isn't particularly small.

The effects of nonlinearity can be quite profound, however, even when the wave amplitude is tiny. Equatorial waves are trapped in latitude (by Coriolis forces) and in depth (by the finite thickness of the ocean) and have the structure of standing waves in both y and z.[1] Consequently, weak nonlinear effects will not significantly change either the latitudinal or vertical structure of the waves. This is very important, and we cannot emphasize it too much. Because of this lack of modifications in y and z, we can still classify equatorial modes in the nonlinear theories discussed below in terms of the same scheme as for their linear counterparts — a breaking first baroclinic mode Kelvin wave, a solitary $n = 3$, second baroclinic mode Rossby wave and so on.

In the *zonal* direction, however, the waves *propagate* freely. The distortions created by nonlinearity gradually accumulate as the wave travels until the x-structure of the pulse has been altered almost beyond recognition, even though the nonlinear terms may remain small in comparison to the linear terms throughout the whole passage.

[1] In the 1-1/2 layer model, the waves have the vertical structure of the first baroclinic mode, which is independent of depth in the thin upper layer and zero in the thick layer below the thermocline. Nonetheless, it is true, just as for the continuously stratified model, that waves can freely propagate only in the zonal direction. Whether the waves are sinusoidal with depth, or instead propagate as a ripple on the thermocline, does not change the argument.

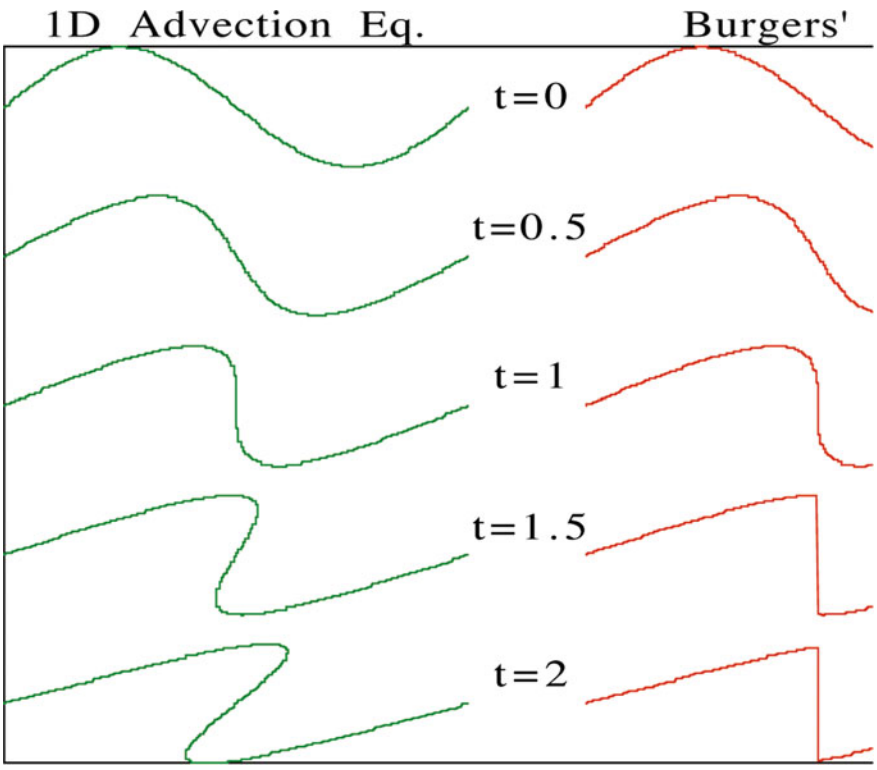

Fig. 16.4 Solutions evolving from the initial condition $u(x, t = 0) = \sin(x)$ when there is no wave dispersion. *Left side* One-dimensional advection equation, $u_t + uu_x = 0$. *Right side* Burgers' equation $u_t + uu_x = \nu u_{xx}$ with $\nu << 1$ so that the frontal zone appears as jump discontinuity to the eye. For both equations, the flow "breaks" at at $t = 1$, that is, develops a point where $u_x = \infty$ [ODA Eq.] or $u_x = O(1/\nu)$ [Burgers' Eq.]. The physics of nonlinear steepening in fluids is well-captured by these onespace-dimensional equations as discussed in Sect. 16.7

Astronomers call this successive addition of small effects "secular" behavior, and the term has been borrowed by applied mathematics in general. For example, if we calculate the orbit of an asteroid, and make an error of only 0.1% in the period on a time interval equal to one orbit, the accumulated error will be small after a few orbits. If we wait 500 periods, however, the small error in the period will be multiplied by a factor of 500, and the true position of the asteroid may be half a revolution out of phase with the prediction.

In a similar way, weak nonlinearity will cause an equatorial Kelvin wave pulse to steepen along its leading edge and flatten and expand on its trailing edge. Because the nonlinearity is assumed to be small, the rate of steepening/expansion is also small. Given enough time, however, (and ignoring damping) the leading edge will inevitably develop an infinite slope and the wave will break.

Thus, it is important to understand nonlinear effects, but the confinement of equatorial waves in both y and z implies that we describe this nonlinearity through *one-dimensional* theories. That is to say, the latitudinal and vertical structure of the waves will be, to lowest order, the same as for linear waves, but the y and z factors will be multiplied by an amplitude $A(x, t)$ which satisfies a *nonlinear* equation in a single spatial dimension (longitude) such as those in Fig. 16.4 or (with dispersion) the KdV and NLS equations discussed below.

There is an amusing irony in this automatic one-dimensionality for equatorial waves. John Scott Russell discovered solitary waves in a canal when generated by a horse drawn barge that stopped abruptly when it hit an underwater obstacle, and Korteweg and deVries showed that the physical confinement of the waves by the sides and bottom of the canal meant that the solitary wave could be described by the one-dimensional partial differential equation that now bears their name. Their model equation has since been applied to a vast range of phenomena, but no one except antiquarians is much interested in barge canals and analysts have had to make a variety of imaginative assumptions to fit their problems into the one-dimensional corset. Although the equatorial beta-plane is one of the most recent applications of the KdV equation, it is also one of the very few for which we can use a channel model without apology.

However, *which* one-dimensional model is appropriate depends upon the relative strength of nonlinearity and dispersion. In the absence of a mean shear, the equatorial Kelvin wave is nondispersive. This implies that there is nothing to counteract the nonlinear steepening, so the fate of the wave — unless it hits a coast first, or is strongly damped — must be frontogenesis and wavebreaking.

When the dispersion is non-zero but *weak*, then this can be balanced directly against the nonlinearity to create single-crested solitons which solve the Korteweg-deVries equation. In the physics literature, these are often called "bell" solitons because the solitary wave has but a single peak, and rather resembles a bell or a Gaussian as illustrated in the upper right of Fig. 16.5. For equatorial Rossby waves and for Kelvin waves in the presence of a dispersion-creating mean shear, the dispersion goes to zero as the zonal scale goes to infinity. Consequently, however weak the amplitude, it is always possible in KdV physics to create a solitary wave provided the pulse is sufficiently wide in longitude.

The third case is *strong* dispersion, which we may define to mean that two waves of wavenumbers k and $2k$ differ in phase speed by an amount which is $0(1)$ [instead of $0(k^2)$ as for long Rossby waves]. Gravity waves, the Yanai or mixed Rossby-gravity mode, and *short* Rossby waves are all strongly dispersive in this sense. Because the dispersion is large, it is not possible to balance it directly against *small* nonlinearity. An initial condition of a single pulse with a broad Fourier spectrum (as shown schematically on the left of Fig. 16.6) would simply be ripped apart by dispersion unless the wave amplitude was so large that perturbation theory fails.

Nonetheless, there is an escape clause: if we consider a wave packet in the narrow sense of a spatially-localized disturbance whose Fourier spectrum is sharply peaked about a particular wavenumber k_c, as shown on the right in Fig. 16.6, then the packet will spread slowly because all its components have approximately the same phase

speed $c(k)$. We can balance this weak linear spreading of the packet against the self-focusing effect of weak nonlinearity to create a solitary wave with many crests and troughs: an "envelope soliton".

The x-dependence of the wave packet is shown schematically in Fig. 16.7 and is of the form

$$\exp[ikx - iw(k)t]A(x,t)$$

The frequency $w(k)$ satisfies the linear dispersion relation; k is simply the peak of the Fourier amplitude as illustrated in Fig. 16.6. The factor $A(x,t)$, which is the dotted line in Fig. 16.7, is called the "envelope" because it bounds the many crests and troughs which make up the packet. The envelope satisfies the Nonlinear Schroedinger (NLS) equation and travels at a rate equal to the linear group velocity plus a small nonlinear correction. The crests and troughs move within the envelope at the linear phase velocity $c(k)$, appearing at one side of the envelope, moving through the packet from front-to-back or vice versa, and then disappearing into the exponential tail on the other side of the envelope like the moving steps of an escalator.

In the next few sections, we will deal in turn with each of these three cases: no dispersion, weak dispersion, and strong dispersion, and outline the mathematics which converts the qualitative concepts described here into quantitative theories.

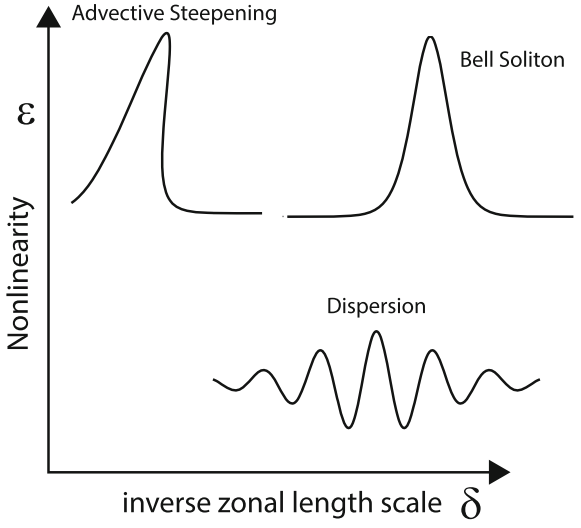

Fig. 16.5 When nonlinearity is strong (large ε) and the wave is a long wave (small δ), the Rossby wave will steepen (frontogenesis) [*upper left*]. When δ is large and ε small, the disturbance will disperse. When nonlinear steepening and dispersive spreading balance, the result is a "bell" solitary wave

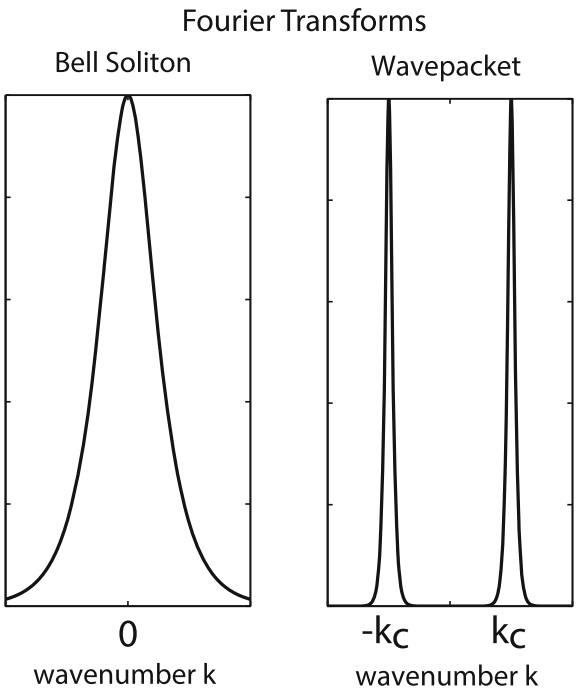

Fig. 16.6 *Left* the absolute value of the Fourier Transform of a bell soliton. Although not marked on the schematic, the envelope of a wave packet, whether linear or nonlinear, is similar. *Right* same but for the entire wavepacket, which is the product of the envelope with a sinusoidal carrier wave. The peaks are at the wavenumber k_c of the carrier wave and $-k_c$

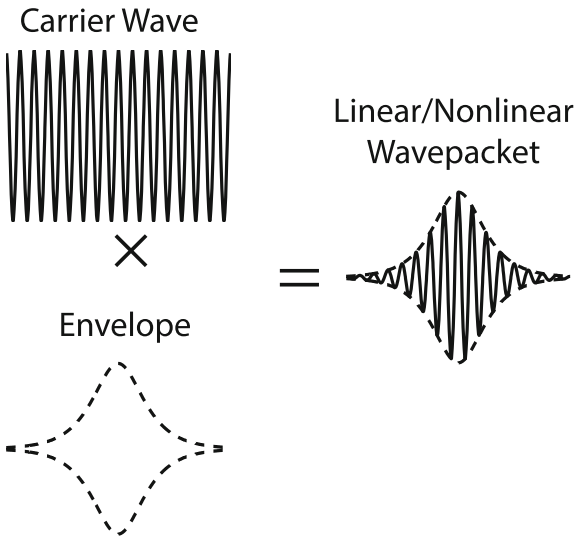

Fig. 16.7 A wavepacket (in the narrow sense used in this book) [*right*] is the product of a sinusoidal "carrier wave" [*upper left*] with a slowly-varying (in both time and space) "envelope" [*dashed*, *left bottom*]

16.4.1 Derivation of the Group Velocity Through the Method of Multiple Scales

Whitham's book gives a marvelously lucid treatment of the *linear* dynamics of a wavepacket using Fourier Transforms followed by stationary phase/steepest descent asymptotics [8]. As a preliminary to the *nonlinear* analysis using multiple scales perturbation theory, it is useful to apply the same methodology to linear, dispersive wave packets.

To illustrate the generality, we shall analyze wavepackets for the one-dimensional PDE

$$u_t + \mathscr{L}u = 0, \qquad \mathscr{L} = \sum_{j=0}^{\infty} a_{2j+1} \frac{\partial^{2j+1}}{\partial x^{2j+1}} = \omega\left(-i\frac{\partial}{\partial x}\right) \tag{16.2}$$

The linear dispersion relation is

$$\omega(k) \equiv \sum_{j=0}^{\infty} a_{2j+1}(-1)^j k^{2j+1} \tag{16.3}$$

Our goal is to understand the further evolution of a slowly-varying plane wave:

$$u(x,t) \sim A(\varepsilon x, \varepsilon t)\, \exp(ikx - iw(k)t) + c.c. \tag{16.4}$$

where "c. c." denotes the complex conjugate. ("Further evolution" is necessary because the multiple scales analysis may not apply for small times, but for sufficiently large time, dispersion will widen the disturbance into one which is varying slowly in space and time (modulo the carrier wave) as also shown by Whitham.)

With $\varepsilon \ll 1$ as the small parameter, the slow variables are

$$\xi \equiv \varepsilon x, \qquad \tau \equiv \varepsilon t \tag{16.5}$$

$$\frac{\partial}{\partial t} \to \frac{\partial}{\partial t} + \varepsilon \frac{\partial}{\partial \tau} \tag{16.6}$$

$$\begin{aligned} \mathscr{L} &\to \sum_{j=0}^{\infty} a_{2j+1} \frac{\partial^{2j+1}}{\partial x^{2j+1}} + \varepsilon \sum_{j=0}^{\infty} (2j+1) a_{2j+1} \frac{\partial^{2j}}{\partial x^{2j}} \frac{\partial}{\partial \xi} \\ &\to i\omega\left(\frac{\partial}{\partial x}\right) + \varepsilon \frac{\partial \omega}{\partial k}\left(-i\frac{\partial}{\partial k}\right)\frac{\partial}{\partial \xi} \end{aligned} \tag{16.7}$$

where we used the identity that the product of $(2j+1)$ derivatives becomes

$$\left(\frac{\partial}{\partial x}+\varepsilon\frac{\partial}{\partial \xi}\right)\ldots\left(\frac{\partial}{\partial x}+\varepsilon\frac{\partial}{\partial \xi}\right)=\frac{\partial^{2j+1}}{\partial x^{2j+1}}+2j+1\varepsilon\frac{\partial^{2j}}{\partial x^{2j}}\frac{\partial}{\partial \xi}+O(\varepsilon^2)$$

$O(\varepsilon^0)$: $w(k)=\omega(k)$

$O(\varepsilon^1)$:

$$\frac{\partial A}{\partial \tau}+\frac{\partial \omega}{\partial k}(k)\frac{\partial A}{\partial \xi}=0 \tag{16.8}$$

whose general solution is $A(\xi,\tau)=A(\xi-[\partial\omega/\partial k]t)$. Using the standard definition of the group velocity, $c_g(k)\equiv\partial\omega/\partial k$, and rewriting in terms of the original coordinates gives

$$A(\xi,\tau)=A(\varepsilon[x-c_g(k)t]) \tag{16.9}$$

where $c_g(k)=\partial\omega/\partial k(k)$.

Thus, the envelope A propagates at the group velocity. Later, applying the same multiple scales methodology, we shall explore how nonlinearity modifies this wavepacket dynamic.

16.5 Integrability, Chaos and the Inverse Scattering Method

Reductive perturbation theory for equatorial dynamics yields a number of one-space-dimensional PDEs that are "exactly integrable" in the sense of Hamiltonian dynamics. Integrable PDEs can be solved by the so-called inverse scattering method" which has the fairly amazing ability of solving *nonlinear* PDEs through a series of steps that require solving only *linear* operations.

Books like Whitham [8] and Yang [9] describe integrability and inverse scattering. Here, we must be content with a catalogue

- Integrable equations *never* have *chaotic* solutions in the sense of dynamical systems theory.
- Classical inverse scattering applies to an unbounded interval; there is a generalization called "finite gap" or "polycnoidal wave" theory that applies with periodic boundary conditions
- N-Polycnoidal wave solutions can be very complicated when N is large where N is the number of independent phase variables each of the form $\zeta_j=k_j[x-c_t t]+\phi_j$. Nevertheless, N-polycnoidal waves are always quasi-periodic, alias multiply-periodic, and therefore always lack the continuous distribution of frequencies which is an essential property of true chaos.
- In the the 60s, there was a brief flurry of optimism that "Inverse Scattering Solves All" but it is now known that this is not true.
- Most nonlinear PDEs and wave systems are *not* integrable and have chaotic solutions in at least some parametric domains.

- Integrable systems and their elliptic function and hyperelliptic function solutions are relatively well understood.
- The inverse scattering method is rarely used to calculate the transient, non-solitonic solutions because of its complexity; a Fourier pseudospectral program, which may contain fewer than twenty lines, is the simplest way to graph KdV solutions.
- Inverse scattering, because its steps are linear, has been very valuable for *proving theorems* and *qualitative analysis*, as we shall amplify below.

16.6 Low Order Spectral Truncation (LOST)

Low Order Spectral Truncation (LOST) is a surprisingly useful tool that underlies much of the theory for nonlinear waves presented later in this chapter. The classic books by Kantorovich and Krylov [10] and Finlayson [11] are mathematical hymns to the power of LOST. Sometimes LOST is effective simply because the solution is very smooth. A remarkably large number of phenomena are well approximated by parabolas.

Ed Lorenz was a great exponent of the philosophy of "maximum simplification" as he called it. His 1960 model of planetary scale instability and vacillation was the resonant triad equation set [12]. Another trio of equations, which made him the founding father of dynamical systems theory and chaos in 1963 [13], was a three-mode truncation of a Fourier Galerkin discretization of two-dimensional convection. The system of five ordinary differential equations in time, the Lorenz–Krishnamurthy Quintet, was a valuable theoretical model of the slow manifold and its exponentially small nonexistence [14–17]. All are rich fruits of this philosophy of "maximum simplification", which is the belief that insights gained from the simplest possible models, even if numerically inaccurate, can extend to much more complex and realistic models.

However, in many other applications including all those which will be discussed later in this chapter, theory provides a plausible rationale for the numerical accuracy of truncating a spectral expansion at very low order. Examples from equatorial dynamics include

1. resonant triads
2. Kelvin frontogenesis
3. Rossby and Kelvin solitary waves
4. nonlinear wavepackets captured by sparse Fourier discretizations as encapsulated in the Nonlinear Schrdinger Equation (NLS)

For example, the central idea of resonant triad theory is that when the wave amplitude ε is small, the time-periodic exchanges of energy between non–resonant waves are tiny. However, waves that are in mutual resonance have secular interactions. That is, the matching of phase speeds, which is the definition of resonance, implies that interactions are not periodic in time, but rather can steadily accumulate change until resonant modes are altered by $O(1)$ amounts over an $O(1/\varepsilon)$ time scale. Truncating

the Fourier expansion to just the three waves in mutual resonance is therefore sensible in the limit that $\varepsilon \to 0$,

Similarly, the One-Dimensional Advection Equation (ODA) for Kelvin waves is the truncation of a Hough-Hermite Galerkin discretization to just the Kelvin waves. Ripa [18], who derives the nonlinear Hough-Hermite discretization of the shallow water equations in great detail, explains in the same article that a truncation to Kelvin waves and only to Kelvin waves is sensible because all Kelvin waves of different zonal wavenumbers are in mutual resonance. For dispersive modes, different wavenumbers k must be treated independently. All Kelvin waves proportional to $\exp(ikx)$ for various k translate eastward at the same phase speed $c = 1$ and have the same latitudinal structure independent of wave number.

Such justifications apply only to *weakly* nonlinear flows. When the amplitude is strong, resonances are not required and wave-wave interactions are as subtle as a besieging army battering through the gate of a castle and storming the fortress.

It is nevertheless remarkable how much of what passes for nonlinear wave theory, celestial mechanics and perturbation theory of all kinds is really a Low Order Spectral Truncation combined with a rationale, either resonance or maximum simplification, for discarding most of the modes of an honest Galerkin discretization.

16.7 Nonlinear Equatorial Kelvin Waves

In the absence of mean currents and mean-flow-induced dispersion, the Kelvin wave is *nondispersive* on the equatorial beta-plane.[2] The zero-dispersion weakly nonlinear dynamics applies to a pulse that is the product of the usual latitudinal structure of the *linear* Kelvin mode — zero north-south velocity, Gaussian u and ϕ with $u = \phi$ — multiplied by the longitude-and-time factor $A(x, t)$:

$$u = \phi = A(x, t)\ \exp(-[1/2]y^2); \qquad v \equiv 0 \tag{16.10}$$

The longitude-and-time factor $A(x, t)$ satisfies the so-called "One-Dimensional Advection" (ODA) equation, also known as the "Inviscid Burgers" equation:

$$A_t + A_x + 1.225AA_x = 0 \qquad \text{[ODA Eqn.]} \tag{16.11}$$

The solution for a continuously stratified flow is the same except for multiplication by a vertical baroclinic mode. This equation has been derived in multiple ways:

[2]The restriction to the equatorial beta-plane is necessary because spherical geometry induces dispersion as shown first by Zhou and Boyd [19]. They discuss the effects of this dispersion on the nonlinear dynamics in [20]. The leading terms of the large-ε series on the sphere are $c \sim 1 + (1/4)\,\varepsilon^{-1/2} + (1/8)\,\varepsilon^{-1} + \left((3/32) - s^2/16\right)\varepsilon^{-3/2}$; only the $-(s^2/16)\varepsilon^{-3/2}$ term is dispersive, so the sphere-induced dissipation is very weak unless the zonal wavenumber $s \sim \varepsilon^{-3/4}$ or larger. Longuet-Higgins [21] gives the first two terms as his (8.27).

1. Strained coordinates perturbation theory (Boyd, 1980) [22]
2. All-harmonics Kelvin self-resonance (Ripa 1982, 1983) [18, 23]
3. Multiple scales perturbation theory [24]
4. Theory-Plausible Low Order Spectral Truncation (LOST) (Ripa [18], and this book).

These different derivations offer different insights into the underlying physics so we shall return to them as we proceed. Since the One-Dimensional Advection equation is also applicable to Rossby waves in the regime where nonlinearity $\gg$ dispersion, we shall generalize (16.11) to

$$u_t + cu_x + buu_x = 0 \tag{16.12}$$

by allowing c and b to be mode-dependent constants.

16.7.1 Physics of the One-Dimensional Advection (ODA) Equation: $u_t + cu_x + buu_x = 0$

Equation (16.12) can be solved exactly via the method of characteristics. The key is to realize that (16.12) is in the form of the conservation equation

$$Du/Dt = 0 \tag{16.13}$$

where D/Dt is the usual "total" or "convective" time derivative of fluid mechanics if we identify the advecting velocity with $(c + bu)$. Equation (16.13) implies that u (and the combination $[c + bu]$) are *constant* following the motion. If u is constant, however, then a fluid particle will simply move in a straight line at whatever was its initial velocity. Each trajectory in the $x - t$ plane will be a contour line of $u(x, t)$. The PDE can be solved graphically by drawing these straight contour lines with a ruler or other straight edge as in Fig. 16.8.

If the initial condition is

$$u(x, t = 0) \equiv Q(x) \tag{16.14}$$

for some arbitrary function $Q(x)$, then a particle which is at $x = \xi$ at $t = 0$ will have $u = Q(\xi)$ forever, and its trajectory is the line $x = \xi + [c + bQ(\xi)]t$.

In the language of computer graphics, the equations

$$x = \xi + [c + bQ(\xi)]t \tag{16.15}$$

$$u(x, t) = Q(\xi) \tag{16.16}$$

are, at a given time, an explicit *parametric* description of the curve in the $x - u$ plane which is the solution. The solution over a finite time interval is a parametric surface in $x - t - u$ space parameterized by $(\xi, \beth)$; the three parametric equations are

$$x = \xi + [c + bQ(\xi)]\beth \tag{16.17}$$

$$u(x,t) = Q(\xi) \tag{16.18}$$

$$t = \beth \tag{16.19}$$

The parametric representation is nonsingular even after developing folds or pleats. A folded surface u implies the function is multi-valued as a function of (x,t).

As a means to evaluate u at a specific space-time point, the parametric solution has the disadvantage that it is only an *implicit* solution. The complication is that (16.20), the trajectory equation $x = \xi + [c + bQ(\xi)]t$ must be solved for $\xi(x,t)$ before u can be evaluated from $u(x,t) = Q(\xi[x,t])$. Platzman [25] showed that for the special case of a sinusoidal initial condition, the characteristic equation is the Kepler equation of celestial mechanics and is solved by an infinite series of Bessel functions due originally to Bessel. (See Fig. 16.9 and its caption.)

The best numerical strategy is to expand the initial condition $Q(\xi)$ as a Chebyshev series, $\sum q_n T_n(\xi)$. This only needs to be done once. Note that $\xi = T_1(\xi)$ and x can be interpreted as xT_0 in the Chebyshev series. It is then trivial to obtain the Chebyshev coefficients of the polynomial $P(\xi; x, t)$ whose zeros give the characteristic coordinate ξ as a function of x and t. For example, when $Q(\xi) = -\sin(\pi\xi)$, $\xi \in [-1, 1]$

$$\begin{aligned} P(\xi[x,t) = &-(x - ct)T_0 + (1 - 0.569t)T_1(\xi) \\ &+ 0.6669\,t\,T_3(\xi) - 0.1043\,t\,T_5(\xi) + 0.00684\,t\,T_7(\xi) + \ldots \end{aligned} \tag{16.20}$$

The roots are the eigenvalues of the Chebyshev companion matrix as described, along with strategies for simple, adaptive Chebyshev interpolation, in Boyd's book [26].

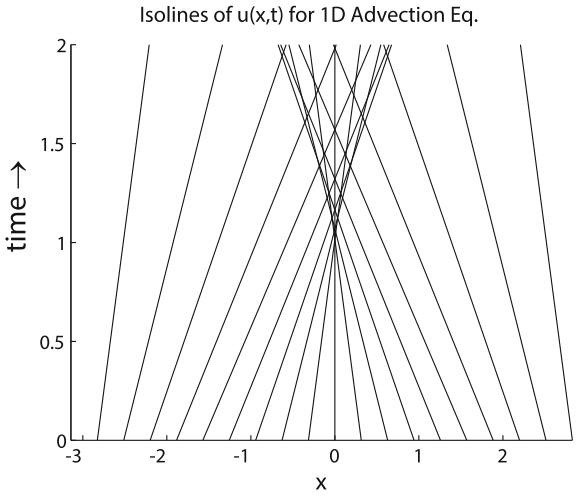

Fig. 16.8 Contour plot of solutions for the One-Dimensional Advection equation, begun from $u(x,0) = -\sin(x)$. The isolines of $u(x,t)$ are straight lines; these isolines are also the particle trajectories. The solution is triple-valued for some x for $t > t_B$ where the time of breaking $t_B = 1$

The implicit solution (16.20) is a full description of the wave before breaking. Figure 16.8 is an $x - t$ plot that shows that the straight lines which represent the velocity-conserving trajectories of fluid particles, again shown in a reference frame moving at the linear velocity c so that all motion seen in the plot is due to nonlinearity. When two or more straight lines intersect so that the ODA solution is multi-valued for larger times and the wave breaks.

Nonetheless, it is possible to deduce much simply from the form of the implicit solution.

Definition 16.3 (*Breaking*) Solutions of the one-dimensional advection equation are said to "break" at a time t_B when $\partial u/\partial x = \infty$ for some x at the time t_B and the slope is finite for all smaller t.

Theorem 16.1 (Time of Breaking)

$$t_B = -\frac{1}{Q'(\xi_{min})} \tag{16.21}$$

where $x = \xi_{min}$ is the point where $Q(x)$ has its most negative slope (Whitham [8])

Proof:

$$u_x = Q'(\xi[x])\frac{\partial \xi}{\partial x} \tag{16.22}$$

$$= Q'(\xi[x])\frac{1}{1 + tQ'(\xi[x])} \tag{16.23}$$

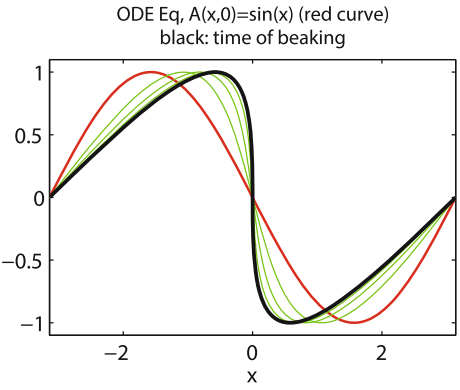

Fig. 16.9 Solutions to the One-Dimensional Advection (ODA) equation for various times beginning with the sinusoidal initial condition. The coordinate system is moving at the linear phase speed so as to keep the developing front centered at the spatial origin. For this special initial condition, Bessel found an infinite series solution which converges for $\bar{t} < 1$ where $\bar{t} = 1.22t$: $A(x, \bar{t}) = -2\sum_{n=1}^{\infty}\left\{J_n(n\bar{t})/(n\bar{t})\right\}\sin(nx)$ (The 1.22 is the coefficient of the nonlinear term in the One-Dimensional Advection equation for the Kelvin wave)

by differentiation of $x[\xi] = \xi + tQ(\xi)$] with respect to ξ. The denominator is zero, implying u_x is infinite, at the time give in the theorem. ■

Figure 16.9 shows different stages in the evolution of a wave that was a sine function at $t = 0$, presented in a frame of reference moving at the linear phase velocity c. Equation (16.20) shows that each part of the wave travels at its own speed, $c + u$, so that the nonlinear disturbance cannot be characterized by a single velocity. The crest travels fastest, and gradually overruns the lower and slower moving parts of the wave in front of it. The trailing edge of the crest, however, lags behind the maximum. The net result is that the leading edge [for a Kelvin wave, the eastward edge] of a crest steepens while the trailing edge is stretched and rarified. At $t = 1.22$, the leading edge has steepened to the point of an infinite slope where $u = 0$ — the wave breaks.

16.7.2 Post-Breaking: Overturning, Taylor Shock or "Soliton Clusters"?

After an infinite slope develops, several fates are possible. An infinity is usually the signature of omitted physics. The breaking Kelvin wave has different futures depending on what new physics becomes important. Figure 16.10 shows three alternatives.

A similar steepening occurs when waves break on a beach and for nondispersive waves in general. However, what happens after the wave breaks varies from one wave species to another. Ordinary water waves continue to steepen, the wave height becomes multi-valued, and the wave eventually collapses into foam. In the conceptual model of the equatorial ocean as a one-and-a-half-layer model, multi-valued breaking is an overturning of the main thermocline, a violent mixing of the layers above and below the main pyncocline.

Later, we shall show that mean shear induces dispersion in the Kelvin wave. The simplest model is the Korteweg-deVries (KdV) equation which appends a third derivative to the One-Dimensional Advection equation. Breaking is impossible because the third derivative scales with a high gradient layer of thickness δx as $O(1/[\delta x]^3)$ whereas the nonlinear term contains only a first derivative and grows more slowly (as $O(1/[\delta x])$). Dispersion always wins as $\delta x \to 0$. Lax and Levermore famously analyzed the zero dispersion limit. First, a triangle-shaped pulse with a near-vertical leading edge evolves, well-described by the ODA equation, and then dispersion finally becomes significantly strong to forestall a discontinuity at the leading edge, and the triangular pulse dissolves into many thin, tall solitary waves.

With all due espect to the eminence of Lax and Levermore, their picture as applied to the ocean is a silly fantasy. The KdV equation is a *long wave* approximation; thevery tall, narrow solitary waves of the Lax-Levermore zero dispersion limit have Fourier spectra that decay slowly with wavenumber and therefore have lots of amplitude for $k \gg 1$. Since the KdV equation is derived by a Taylor expansion of the phase speed $c(k)$, the reduction of three-dimensional fluid dynamics to the one-dimensional KdV

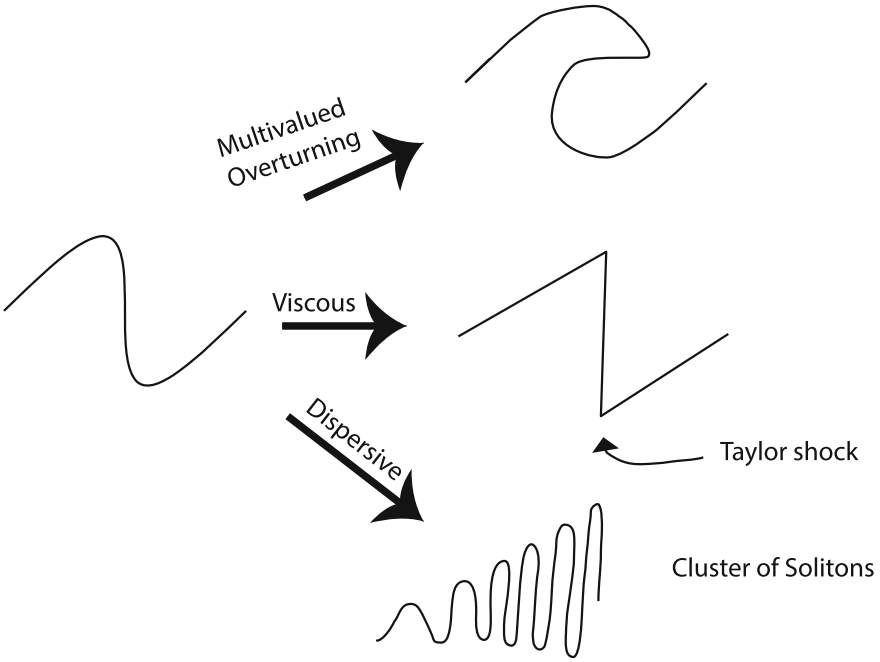

Fig. 16.10 The One-Dimensional Advection equation predicts an infinite slope in finite time, but in physics, predictions of infinity usually indicate that additional physical mechanisms are activated near the time of unboundedness. The figure shows three such additional mechanisms

equation must fail. Dispersion may indeed limit the slope of the Kelvin wave but a more complicated model than the KdV equation is necessary.

16.7.3 Viscous Regularization of Kelvin Fronts: Burgers' Equation And Matched Asymptotic Perturbation Theory

Adding a little viscosity to the One-Dimensional Advection equation changes it into Burgers equation whose solutions are always single-valued (Fig. 16.11).

For Burgers equation, as opposed to the full three-dimensional hydrodynamic equations, D.G. Crighton and colleagues in acoustics [27] and Boyd showed through singular perturbation theory that the frontal zone is well-approximated by the hyperbolic tangent, the so-called "Taylor shock" [28]. Figure 16.12 shows the triple-valued post-breaking solution of the One-Dimensional Advection equation, the single-valued Burgers solution and the Taylor shock that describes the frontal zone. Because the frontal zone has width proportional to the (small) viscosity coefficient ν, graph-

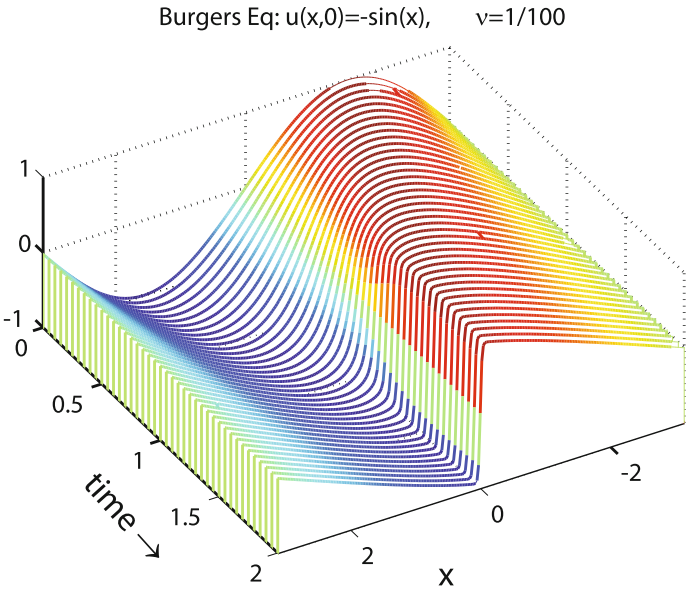

Fig. 16.11 Burgers' equation: $u(x, 0) = -\sin(x)$ with the viscosity $\nu = 1/100$. Same as Fig. 16.9 but as a waterfall plot, extended beyond the time of breaking with viscous regularization of the singularity (i.e., as a solution to Burgers' equation)

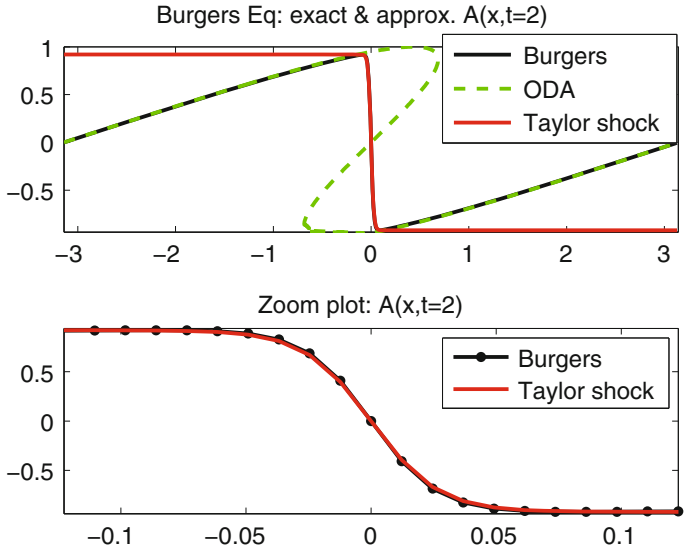

Fig. 16.12 The post-breaking solution of the One-Dimensional Advection equation (*green dashes*) is compared with the Taylor shock (*red curve*, $-\tanh(x/(2\nu))$) and with the solution to Burgers equation (*black*). The latter is well-approximated outside the frontal zone by one branch of the One-Dimensional Advection equation (*green dashes*) and well-approximated in the neighborhood of the front by the Taylor shock as shown more clearly in the zoom plot in the lower figure

ical clarity requires the zoom plot [lower graph] that shows only the neighborhood of the front. The Taylor shock is indeed a remarkably good approximation to Burgers' solution in the frontal zone. Away from the front, however, Burgers' solution is well-approximated by the appropriate branch of the solution to the One-Dimensional Advection equation [8, 28].

Greatbatch [3] finds that post-breaking equatorial Kelvin waves propagate as a kind of hydraulic jump [3]. Fifteen years later, Fedorov and Melville performed more numerical studies and also developed a hydraulic jump/undular bore model of steadily-propagating equatorial Kevin waves [29]. However, there is no obvious reason why the Kelvin wave could not develop an overhanging tongue and collapse as a plunging breaker. The single-valued solution to the shallow water wave equations cannot represent a triple-valued reality. An accurate post-breaking description of the Kelvin wave is an Open Problem.

16.8 Kelvin-Gravity Wave Shortwave Resonance: Curving Fronts and Undulations

Figure 16.13, which shows the contours of the height field near but *before* the time of breaking, shows that front has a strong curvature. This effect is completely missed by the perturbative analyses of Boyd [22] and Ripa [18]. As stressed earlier, the lowest order perturbative solutions are always in "separable form", being the product of a factor of depth times a factor of latitude times a bivariate function $A(x, t)$ which satisfies a one-space dimensional partial differential equation such as the One-Dimensional Advection equation or Burgers' equation. This constraint is relaxed at first order, but curvature due to the first order solution should be almost invisible. The curvature looks as big as a leading order effect, which is possible only if the perturation theory of Boyd [22] is wrong or incomplete somehow.

In fact, Boyd's early theory is deficient. An implicit assumption of his analysis was the lack of resonance except for the interaction of the Kelvin wave with itself. In the long wave limit, Kelvin waves are indeed not resonant with any other wave species. Unfortunately, in the short wave limit, $k \to \infty$, the phase speed of eastward-propagating gravity waves tends to one from above. Using a simple mode that retained just four latitudinal modes, Boyd [30] showed that *short wave* Kelvin-Gravity

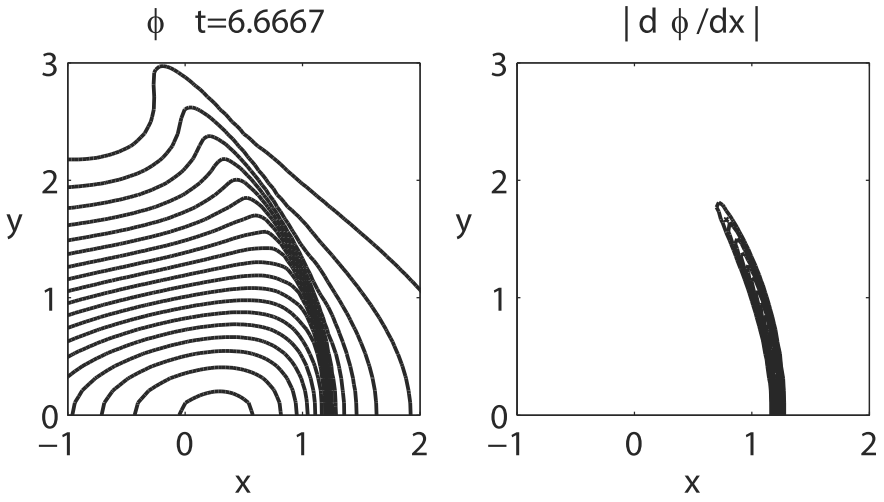

Fig. 16.13 *Left* Contours of the height/pressure field at $t = 6.67$ from the initial condition $u = \phi = \sum_{m=-\infty}^{\infty} \frac{1}{3}\,\mathrm{sech}^2((1/2)[x - 20\,m])$, $v \equiv 0$. *Right* Same but showing the contours of the x-derivative of ϕ. Numerical solution to the nonlinear shallow water equations on the equatorial beta-plane

resonance was responsible for both the westward curvature of the Kelvin front and also for undulations westward of the front [not visible in the figure.]

Fedorov and Melville [29] independently arrived at the same conclusions. Their article deserves some criticism. They asserted twice that both Boyd and Ripa assumed meridional geostrophy a priori. This is not true; rather, perturbation forces the lowest order solution to be in geostrophic balance. Meridional geostrophy is relaxed in higher order. Their paper omits all mention of Boyd [30], who had scooped their main conclusion that gravity wave resonance cause frontal curvature and undulations westward of the front. Their theory of Kelvin wave hydraulic jumps is unobjectionable, but implicitly assumes that viscosity arrests the steepening of the Kelvin front. This is by no means certain or even likely. Boyd dodged this problem by focusing on time-evolution of the front *near* but *before* the time of breaking.

A four-mode analysis of resonance and undulations is not the same as a theoretical analysis for the shallow water equations and a steadily-translating undular bore is not a description of the complex frontogenesis that preceded it. A full analysis of the Kelvin/short-gravity-wave resonance is an Open Problem (Table 16.2).

16.9 Kelvin Solitary and Cnoidal Waves

The Kelvin wave becomes dispersive in two ways. First, replacement of the equatorial beta-plane by spherical geometry will induce dispersion: $c_{Kelvin} \sim 1 + (1/4)\,\varepsilon^{-1/2} +$

Table 16.2 Nonlinear Kelvin waves: selected references

Reference	Remark
Bouchut, LeSommer and Zeitlin [31]	Equatorial modons colliding elasticallywith Kelvin waves
Boyd [22]	Derivation of inviscid Burgers equation
Boyd [32]	Kelvin solitons; mean-current-induced dispersion
Boyd [30]	Numerical study of Kelvin frontogenesis
Boyd [33]	4-mode model
Chen and Boyd [34]	Nonlinear Kelvin wavepackets
Fedorov and Melville [29]	Kelvin fronts and hydraulic jump theory
Fujiwara, Yamamoto, Hashiguchi, Horinouchi and Fukao [35]	Radar observations of turbulence due to breaking Kelvin waves In tropical tropopause
Greatbatch [3]	Kelvin wave fronts and nonlinear spin-up
LeSommer, Reznik and Zeitlin [36]	Numerical study of response to localized perturbation
Long and Chang [37]	Propagation of an equatorial Kelvin wave in a varying thermocline
Matsuura and Iizuka [38]	Nonlinear Kelvin waves via a pair of equations like Stern's treatment of coastal waves
Milewski and Tabak [39]	Dispersion due to finite depth; Kelvin solitons
Ripa [18]	Kelvin wave frontogenesis as simultaneous resonance with all its harmonics
Ripa [4]	Wave-wave interactions and chaos
Ripa [2]	Nonlinear Kelvin pulses across the Pacific Ocean
Ripa [40]	Ray-tracing
Stepanov and Novotryasov [41]	Kelvin frontogensis
Zheng and Susanto Yan, Liu and Ho [42]	Observation of equatorial Kelvin solitary waves in a slowly varying thermocline

$(1/8)\,\varepsilon^{-1} + \left((3/32) - s^2/16\right)\varepsilon^{-3/2} + O(\varepsilon^{-2})$; the $\varepsilon^{-3/2}$ correction includes a term depends on the wavenumber s.

Second, Boyd [32] noted that latitudinal shear of the mean zonal current will induce dispersion. Boyd [32], Greatbatch [3], G.-Y. Chen and Boyd [34] have all analyzed shear-induced Kelvin mode dispersion using perturbation theory with shear strength as the small parameter.

Figure 16.14 is a numerical solution of the full nonlinear shallow water equations on the equatorial beta-plane that qualitatively verifies these perturbative analyses. A Kelvin mode with the longitudinal structure of a KdV solitary wave was superimposed on a mean current that at least crudely mimics the South Equatorial Current. Although there is a little dispersion during the early stages because it was not possible

to initialize with the exact soliton, Fig. 16.14 shows the initial structure is remarkably persistent.

The essential role of the zonally-independent mean current in inducing dispersion can be demonstrated by repeating the integration with all the same except for omission of the mean flow. Figure 16.15 shows that within fifteen nondimensional time units — about 1/7 of the time interval of the previous plot — the wave has broken.

The leading order effect of mean shear is to modify the One-Dimensional Advection equation by adding a linear third derivative term to convert it into the Korteweg-deVries equation. However, the shear-induced dispersion is weak. Experimentally, all sufficiently large amplitude Kelvin waves break. It follows that it is useful — indeed, essential - to analyze the solutions of (16.11) before proceeding to the more complicated KdV equation which governs Kelvin waves of small and moderate amplitude.

The derivation of the Korteweg-deVries equation for Kelvin waves is similar to that for Rossby waves given in detail below. However, the details are more complicated because of the necessity of including mean flow and/or sphericity, so the interested reader is referred to [32].

One issue is worth a comment here. The KdV perturbation theory is always *in principle* a *longwave* approximation, that is, the zonal scale is large compared to

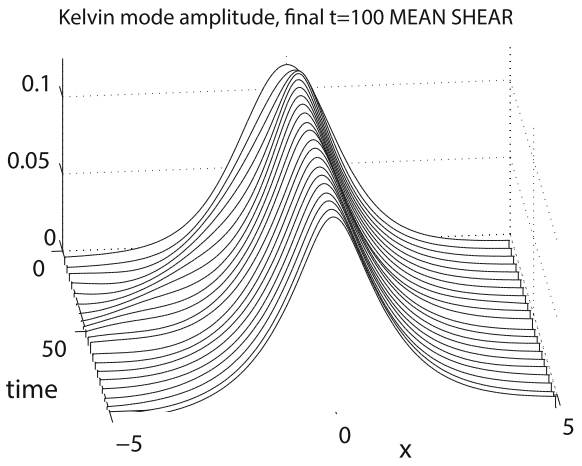

Fig. 16.14 Amplitude of the Kelvin mode versus time for numerical integration of the nonlinear shallow water equations on the equatorial beta-plane. (Note that a coordinate system moving at a constant velocity was used to keep the solitary wave centered on the graphs.) The zonal-mean (x-independent current) was $U(y) = -(1/2)\exp(-(25/64)y^2)\,(1 - (25/64)\,y^2)(1 + (25/36)y^2)$ in geostrophic balance with the mean height field $\Phi(y) = (1/2)\exp(-(25/64)y^2)\,((512/225) + (25/18)y^2 + (25/72)y^4)$. (Note that total column height is $h = 1 + \phi(x, y)$.) Superimposed on this was a Kelvin wave with the longitudinal shape of a KdV soliton: $v \equiv 0$, $u = \phi = 0.12\mathrm{sech}^2(0.7x)\exp(-(1/2)y^2)$. The computational domain, $x \in [-50, 50]$, was ten times larger than what is shown in the graph to minimize endpoint effects. The time interval, which is roughly two-and-a-half times the Pacific-crossing time for the Kelvin mode, was chosen to show the stability and persistence of the Kelvin solitary wave on a long time scale

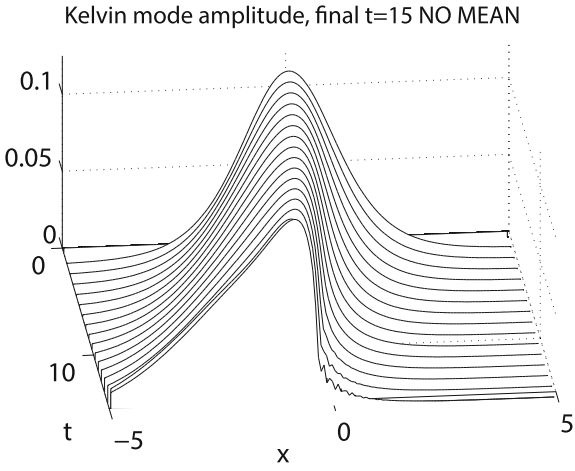

Fig. 16.15 Same as the previous figure except that zonal-mean current is omitted, and the length of the time integration is shortened from 100 to 15

one when the perturbation parameter is small. For solitary waves, the amplitude is proportional to the inverse of the square of the zonal scale; this scaling allows nonlinear steepening and dispersive spreading to balance so that the soliton neither steepens nor spreads but merely translates. Since the latitudinal scale is $O(1)$ [the decay scale of $\exp(-(1/2)y^2)$ for the Kelvin mode], the KdV theory in principle applies only to *zonally-elongated* disturbances.

Figure 16.16 shows the contours of the zonal velocity of the Kelvin solitary wave depicted in Fig. 16.14. The near-circularity of the contours is not an illusion; this bivariate function can be approximated to within a relative error of less than 8% by the radially-symmetric function $u_{sym}(x, y) = 0.12\exp(-(1/2)[x^2 + y^2])$.

In later sections, we shall exhibit roundish, large amplitude Rossby solitary waves that are nevertheless well described by perturbation theory. Because the Kelvin wave is non-dispersive except for the weak dispersion induced by the mean shear flow, this is especially true of this mode. Small amplitude Kelvin mode solitary waves and cnoidal waves really do have zonal elongation with east-west scales larger than the north-south decay scale; the point is simply that large zonal scales and elliptically-shaped contours are not necessary for conceptual and even quantitative accuracy of the KdV model for Kelvin waves.

As explained most clearly by Bender and Orszag [43] through a plethora of well-chosen examples, singular perturbation methods are often qualitatively accurate, and sometimes quantitatively, too, *even when the small parameter is not small*. Large, round Kelvin and Rossby solitons are thus not exceptional.

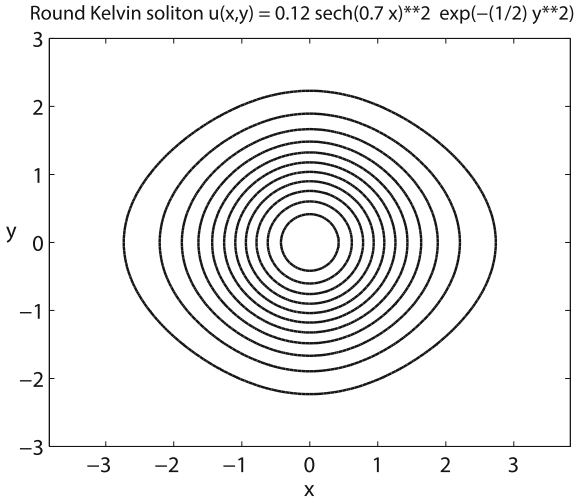

Fig. 16.16 Contours of the initial zonal velocity u for the Kelvin solitary wave whose projection on the Kelvin mode is shown two figures previously, $u = 0.12\text{sech}^2(0.7x)\exp(-(1/2)y^2)$. (Strictly speaking, the Kelvin soliton is the final flow after the initial, non-solitonic transients have radiated away. These transients are negligibly small so that the initial, factored Kelvin height field is a very good approximation to that of the final Kelvin soliton.) The zonal factor, which has the shape of a KdV solitary wave, can be approximated with a maximum relative error of only 7.6% by the Gaussian function that is the latitudinal structure, that is, $\text{sech}^2(0.7x) - 1.032\exp(-(1/2)x^2)| \leq 0.076$

16.10 Corner Waves and the Cnoidal-Corner-Breaking Scenario

The KdV equation is misleading and atypical in that it allows solitary waves of arbitrarily large height and arbitrarily narrow width. In contrast, surface gravity waves, cnoidal waves and solitons in water have a maximum height, and all taller waves break.

The boundary between breaking and nonbreaking, steadily-propagating waves usually is a wave with the following structure.

Definition 16.4 (*Corner Wave*) A corner wave is a steadily traveling nonlinear wave in which the wave height function $u(x - ct)$ has a peak or trough which is also a slope discontinuity.

More than a century and a half ago, Stokes demonstrated that the largest surface water wave was a corner wave with the sides making an angle of 120°. Table 16.3 catalogues a considerable body of work on corner waves in various wave equations, most of which are intended to model water waves.

Chen and Boyd showed by numerically solving the nonlinear shallow water equations that large amplitude equatorial Kelvin waves always break [34]. However,

their studies clearly showed Kelvin solitary and cnoidal waves, stable and steadily-propagating, for smaller amplitude. (See also Fig. 16.14.)

Boyd confirmed the existence of distinct small amplitude and large amplitude regimes by specializing the Equatorial Four-Model Model of [30] to traveling waves. This model is LOST, that is, a drastic truncation of a Hermite-Galerkin discretization of the equatorial shallow water wave equations. For a steadily-propagating wave with phase speed c, the Four-Mode Model reduces to a quartet of coupled, quadratically nonlinear ordinary differential equations in the coordinate $X \equiv x - ct$.

$$\begin{aligned} &(1-c)S0_X \quad + \quad 0.6124\ S0\ S0_X \\ &-0.4083\ (S0\ S2)_X \quad +2.4495\ S2\ S2_X \quad +0.5443\ \aleph_X\ D0 + 2.1773\,\aleph_X\ S2 \\ &-0.2041\ D0\ D0_X \quad -0.2041\ (S0\ D0)_X \quad +0.1361\ (D0\ S2)_X = 0 \end{aligned} \tag{16.24}$$

Table 16.3 Examples of systems with corner waves and the CCB scenario

Equation or wave name	Equations	Sources
Equatorial		
Equatorial waves Barotropic Mode	$K_t + K K_x = \gamma \left\{ Y(t)\mathrm{e}^{ix} + \overline{Y}(t)\mathrm{e}^{-ix} \right\}$; $Y_t = -\gamma \hat{K}(x=1,t)$	[24]
Equatorial Waves Baroclinic Mode	3 coupled PDES in (x,t)	[24]
Resonant Triads, One Nondispersive	$u_t + uu_x = 2\mathrm{Re}\,(ikab\exp(-ikx))$; $a_t = -i\omega_a b\overline{\hat{u}_k}$; $b_t = -i\omega_b a\overline{\hat{u}_k}$	[44]
Equatorial Kelvin (4-mode Model)	4 coupled PDEs in x, t	[33]
Equatorial Kelvin (shallow water)	3 coupled PDEs in x, y, t (shallow water Eqs.)	[20]
Non-equatorial		
Surface Irrotational Water Waves	Euler equations in x, z	[45–48]
Edge Waves on Vortex Patches	Two-space-dimensional Euler equations (x, y)	[49]
Camassa–Holm	$u_t - u_{xxt} + (2\kappa + 3u - 2u_{xx})u_x - uu_{xxx} = 0$	[50–52]
Ostrovsky–Hunter	$(u_t + uu_x)_x = u$	[53, 54]
Gabov/Shefter–Rosales	$(u_t + uu_x)_x = \int_0^{2\pi} \cos(x-y)\,u(y)\,dy$	[55, 56]
Whitham	$(u_t + uu_x)_x = pb^2 \times \left\{ u - \int_0^{2\pi} \frac{b\cosh(b\{\lvert X-y\rvert - \pi\})}{2\sinh(\pi b)}\,u(y)dy \right\}$	[57, 58] [8, 59]

$$\begin{aligned}
&(1-c)S2_X \quad -\aleph_X \quad -0.051031\ S0\ S0_X \\
&+0.3062\ (S0\ S2)_X \quad +0.8845\ S2\ S2_X \quad -0.2722\ \aleph_X\ S0 - 0.1361\ \aleph_X\ D0 \\
&+0.01701\ D0\ D0_X \quad -0.1021\ (S2\ D0)_X \quad +0.017010\ (S0\ D0)_X = 0
\end{aligned} \tag{16.25}$$

$$\begin{aligned}
-c\aleph_{XXX} + 2S2_X - \frac{1}{2}D0_X + 0.2722\aleph_{XX}\,(\,S0_X\, + 2S2_X - D0_X) \\
+0.2722\aleph_{XXX}\,(\,S0\, + 2S2 - D0) = 0
\end{aligned} \tag{16.26}$$

$$\begin{aligned}
&-(1+c)D0_X \quad +2\,\aleph_X \\
&+0.2041\ (S0\ S0_X \quad +(S0\ D0)_X) \quad -0.1361\ (S2\ [D0+S0]\,)_X \\
&+0.8165\ S2\ S2_X \quad +0.5443\ \aleph_X\ (S0-2\ S2) \quad -0.6124\ D0\ D0_X = 0
\end{aligned} \tag{16.27}$$

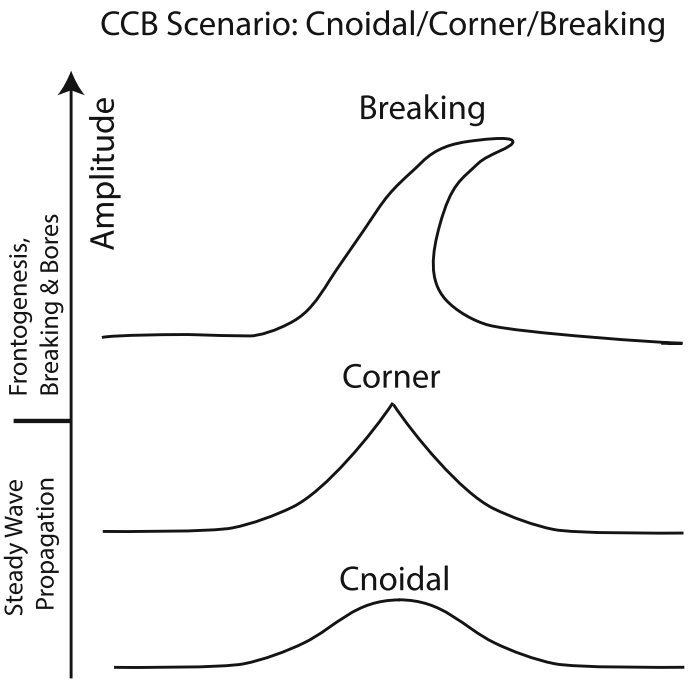

Fig. 16.17 CCB Scenario. The corner wave is the largest nonbreaking wave. When the amplitude is less than the amplitude of the corner wave, steadily-propagating solitary waves and their spatially periodic generalizations, cnoidal waves, are possible. When the amplitude is larger than the corner wave amplitude, marked by the horizontal tic mark on the axis, all waves steepen and break or steepen and propagate as a bore or hydraulic jump

$S2$ is for example the coefficient of $\psi_2(y)$ in the Hermite series for the sum variable S; the north-south velocity variable v_1 is replaced by the new unknown $\aleph = \int^X v_1$ so that all four unknowns are symmetric in X. This system is just simple enough that the formidable numerical challenges of computing singular traveling waves — the corner wave is not only discontinuous in its first derivative, but is also the terminus of a branch of solutions that simply stops in a fashion that has no counterpart in bifurcations in finite-dimensional systems — can be overcome as detailed in [33].

In summary, neither the breaking-but-no-traveling-waves physics of the One-Dimensional Advection equation nor the traveling-waves-but-no-breaking fluid mechanics of the Korteweg-deVries equation are complete pictures of Kelvin wave dynamics. Rather, the equatorial Kelvin wave is but one of many wave species that exhibit the "Cnoidal/Corner/Breaking (CCB) Scenario". For small amplitude, there are traveling waves analogous to the cnoidal waves of the Korteweg-deVries equation, culminating in a limiting wave of maximum amplitude which has a discontinuous slope ("corner") while all waves larger than the corner wave break, that is, develop infinite slopes in the absence of viscosity as illustrated schematically in Fig. 16.17. Surface water waves with purely gravitational forces are the most familiar example; the CCB Scenario was clearly delineated by Stokes in 1847 [45]. However, all the wave species studied in the articles in Table 16.3 have been shown to be described by the Cnoidal/Corner Wave/Breaking Scenario also (Table 16.4).

16.11 Rossby Solitary Waves

"... if a physical system is capable of supporting solitary wave motions then such motions will invariably arise from quite general excitations."

— T. Maxworthy [72], p. 52.

Constructing a model that incorporates both nonlinear effects and dispersion requires a double perturbative expansion. Nonlinearity vanishes as the wave amplitude $A(x,t)$ goes to zero, so A (strictly speaking, $\max |A(x,t)|$) is one perturbation parameter. The linear dispersion relation for Rossby waves is approximately

$$c = -1/(2n + 1 + k^2) \tag{16.28}$$

which shows that all Rossby waves of a given latitudinal mode n have the same phase speed – the definition of non-dispersive – to within $0(k^2)$ if k is small. Thus, k^2, or equivalently, $1/L^2$ where L is the zonal length scale of the disturbance, is our second, independent perturbation parameter.

We obtain the Korteweg-deVries equation merely by computing both expansions to lowest non-trivial order and then adding the corrections. Because we keep only $O(A)$ and $0(1/L^2)$ and neglect all cross-terms, which are of the much smaller magnitude $O(A/L^2)$, the contributions of the two *different* perturbations are completely

Table 16.4 Rossby solitons: selected references

Reference	Remark
Bouchut, LeSommer and Zeitlin [31]	Equatorial modons colliding elastically with Kelvin waves
Boyd [60]	Derivation of KdV and [antisymmetric modes] MKdV
Boyd [61]	Derivation of NLS Eq., which admits envelope solitons
Boyd [62]	Large amplitude (modons)
Williams and Wilson [63]	Numerical confirmation of (i) Ease of exciting solitons (ii) Near-elastic Rossby soliton collisions (iii) Radiative decay of higher n mode [nonlocal] solitons
Boyd [64]	Radiative decay of nonlocal Rossby solitons
Boyd [65]	Review of nonlinear equatorial waves
Boyd [66]	Monopoles and dipoles in AEW approx
Boyd [67]	$n = 3$ nonlocal Rossby soliton
Boyd [33]	4-mode model
Boyd [68]	Numerical experiments: tilted modons and nonlocal soliton radiation
Greatbatch [3]	Kelvin wave fronts and nonlinear spin-up
LeSommer, Reznik and Zeitlin [36]	Numerical study of response to localized perturbation
Long and Chang [37]	Propagation of Kelvin & Rossby waves in a varying thermocline
Ma [69, 70]	Numerical studies of wave packets and North Brazil Current retroflection eddies
Matsuura and Iizuka [38]	Nonlinear Kelvin waves via a pair of equations like Stern's treatment of coastal waves
Ripa [40]	Ray-tracing
Zhao, Fu and Liu [71]	Envelope soliton in shear flow

independent. We can compute the dispersive correction from the *linear* shallow water wave equations, the same model that we solved exactly in Chap. 3. Similarly, the nonlinear contribution can be evaluated quite independently of the dispersion.[3]

Since solitons are waves for which the dispersion and nonlinearity are of comparable magnitude, we shall take a shortcut by defining a *single* perturbation parameter ε and impose the conditions that

$$A \sim O(\varepsilon) \tag{16.29}$$

$$L \sim O(1/\varepsilon^{l/2}) \tag{16.30}$$

[3]Boyd [32] is an amusing illustration of this independence; the nonlinear coefficient of the KdV equation for Kelvin waves in a shear flow was calculated four years earlier in Boyd [60]! This old computation could be recycled because the dispersion modifies the nonlinear term only if we proceed in the double expansion to a *higher* order than that which gives the Korteweg-deVries equation.

which insure that these two effects are both included. The derivation closely follows Clarke [73] and Boyd [60] and uses the singular perturbation technique known as the "method of multiple scales". For those unfamiliar with this, we refer the reader to Nayfeh [74] and Bender and Orszag [43]. Other references given in Table 16.4.

The equations we solve are the usual nonlinear shallow water set derived in Chap. 3:

$$u_t + uu_x + vu_y - yv + \phi_x = \tag{16.31a}$$
$$v_t + uv_x + vv_y + yu + \phi_y = 0 \tag{16.31b}$$
$$\phi_t + u_x + v_y + (u\phi)_x + (v\phi)_y = 0 \tag{16.31c}$$

The phase speed of solitary waves is the sum of two components: the linear, nondispersive phase velocity for the mode in question plus the (small) correction due to the perturbations. It is convenient to shift into a coordinate system moving with the linear, nondispersive phase velocity for the chosen wave mode so that all time variations will be due to the perturbing effects of nonlinearity and dispersion. Defining

$$s = x - c_0 t \tag{16.32}$$

where c_0 is given by (16.41) below, the equations of motion become

$$u_t + (u - c_0)u_s + vu_y - yv + \phi_s = 0 \tag{16.33a}$$
$$v_t + (u - c_0)v_s + vv_y + yu + \phi_y = 0 \tag{16.33b}$$
$$\phi_t - c_0\phi_s + u_s + v_y + (u\phi)_s + (v\phi)_y = 0 \tag{16.33c}$$

In the method of multiple scales, we normally define both "fast" and "slow" variables. Because of the change of coordinate (16.32), however, we have no "fast" variables, but only the "slow" variables[4]

$$\xi = \varepsilon^{1/2} s \tag{16.34}$$
$$\tau = \varepsilon^{3/2} t \tag{16.35}$$

The unknowns are expanded as

$$u = \varepsilon[u^0(\xi, y, \tau) + \varepsilon u^1(\xi, y, \tau) + \ldots] \tag{16.36}$$
$$v = \varepsilon^{3/2}[v^0(\xi, y, \tau) + \varepsilon v^1(\xi, y, \tau) + \ldots] \tag{16.37}$$
$$\phi = \varepsilon[\phi^0(\xi, y, \tau) + \varepsilon\phi^1(\xi, y, \tau) + \ldots] \tag{16.38}$$

The formal way of deriving the exponents of ε in (16.34)–(16.39) is to allow each unknown to have a symbolic exponent, substitute in (16.33), and then match powers.

[4]Notational warning: ε in this section is not Lamb's parameter but is the perturbation parameter.

The conditions that (i) the lowest order set of equations be the linear, "long wave" or "meridional geostrophy" approximation and (ii) that the dispersive and nonlinear terms all appear at first order then force us to choose the exponents as given above.

The physical explanation of (16.34) is that we choose the "slow" zonal variable ξ so that a unit length scale in ξ corresponds to a scale of $L \sim 1/\varepsilon^{1/2}$, which in turn implies a dispersion that is $0(\varepsilon)$. The scaling of τ implies that the phase speed will be altered by an amount of $0(\varepsilon)$ by the nonlinear-and-dispersive correction; the extra $\varepsilon^{1/2}$ is needed because the phase speed correction must be multiplied by the inverse of the zonal length scale to convert it to a frequency correction. Finally, the reason for the extra factor of $\varepsilon^{1/2}$ multiplying v^0 is that the meridional velocity for an ultra-long Rossby wave is $O(1/L)$ smaller than u and ϕ.

The lowest order set, the unperturbed equations, are

$$-c_0 u^0_\xi - y v^0 + \phi^0_\xi = 0 \tag{16.39a}$$

$$y u^0 + \phi^0_y = 0 \tag{16.39b}$$

$$-c_0 \phi^0_\xi + u^0_\xi + v^0_y = 0 \tag{16.39c}$$

which are identical to the linear, "long wave" or "meridional geostrophy" approximation. The only change from the full linear equations is the neglect of the time derivative in the north-south momentum equation, but this one omission is sufficient to (i) filter out all gravity waves and the mixed Rossby-gravity wave and (ii) make Rossby waves nondispersive as explained in Chap. 4. The solutions derived there are

$$v^0 = N_n A_\xi(\xi, \tau)\psi_n(y) \tag{16.40a}$$

$$u^0 = N_n A(\xi, \tau)\left\{\frac{1}{1-c}\left[\frac{n+1}{2}\right]^{1/2}\psi_{n+1}(y) + \frac{1}{1+c}\left[\frac{n}{2}\right]^{1/2}\psi_{n-1}(y)\right\} \tag{16.40b}$$

$$\phi^0 = N_n A(\xi, \tau)\left\{\frac{1}{1-c}\left[\frac{n+1}{2}\right]^{1/2}\psi_{n+1}(y) - \frac{1}{1+c}\left[\frac{n}{2}\right]^{1/2}\psi_{n-1}(y)\right\} \tag{16.40c}$$

where the N_n are the normalization constants and where the phase speed is

$$c_0 = -1/(2n+1) \qquad n \geq 1 \tag{16.41}$$

The major difference from the "long wave" or "meridional geostrophy" solutions given in Chap. 4 is that the wave amplitude is no longer a constant, but rather a function $A(\xi, \tau)$ of the "slow" variables which will be determined only at the next order in perturbation theory.

Note also that v depends on the ξ-derivative of $A(\xi, \tau)$ whereas u and ϕ depend on $A(\xi, \tau)$ itself. This ξ-differentiation is equivalent to the factor of (ik) that would

multiply v in the analogous solutions for a disturbance with a single zonal wavenumber.

Matching powers of ε gives the first order set:

$$-c_0 u^1_\xi + yv^1 + \phi^1_\xi = -u^0_\tau - u^0 u^0_\xi - v^0 u^0_y \tag{16.42a}$$

$$yv^1 + \phi^1_y = c_0 v^1_\xi \tag{16.42b}$$

$$-c_0 \phi^1_\tau + u^1_\xi + v^1_y = -\phi^0_\tau - (u^0\phi^0)_\xi - (v^0\phi^0)_y \tag{16.42c}$$

As in all perturbation theories, the left-hand side of the first order set is identical with that of the zeroth-order set. We can reduce this system down to the single equation

$$v_{yy} + (-1/c_0 - y^2)v = F_4 \tag{16.43}$$

where, identifying the R. H. S.'s of (5.14) as F_1, F_2, and F_3, respectively,

$$F_4 = [1 - 1/c_0^2](yF_1 + c_0 F_{2\xi}) + yF_1/c_0^2 + yF_3/c_0 + F_{1y}/c_0 + F_{2y} \tag{16.44}$$

Because (16.43) has a homogeneous solution, the forced boundary value problem has a finite solution if and only if F_4 is orthogonal to the homogeneous solution, i.e.,

$$\int_{-\infty}^{\infty} \mathrm{d}\, y \psi_n(y)\, F_4(\xi, y, \tau) = 0 \tag{16.45}$$

The integration removes the y-dependence, but F_4 depends on ξ and τ through $A(\xi, \tau)$ and its *derivatives*. Since (16.45) must hold for *all* ξ and τ, this condition demands that $A(\xi, \tau)$ must satisfy a *differential* equation which turns out to be the Korteweg-deVries equation:

$$A_\tau + a(n) A A_\xi + b(n) A_{\xi\xi\xi} = 0 \tag{16.46}$$

Now that the dummy scaling variable ε has served its purpose, we can set it equal to 1 and also change variables back to the earth-fixed coordinate system to obtain the equivalent equation

$$A_t + c_0(n) A_x + a(n) A A_x + b(n) A_{xxx} = 0 \tag{16.47}$$

One striking feature of the KdV coefficients and also the latitudinal structure functions given by (4.14) — is that they are functions *only* of the *latitudinal mode number* n. Consequently, $a(n)$ and $b(n)$ for all possible cases can be summarized for the first 20 modes by Table 1 of Boyd (1980c).

16.12 Antisymmetric Latitudinal Modes and the Modified Korteweg-deVries (MKdV) Equation

There is one modest complication. If we neglect the equatorial currents, we find that the *nonlinear coefficient* of the *KdV equation* is *zero* for all even n, that is to say, for all modes which have u and ϕ that are antisymmetric about the equator. This does not imply that nonlinear effects are nonexistent for these modes, merely weaker. Boyd [60] shows that by extending the perturbation theory to one higher order, one can derive the so-called "Modified" Korteweg-deVries (MKdV) equation [60]. This is identical in form to (16.47) except that the nonlinear term is $A^2 A_x$, that is, a *cubic* nonlinearity instead of quadratic.

For several reasons, we will omit further discussion of the antisymmetric modes. First, nonlinear effects are weaker for these waves than for the symmetric modes. Second, explicit calculation shows that these equatorial waves do not form solitons. Third, the symmetric component of the equatorial forcing is considerably stronger than the antisymmetric component (except in the Indian Ocean). Fourth, mean zonal currents — unless symmetric with respect to the equator — will invalidate the MKdV analysis by creating a small, non-zero coefficient of quadratic nonlinearity. Unfortunately, this shear-induced current is sensitive to the mean flow profile.

The mixed cubic-and-quadratic KdV/MKdV equation does have a solitary wave known in analytical form. Let f and g be numerical coefficients. The amplitude S parameterizes the entire branch of solutions [actually, *two* branches] which are most easily expressed in terms of the derived parameter

$$q \equiv 12 g S^2 \tag{16.48}$$

Then

$$u_{XXX} + (f^2 g\, u^2 + f u - c) u_X = 0 \tag{16.49}$$

is solved by

$$c = 4S^2 \tag{16.50}$$

$$u = 12S^2(\mu + 1)/(\cosh^2(SX) - \mu \sinh^2(SX)) \tag{16.51}$$

provided that

$$\mu = -1 - 1/q \pm \sqrt{1 + 2q}/q) \tag{16.52}$$

16.13 Shear Effects on Nonlinear Equatorial Waves

The effects of shear on nonlinear equatorial waves have not been thoroughly studied. Boyd [32] shows one method in calculating the shear-induced dispersion for equatorial Kelvin waves. Greatbach [3] illustrates an alternative scheme for calculating the correction to the linear, nondispersive phase speed. Both these articles treated the shear using a form of perturbation theory, but other studies such as Redekopp [75] have demonstrated that the nonlinear/dispersive perturbation theory goes through with only minor changes even when the shear is incorporated directly into the zeroth order equations and these are solved numerically.

The *qualitative* behavior of the waves, however, will not be altered by the mean current because this merely alters the numerical coefficients $a(n)$ and $b(n)$. We note that at least in the absence of shear, the signs of the coefficients are such that equatorial Rossby and Kelvin solitary waves are *always positive*. In other words, these solitons are crests in the language of the shallow water wave equations and thicken the upper layer by lowering the thermocline in the context of the 1-1/2 layer model.

16.14 Equatorial Modons

Independently, Stern [76] and Larichev and Reznik [77] showed that the quasi-geostrophic equations on the midlatitude beta-plane had a class of nonlinear exact solutions which Stern named "modons". Flierl et al. [78], McWilliams et al. [79], Boyd [80], Shen [81], McWilliams and Zabusky [82], Mied and Lindemann [83, 84], McWilliams [85] and Boyd [80, 86] were but the beginning of a new subfield of geophysical fluid dynamics.

The analytic theory of modons is most easily derived in a frame of reference *moving* with the nonlinear phase velocity of the vortex. Flierl et al. [78] label the streamlines in this moving coordinate system as "streaklines", and show that modons must always have a patch of closed streaklines. The outermost closed streakline plays a pivotal role in the theory. The streaklines outside this curve extend to infinity where the isolated vortex has decayed to infinitesimal amplitude. Because the dynamics must be linear at infinity, the potential vorticity must be a *linear function* of the *streamfunction* along all these *open streaklines*. Within the region bounded by the outermost closed streakline, however, the relationship between the potential vorticity and streamfunction cannot be determined by the "far field" behavior of the vortex and in general is very complicated.

When the outermost closed streakline is a circle of radius a on the midlatitude beta-plane, however, one can show that the ad hoc assumption that the potential vorticity is a linear function of the streakfunction Ψ on the closed streaklines, too, gives a consistent solution. The streamfunction is of the form

$$\psi = f(r)\ \sin(\theta) \tag{16.53}$$

in terms of polar coordinates (r, θ) centered on the modon. Tribbia [87] and Verkely [88–91] have added corrections for the earth's sphericity to modon theory, but still within the framework of quasigeostrophy and an outermost closed streakline that is a circle. No attempts to extend the potential vorticity ideas to the shallow water wave equations have yet succeeded. Until such a generalization works, labelling an equatorial beta-plane solution as a "modon" will probably be controversial. Nonetheless, when the Korteweg-deVries theory of Boyd [60] is cranked up to moderate amplitude, the solitary waves bear striking similarities to modons.

Boyd [62] extended the earlier analysis by explicitly solving the first order Eq. (16.40) for the special case of solitons in the $n = 1$ Rossby mode to obtain

$$u = A[-9 + 6y^2]e^{-(1/2)y^2} + c^1\tilde{U}^1(y) + A^2U^1(y) \tag{16.54a}$$

$$\phi = A[3 + 6y^2]e^{-(1/2)y^2} + c^1A\tilde{\phi}^1(y) + A^2\phi^1(y) \tag{16.54b}$$

where

$$c^1 = -0.395B^2 \qquad \text{[First order correction to phase speed]} \tag{16.55}$$

and

$$A(x, t) = 0.771B^2\text{sech}^2\{B[x - (c_0 + c^1)t]\} \tag{16.56}$$

(Note that $A(x, t)$ is not energy-normalized, unlike the previous section.) The functions $V^1(y)$, $U^1(y)$ and $\phi^1(y)$, the latter pair both with and without tildes, are given in in Boyd [62]. The parameter B will be called the "pseudowavenumber" since its reciprocal is the zonal length scale of the soliton, but note that the square of B also gives the amplitude of the solitary wave. This coupling between zonal scale and the amplitude is essential so that the effects of dispersion and nonlinearity are in balance for the soliton.

The first order solutions show that whenever $B \geq 0.53$, a region of closed streaklines will form. Figure 16.18 illustrates the streaklines for the $n = 1$ Rossby soliton for $B = 0.6$. Boyd [62] explains that the first order corrections are only about 12% of the magnitude of the zeroth order soliton computed in Boyd [60], so the graph, although approximate, should be fairly accurate.

The similarities between an equatorial soliton-with-closed-streaklines like that shown in Fig. 16.18 and a modon include:

1. dipolar vorticity: two equal and opposite centers of rotation
2. streaklines are symmetric about the north-south axis
3. streaklines are antisymmetric with respect to the east-west axis
4. the solitary wave propagates east-west only and remains centered on the same latitude circle
5. the fluid inside the closed streaklines is bound to the wave, and must forever recirculate within the region circumscribed by the outermost closed streakline

6. the phase speed is outside the range of speeds for a *linear* Rossby wave of any zonal scale
7. the outermost closed streakline is a circle (or nearly circular).

The first five similarities are obvious from Fig. 16.18. The last two require some comment.

The dynamics of the "far field", that is to say, of regions far from the center of the vortex where the solitary wave has decayed to very small amplitude, are always governed by the *linear* equations of motion. This in turn implies that given the phase speed c of the disturbance, we can expand the soliton in terms of the linear normal modes and then use the *linear* dispersion relation to determine the asymptotic zonal structure for each mode. Of course, we cannot know the coefficients of the modal expansion of the soliton for large radii until we have computed the solitary wave everywhere. Nevertheless, if the linear dispersion relation predicts that some normal modes have a *real* wavenumber k for c equal to the phase speed of the soliton, then this implies that the solitary wave – or at least the part which projects onto those linear modes – will *oscillate* instead of decay. Unless c is such that the linear wavenumbers k are *complex* for *all latitudinal* modes, the soliton will extend, at

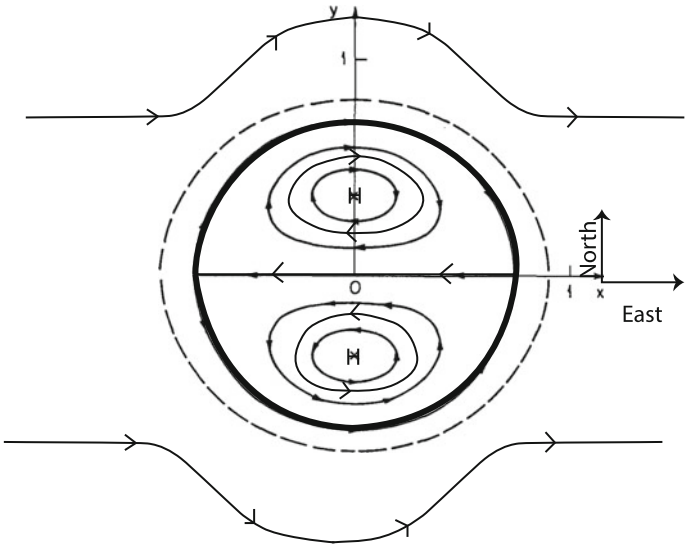

Fig. 16.18 Streaklines [streamlines in the coordinate system moving with the soliton] for $n = 1$ equatorial Rossby soliton for an amplitude about 20% larger than the threshold for closed recirculation. The outermost closed streakline, the bound between the pool of recirculating fluid and the exterior flow, is the *thick curve*. The *top* and *bottom* streaklines are representative of streaklines exterior to the recirculating core. In the moving reference frame, the flow on exterior streaklines is everywhere eastward, the zonal velocity asymptoting as $|x| \to \infty$ to $-c$, the negative of the westward phase velocity of the modon in the earth-fixed coordinate

least with small amplitude, to infinity instead of being an isolated vortex.[5] This requirement for complex wavenumbers means that c must be outside the range we obtain by varying k over the real numbers and computing the phase speed. Thus, (vi) is not an accidental similarity between modons and equatorial Rossby solitons, but rather the consequence of an argument that applies with equal force to many classes of isolated disturbances.[6]

The outermost closed streakline in Fig. 16.18 differs from a perfect circle by only about 2%, but (vii) nonetheless represents a heuristic concept rather than a theorem. The reason that the recirculating regions for modons is always bounded by a circle in modon theory is a limitation of the analysis, not the physics. Ma and Boyd [92] numerically computed modons whose outermost closed streakline was an ellipse extending to rather high eccentricities. McWilliams [85] has shown that numerically-generated modons have a much richer structure than described by the theory (including small oscillations in time), but the shape of the patch of fluid which is bound to the wave is disk-like.

To reiterate this point, we did something rather peculiar: computed the *equatorial* soliton in *polar* coordinates centered on the equator, and then computed the coefficients of the Fourier expansion of the latitudinal and zonal velocities (unpublished). A little calculus will show that for a streamfunction in modon form – a function of r multiplied by $\sin\theta$ – the derivatives in Cartesian coordinates, which give the x and y velocities, should have Fourier coefficients in θ which are all zero except for (a) the coefficient of $\sin(\theta)$ for the north-south current and (b) the constant and $\cos(2\theta)$ terms for the zonal flow. The graphs show that these coefficients completely dominate the expansion. The maximum of $b_4(r)$ for v is only 12% of the coefficient of $\sin(\theta)$ and the maximum of $a_4(r)$ for the x velocity is only 5.5% of the largest value of $a_0(r)$, the component of the zonal flow which is a function only of radius. Although we have slightly mixed our metaphors – or at least coordinate systems – by expanding Cartesian velocities in polar coordinates, the conclusion is straightforward: in spite of the spatial homogeneity created by the variable Coriolis parameter, the equatorial soliton has a structure that closely resembles that of a modon (see also [68]).

The Korteweg-deVries theory implies that modon-like solitons should be readily generated by strong impulsive events like El Nino; this is born out by the numerical models. However, modon theory predicts the existence of *eastward-traveling* vortex pairs and the KdV analysis does not. Before we turn to modons in numerical models, we will first consider an improvement of the KdV theory that predicts solitary waves traveling both east and west.

---

[5] Tribbia [87], Verkely [88–91] and [on the midlatitude beta-plane (Boyd, [86])] generalize the notion of a "solitary" wave by considering modons which will violate this constraint that c outside the linear range [86–91]. These quasi-solitons are global, but the amplitude outside a small circle is less than 10% of the maximum of the vortex, so the modon has a well-defined local peak in spite of the fact that it extends in the form of small ripples over the whole planet. Such "weakly nonlocal solitons" are discussed in length in Sect. 16.18.

[6] The "envelope" solitons that solve the Nonlinear Schroedinger (NLS) equation are an exception that we will explain in later sections.

16.15 A KdV Alternative: The Regularized Long Wave (RLW) Equation

16.15.1 The Useful Non-uniqueness of Perturbation Theory

Peregrine [93] and Benjamin, Bona, and Mahony [94] independently pointed out that the Korteweg-deVries equation has poor numerical properties in the sense that the time step must decrease as the third power of the spatial grid size. The problem is that the dispersion relation for the linearized KdV equation is

$$c = c_0 - b(n)k^2, \qquad b = \frac{4n(n+1)}{(2n+1)^4} \tag{16.57}$$

which shows that the phase speed increases without bound as $k \to \infty$. (Water waves traveling faster than the speed of light?) However, the KdV equation is derived via perturbation theory. Equation (16.57) is an approximation to the true linear dispersion relation with an error $O(k^4)$. Therefore, it is equally consistent with perturbation theory to *replace* the KdV equation by another equation which approximates the dispersion relation to the same order of approximation in the zonal wavenumber k. In particular, they suggested the Regularized Long Wave (RLW) equation:

$$A_t + c_0 A_x + a(n) A A_x - [b(n)/c_0] A_{xxt} = 0 \qquad \text{[RLW Eq.]} \tag{16.58}$$

which is *identical* with the KdV equation except for the linear dispersive term, the triple derivative.

A long wave approximation is a power series approximation to a transcendental dispersion relation $\omega(k)$. Padé approximants are alternative approximations which are rational functions $P_{M/N}$ whose numerator and denominator polynomials of degree M and N, respectively, are chosen so that the first $(M + N)$ terms of the rational function match the first $(M + N)$ terms of the function being approximated, $\omega(k)$ (Bender and Orszag [43]). The KdV linearized dispersion relation and the RLW dispersion relation are the $[2/0]$ and $[0/2]$ approximations to $\omega(k)$. The altered linear dispersion relation profoundly changes the CFL limit from proportional to N^3 for KdV to linear in N for RLW.

Before discussing the phenomenology of the KdV equation, it is useful to note that the apparent uniqueness of this reductive model, a one-dimensional window into three-dimensional reality, is only apparent. Because the perturbative theory is a *long wave* approximation, it does not pretend to describe what happens to the short zonal wave part of the Fourier spectrum in longitude.

The advantages and disadvantages of the RLW equation, discussed in more detail below, are the following:

Virtues

1. Much larger CFL timestep limit for explicit time-marching schemes [two orders of magnitude, a factor of $O(N^2)$]
2. Much greater fidelity to the linear dispersion relation for Rossby waves
3. Explicit approximation to eastward-propagating modon-like solitons

Flaws

1. Not an integrable system and therefore lacking the conservation laws and freedom from chaos of an integrable model like KdV
2. Inelastic soliton-soliton collisions
3. Not solvable by the inverse scattering method.

For *numerical* purposes, the RLW equation is greatly preferable to the Korteweg-deVries equation because the largest stable time step is directly proportional to the spatial grid size rather than to its cube, as true of the KdV equation.

For *analytical* purposes, however, the KdV equation is preferable because it can be solved exactly via the so-called "inverse scattering" method whereas the RLW equation is not, and the mathematical properties of its solutions are much less well understood.

However, for Rossby waves, there is an extra *physical* reason for preferring the RLW equation; the dispersion relation for Rossby waves *is* a rational function to a high degree of approximation. If we substitute the explicit forms for c_2 and c_0 into the RLW equation,we obtain as the linear RLW dispersion relation

$$c = -\frac{1}{(2n+1+k^2[1-1/(2n+1)^2])}$$

which is identical, except for the small correction term proportional to $1/(2n+1)^2$ which multiplies k^2, to the approximate Rossby wave dispersion given in Chap. 3, $c_{Rossby} \approx -1/(2n+1+k^2)$.

KdV soliton-soliton collisions are "elastic' in physics jargon in the sense that the entire energy is recycled into the solitons themselves; no debris or decaying, dispersing transient is left behind. RLW collisions, alas, leave tiny small-scale ripples in the wake of collisions.

Bona, Pritchard and Scott [95] resolved a controversy through a very careful numerical study. Zoom plots were needed to shown the remnants, which are so small as to be invisible on graphs that show the full height of the solitary waves themselves.

The inelasticity implies that the RLW is not integrable, the inverse scattering fails, and chaotic behavior is possible.

16.15.2 Eastward-Traveling Modons and Other Cryptozoa

One conceptual value of the RLW model is that it allows solitons traveling both to the east and the west – just like midlatitude modon theory. To date, the RLW equation is the only equatorial model that predicts eastward-traveling modons.

For the $n = 1$ Rossby mode, Boyd [62] shows that the RLW solitons are

$$A(x,t) = -2.31B^2\, c\, \mathrm{sech}^2[B(x-ct)] \tag{16.59}$$

where

$$c = -(1/3)/(1 - 1.184B^2) \tag{16.60}$$

When $B > 0.92$, the solitary wave has a ϕ field that is everywhere negative and the disturbance propagates eastward with $c > 0$ — again a phase speed outside the linear range, which is $-1/3 \le c < 0$.

Modons of either sign are a pair of equal but opposite vortices; for westward-traveling modons (and $n = 1$ equatorial solitary waves) the northern vortex spins clockwise while the southern vortex is counterclockwise so that their mutual interaction propels the pair at a faster-than-linear speed to the west.

Eastward-traveling vortex pairs must be of the opposite sign, but there is an additional constraint: the vortices must be sufficiently strong to drive the pair eastward in spite of the linear tendency for phase propagation to the west. Otherwise, the net phase speed will fall within the linear range and the disturbance will radiate, at best a "nonlocal solitary wave" in the terminology of Sect. 16.18. This need for *strong* nonlinearity explains why eastward-traveling solitons exist only for *large* values of B.

Equations (16.59) and (16.60) must be taken with many grains of salt because the perturbative treatment of nonlinear effects for such *large* amplitude ($B^2 > .85$) is unreliable.

The Madden–Julian Oscillation in the tropical atmosphere has a dipole structure with eastward propagation. In hopes of imitating the MJO, Wedi and Smolarkiewicz employed a meandering northern boundary as forcing for their numerical model in a rectangular domain. For a limited range of parameters, they did coax vaguely dipolar eastward-moving structures. Convectively-modified eastward modons? Perhaps [96]. But the ocean does not have a moving corrugated wall in the subtropics, and they were tireless tweakers.

In cryptozoology, a "cryptid" is a species whose existence is conjecture, folklore or speculation, as yet unconfirmed by highly credible observations. Eastward modons should exist as unforced free modes. Despite independent searches by Boyd and Jun-ichi Yano (private communication), eastward-traveling equatorial vortex pairs remain an intriguing cryptid.

16.16 Phenomenology of the Korteweg-deVries Equation on an Unbounded Domain

16.16.1 Standard Form/Group Invariance

The KdV equation with arbitrary coefficients can always be reduced to a convenient standard form through the following group invariance theorem.

Theorem 16.2 (KdV Reduction to Canonical Form) *The KdV equation with arbitrary coefficients, but written in a frame of reference moving at the linear long wave speed* c_0,

$$A_\tau + FA_{\xi\xi\xi} + GAA_\xi = 0, \tag{16.61}$$

is related to the solution u *of the canonical form*

$$u_t + u_{xxx} + \mho u u_x = 0 \tag{16.62}$$

by

$$t = \tau \tag{16.63}$$

$$x = \frac{1}{F^{1/3}}\xi \tag{16.64}$$

$$u = \frac{G}{\mho F^{1/3}} A \tag{16.65}$$

In most of our numerical simulations, we choose $\mho = 1$, but to describe the computation of soliton parameters it is convenient to use Whitham's convention, $\mho = -6$.

16.16.2 The KdV Equation and Longitudinal Boundaries

The perceptive reader may be somewhat perplexed by boundary conditions for an equation which is third order in space. These concerns are physically misplaced because the KdV and RLW equations are both "one-way wave equations". Each equation has only a single time derivative and therefore the linear dispersion is single-branched; the waves propagate in one direction only for a given longitudinal wavenumber k. Dennis Moore showed in his Harvard thesis half a century ago that long equatorial waves reflect from the western boundary of a tropical ocean as a Kelvin wave plus short (small longitudinal wavelength) Rossby waves with eastward group velocities. Such a complete change of modal identity can only be

accommodated by a *system* of equations rich enough to allow both two-way wave propagation and a plethora of latitudinal modes.

It follows that it is sensible to discuss KdV (and RLW) phenomenology only on an unbounded domain or a periodic interval. In mid-ocean, far from coastal boundaries, the choice of boundary conditions is irrelevant and one may use whatever boundary conditions are mathematically convenient and non-reflective. Such mid-ocean models are often called "process" models because simulations that are free of coastlines cannot imitate a complete ocean, but are useful in studying processes (such as nonlinear frontogenesis and soliton formation).

A related question is: Does the neglect of boundaries in the inverse scattering method compromise the usefulness of the KdV model? No. The appendix of Boyd [62] explains that the neglect of boundaries is consistent with the perturbation theory provided that one uses the shape of the wave after it has left the coast to initialize the KdV model.

Boyd [60], recent monographs like Ablowitz and Segur [97] and geophysica-11y-oriented reviews like Malanotte-Rizzoli [98] and Hammack [99–101] all give good discussions of the solutions of the KdV equations, so in this section, we shal1 describe only the major conclusions.

The Korteweg-deVries equation and the Nonlinear Schroedinger equation, described later, can be solved *exactly* for *general* initial conditions by the so-called "inverse scattering" method. The simplest case is that in which the boundaries are ignored and the spatial interval is idealized as $[-\infty, \infty]$, but a later extension, usually called the "finite gap" or "N-band" method (Osborne [102], Haupt and Boyd [103, 104]) does the same job when the boundary conditions are spatial periodicity [102–104].

The great usefulness of "inverse scattering" is that it shows that solitons — far from being "curiosities" — are created by quite *arbitrary* initial conditions. Indeed, rather special initial data (such as $A(x, t = 0) \leq 0$ for all x) are required to prevent the formation of at least one soliton. What happens is that the flow evolves in such a way that part of the disturbance settles into a balance between nonlinearity and dispersion (as $t \to \infty$) even if the initial condition was very far from such a balance.

The general solution on $x \in [-\infty, \infty]$ is the sum of two components: (i) a finite number of solitary waves and (ii) an oscillatory wavetrain, which gradually disperses with the amplitude decreasing as $1/t^{1/2}$. If we use the coordinate system moving at the unperturbed phase speed c_0, then the solitons drift westward of the origin – that is, the solitons propagate faster westward than any linear Rossby wave – while the oscillatory wavetrain expands eastward as it disperses. Thus, the two parts of the solution separate as $t \to \infty$, which makes it easy to identify them as shown in Fig. 16.19.

The oscillatory wavetrain is much like the general solution of the *linearized* KdV equation. The usual large-time theory of dispersive linear waves, ably explained in Whitham [8], predicts that most of the wavetrain will decay as $1/\sqrt{t}$. Closest to the initial pulse and traveling at c_0, the linear long wave speed, is a caustic proportional to $t^{-1/3}\mathrm{Ai}(\text{constant}[x - c_0]/t^{1/3})$ where $\mathrm{Ai}(z)$ is the usual Airy function. Ablowitz and Segur ([105]) show that nonlinearity modifies the wave train in a variety of ways,

some subtle, but the most important is the replacement of the parameter-free $\mathrm{Ai}(z)$, which solves $\mathrm{Ai}_{zz} - z\mathrm{Ai} = 0$, by the second Painleve Transcendent $P_{II}(z)$, which solves $P_{II,zz} - zP_{II} - 2P_{II}^3 = 0$ [106]. Remarkably, the nonlinear first trough decays as $t^{-2/3}$ whereas the linear caustic decays as $t^{-1/3}$ as already noted.

In contrast, the soliton is something completely new: a species that has no counterpart in a linear calculation. Figure 16.20 compares three KdV solutions to show this contrast. The positive initial condition has the exact shape of a solitary wave, so the nonlinear solution merely propagates at constant speed without change in form. The initial shape is the final shape. Dispersion opposes frontogenesis to produce a wave of permanent, unvarying form.

The second and third solutions were both initialized by the *negative* of the solitary wave, an "anti-soliton" initialization. The linear and nonlinear troughs both disperse. The red curve of intermediate thickness is the linear solution, obtained by omitting the advective term in the KdV equation. The thick green curve is the nonlinear solution. It resembles the linear solution at a later time; with sign reversed from the soliton, nonlinearity is now working with, rather than against the dispersion to produce an accelerated *"superlinear dispersion"*.

For special initial conditions, either the solitons or the wavetrain may be missing, but unless $A(x, 0)$ is carefully contrived, both will be present. However, for a rather large class of initial conditions – a smooth, positive semi-definite initial condition of

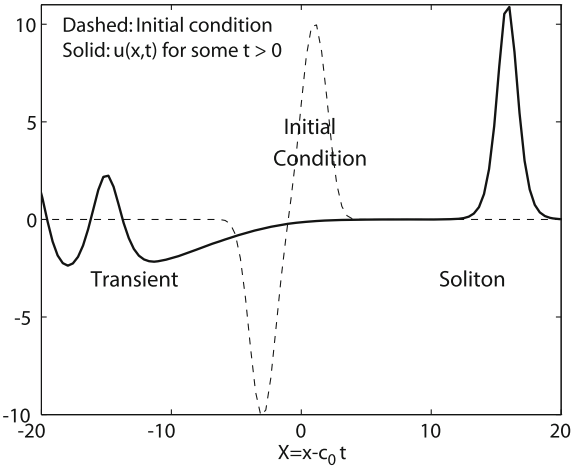

Fig. 16.19 The general solution of KdV consists of one or more solitary waves propagating faster than c_0 plus a dispersing transient which moves slower than c_0. The overlap between the solitary wave and the transient decreases exponentially fast with time because of the non-zero phase speed difference between the transient and the solitons and the exponential decay of the solitary wave with increasing distance from its peak. The solution is shown in a frame of reference that moves at the linear long wave speed c_0, i.e., $X \equiv x - c_0 t$. Warning: in this figure and the rest of this section, the usual KdV convention is employed in which the solitary waves move to the right, and the transients disperse leftward. For Rossby waves, rightward=westward

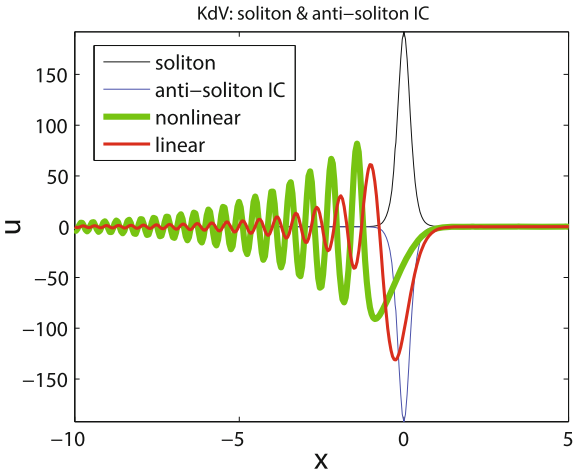

Fig. 16.20 Three solutions to $A_t + AA_x + A_{xxx} = 0$. The thin positive curve is the initial position of the solitary wave $A(x,t) = 192\,\mathrm{sech}^2(16[x-64t])$. The reference frame is moving so the soliton is stationary. In contrast, the initial condition of the negative of the soliton labeled "initial anti-soliton", evolved with time. If the nonlinear terms are omitted, the *thick red curve* results. Nonlinearity generates a rather different solution shown as the *thin green line*; nonlinearity accelerates dispersion of a *negative* initial feature

moderate or large amplitude is one such class - - the wavetrain is weak and the solitons inherit more than 90% of the initial energy. This class includes initial conditions for which nonlinearity dominates dispersion so that the flow *initially* evolves according to the nondispersive One-Dimensional Advection (ODA) equation.

To emphasize the interchangeability of the KdV and RLW models, we here mix KdV and RLW examples at random.

Figure 16.21 shows an RLW integration from an initial condition which was a soliton (shown as the thin, dashed green curve) multiplied by a factor of one-half. This destroys the balance between nonlinearity and dispersion, but the flow *evolves* to an exact nonlinear/dispersion balance. The initial disturbance sheds a transient whose caustic is the rightmost trough, and the tall peak widens slightly to weaken dispersion and then propagates rightward, separating farther and farther from the transient. (At the time shown, the soliton is still close to the caustic, but continuing the integration to larger time shows that the right peak is a soliton.) Most of the energy of the initial condition, even though the initial state is not a soliton, is carried away by the short, fat solitary wave that develops.

An initial condition with much larger amplitude will be equally imbalanced, but nonlinearity initially dominates dispersion as shown in Fig. 16.22. The leading (rightward) edge steepens. Dispersion finally comes into play when the disturbance has developed a steep front: the crest abruptly dissolves into solitons at the very moment we expect the front to break, but then dispersion becomes important when the length scale in the frontal zone has become very narrow. The monotonic wave profile breaks

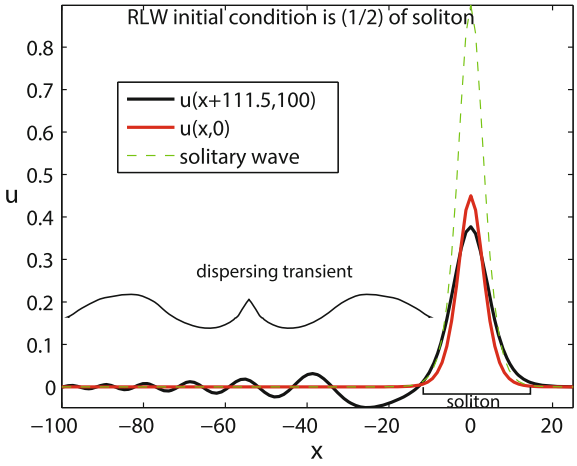

Fig. 16.21 The tall *thin green curve*: an RLW solitary wave. *Thick red curve*: the actual initial condition, which is *half* the soliton. The *thick black curve* shows the shape of the solution at a later time. Note that the RLW solution, which actually propagated far to the right, has been shifted leftward for visualization to align the short, fat evolved soliton [*black*] with the taller, narrower non-solitonic initial condition [*red*]. The dispersive transient is also shown shifted and in *black*

up into many solitons. The envelope of the peaks resembles a right triangle. It can be proved through inverse scattering that maximum height can be no more than double the highest point on the initial condition, but a near-doubling is evident. The five peaks still weakly overlap at $t = 50$. The inset shows that the peaks are solitary waves; because they are ordered tallest-and-fastest to the right, the five solitons will become farther and farther apart as $t \to \infty$.

For a smooth, positive definite initial condition, almost all the initial energy goes into its solitons even from a far-from-solitary-wave initial profile.

Boyd [60] dubbed this"soliton dominance": solitary waves are not special or exotic, but sometimes are almost all that there is. A mathematician would shudder at such a heuristic concept because it is easy to contrive counter-examples[7]; the windowed sine initial condition illustrated in Fig. 16.23 generates both a large transient as well as a pair of solitons. Nonetheless, the inverse scattering method has made it clear that for the Korteweg-deVries equation, solitons are not a curiosity, but often the only thing.

[7] A negative definite initial condition or a multi-signed $A(x, 0)$ of very small amplitude or short zonal scale will disperse almost completely as $t \to \infty$.

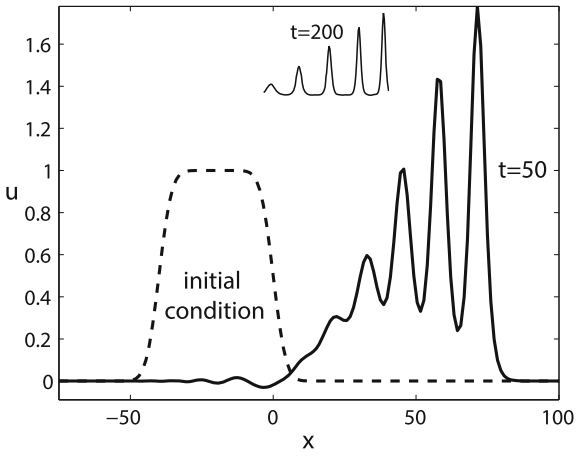

Fig. 16.22 The solitons that emerge from an arbitrary initial condition, such as the "top hat" shown as the *dashed lines* here, can be predicted by calculating the bound states of the Schrödinger equation using the initial condition as the Schrödinger potential energy. Even when the peaks are still overlapping, as at $t = 50$, here the asymptotic state (insert) is foreordained

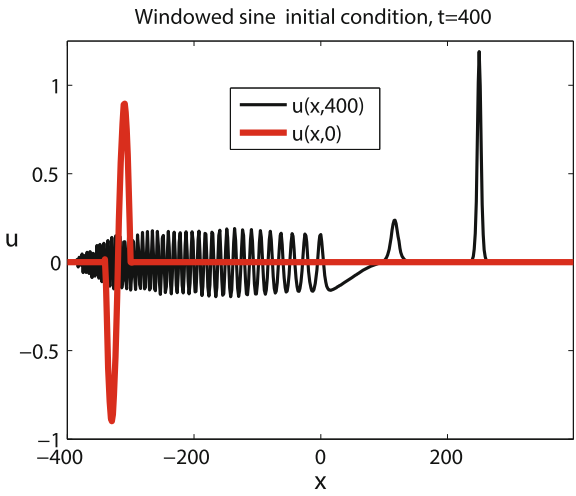

Fig. 16.23 $N = 512$ grid points on $x \in [-400, 400]$ for $u(x, 0) = -0.9 \sin([\pi/20][x + 300])(1/2)\{\text{erf}(x + 320 + 20) - \text{erf}(x + 320 - 20)\}$

16.16.3 Calculating the Solitons Only

Calculating the dispersing, transient part of the KdV solution by the inverse scattering method is very difficult, even though all the steps are linear. Computing just the amplitudes, widths, speeds and energies of the solitary waves that will emerge from a given initial condition is relatively easy. Following Whitham [8], assume that the KdV equation has been reduced to the canonical form

$$u_t + u_{xxx} + uu_x = 0 \tag{16.66}$$

with the initial condition

$$u(x,0) = Q(x) \tag{16.67}$$

The linear eigenvalue problem

$$u_{xx} - \left\{ E - \frac{1}{6} Q(x) \right\} u = 0, \qquad x \in [-\infty, \infty] \tag{16.68}$$

is identical with the stationary Schroedinger equation of quantum mechanics with E the eigenvalue equal to the energy of an eigenmode and the KdV initial condition furnishing the potential energy. The complete spectrum of the Schroedinger solutions consists of a continuous spectrum, which determines the dispersing transient in the KdV solution, plus a countable number of discrete modes, which physicists call "bound states". If $Q(x)$ is negative definite, a "potential barrier", there are no bound states and it can be proved that the initial condition generates no solitary waves. If $Q(x) \geq 0$, a "potential well", there is always at least one bound state. If the energy eigenvalues are written in terms of a new variable B_n as $E_n = -B_n^2$, then the asymptotic solution of the KdV equation from an initial condition that generates N Schroedinger bound states will consist of N solitary waves each of the form

$$u_n(x,t) = 12 B_n^2 \mathrm{sech}^2(B_n[x - 4B_n^2 t + \varphi_n]) \tag{16.69}$$

where φ_n is a phase constant that requires knowledge of the continuous spectrum also.

Thus, by solving a *one-dimensional, linear eigenvalue problem*, one can determine the *number*, *amplitude*, *width* and *speed* of all the solitary waves that will emerge from a given KdV initial condition $Q(x)$.

Calculating the dispersing transient [and soliton positions] from the continuous spectrum is much harder; the integral equation procedure has now been replaced in modern practice by a Riemann-Hilbert problem, but this is still much more difficult than a one-dimensional eigenvalue problem. In practice, the non-soliton parts are usually calculated by a Fourier pseudospectral/Runge–Kutta method that makes no use of the inverse scattering method whatsoever.

Fortunately, the dispersing transient is as ephemeral as smoke.

16.16.4 Elastic Soliton Collisions

The speed of a solitary wave is proportional to the *square* of the height of pulse. Parenthetically, note that the width is inversely proportional to the square root of the amplitude. A tall, narrow soliton will therefore overtake all shorter, broader solitary waves eventually. What then?

Zabusky and Kruskal discovered more than half a century ago that KdV solitay waves collide *elastically*, that is, the solitary waves nvariably emerge from a

collision *unchanged in shape, speed* and *energy*, altered only by a *phase shift*. A KdV solitary wave can no more be collisionally reduced to two-thirds of a solitary wave or a collection of dispersing waves or wave debris than an electron can be transformed by collisions into two-thirds of an electron. Zabusky and Kruskal coined the term "soliton" as an appelation for the solitary wave to emphasize this particle-like property.

The inverse scattering method proves that KdV soliton-soliton collisions are always "elastic" in this sense. Figure 16.24 illustrates a typical case.

There is an explicit, closed-form solution which describes all possible KdV solutions that begin and end with N solitary waves. Unfortunately, the N-soliton is a messy ratio of the determinants of matrices whose elements are exponential functions of time and space with N different arguments. We therefore leave the details to Whitham [8].

16.16.5 Periodic BC

The *linear* dynamics of the KdV equation with periodic boundary conditions with period P is easy. Expand the initial condition as a Fourier series:

$$u(x,0) = Q(x) = \sum_{n=1}^{\infty} a_n \cos(n[2\pi/P]x) + \sum_{n=1}^{\infty} b_n \sin(n[2\pi/P]x) \quad (16.70)$$

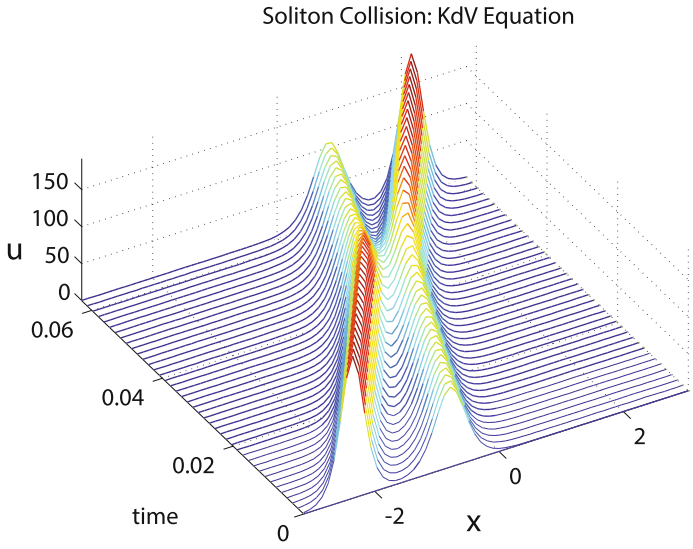

Fig. 16.24 Waterfall plot in the $x - t$ plane of two KdV solitary waves with collision.

Then the general solution to $u_t + u_{xxx} = 0$ is

$$u(x,t) = \sum_{n=1}^{\infty} a_n \cos(n[2\pi/P][x - c_n t]) + \sum_{n=1}^{\infty} b_n \sin(n[2\pi/P][x - c_n t]) \tag{16.71}$$

where

$$c_n = -\frac{4\pi^2}{P^2} n^2, \qquad n = 1, 2, \ldots \tag{16.72}$$

The sine and cosine of the same wavenumber can be combined into a single term:

$$\begin{aligned} a_n \cos(n[2\pi/P][x - c_n t]) + b_n \sin(n[2\pi/P][x - ct]) &= A_n \cos(n[2\pi/P][x - c_n t] + \varphi_n) \\ A_n &= \sqrt{a_n^2 + b_n^2} \\ \varphi_n &= \arctan(b_n/a_n) \end{aligned} \tag{16.73}$$

In practice this series must be truncated at $n = N$ for some user-chosen N. Although there is only a single spatial variable, the linear solution can also be regarded as infinite-dimensional (or with truncation, N-dimensional) function of the phase variables

$$\zeta_n \equiv n\frac{2\pi}{P}(x - c_n t) + \varphi_n \tag{16.74}$$

There is a spatially-periodic generalization of the inverse scattering method which is known variously as the "method of finite gap solutions", the "method of finite band solutions", the "periodic Nonlinear Fourier Transform" and "periodic inverse scattering". The KdV equation is fully integrable in the Hamiltonian dynamics sense with an infinite number of discrete conservation laws. The theory asserts that in the nonlinear case as for the dynamics of its linear, infinitesimal amplitude solutions, (i) There exists solutions with N phase variables ζ_n and (ii) The solution evolving from an arbitrary initial condition can be approximated arbitrarily well by an N-phase solution for sufficiently large N and wisely chosen solution parameters.

This is an astoundingly bold statement. It is well-known that chaos, which is not periodic and cannot be represented as an N-phase function, is ubiquitous in realistic fluid systems. One would imagine that, even with the neglect of explicit forcing and dissipation, the interaction of one solitary wave with another would somehow accelerate and deaccelerate interacting pulses in such a way as to destroy any simple dependence on steadily translating phase variables. The assertion is nonetheless true.

The N-phase solutions have a multiplicity of names, but here they will be labeled "N-polycnoidal waves" with "cnoidal", "double cnoidal" and "triple cnoidal" for the $N = 1$, $N = 2$ and $N = 3$ cases.

The ordinary cnoidal wave is a combination of elliptic functions and can also be written as a ratio of theta functions. N-polycnoidal waves with $N > 1$ are N-variate hyperelliptic functions that can be written as ratios of N-variate theta functions.

Alfred Osborne's *Nonlinear Ocean Waves and the Inverse Scattering Transform* is the definitive treatment [102]. Unfortunately, although his book is very readable, it is always a bad sign when the chapters on applications begin on pg. 570. He shows that one can analyze nonlinear water waves as thoroughly as linear waves and demonstrates that the Nonlinear Fourier Transform is as useful as the usual Fourier Transform. However, one must take a one-semester reading course to follow Osborne's path. A good introduction which requires no knowledge except a general grasp of wave theory and a feeling for ordinary Fourier series is Boyd and Haupt [103, 104]. Although the author has written half a dozen papers on polycnoidal wave theory, [103, 104, 107–111], it is necessary to leave most details to the references. There are, however, some basics that are both easily explained and significant for understanding nonlinear waves in the tropical ocean.

16.16.6 The KdV Cnoidal Wave

Korteweg and de Vries themselves discovered single-phase exact nonlinear solutions to the KdV equation. Because the solutions were a simple function of the elliptic

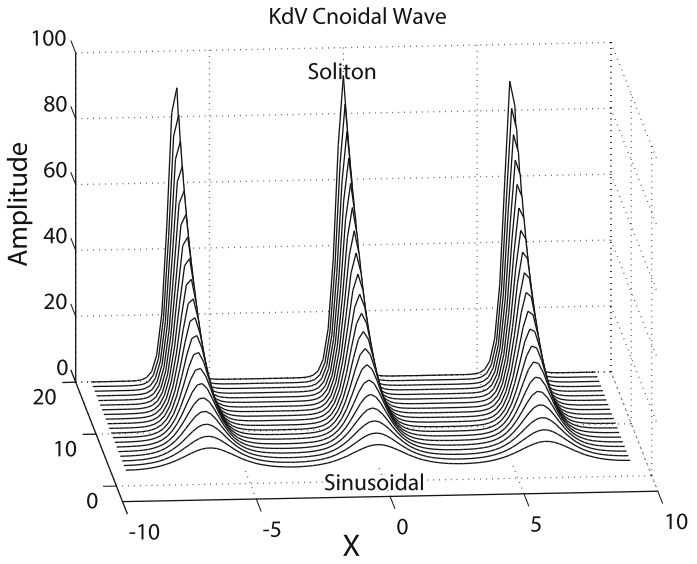

Fig. 16.25 The KdV cnoidal wave as a function of x (for fixed time) and amplitude a. For small amplitude (foreground), the cnoidal wave is sinusoidal. As the amplitude increases, the peaks become taller, narrower, and more soliton-like

function whose symbol is "cn", they recalled the Greek termination "-oidal" meaning "similar to" or "like" and dubbed these spatially periodic traveling undulations "cnoidal waves".

These form a one-parameter family of shapes where one widely used parameter is the "elliptic modulus", $m \in [0, 1]$. Cnoidal waves are infinitesimal amplitude linear sine waves for small m and are arbitrarily tall and narrow peaks, well-approximated by shape of a KdV soliton, in the limit $m \to 1$ as illustrated in Fig. 16.25.

16.17 Soliton Myths and Amazements

16.17.1 Imbricate Series and the Nonlinear Superposition Principle

Assertion 16.1 (*Nonlinear Superposition)* For many types of cnoidal waves described by elliptic functions, the cnoidal wave is the exact superposition of solitary waves.

A function $G(x)$ which is spatially localized and decays sufficiently fast as $|x| \to \infty$ such as solitary wave, can always be "periodized" or "imbricated" to generate a spatially periodic function with arbitrary period P. First, make an infinite number of copies of $G(x)$. Second, space the copies a distance apart over the entire real axis

$$f_P(x) \equiv \sum_{m=-\infty}^{\infty} G(x - mP) \tag{16.75}$$

The "imbricate series" clearly has the required spatial periodicity, that $f_P(x + P) = f_P(x)$ for all real x.

One important but rarely discussed piece of mathematical folklore is that *every* periodic function has an imbricate series. The justification for this claim is the Poisson Summation Theorem, which asserts that the pattern function is the Fourier Transform of the function which, evaluated at the integers, gives the Fourier coefficients.

When the solitons are narrow compared to the spatial period, P, it is obvious that the periodization of the solitary wave will be an *approximate* solution to the wave equation with period P. The solitary waves have an overlap which is exponentially small in $1/P$; when $P \gg 1$, the overlap is negligible and the periodic solution may be approximated on an interval around each peak by just the solitary wave itself.

The imbrication of the KdV solitary wave is both banal and bizarre. It is banal because every localized lump has an imbricate series, and the imbrication of solitons must be a good approximation to a periodic function when the spatial period P is large compared to the soliton width. It is bizarre because the KdV cnoidal equation obeys a nonlinear superposition principle: the KdV cnoidal wave is not *approximately* the

sum of an infinite series of evenly spaced solitary waves; it is *exactly* the sum of an imbricate series of solitary waves for all amplitudes!

This remarkable fact was discovered by Toda in 1975 [112] using the infinite product for the theta function whose second logarithmic derivative is the cnoidal wave; Gerald Whitham in his next-to-last published paper gave an elementary proof in 1984 [113] of what was true "to my amazement, at least". Boyd [108] and Zaitsev [108, 114] also gave different proofs of the "Toda Miracle".[8]

Denote the spatial period by P and normalize the KdV equation to

$$u_{xxx} - cu_x + uu_x = 0 \tag{16.76}$$

The phase speed is

$$c = M - 24\frac{\varepsilon}{P} + 4\varepsilon^2 - 24\varepsilon^2 \sum_{n=1}^{\infty} \text{cosech}^2(n\,P\,\varepsilon) \tag{16.77}$$

and $u(X)$ [where $X \equiv x - ct$] is given by either the imbricate series [top] or Fourier expansion [bottom]

$$\begin{aligned} u(x) &= M - 24\frac{\varepsilon}{P} + 12\varepsilon^2 \sum_{j=-\infty}^{\infty} \text{sech}^2(\varepsilon(x - Pj)) \\ &= M + 12\frac{4\pi^2}{P^2} \sum_{j=1}^{\infty} \frac{n}{\sinh\left(\frac{n\pi^2}{\varepsilon P}\right)} \cos([2\pi/P]nx) \end{aligned} \tag{16.78}$$

where M and $\varepsilon > 0$ are arbitrary constants. (The solution is given in terms of elliptic functions in [108], Appendix D of [116] and most books on nonlinear waves.)

Note that M is merely an additive shift added simultaneously to both $u(x)$ and c. The statement that the cnoidal wave is the exact superposition of solitary waves is true only with a slight qualification. The Fourier series has zero mean, that is, $\int_0^P u(X)dX = 0$. However, solitary waves are positive definite; the imbrication of solitons has a nonzero mean. A group invariance theorem states that the sum of a solution to the KdV equation with an arbitrary constant M is also a KdV solution if the same constant is added to the phase speed, as easily proved by substitution into the differential equation. The additive factor of $-24\frac{\varepsilon}{P}$ converts the imbricate series of solitary waves into the zero mean cnoidal wave.

One reason that the imbricate series is useful is that it converges rapidly when its pattern function is narrow and the Fourier coefficients decay slowly while the Fourier series is fast when the solitons are very wide and strongly overlap. The elliptic modulus $m = 1/2$ is where the Fourier series and imbricate series converge

[8] I was as surprised as Whitham, but made heavy use of the Nonlinear Superposition Principle in my work on polycnoidal waves [108–111, 115].

at equal rates. This parameter value is thus a boundary between the soliton and sine wave regimes.

16.17.2 The Lemniscate Cnoidal Wave: Strong Overlap of the Soliton and Sine Wave Regimes

Assertion 16.2 (*Strong Overlap of Soliton and Sine Wave Regimes*) A periodic structure that is well-approximated by the pattern function of its imbricate series may also be well-approximated by a rapidly convergent Fourier series.

The case of equal convergence rates for Fourier and imbricate series has a special name, the "lemniscate" case. Gauss coined the names "cosine-lemniscate" and "sine-lemniscate" in 1797 [117].

The lemniscate cnoidal wave is $u(X) = -0.95493 + 2.08975\,\mathrm{cn}^2(.5902X; m = 1/2)$ where $X = x - ct$ is the spatial coordinate in a frame of reference moving with the wave. For period 2π, the Fourier series is

$$u(X; m = 1/2) = 1.039\cos(X) + 0.090\cos(2X) + 0.0058\cos(3X) + \ldots \quad (16.79)$$

which shows the very rapid convergence of the Fourier series; the second harmonic is less than one-twelfth the amplitude of $\cos(X)$. For this special case, the zero-mean cnoidal wave has an exact phase speed of -0.95493. The phase speed of an infinitesimal amplitude cosine wave, $c_{cos} = -1$, is in error by only 4.7%, consistent with the fact that its Fourier series is dominated by just one term. The phase speed of a solitary wave (with correction so that the mean of the imbricate series is zero) is $c_{sol} = 1 - 6/\pi = -0.90986$, which is in error by the same 4.7% but in the opposite direction. The imbricate series is

$$u(X) = -\frac{6}{\pi} + 3\sum_{m=-\infty}^{\infty} \mathrm{sech}^2\left((x - Pm)/2\right) \quad (16.80)$$

as illustrated in Fig. 16.26. The KdV cnoidal wave for this value of the elliptic modulus is equally cosine and soliton.

When John Scott Russell discovered solitary water waves in 1834, most scientists of the time thought he was crazy. G. G. Stokes calculated a perturbation series for the cnoidal wave to prove that nonlinear waves must be periodic rather than the single, isolated crest as asserted by Russell. However, Toda and later Boyd, Whitham and Zaitsev [108, 112–114] by different means all proved that the cnoidal wave of the Korteweg-deVries equation can be written – without approximation and for all amplitudes – as a sum of evenly spaced solitary waves plus a constant:

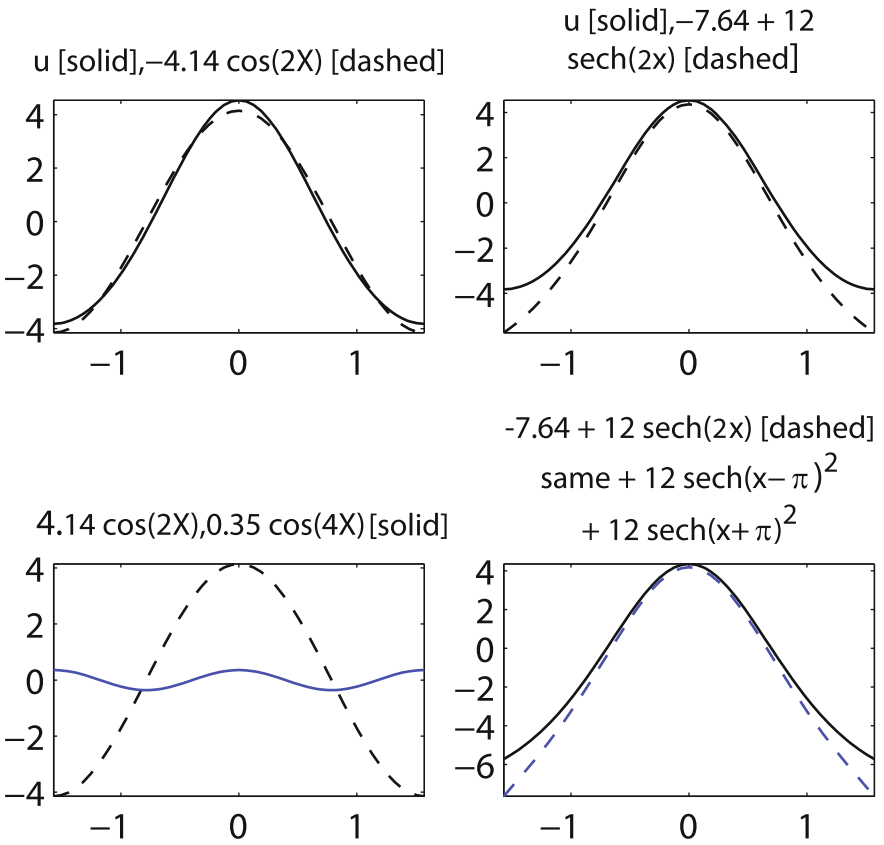

Fig. 16.26 The KdV lemniscate-cnoidal wave. The *left two panels* compare the cnoidal wave [solid in the *upper panel* only] with its one-term Fourier approximation, $-4.14\cos(2X)$ [*dashed* in both *left panels*] and the Fourier term $0.35\cos(4X)$ [solid in *lower left panel*]. *Upper right* the cnoidal wave is solid and its lowest imbricate approximation, $-7.64 + 12\text{sech}^2(X)$, is *dashed* in *both right panels*. The second imbricate approximation [including the first order approximation] is solid in the *lower right panel*

$$A(x,t) = -24s/\pi + \sum_{n=-\infty}^{\infty} 12s^2\,\text{sech}^2[s(x - ct - n\pi)] \tag{16.81}$$

And yet the lemniscate cnoidal wave is only a *small perturbation* of a *linear* sine wave.

16.17.3 Solitary Waves Are Not Special

Assertion 16.3 (*Solitons from Special Initial Conditions Only*) It is false that special initial conditions, that is, initial conditions that are a set of measure zero in the parameter space, are necessary for solitary wave generation.

First, since the solitary wave requires an *exact* balance between two competing influences — dispersion, which broadens a wave disturbance, and nonlinearity, which can steepen and focus it — it seemed that the soliton was merely an oddity, a laboratory card-trick. In the words of Scott et al. [118], the soliton was

> long considered a rather unimportant curiosity. . . Since it is clearly a special solution to the partial differential equation, many have assumed that somewhat special initial conditions would be required to launch it and, therefore, that its role in the (general) initial value problem would be a minor one at best.

The "soliton paradox" is that solitary waves were first discovered in nature, and solitons have since been observed in an enormous variety of fields — plasmas, nonlinear optics, condensed matter physics, and many others. If solitary waves are just "curiosities", why are they so readily observed? The answer is that many physical systems *evolve* to an exact balance between nonlinearity and dispersion even though the initial condition may have been far from such a balance.

A solution to the Korteweg-deVries (KdV) equation in which nonlinearity is initially stronger than dispersion will steepen. But when the shortest spatial scale is $L(t)$, nonlinearity is $O(amplitude/L)$ but dispersion is $O(1/L^3(t))$. Frontogensis will reduce the spatial scale until dispersion and nonlinearity balance — solitons.

If dispersion is initially the stronger, then $L(t)$ will increase as the disturbance becomes smoother. With its inverse cubic dependence on scale, dispersion diminishes more rapidly than nonlinearity until the two are in balance – small solitons! Theory and numerical experiment show that in many situations, the solitons dominate the solution, even from spectacularly unsolitonic initial conditions.

16.17.4 Why "Solitary Wave" Is the Most Misleading Term in Oceanography

Assertion 16.4 (*Soliton Loneliness*) It is false that solitary waves must be far from any other disturbance to merit the label of "soliton".

"The label "solitary wave" is a terminological atrocity. A flow with many closely-spaced peaks may be the superposition of solitons which have only a weak mutual interaction in spite of close proximity: for solitary waves that decay exponentially as $|X| \to \infty$, the interaction between two solitons separated by P must fall *exponentially fast* with P. Therefore, restricting the genera "solitary waves" to the class of isolated peaks, "alone on a wide, wide sea", is far too narrow.

"Solitary wave" is a translation of 'onde solitaire" which was introduced by Joseph Boussinesq (1842–1929) in 1871 in [119]. What wonderful sounds! "Onde solitaire". It is poetry disguised as terminology.

Water wave theorists were not misled. Walter Munk and Joseph Keller showed their viewpoint in their titles: "The solitary wave theory and its application to surf problems" [120] and "The solitary wave and periodic waves in shallow water" [121]. Michael Stiassnie and D. Howell Peregrine analyzed water waves propagating into shallower water using Trains of Solitary Waves (TSW's) as the major theoretical construct: "Thus periodic waves can be accurately represented by a train of solitary waves," as Peregrine wrote in a later invited review [pp. 154] of [122]. Costa, Osborne et al. take this a step further with observations of "soliton turbulence" in shallow water waves observed in Currituck Sound, North Carolina; the periodic version of the inverse scattering method, described in Osborne's book [102], allows an unambiguous identification of the solitary waves as solitons.

However, John Kindle's fine study of Trains of Equatorial Rossby Solitary Waves generated by fluctuating wind stresses in the Pacific was published only as a book chapter [123]. A longer version submitted to *Journal of Physical Oceangraphy* was not accepted because most blue water oceanographers could not accept the notion of a Train of Solitary Waves.

16.17.5 Scotomas and Discovery: The Lonely Crowd

Assertion 16.5 (*The Lonely Crowd*) Solitary waves are as ubiquitous as beetles and cockroaches, hiding in plain sight. Solitons are not discovered in new, exotic parameter ranges but rather are seen in familiar environments by "prepared minds".

John Scott Russell discovered solitary waves in 1832 with no technology except his own eyes. For how many thousands of years have men and women looked at the sea, incognizant? Indeed, *homo erectus* must have seen solitary, isolated wave peaks rolling down a brook or rushing floodwaters more than a million years ago, but it is easy to forgive those small-brained, illiterate hunter-gathers for their failure of recognition. It is much more remarkable that even in the Age of Technology solitary waves of various species have continued to be cognitively invisible.

In medicine, a "scotoma" is a blind spot, such that some objects in the field of vision are there but invisible. Oliver Sacks applied this medical term to a "knowledge gap", a selective blindness to a part of the scientific literature. For example, Tourette's syndrome, though discovered in the 1880s, was a scotoma in medical thinking until interest was revived by Sacks' work with L-dopa. Similarly, the Fast Fourier Transform (FFT) was a numerical analysis scotoma. Runge's article of 1903 [124] and the massive Fourier analysis efforts at Sir Edmund Whitaker's mathematics lab in Edinburgh were well summarized in 1913 by Carse and Urquart [125], but somehow faded into invisibility, forgotten, until the FFT was rediscovered quite independently of all earlier work by Cooley and Tukey in 1967.

Mark Ablowitz is a distinguished kymologist[9] who, among a lifetime of wave studies, examined the mathematics of polycnoidal waves. Double cnoidal waves have, in some parameter ranges, a characteristic Y-shaped pattern of ridges. To his great astonishment, on a family visit to the ocean, he observed and photographed precisely the shapes he and his collaborators had foretold, hyperelliptic functions incarnate in salt water. Subsequently he and Douglas E. Baldwin have made many photographs and videos that can be observed on their websites and their article [126]. Careful research turned up only a single earlier photograph of such ridges, published nearly forty years prior to [126]. In their conclusions, they write

> We report that X- and Y-type shallow water wave interactions on a flat beach are frequent, not rare, events. Casual observers can see these fundamental wave structures once they know what to look for.

The title of the classic sociology treatise, *The Lonely Crowd*, is a sort of metaphor for solitary waves. We understand that the throngs jostling on a busy street in Beijing or New York are close but unrelated and indifferent. Jostling dense wavepackets may be equally noninteracting and dynamically isolated.

The Sapir-Whorf Hypothesis is that language shapes our thinking and leads to blind spots, scotomas of the mind. Benjamin Whorf was an MIT-educated chemical engineer who worked as a fire inspector for the Hartford Insurance company while simultaneously studying and collaborating with Yale linguistics professor Robert Sapir. His two worlds collided in an industrial waste holding tank. The factory workers called it "The Pool." Connotations of ponds, harmless. The workers often surreptiously flicked spent cigarette butts into the pond. A more accurate name would have been," A tank of highly flammable industrial waste which will erupt in a fireball when touched by even a tiny ignition source, such as a cigarette that was not quite extinguished." The Hartford had to pay millions of dollars when "The Pool" burned down the factory; hundreds of workers lost their job.

The connotations of "solitary" are equally misleading. The very word seems to convey an image of a single, isolated peak, surrounded on all sides and for vast distances by flat, undisturbed water. However, numerical experiments have shown that solitons can quite happily coexist with a background of small amplitude waves and that they also retain their identity, shape, and phase speed when cheek-to-jowl with other solitary waves.

Keller, Munk, Peregrine and Ablowitz saw solitary waves in multipeaked, ever-fluctuating wave trains because their "prepared minds" discarded the connotations of that badly chosen label, "solitary wave", and saw the true dynamics in the water (Table 16.5).

[9] "Kymology" is the study of waves, from the Greek "$\kappa\upsilon\mu\alpha$", "wave".

Table 16.5 A bibliography of weakly nonlocal solitary waves

Reference	Remark
Williams and Wilson [63]	Numerical confirmation of (i) ease of exciting solitons (ii) near-elastic collisions of Rossby soliton collisions (iii) radiative decay of higher n mode [nonlocal] solitons
Boyd [64]	Radiative decay of nonlocal Rossby solitons
Segur, Tanveer and Levine [127]	Collection on exponential asymptotics
Boyd [65]	Review of nonlinear equatorial waves
Boyd [128]	Review of nonlocal solitary waves
Boyd [66]	Monopoles and dipoles in AEW approx.
Boyd [116]	Far-ranging book on all forms of exponential smallness
Boyd [67]	$n = 3$ nonlocal Rossby soliton
Boyd [68]	Numerical experiments: tilted modons and nonlocal soliton radiation
Boyd [33]	4-mode model
Yang [9]	Monograph on nonlinear waves good treatment of "embedded solitons"

16.18 Weakly Nonlocal Solitary Waves

16.18.1 Background

Williams and Wilson [63] numerically confirmed *most* of the KdV/Rossby soliton theory of Boyd [60, 61]. For the $n = 1$ latitudinal mode, there was good qualitative agreement even at large amplitude. Solitary waves of different amplitudes collided elastically, just as predicted.[10]

Initialization of the flow with the lowest order perturbative approximation to the $n = 3$ solitary wave yielded, for *small* amplitude, a very robust coherent structure. After a transient adjustment in which the quadrapole vortex sheds a little radiation, the equilibrated soliton translated steadily without change in shape or amplitude for a very long time.

At *moderate* amplitude, however, the solitary waves in latitudinal modes $n = 3$ and higher radiated sinusoidal Rossby waves of mode number $(n - 2)$ and rapidly decayed. Boyd [64, 116] noted that the nonlinear phase speed of the $n = 3$ solitary wave matches the *linear* phase speed of small amplitude sine waves in the $n = 1$ mode of zonal wavenumber $k \approx 2$. The resulting resonance forces the radiative decay of the solitary wave through emission of the resonant waves of lower latitudinal mode.

The amplitude of the emitted waves is an *exponential* function of the *reciprocal* of the soliton amplitude parameter ε. The reason the radiation is so small is that

[10] In this context, "elastic" means that no energy is lost to transient radiation; rather, the solitary waves colllectively have the same energy both before and after all soliton-soliton collisions.

its strength depends on the amplitude of the solitary wave at $k = 2$. The Fourier Transform of $\text{sech}^2(\varepsilon X)$ is

$$\int_{-\infty}^{\infty} \exp(ikx)\,\text{sech}^2(\varepsilon\, x)dx = \frac{\pi}{\varepsilon^2}\,\frac{k}{\sinh\left(\pi k/(2\varepsilon)\right)} \tag{16.82}$$

This is proportional to $\exp(-q/\varepsilon)$ for $\varepsilon \ll 1$ where $q = \pi/2$. Resonance amplifies forcing but does not create a response *ex nihilo*; the resonant radiation must be proportional to this tiny portion of the core of the soliton which has the same wavenumber as the resonant radiation.

The solitary wave is radiating for *all* amplitudes and there is *no stability threshold*. However, the rate of emission is so small for *small* ε as to be *invisible* in Williams and Wilson's numerical solution.

The leakage includes Rossby waves of all lower odd n, but the amplitude of modes lower than $(n-2)$ is exponentially small compared to emission in the $(n-2)$ mode, so only a single mode of radiation was detectable.

Such radiatively-decaying soliton-like structures have a common appellation.

Definition 16.5 (*Weakly Nonlocal Solitary Wave*) A steadily-translating, finite amplitude wave that decays to a small amplitude oscillation (rather to zero) as one moves away from the core of the disturbance.

The author's earlier book [116] and reviews [107, 128] and the book edited by Segur, Tanveer and Levine [127] and other references in Table 16.5 show that such generalized solitary waves are at least as common as classical solitary waves.

Some are so-called "micropterons" for which the far field resonant radiation is a finite power of the amplitude of the core of the solitary wave. The equatorial Rossby nonlocal solitary waves are "nanopterons" because the sinusoidal radiation in latitudinal mode $(n-2)$ has an amplitude which is an *exponential* function of the *reciprocal* of the amplitude of the core of the solitary wave, a KdV soliton in latitudinal mode n.

Williams [5] has shown through numerical solutions of a baroclinic model that solitary waves in the $n = 1$ latitudinal mode and the first baroclinic mode are weakly nonlocal through *radiative leakage* into the *barotropic* mode [5]. Higher latitudinal modes should similarly leak to the barotropic mode, and thence escape to the extratropics.

The radiation is emitted by the soliton whenever its phase speed matches that of a small amplitude wave of any species. Figure 16.27 shows as unshaded the phase speeds which allow classical, nonradiating solitary waves.

16.18.2 Initial Value Experiments

Boyd ran his equatorial beta-plane model with an initial condition equal to the $n = 3$ zeroth-order solitary wave (Boyd [22]) for $\varepsilon = 1/3$ [$B = 1/3$ in the notation of

the earlier article]. Results are shown in a frame of reference which moves at the predicted phase speed of the solitary wave, $c_{sol} = -1/7 - 0.08\varepsilon^2$, so that the core of the coherent structure remains centered at $x = 0$.

Figure 16.28 shows that the expected wavepacket does indeed emerge. The amplitude of the radiation is very small, however, and only the north-south current is readily visible. The radiation is noticeably more confined in latitude than the core of the solitary wave, which has the structure of the second lowest second symmetric mode. It follows that the radiation lies almost entirely in the $n = 1$ mode as predicted.

The resonance condition is

$$c_{sol} = \frac{-1}{k^2 + 3} \tag{16.83}$$

where the right-hand side is the (approximate but very accurate) dispersion relation for infinitesimal amplitude $n = 1$ Rossby waves of zonal wavenumber k. This predicts the wavelength of the emission to be

$$k_{res} = 2\,\frac{\sqrt{1 + 0.14\varepsilon^2 - 0.235\varepsilon^4}}{1 + 0.56\varepsilon^2} \tag{16.84}$$

which is $k_{res} \approx 1.895$ for $\varepsilon = 1/3$.

By multiplying the wavepacket by a windowing function, it is possible to take the Fourier transform of the packet with minimal contamination from the core of the solitary wave. When the north-south current is expanded in Hermite functions, the j-th coefficient is the sum of three latitudinal modes. However, because our initial condition is a nonlinear Rossby wave, the gravity wave modes are only weakly excited. It is thus a good approximation to take the amplitude of the $n = 1$ Hermite coefficient of v as the amplitude of the $n = 1$ Rossby mode. Figure 16.29 shows

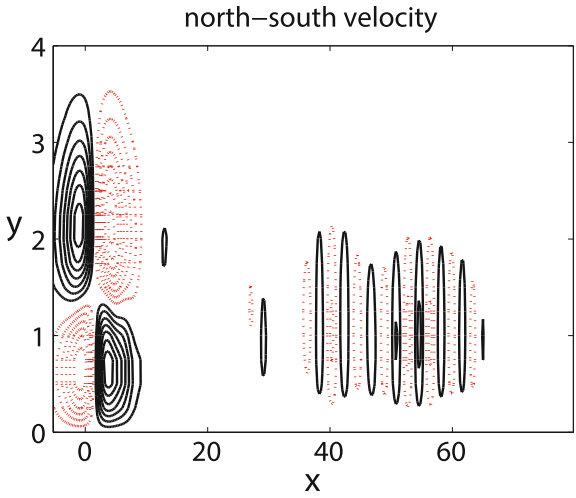

Fig. 16.27 Isolines of the north-south velocity at $t = 400$. Negative-valued contours are *dashed*. Computations were performed in a frame-of-reference moving at the perturbatively-predicted phase speed of the solitary wave, so the quadrapole pattern which is the core of the soliton remains centered at $x = 0$. The computational interval extended from $x \in [-80, 80]$, but only part of this domain is illustrated

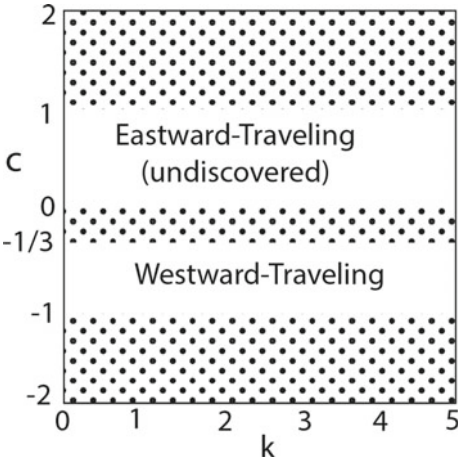

Fig. 16.28 Classical Rossby solitary waves are possible only when their phase speed lies in one of the unshaded regions, that is, ouside the range of infinitesimal amplitude waves. Coherent structures with phase speeds in the shaded regions will be weakly nonlocal and radiatively decaying

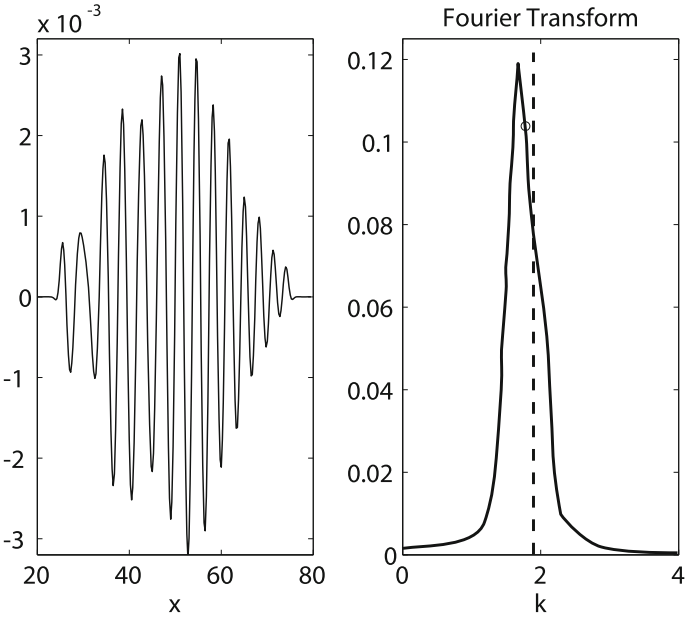

Fig. 16.29 *Left* coefficient of $\psi_1(y)$ in the expansions of $v(x, y, t = 400)$ after windowing; the window function was $W(x) \equiv \{\text{erf}(x - 25) - \text{erf}(x - 75)\}$. *Right panel* Fourier Transform in longitude of the windowed waveacket. The *vertical dashed line* marks the resonant wavenumber, $k = 1.9$

the Fourier transform of the windowed wavepacket indeed has a peak close to the predicted $k = 1.895$.

The group velocity in an earth-fixed frame of reference is $c_g = (3-k^2)/(3+k^2)^2$. The group velocity relative to the core of the solitary wave, and therefore in the

moving reference frame,

$$c_g^{relative} \approx \frac{1}{7} + 0.08\varepsilon^2 + \frac{k_{res}^2 - 3}{(k_{res}^2 + 3)^2} \tag{16.85}$$

which is approximately 1/6. Thus, over the length of the integration, which is 400 nondimensional time units, the emitted radiation should spread eastward relative to the solitary wave about 66 spatial units, which is only a slight underestimate.

The one puzzle is that the envelope of the wavepacket does not show a constant amplitude, as expected, but instead diminishes to almost nothing close to the solitary wave. Similar fluctuations have been seen in other weakly nonlocal solitary waves such as the ϕ^4 breather [116], but the cause is not understood.

16.18.3 Nonlinear Eigenvalue Solutions

It is also possible to solve for the $n = 3$ mode as the steadily-translating solution to a nonlinear eigenvalue problem. In an ocean basin, this is rather artificial: such steadily-propagating coherent structures are east-west symmetric with respect to the center of the core, and thus have small amplitude $n = 1$ sine waves *west* of the core. This is possible only in a land-free ocean where the eastward-radiating $n = 1$ waves are free to expand around the globe until they reach the core from the west.

In spite of its artificiality, such nonlinear eigensolutions (with periodic boundary conditions in the east-west) have been very useful in the general theory of nonlocal coherent structures as reviewed in Boyd's monograph [116]. The outside-the-core oscillations of the nonlocal solitary waves confirm that the radiation is an intrinsic part of the dynamics, and not merely a transient feature that might disappear from an initial-value calculation if only it were run longer.

Figure 16.30 shows the result of such a computation. (The numerical intricacies are described in Boyd [67, 129, 130].) The rather eccentric value of the phase speed was the result of initializing the iteration with the perturbative solution of Boyd [60] for the $n = 3$ mode with $B = 3/8$, that is, the east-west structure proportional to $\mathrm{sech}^2(0.375[x - ct])$. Other calculations, not shown, demonstrate that the amplitude of the $n = 1$ oscillations falls exponentially fast with decreasing amplitude in the core of the structure. For sufficiently large core amplitude (or negative c), the far field oscillations are as big or bigger as the core, at least in the north-south flow, and the solitary wave is no longer "weakly" nonlocal.

For this moderate amplitude, however, we have confirmation that this latitudinal mode indeed has weakly nonlocal solitary waves. The eigensolutions are fully consistent with the initial value experiments of Williams and Wilson [63] and our own initial-value experiments here.

The general theory of weakly nonlocal waves has many subtleties, such as those associated with the quantization effects of periodic boundary conditions. These issues are fully explained in the monograph [116] and the review articles [107, 128].

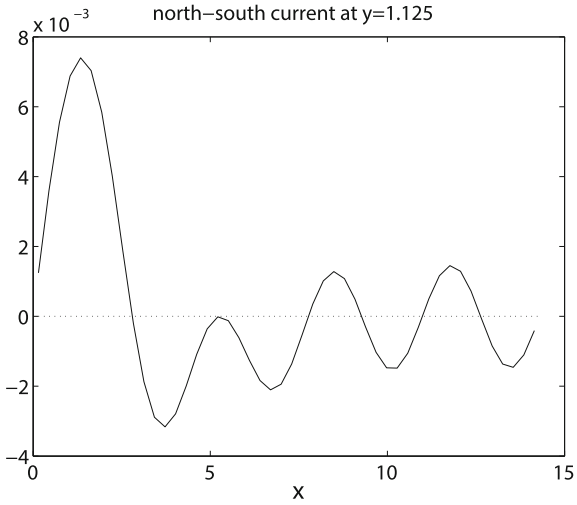

Fig. 16.30 Nonlinear eigenvalue solution for $c = -0.15411$: the north-south current at a particular latitude, $y = 1.125$, is plotted versus longitude. The calculation used a 48×24 grid with centered eighth-order differences on the domain $x \in [0, 4.25W_f] \otimes y \in [0, 6]$ where $W_f = 2\pi/k_{res}$ is the predicted wavelength of the far field oscillations in the $n = 1$ latitudinal Rossby mode. u and ϕ were assumed symmetric with respect to both the x and y axes; v was doubly antisymmetric

16.19 Tropical Instability Vortices

Tropical instability vortices (TIVs) are strong oceanic anticyclones as much as 500 km in diameter that are centered 4–5° north of the equator, embedded between the NECC (eastward) and the SEC (westward) (Fig. 16.31). TIVs are largely confined above the thermocline and have little amplitude below one hundred and fifty meters. Kennan and Flament [131] are a very careful observational study [131] from which Fig. 16.31 was derived. TIVs were first discovered in the Pacific, but are seen in the Atlantic also.

Recall that the mean zonal currents in the tropics are alternating zonal jets: westward South Equatorial and North Equatorial currents (SEC and NEC) and the eastward Equatorial Undercurrent (EUC) and eastward North Equatorial Countercurrent (NECC). Each June, after the southeast trade winds strengthen, the equatorial currents accelerate and upwelling intensifies. "Tropical Instability Waves" (TIWs) with a wavelength about 1000 km spontaneously appear and persist through the boreal fall and winter to fade out in March. Tropical Instability Waves were first detected in current meter records as meanderings of the Atlantic South Equatorial Current (SEC) [133] and in satellite infrared images as cusplike deformations of the North Equatorial Front (NEF) in the Pacific ocean by Legeckis [134]. They were also subsequently identified in various observations, including in situ measurement of velocity [135, 136], in situ measured sea temperature [137], remote satellite measured SST [138, 139], surface dynamic height [140], surface wind [139], as well as sea colors [141].

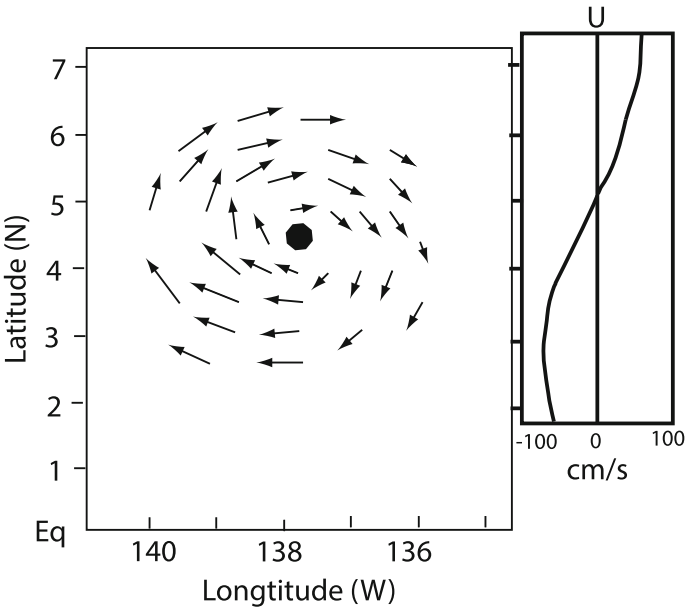

Fig. 16.31 Currents in a Tropical Instability Vortex (*left*) and the zonally-averaged mean flow $U(y)$ (*right*). Sketched using the data of Flament et al. [132] to constrain the centroid, maximum velocity of the vortex and mean current, etc

These measurements gave various wavelengths, and periods, in ranges of 600–2000 km and 16–40 days, respectively [136].

In the Pacific Ocean, cusplike deformations appear on both north and south boundaries of the cold tongue. But in the Atlantic Ocean, they usually only show on the north boundary. The cusplike deformations on the *north* boundary are related to coherent anticyclonic vortices centered about $5°N$ which are labeled as "Tropical Instability Vortices (TIVs)" [131, 132, 142].

Despite an extensive multi-decadal literature many fundamental questions about TIW/TIV kinematics and generating mechanisms still remain unclear. Rather than one broadbanded process, the TIWs are bichromatic. The first spectral peak has periods around 15–23 days, is most prominent in meridional velocity, and has been observed within a few degrees of the equator [131, 132, 135–137]. The second species has a period of roughly one month and has been observed in sea surface height, thermocline depth, velocity, and subsurface temperature centered about $5°N$ [131, 132, 137, 140, 143]. These two types of TIWs can appear in the ocean simultaneously. Flament et al. [131, 132] observed two drastically different propagation speeds of TIWs at the same time, 0.8 ms^{-1} along the the equator and 0.3 ms^{-1} along $4.5°N$.

The generation of TIWs is generally thought to be associated with the instability of the intense zonal currents present in the equatorial ocean. However, there is no consensus. TIWs may be spawned by barotropic instability of the horizonal

shear between South Equatorial Current (SEC) and North Equatorial Countercurrent (NECC)/Equatorial Undercurrent (EUC) [136, 141, 144, 145]. On the other hand, vertical shear in the equatorial jets is also strong. TIWs may be generated by baroclinic instability from the vertical shear of the South Equatorial Current or by a mixed barotropic-baroclinic instability [146–148]. The apparent existence of two species of TIWs does not make it easier to explain their origins.

In an effort to explain the coexistence of these two very different types, especially the surface trapped Yanai-type TIWs, Zhou and Boyd [149] studied nonlinear TIW development in a shallow-water model with initial mean states from the Pacific Ocean. During the early linear stage, the most unstable mode determined by the background mean currents dominates from $8°S$ to $8°N$. As this mode grows, the wave component centered $5°N$ rolls up into nonlinear vortices and then stabilizes the background mean substantially. Meanwhile the wave component near the equator assumes its own identity as a fast-moving neutral Yanai wave. The wave center near $6°S$ becomes distinct from the wave centered at $5°N$ with a slower phase speed. So instead of a single instability, now there are three waves with three different phase speeds.

Yet this theory still remains to be tested in a realistic 3-D model and in observations. The variance of wind stress was not included. In the real equatorial ocean, the primary cause of the change of zonal current is the change of wind stress and nonlinearity may only be secondary. However, the essential point of [149] is that the wave components at different latitudes of an initially single TIW may decouple if the background mean changes no matter whether it is due to the change of wind stress or due to the nonlinearity. The faster Yanai-type waves near the equator are not necessarily directly arising from the instability of the background currents. Such decoupling process may explain why the cross-equatorial structure of the SST front or SSH is sometimes phase locked and sometime not.

Using Lindzen's wave over-reflection theory, Proehl [150] argued that TIWs arise by extracting energy from the background state through varying mixes of baroclinic, barotropic, and Kelvin–Helmholtz mechanism rooted in the critical layer and therefore attempting to classify the instability through energy conversion was misleading. However, theory lags. Part of the reason is that five different paradigms or perspectives on barotropic and baroclinic instability compete in the literature.

The last interesting question regards the coherent anticyclonic TIVs in the shear between SEC and NECC. How do they remain quasi-steady (except for a steady westward phase propagation) for months before dying out in the western Pacific ocean? Flament et al. [132] found that the inteior potential vorticity is almost zero, and suggested this may be fossilized inertial instability. Examining 3-D Lagrangian trajectories in the Atlantic, Dutrieux et al. [151] found that while TIVs are highly coherent throughout their lifetime, up to 50% of their water may be renewed over one rotation of the vortex. They suggested that TIVs are a hybrid of an isolated vortex with trapped fluid and a dispersive linear wave that propagates without recirculation.

Mature TIVs are like solitary waves. Theories of vortices in a shear zone, originally developed for the Great Red Spot of Jupiter,[11] are applicable to TIVs [75, 152–156].

16.20 The Missing Soliton Problem

> Rossby solitons turn out to be robust solutions in a non-linear 1.5-layer equatorial ocean model. Not only are they maintained when explicitly initiated, but they can also be caused by reflection of any Kelvin wave, irrespective of how it was generated. In particular, a temporal weakening of easterly trade winds, modelled here by a temporal patch of westerly winds, generates such a Kelvin signal. …
>
> They have however never been observed so far although observations of internal gravity solitons are common [157, 158]. It might be that the signal is so weak in the real ocean that it disappears in other wave signals. Another explanation may be that nobody has ever attempted to extract a Rossby soliton signal from observational data.

T.R.F. Feitsma and H.A. Dijstra [159].

The $n = 1$ equatorial Rossby solitary wave is a dipole, two contra-rotating vortices of equal strength, propelling each other westward at a faster-than-linear speed. When mean currents are omitted, the dipole soliton is readily generated in a numerical model. Kindle [123] and Greatbatch [3] showed that changes in wind stress, even uniform changes over a huge area whose shape bears not the slightest resemblance to a Kelvin or Rossby wave will nevertheless excite one or more dipole Rossby solitary waves. Williams and Wilson applied forcing at twenty degrees of latitude, expecting to generate an imitation of the Great Red Sot of Jupiter, but excited equatorial solitons instead [63]! This accident led them to investigate equatorial soliton collisions and experimentally discover that higher latitudnal mode solitons are weakly nonlocal and radiatively decaying as discussed in Sect. 16.18.

Indeed, the $n = 1$ Rossby solitary is so easily generated, and the perturbation series so accurate, that the equatorial soliton is now widely used as a test case to validate numerical models [160–163],

Williams and Wilson made the equatorial solitons go away by adding the alernating zonal jets that dominate the Jovian atmosphere. It seems likely the alternating equatorial jets are equally villanous for dipole solitons straddling the terrestial equator, but other factors are important also. There has not as yet been a thorough study.

However, Vaid, Gnanaseelan and Salvekar [164], validating earlier numerical and observational studies they review, found a dipole of Rossby waves appearing seaonally in the equatorial Indian Ocean [164]. These authors note that these Rossby waves are strongly modulated by mean currents, reinforcing the remarks made earlier here that future Rossby soliton studies have a desperate need to include mean flow.

[11] The Great Red Spot of Jupiter is an anticyclonic vortex embedded between alternating jets at about twenty degrees S. latitude. The vortex has been spinning for as long as there have been telescopes to observe it, nearly four centuries.

References

1. Marshall HG, Boyd JP (1987) Solitons in a continuously stratified equatorial ocean. J Phys Oceanogr 17:1016–1031
2. Ripa P (1985) Nonlinear effects in the propagation of Kelvin pulses across the Pacific Ocean. In: Debnath L (ed) Advances in nonlinear waves. Pitman, New York, pp 43–56
3. Greatbatch RJ (1985) Kelvin wave fronts, Rossby solitary waves and nonlinear spinup of the equatorial oceans. J Geophys Res 90:9097–9107
4. Ripa P (1984) Harmonics resonance and chaos in the equatorial waveguide. AIP conference proceedings. American Institute of Physics/Woodbury, Ensenada, Mexico, pp 537–545
5. Williams GP (1996) Jovian dynamics. Part I. Vortex stability, structure and genesis. J Atmos Sci 53(18):2685–2734
6. Boyd JP (2002) Solitons. In: Holton JR, Pyle J, Curry JA (eds) Encyclopedia of the atmospheric sciences. Elsevier, New York
7. Boyd JP (2007) Planetary-scale solitary waves. In: Grimshaw R (ed) Solitary waves in fluids. WIT Press, United Kingdom, Southhampton, pp 125–158
8. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
9. Yang J (2010) Nonlinear waves in integrable and nonintegrable systems, Mathematical modeling and computation, vol 16. SIAM, Philadelphia, p 430
10. Kantorovich LV, Krylov VI (1958) Approximate methods of higher analysis. Interscience, New York, trans. by Curtis D. Benster, p 681
11. Finlayson BA (2013) The method of weighted residuals and variational principles, 2nd edn. SIAM, New York
12. Lorenz EN (1960) Maximum simplification of the dynamic equations. Tellus 12:243–254
13. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141
14. Lorenz EN, Krishnamurthy V (1987) On the nonexistence of a slow manifold. J Atmos Sci 44:2940–2950
15. Boyd JP (1995) Eight definitions of the slow manifold: seiches, pseudoseiches and exponential smallness. Dyn Atmos Oceans 22:49–75
16. Boyd JP (1994) The slow manifold of a five mode model. J Atmos Sci 51:1057–1064
17. Camassa R, Tin SK (1996) The global geometry of the slow manifold in the Lorenz-Krishnamurthy model. J Atmos Sci 53:3251–3264
18. Ripa P (1982) Nonlinear wave-wave interactions in a one-layer reduced-gravity model on the equatorial beta-plane. J Phys Oceangr 12:97–111
19. Boyd JP, Zhou C (2008) Uniform asymptotics for the linear Kelvin wave in spherical geometry. J Atmos Sci 65(2):655–660
20. Boyd JP, Zhou C (2008) Kelvin waves in the nonlinear shallow water equations on the sphere: nonlinear traveling waves and the corner wave bifurcation. J Fluid Mech 617:185–205
21. Longuet-Higgins MS (1968) The eigenfunctions of Laplace's tidal equations over a sphere. Phil Trans Roy Soc London A 262:511–607
22. Boyd JP (1980) The nonlinear equatorial Kelvin wave. J Phys Oceangr 10:1–11
23. Ripa P (1983) Resonant triads of equatorial waves in a one-layer model. Part I. Non-local triads and triads of waves with the same speed. J Phys Oceangr 13:1208–1226
24. Majda AJ, Rosales RR, Tabak EG, Turner CV (1999) Interaction of large-scale equatorial waves and dispersion of Kelvin waves through topographic resonances. J Atmos Sci 56(24):4118–4133
25. Platzman GW (1964) An exact integral of complete spectral equations for unsteady one-dimensional flow. Tellus 16:422–431, corrigendum, 17, 150 (1965)
26. Boyd JP (2014) Solving Transcendental Equations: The Chebyshev Polynomial Proxy and Other Numerical Rootfinders. Perturbation Series and Oracles, SIAM, Philadelphia, p 460
27. Lesser MB, Crighton DG (1975) Physical acoustics and the method of matched asymptotic expansions. Phys Acoust 11:69–149
28. Boyd JP (1992) The energy spectrum of fronts: the time evolution of shocks in Burgers' equation. J Atmos Sci 49:128–139

29. Fedorov AV, Melville WK (2000) Kelvin fronts on the equatorial thermocline. J Phys Oceangr 30(7):1692–1705
30. Boyd JP (1998) High order models for the nonlinear shallow water wave equations on the equatorial beta-plane with application to Kelvin wave frontogenesis. Dyn Atmos Oceas 28(2):69–91
31. Bouchut F, Le Sommer J, Zeitlin V (2005) Breaking of balanced and unbalanced equatorial waves. Chaos 15(013), 503–1–013, 503–19
32. Boyd JP (1984) Equatorial solitary waves, Part IV: Kelvin solitons in a shear flow. Dyn Atmos Oceans 8:173–184
33. Boyd JP (2006) Fourier pseudospectral method with Kepler mapping for traveling waves with discontinuous slope: application to corner waves of the Ostrovsky-Hunter equation and equatorial Kelvin waves in the four-mode approximation. Appl Math Comput 177(1):289–299
34. Chen GY, Boyd JP (2002) Nonlinear wave packets of equatorial Kelvin waves. Geophys Astrophys Fluid Dyn 96(5):357–379
35. Fujiwara M, Yamamoto MK, Hashiguchi H, Horinouchi T, Fukao S (2003) Turbulence at the tropopause due to breaking Kelvin waves observed by the Equatorial Atmosphere Radar. Geophys Res Lett 30(4):1171
36. Le Sommer J, Reznik GM, Zeitlin V (2004) Nonlinear geostrophic adjustment of long-wave disturbances in the shallow-water model on the equatorial beta-plane. J Fluid Mech 515:135–170
37. Long B, Chang P (1990) Propagation of an equatorial Kelvin wave in a varying thermocline. J Phys Oceanogr 20:1826–1841
38. Matsuura T, Iizuka S (2000) Zonal migration of the Pacific warm-pool tongue during El Niño events. J Phys Oceangr 30(7):1582–1600
39. Milewski PA, Tabak EG (1999) A reduced model for nonlinear dispersive waves in a rotating environment. Geophys Astrophys Fluid Dyn 90:139–159
40. Ripa P (1994) Horizontal wave-propagation in the equatorial wave-guide. J Fluid Mech 271:267–284
41. Stepanov DV, Novotryasov V (2011) Internal Kelvin wave frontogenesis on the equatorial pycnocline. Geophys Astrophys Fluid Dyn 105(4–5):
42. Zheng Q, Susanto RD, Yan XH, Liu WT, Ho CR (1998) Observation of equatorial Kelvin solitary waves in a slowly varying thermocline. Nonlinear Process Geophys 5(3):153–165
43. Bender CM, Orszag SA (1978) Advanced Mathematical Methods for Scientists and Engineers. McGraw-Hill, New York, p 594
44. Rosales RR, Tabak EG, Turner CV (2002) Resonant triads involving a nondispersive wave. Stud Appl Math 108(1):105–122
45. Stokes GG (1847) On the theory of oscillatory waves. Camb Trans 8:441–473, also published in vol. 1 of *Mathematical and Physical Papers* (Cambridge University Press, Cambridge, 1880), pgs. 197–229
46. Longuet-Higgins MS, Fox MJH (1977) Theory of the almost-highest wave: the linear solution. J Fluid Mech 80:721–741
47. Longuet-Higgins MS, Fox MJH (1978) Theory of the almost-highest wave: Part 2, Matching and analytical extension. J Fluid Mech 85:769–786
48. Longuet-Higgins MS, Fox MJH (1996) Asymptotic theory for the almost-highest wave. J Fluid Mech 317:1–19
49. Pullin DI (1992) Contour dynamics methods. Ann Revs Fluid Mech 24:89–115
50. Camassa R, Holm DD (1993) An integrable shallow water equation with peaked solitons. Phys Rev Lett 71(11):1661–1664
51. Liu ZR, Wang RQ, Jing ZJ (2004) Peaked wave solutions of Camassa-Holm equation. Chaos, Solitons and Fractals 19(1):77–92
52. Boyd JP (1997) The periodic generalization of Camassa-Holm "peakons": An exact superposition of solitary waves. Appl Math Comput 81(2):173–187
53. Hunter JK (1990) Numerical solutions of some nonlinear dispersive wave equations. Lect Appl Math 26:301–316

54. Grimshaw RHJ, He JM, Ostrovsky LA (1998) Terminal damping of a solitary wave due to radiation in rotational systems. Stud Appl Math 101(2):197–210
55. Gabov S (1978) On Whitham's equation. Soviet Math Dokl 19:1225–1229
56. Shefter M, Rosales RR (1999) Quasiperiodic solutions in weakly nonlinear gas dynamics. Part I. Numerical results in the inviscid case. Stud Appl Math 103(4), 279–337
57. Whitham GB (1967) Variational methods and applications to water waves. Proc R Soc Lond A 299(6):6–25
58. Seliger RL (1968) A note on the breaking of waves. Proc R Soc Lond A 303:493–496
59. Zaitsev AA (1986) Whitham stationary waves and their dispersion relation. Sov Phys Dokl 31(2):118–120
60. Boyd JP (1980) Equatorial solitary waves, Part I: Rossby solitons. J Phys Oceanogr 10:1699–1718
61. Boyd JP (1983) Equatorial solitary waves, Part II: envelope solitons. J Phys. Oceanogr 13:428–449
62. Boyd JP (1985) Equatorial solitary waves, Part 3: modons. J Phys. Oceanogr 15:46–54
63. Williams GP, Wilson RJ (1988) The stability and genesis of Rossby vortices. J Atmos Sci 45:207–249
64. Boyd JP (1989) Non-local equatorial solitary waves. In: Nihoul JCJ, Jamart BM (eds) Mesoscale/synoptic coherent structures in geophysical turbulence: proceedings of 20th liege colloquium on hydrodynamics. Elsevier, Amsterdam, pp 103–112
65. Boyd JP (1991) Nonlinear equatorial waves. In: Osborne AR (ed) Nonlinear topics of ocean physics: fermi summer school. Course LIX, North-Holland, Amsterdam, pp 51–97
66. Boyd JP (1991) Monopolar and dipolar vortex solitons in two space dimensions. Wave Motion 57:223–243
67. Boyd JP (2002) Deleted residuals, the QR-factored Newton iteration, and other methods for formally overdetermined determinate discretizations of nonlinear eigenproblems for solitary, cnoidal, and shock waves. J Comput Phys 179(1):216–237
68. Boyd JP (2002) Equatorial solitary waves. Part 5. Initial value experiments, co-existing branches and tilted-pair instability. J Phys Oceangr, 32(9), 2589–2602
69. Ma H (1992) The equatorial basin response to a Rossby wave packet: the effects of nonlinear mechanism. J Marine Res 50:567–609
70. Ma H (1996) The dynamics of North Brazil current retroflection eddies. J Marine Res 54:35–53
71. Zhao Q, Fu ZT, Liu SK (2001) Equatorial envelope Rossby solitons in a shear flow. Adv Atmos Sci 18(3):418–428
72. Maxworthy T (1980) On the formation of nonlinear internal waves from the gravitational collapse of mixed regions in two and three dimensions. J Fluid Mech 96:47–64
73. Clarke RA (1971) Solitary and cnoidal planetary waves. Geophys Fluid Dyn 2:343–354
74. Nayfeh AH (1973) Perturbation Methods. Wiley, New York
75. Redekopp LG (1977) On the theory of solitary Rossby waves. J Fluid Mech 82:725–745
76. Stern ME (1975) Minimal properties of planetary eddies. J Marine Res 33:1–13
77. Larichev VD, Reznik GM (1976) Two-dimensional Rossby soliton: an exact solution. Rep USSR Acad Sci 231:1077–1079
78. Flierl GR, Larichev VD, McWilliams JC, Reznik GM (1980) The dynamics of baroclinic and barotropic solitary eddies. Dyn Atmos Oceans 5:1–41
79. McWilliams JC, Flierl GR, Larichev VD, Reznik GM (1981) Numerical studies of barotropic modons. Dyn Atmos Oceans 5:219–238
80. Boyd JP (1981) Analytical approximations to the modon dispersion relation. Dynamics Atmos Oceans 6:97–101
81. Shen CY (1981) On the dynamics of a solitary vortex. Dynam Atmos Oceans 5:239–267
82. McWilliams JC, Zabusky NJ (1982) Interactions of isolated vortices. 1. Modons colliding with modons. Geophys Astro. Fluid Dyn 19(3–4):207–227
83. Mied RP, Lindemann GJ (1979) The propagation and evolution of cyclonic Gulf Stream rings. J Phys Oceangr 9:1183–1206

84. Mied RP, Lindemann GJ (1982) The birth and evolution of eastward propagating modons. J Phys Oceangr 12:213–230
85. McWilliams JC (1983) Interactions of isolated vortices. 2. Modon generation by monopole collision. Geophys Astro. Fluid Dyn 24(1):1–22
86. Boyd JP (1994) Nonlocal modons on the beta-plane. Geophys Astrophys Fluid Dyn 75:163–182
87. Tribbia JJ (1984) Modons in spherical geometry. Geophys Astrophys Fluid Dyn 30:131–168
88. Verkley WTM (1984) The construction of barotropic modons on a sphere. J Atmos Sci 41:2492–2504
89. Verkley WTM (1987) Stationary barotropic modons in westerly background flows. J Atmos Sci 44:2383–2398
90. Verkley WTM (1990) Modons with uniform absolute vorticity. J Atmos Sci 47:727–745
91. Verkley WTM (1993) A numerical method to find form-preserving free solutions of the barotropic vorticity equation on a sphere. J Atmos Sci 50:1488–1503
92. Boyd JP, Ma H (1990) Numerical study of elliptical modons by a spectral method. J Fluid Mech 221:597–611
93. Peregrine DH (1966) Calculations of the development of an undular bore. J Fluid Mech 25:321–330
94. Benjamin TB, Bona JL, Mahony JJ (1972) Model equations for long waves in nonlinear dispersive systems. Phil Trans Roy Soc Lond A 272:47–78
95. Bona JL, Pritchard WG, Scott LR (1981) An evaluation of a model equation for water waves. Phil Trans Roy Soc Lond A 302(1471):457–510
96. Wedi NP, Smolarkiewicz PK (2010) A non linear persective on the dynamics of the MJO: idealized large-eddy simulations. J Atmos Sci 67:1202–1217
97. Ablowitz MJ, Segur H (1981) Solitons and the Inverse Scattering Transform. SIAM, Philadelphia, Pennsylvania
98. Rizzoli PM (1982) Planetary solitary waves in geophysical flows. Adv Geophys 24:147–224
99. Harvey Segur, Hammack JL (1982) Soliton models of long internal waves. J Fluid Mech 118:285–304
100. Hammack JL, Segur H (1974) Korteweg-deVries equation and water-waves. 2. Comparisons with Experiments. J Fluid Mech 65:289–314
101. Hammack JL, Segur H (1978) Korteweg-deVries equation and water-waves. 3. Oscillatory waves. J Fluid Mech 84:337–358
102. Osborne A (2010) Nonlinear Ocean Waves & the Inverse Scattering Transform, International Geophysics, vol 67. Academic, New York, p 944
103. Haupt SE, Boyd JP (1991) Double cnoidal waves of the Korteweg-deVries equation: the boundary value approach. Phys D 50:117–134
104. Boyd JP, Haupt SE (1991) Polycnoidal waves: spatially periodic generalizations of multiple solitary waves. In: Osborne AR (ed) Nonlinear topics of ocean physics: Fermi Summer School. Course LIX, North-Holland, Amsterdam, pp 827–856
105. Ablowitz MJ, Segur H (1977) Asymptotic solutions of Korteweg-deVries equation. Stud Appl Math 37(1):13–44
106. Olver FWJ, Lozier DW, Boisvert RF, Clark CW (eds) (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, New York
107. Boyd JP (1989) New directions in solitons and nonlinear periodic waves: polycnoidal waves, imbricated solitons, weakly non-local solitary waves and numerical boundary value algorithms. In: Wu TY, Hutchinson JW (eds) Advances in Applied Mechanics, advances in applied mechanics, vol 27. Academic Press, New York, pp 1–82
108. Boyd JP (1982) Theta functions, Gaussian series, and spatially periodic solutions of the Korteweg-de Vries equation. J Math Phys 23:375–387
109. Boyd JP (1984) The double cnoidal wave of the Korteweg- de Vries equation: An overview. J Math Phys 25:3390–3401
110. Boyd JP (1984) Perturbation theory for the double cnoidal wave of the Korteweg-de Vries equation. J Math Phys 25:3402–3414

111. Boyd JP (1984) The special modular transformation for the polycnoidal waves of the Korteweg-de Vries equation. J Math Phys 25:3390–3401
112. Toda M (1975) Studies of a non-linear lattice. Phys Rep 18:1–124
113. Whitham GB (1984) Comments on periodic waves and solitons. IMA J Appl Math 32:353–366
114. Zaitsev AA (1983) Formation of stationary nonlinear waves by superposition of solitons. Sov Phys Dokl 28:720–722
115. Boyd JP (1984) Cnoidal waves as exact sums of repeated solitary waves: new series for elliptic functions. SIAM J Appl Math 44:952–955
116. Boyd JP (1998) Weakly nonlocal solitary waves and beyond- all-orders asymptotics: generalized solitons and hyperasymptotic perturbation theory, mathematics and its applications, vol 442. Kluwer, Amsterdam, p 608
117. Boyd JP (2011) New series for the cosine lemniscate function and the polynomialization of the lemniscate integral. J Comput Appl Math 235:1941–1955
118. Scott AC, Chu FYF, McLaughlin DW (1973) The soliton: a new concept in applied science. Proc IEEE 61(10):1443–1483
119. Boussinesq JV (1871) Thorie de l'intumescence liquide, applele onde solitaire ou de translation, se propageant dans un canal rectangulaire. Comptes Rendus de l'Academie des Sciences 72:755–759
120. Munk W (1949) The solitary wave theory and its application to surf problems. Ann N Y Acad Sci 51(3):376–424
121. Keller JB (1948) The solitary wave and periodic waves in shallow water. Ann N Y Acad Sci 51(3):345–350
122. Peregine DH (1983) Breaking of waves on beaches. Ann Rev Fluid Mech 15(1):149–178
123. Kindle J (1983) On the generation of Rossby solitons during El Nino. In: Nihoul JCJ (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 353–368
124. Runge C (1903) Über die Zerlegung empirisch gegebener periodischer Funktionen in sinuswellen. Zeit f Math u Phys 48:443–456
125. Carse GA, Urquhart J (1914) Harmonic analysis. In: Horsburgh EM (ed) Modern Instruments and methods of calculation: a handbook of the napier tercentenary exhibition, G. Bell and Sons, in cooperation with the Royal Society of Edinburgh, London, pp 220–248, 300 pp
126. Ablowitz MJ, Baldwin DE (2012) Nonlinear shallow ocean-wave soliton interactions on flat beaches. Phys Rev E 86(036):305
127. Segur H, Tanveer S, Levine H (eds) (1991) Asymptotics beyond all orders. Plenum, New York, p 389
128. Boyd JP (1991) Weakly nonlocal solitary waves. In: Osborne AR (ed) Nonlinear topics of ocean physics: Fermi Summer School. Course LIX, North-Holland, Amsterdam, pp 527–556
129. Boyd JP (2001) Chebyshev and Fourier Spectral Methods. Dover, New York, p 680
130. Boyd JP (2007) Why Newton's method is hard for travelling waves: Small denominators, KAM theory, Arnold's linear Fourier problem, non-uniqueness, constraints and erratic failure. Math Comput Simul 74(2–3):72–81
131. Kennan SC, Flament PJ (2000) Observations of a Tropical Instability Vortex. J Phys Oceangr 30(9):2277–2301
132. Flament PJ, Kennan SC, Knox RA, Niiler PP, Bernstein RL (1996) The three-dimensional structure of an upper ocean vortex in the tropical Pacific Ocean. Nature 382:610–613
133. Duing W, Hisard P, Katz E, Meincke J, Miller L, Moroshkin K, Philander G, Kibnikov A, Voigt K, Weisberg R (1975) Meanders and long waves in equatorial Atlantic. Nature 257(6624):280–284
134. Legeckis R (1977) Long waves in eastern equatorial Pacific Ocean - view from a geostationay satellite. Science 197(4309):1179–1181
135. Halpern D, Knox R, Luther D (1988) Observations of 20-day period meridional current oscillations in the upper ocean along the Pacific equator. J Phys Oceanogr 18(11):1514–1534
136. Qiao L, Weisberg RH (1995) Tropical instability wave kinematics: Observations from the Tropical Instability Wave Experiment TIWE. J Geophys Res 100:8677–8693

137. Lyman JM, Johnson GC, Kessler WS (2007) Distinct 17- and 33-day Tropical Instability Waves in subsurface observations. J Phys Oceangr 37:855–872
138. Chelton DB, Wentz FJ, Gentemann CL, de Szoeke RA, Schax MG (2000) Satellie microwave SST observations of transequatorial tropical instability waves. Geophys Res Lett 27(9):1239–1242
139. Chelton DB, Esbensen S, Schlax MG, Thum N, Freilich M, Wentz F, Gentemann C, McPhaden M, Schopf P (2001) Observations of coupling between surface wind stress and sea surface temperature in the eastern tropical Pacific. J Climate 14(7):1479–1498
140. Miller L, Watts D, Wimbush M (1985) Oscillations of dynamic topography in the eastern equatorial Pacific. J Phys Oceanogr 15(12):1759–1770
141. Jochum M, Malanotte-Rizzoli P, Busalacchi A (2004) Tropical instability waves in the Atlantic Ocean. Ocean Modelling 7(1–2):145–163
142. Menkes C, Kennan S, Flament P, Dandonneau Y, Masson S, Biessy B, Marchal E, Eldin G, Grelet J, Montel Y, Morliere A, Lebourges-Dhaussy A, Moulin C, Champalbert G, Herbland A (2002) A whirling ecosystem in the equatorial Atlantic. Geophys Res Lett 29(11):
143. Lyman JM, Chelton DB, de Szoekes RA, Samelson RA (2005) Tropical instability waves as a resonance between equatorial Rossby waves. J Phys Oceangr 35(2):232–254
144. Philander SGH (1976) Instabilities of zonal equatorial currents. J Geophys Res 81(21):3725–3735
145. Philander SGH (1978) Instabilities of equatorial currents. In: Moore DW, Witte J (eds) Review Papers of Equatorial Oceanography FINE Workshop Proceedings, FGGE/Index/NORPAX Equatorial, Nova University/N. Press, Dania, Florida, Y. I. T, pp 1–14
146. Cox MD (1980) Generation and propagaion of 30-day waves in a numerical model of the Pacific. J Phys Oceangr 10(8):1168–1186
147. Donohue KA, Wimbush M (1998) Model results of flow instabilities in the tropical Pacific Ocean. J Geophys Res 103(C10):21,401–21,412
148. Masina S, Philander SGH, Bush ABG (1999) An analysis of tropical instability waves in a numerical model of the Pacific Ocean - 2. generation and energetics of the waves. J Geophys Res-Oceans 104(C12):29,637–29,661
149. Zhou C, Boyd JP (2010) Nonlinear shallow water tropical instability waves on the equatorial beta-plane: Genesis of two distinct types of waves. Geophys Res Lett 36(23):L23,605, 4 pp
150. Proehl JA (1996) Linear stability of equatorial zonal flows. J Phys Oceangr 26:601–621
151. Dutrieux P, Menkes CE, Vialard J, Flament P, Blanke B (2008) Lagrangian study of Tropical Instability Vortices in the Atlantic. J Phys Oceangr 38(2):400–417
152. Ingersoll AP (1973) Jupiter's Great Red Spot: A free atmospheric vortex? Science 182:1346–1348
153. Maxworthy T, Redekopp LG (1976) A solitary wave theory of the great red spot and other observed features in the Jovian atmosphere. Icarus 29(2):261–271
154. Maxworthy T, Redekopp LG, Weidman PD (1978) On the production and interaction of planetary solitary waves: Applications to the Jovian atmosphere. Icarus 33:388–409
155. Ingersoll AP, Cuong PG (1981) Numerical model of long-lived Jovian vortices. J Atmos Sci 38:2067–2076
156. Showman AP, Dowling TE (2000) Nonlinear simulations of Jupiter's $5-\mu m$ hot spots. Science 289(5485):1737–1740
157. Osborne AR, Burch TL (1980) Internal solitons in the Andaman sea. Science 208(4443):451–460
158. Chen G, Su FC, Wang CM, Liu CT, Tseng RS (2011) Derivation of internal solitary wave amplitude in the South China Sea deep basin from satellite images. J Oceanography 67(6):689–697
159. Feitsma TRF, Dijkstra HA (1995) The Rossby soliton: A robust non-linear structure in the equatorial ocean. In: Zwerver S, van Rompaey RSA (eds) Climate Change Research: Evaluation and Policy Implications, A & B: International Conference on Climate Change Research - Evaluation and Policy Implications, Maastrict, Netherlands, Springer, no. 65 in Studies in Environmental Science, pp 139–158

160. Han L (2014) A two-time-level split-explicit ocean circulation model MASNUM and its validation. Acta Oceanologica Sinica 11:11–35
161. Bousquet A, Marion M, Petcu M, Temam R (2013) Multilevel finite volume methods and boundary conditions for geophysical flows. Computers and Fluids 74:66–90
162. Giraldo FX, Restelli M (2010) High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model. Internat J Numer Meths Fluids 63(9):1077–1102
163. Chu PC, Fan C (2010) Space-time transformation in flux-form Semi-Lagrangian schemes. Terrestial Atmos Oceanic Sci 21(1):17–26
164. Vaid BH, Gnanaseelan C, Salvekar PS (2008) Observed twin gyres and their interannual variability in the equatorial Indian Ocean using Topex/Poseidoon altimetry. Current Sci 98(1):30–35

Chapter 17
Nonlinear Wavepackets and Nonlinear Schroedinger Equation

Abstract Wave packets in the narrow meaning of the word denotes pulses that are the product of a "carrier wave" (with the three-dimensional structure of an infinitesimal amplitude sine wave) multiplied by an "envelope" or "modulation" function that slowly varies in both space and time and satisfies the cubically-Nonlinear Schroedinger (NLS) equation. The coefficient of the nonlinear terms, the "Landau constant", has many discrete singularities at various carrier wave wavenumbers k; each pole is the mark of a resonance. Long wave/short wave resonance, second harmonic resonance and resonant triad interactions are all discussed. To illustrate the connections between different PDEs, the KdV equation for a given weakly dispersive mode is derived from the NLS equation. Lastly, selected numerical studies of idealized nonlinear wave dynamics, not restricted to small amplitude, are reviewed.

> It is a common experience that asymptotic calculations tend to give errors which are far smaller than one might have reasonably expected.
>
> Carl M. Bender and Steven A. Orszag on p. 340 of their book [1]. The context is a matched asymptotics treatment of an eigenvalue problem which becomes increasingly accurate as the mode number increases; for this example, even the ground state eigenvalue is accurate to within about 3%.

17.1 The Nonlinear Schroedinger Equation for Weakly Nonlinear Wavepackets: Envelope Solitons, FPU Recurrence and Sideband Instability

Wave packets are pulses that are the product of a "carrier wave" (with the three-dimensional structure of an infinitesimal amplitude plane wave) multiplied by an "envelope" or "modulation" function that slowly varies in both space and time and satisfies the cubically-Nonlinear Schroedinger (NLS) equation. The Nonlinear Schroedinger (NLS) equation is second order in space, first order in time and cubically nonlinear. It is not an equation for the longitude-and-time dependence of a wave, but rather is an equation for the *envelope* of a wave packet. Its solution is

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0_17

usually *complex-valued* because the complete $x - t$ factor is not $A(x,t)$ but rather $A(x,t)\mathfrak{C}$ + complex conjugate where $\mathfrak{C} = \exp(ikx + i\omega t)$ is the usual carrier wave of a wavepacket.

In the next subsection, we show how very general initial conditions evolve under *linear* wave dispersion to the form of a "carrier wave" multiplied by a slowly-varying "envelope" or "modulation". Generically, the NLS equation describes how nonlinearity modifies the evolution of the envelope. In purely linear dynamics, dispersion always causes the envelope to widen in space, varying more and more slowly with increasing time, and simultaneously shrinking in amplitude. Nonlinearity can either accelerate linear dispersion ("defocusing") or oppose it, allowing in the "focusing" case the formation of envelope solitary waves.

The NLS approximation fails when the "carrier wave" is resonant with other waves. Later, we shall discuss each of these types of resonances in turn. The nonlinear coefficient of the Nonlinear Schroedinger equation is called the "Landau constant" and is a function $\nu(k)$ of the wavenumber k of the carrier wave. Each resonance creates a pole in $\nu(k)$. Each latitudinal mode/baroclinic mode has its own NLS dynamics; thus, the Landau constant is really a countable double infinity of functions.

Marshall and Boyd [2] derive the NLS for equatorial waves. Multiple scales perturbation theory is applied in a fashion quite similar to the KdV derivation given earlier. Instead of copying their analysis here, we show the conceptual skeleton in a much simpler context by deriving the NLS equation from the KdV equation.

Our derivation will follow Johnson [3, 4], who showed that in the long wave limit, the dynamics of wavepackets in the Korteweg-deVries (KdV) equation could be described by the Nonlinear Schrödinger (NLS) equation. Marshall and Boyd [2] wrote a program to compute the NLS coefficients for equatorially-trapped planetary-scale ocean waves and then derived the KdV coefficients directly from those of the NLS equation. Tracy, Larson, Osborne and Bergamasco [5, 6] and also Osborne's book give a very readable analysis of the connection between the NLS and KdV periodic inverse scattering methods for wavepackets.

17.2 Linear Wavepackets

Before analyzing nonlinear wavepackets, it is helpful to first understand the simpler case of infinitesimal amplitude waves. In a spatially-uniform medium without boundaries, a one-dimensional wave disturbance can always be represented as a sum of Fourier integrals, one for each branch of the dispersion relation, of the form [7]

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(\tilde{k})\exp(i\tilde{k}x - iw(\tilde{k})t)\,d\tilde{k} \tag{17.1}$$

where $w(\tilde{k})$ is the dispersion relation for a given branch. For a wavepacket in the narrow sense, as used here, the Fourier transform amplitude is sharply peaked about

some wavenumber $\tilde{k} = k$. Without approximation, the integral can be rewritten as the product of a "carrier wave" factor,[1]

$$E(x,t) \equiv \exp(i[kx - \omega(k)t]) \tag{17.2}$$

and an "envelope" $A(x - c_g t, t)$ where

$$c_g(k) \equiv \frac{\partial w}{\partial k}(k), \qquad \text{[Group velocity]} \tag{17.3}$$

Introducing a spatial coordinate in a frame of reference moving at the group velocity,

$$\zeta \equiv x - c_g t, \tag{17.4}$$

the wavepacket is

$$u(x,t) = E(x,t)A(\zeta,t) \tag{17.5}$$

where $A(\zeta, t)$ is given *exactly* by, replacing $\bar{k}$ by $\kappa \equiv \bar{k} - k$,

$$A(\zeta,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \alpha(\kappa)\, \exp(i\kappa\zeta - i\Omega(\kappa)t)\, d\kappa \tag{17.6}$$

where the Fourier amplitude is

$$\alpha(\kappa) \equiv U(k+\kappa) \tag{17.7}$$

and the new frequency is

$$\Omega(\kappa) \equiv w(k+\kappa) - w(k) - c_g(k)\kappa \tag{17.8}$$

The derivation is merely a matter of splitting off the first two terms in the Taylor series of $w(k + \kappa)$ about k.

The crucial point is that the envelope is described by a Fourier integral of exactly the same form as the integral for $u(x, t)$ itself except that the spatial coordinate is now in a frame of reference which is traveling at the group velocity and the amplitude and frequency functions have been redefined.

Because $\alpha(\kappa)$ is sharply peaked about $\kappa = 0$ if the envelope is sufficiently wide in ζ, it is legitimate to expand the frequency in a power series in κ and retain only the first few terms. This is equivalent to the multiple scales expansion of the next section where the perturbation parameter is the ratio of the narrow scale of the carrier wave, $1/k$, with the broad scale of the envelope. The approximation of Taylor-expanding Ω up to a finite order is equivalent, without additional approximations, to replacing the true wave equation satisfied by $A(\zeta, t)$ by the approximation

[1] E is used for the carrier wave in this chapter only to be consistent with Johnson.

$$A_t - i\,\frac{1}{2}w_{kk}(k)\,A_{\zeta\zeta} - \frac{1}{6}w_{kkk}(k)\,A_{\zeta\zeta\zeta} + i\,\frac{1}{24}w_{kkkk}(k)\,A_{\zeta\zeta\zeta\zeta} + \frac{1}{120}w_{kkkkk}(k)\,A_{\zeta\zeta\zeta\zeta\zeta} + \ldots = 0 \tag{17.9}$$

Example: Linearized KdV

$$u_t + cu_x + bu_{xxx} = 0 \tag{17.10}$$

where b and c are constants.

$$w = kc - bk^3 \qquad w_k = c - 3bk^2, \qquad w_{kk} = -6bk, \qquad w_{kkk} = -6b \tag{17.11}$$

$$\Omega(\kappa) = -3bk\kappa^2 - b\kappa^3 \tag{17.12}$$

Thus, the envelope of a linearized KdV wave packet satisfies the equation

$$A_t + i3bkA_{\zeta\zeta} + bA_{\zeta\zeta\zeta} = 0 \tag{17.13}$$

The partial differential equation for the envelope of a weakly *nonlinear* wavepacket in the KdV equation will be identical except for additional terms, nonlinear in the envelope, which will be derived in the next section.

17.2.1 Perturbation Parameters

There is some arbitrariness in the definitions of the perturbation parameters because phrases like "width of the envelope" are meaningful only in order-of-magnitude. Nevertheless, the following choices are reasonable. The dispersion parameter λ is the ratio of the length scale of the carrier wave, $1/k$, to the width $1/\beta(t)$ of the envelope:

$$\lambda(t) \equiv \beta(t)/k \qquad \text{[Dispersion Parameter]} \tag{17.14}$$

The nonlinearity parameter will be defined to be

$$\mu(t) \equiv \max_x |A(x,t)| \tag{17.15}$$

For sufficiently large time, as discussed below, the wavepackets of the KdV equation are always widening, implying that as $t \to \infty$, $\beta(t) \sim O(1/t)$ is a decreasing function of time. To conserve the total energy of the wavepacket, the envelope must simultaneously decrease as $O(1/\sqrt{t})$. [This square root is counterintuitive, but recall that the total, integrated energy, which must be constant in time, is proportional to the *square* of the amplitude but is linear in the *width* of the envelope.] The ratio of the nonlinear to dispersive terms in the NLS equation is

$$\frac{\text{nonlinear}}{\text{dispersive}} \sim \left(\frac{\mu^2}{\lambda^2}\right) \sim O\,(t)\,, \qquad t \to \infty \tag{17.16}$$

Thus, the envelope evolves more and more *slowly* as time increases, but such evolution as there is becomes increasingly dominated by *nonlinearity*.

17.3 Derivation of the NLS Equation from the KdV Equation

The most general perturbation expansion is a *double* series in which both the nonlinear parameter and dispersion parameter are allowed to vary independently. However, the most interesting regime [arguably the *only* interesting regime] is when the nonlinearity and dispersion are of comparable importance. We shall henceforth assume this, thereby collapsing a double series into an expansion in a single parameter ε. However, the resulting series will still be accurate when nonlinearity $\gg$ dispersion or vice versa so as long as the larger effect is sufficiently small.

The KdV equation is

$$u_t + cu_x + auu_x + bu_{xxx} = 0 \tag{17.17}$$

where a, b, and c are constants. To focus on weakly nonlinear wavepackets, assume that the lowest order approximation is

$$u(x,t) \sim \varepsilon\, A(\zeta,\tau)\, \exp(i[kx - \omega(k)t]) + \text{C. C.} \tag{17.18}$$

where $\varepsilon << 1$ is the perturbation parameter and k is the zonal wavenumber of the "carrier wave" and where C. C. denotes the complex conjugate of the preceding term. The frequency, phase and group velocities of the carrier wave are specified by the *linear* KdV dispersion relation as

$$\omega = kc - bk^3 \quad \rightarrow \quad c_p(k) = c - bk^2 \quad \& \quad c_g(k) = c - 3bk^2 \tag{17.19}$$

where the "slow" space and time variables are

$$\zeta \equiv \varepsilon\,(x - c_g(k)t), \qquad \tau = \varepsilon^2 t, \quad \tilde{\tau} = \varepsilon^3 t, \quad \overline{\tau} = \varepsilon^4 t \tag{17.20}$$

The factor $A(\zeta,\tau)$ is the called the "envelope" and will evolve according to the NLS equation or its generalization. Different powers of ε (from the scalings of Eqs. 17.18 and 17.20) also give legitimate limits, but a posteriori our choices turn out to be the ones such that nonlinearity and dispersion are of equal magnitude.

Through the method of multiple scales [3, 8, 9], one can show that the weakly nonlinear wavepacket has an asymptotic expansion of the form

$$u = \varepsilon \sum_{j=0}^{\infty} \varepsilon^j \sum_{m=0}^{j+1} A^{jm}(\zeta, \tau, \tilde{\tau}, \overline{\tau})\, E^m + \text{C. C.} \tag{17.21}$$

where E is the "carrier wave":

$$E(x,t) \equiv \exp(i[kx - \omega(k)t]) \tag{17.22}$$

For notational simplicity, write $A^{01} = A(x,t)$. The terms which are independent of the carrier wave are known collectively as the "long wave" components; the terms proportional to $E^{\pm m}$ are the m-th harmonics except for $m = 1$, which is the "fundamental" $A(x,t)$.

Derivatives are interpreted by applying the chain rule:

$$\frac{\partial}{\partial t} \to \frac{\partial}{\partial t} - \varepsilon\, c_g \frac{\partial}{\partial \zeta} + \varepsilon^2 \frac{\partial}{\partial \tau} + \varepsilon^3 \frac{\partial}{\partial \tilde{\tau}} + \varepsilon^4 \frac{\partial}{\partial \overline{\tau}}, \qquad \frac{\partial}{\partial x} \to \frac{\partial}{\partial x} + \varepsilon \frac{\partial}{\partial \zeta} \tag{17.23}$$

In a minor (and common) abuse of notation, the same symbols x and t are used for the "fast" space and time variables as for the original coordinates.

Matching powers of ε and the carrier wave exponential E shows that at $O(\varepsilon^{j+1} E^m)$, A^{jm} solves an equation of the form

$$\left\{ ik^3 bm\,[1 - m^2] \right\} A^{jm}(\zeta, \tau) = R^{j,m} \tag{17.24}$$

where the expression in braces is $[(\partial_t + c\partial_x + b\partial_{xxx})E^m]\,/E^m$ and where $R^{j+1,m}$ denotes the $O(\varepsilon^j E^m)$ part of the "residual" function, which is the result of substituting the expansion into the KdV equation and then dividing by the common factor of ε.

For $m = 1$, i.e., the fundamental, the left-hand side of Eq. 17.24 is zero. Therefore, the perturbation scheme fails unless $R^{j,1} = 0$ for all $j \geq 0$. This set of "nonsecularity" conditions determines the linear dispersion relation at lowest order $j = 0$, regurgitates the usual linear definition of group velocity ($c_g \equiv \partial\omega(k)/\partial k$) for $j = 1$, and demands that A evolve according to the NLS equation to satisfy $R^{2,1} = 0$. Figure 17.1 is a visual summary. The next two orders give expressions for the time derivatives of A on the very slow ($\tilde{\tau}$) and ultra-slow ($\overline{\tau}$) time scales.

The longwave components are independent of the "fast" variables; because derivatives with respect to the slow variable ζ are multiplied by ε, A^{j0} appears in the residual only at one order higher in ε than the harmonics of the same magnitude in ε, A^{jm}, $m > 1$. The longwave components satisfy equations of the form

$$3bk^2 A^{j0}_{\zeta} = R^{j,0} \tag{17.25}$$

It turns out that after some manipulation, the longwave component of the residual can always be written as the derivative of an expression with respect to ζ. This allows us to integrate with respect to ζ (trivially) and obtain an explicit expression for A^{j0}.

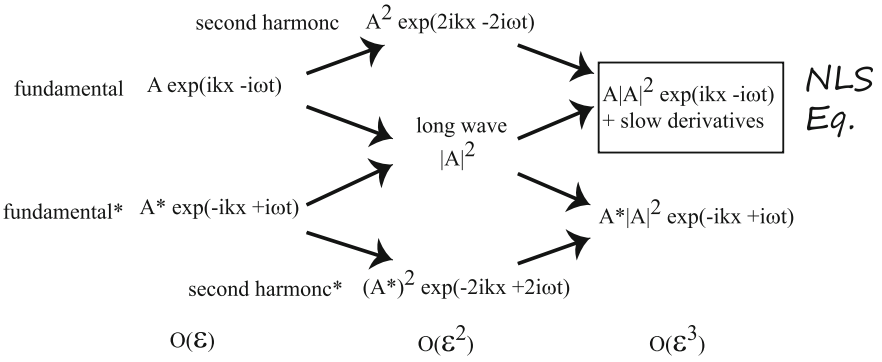

Fig. 17.1 Schematic of NLS perturbation theory. The $O(\varepsilon)$ solution, the "fundamental", and its complex conjugate, interact to produce three terms at $O(\varepsilon^3)$. The interaction of the fundamental with itself generates the second harmonic; similarly, the self-interaction of the complex conjugate of the fundamental yields the complex conjgate of the second harmonic. The interaction of the fundamental with its complex conjugate gives the long wave component. This lacks the rapidly-varying carrier wave [(hence the label "long wave"] and is pure envelope, the absolute value of the square of the envelope A of the envelope. None of the $O(\varepsilon^2)$ terms is resonant with the fundamental. At next order, the interaction of the fundamental with its second harmonic gives a third harmonic (proportional to $\exp(3ikx - 3iw(k))$) [not shown] plus terms which have the same dependence on the fast carrier wave as the fundamental, and therefore are resonant. Linear operators that are partial derivatives of the slow time and space derivatives are also resonant at this order. The resonance disappears, and secular terms therefore are absent, only if the sum of the resonant terms is zero. This sum is the NLS equation

The lowest non-trivial solution is to compute the $O(\varepsilon^2)$ longwave and second harmonic components and then $R^{2,1} = 0$ yields the NLS equation. This was previously derived in Appendix C of [8], closely following [3]. Here, we extend the analysis two orders higher.

This introduces two minor technical complications. First, it is very helpful to remove time derivatives from the higher order residuals by replacing A_τ by invoking the NLS equation it solves. Second, for the longwave components, it is necessary to manipulate the simplified residual into the form of a ζ-derivative so that the expression for A_ζ^{j0} can be integrated. This requires repeated use of identities such as $AA^*_{\zeta\zeta} = (AA^*_\zeta)_\zeta - A_\zeta A^*_\zeta$ where the asterisk denotes complex conjugation; formally integrating the residual in ζ, these identities can be equivalently interpreted as repeated integration-by-parts.

A^* will denote the complex conjugate of A. Then

$$A^{10} = -\frac{a}{3bk^2}\,|A|^2, \qquad A^{12} = \frac{a}{6bk^2}\,A^2 \tag{17.26}$$

$$A^{20} = i\,\frac{a}{3\,b\,k^3}\,\left\{AA^*_\zeta - A^*A_\zeta\right\}, \quad A^{22} = i\,\frac{a}{3\,b\,k^3}\,A\,A_\zeta, \quad A^{23} = \frac{a^2}{48\,b^2\,k^4}\,A^3 \tag{17.27}$$

$$A^{30} = \frac{a}{3\,b\,k^4}\left\{A A^*_{\zeta\zeta} + A^* A_{\zeta\zeta} - A_\zeta\, A^*_\zeta\right\} \tag{17.28}$$

$$A^{32} = -\frac{a^3}{432 b^3\, k^6} A^3\, A^* - \frac{a}{6bk^4} A^2_\zeta - \frac{a}{3bk^4} A A_{\zeta\zeta} \tag{17.29}$$

$$A^{33} = i\,\frac{a^2}{12\,b^2\,k^5} A^2\, A_\zeta \tag{17.30}$$

$$A^{34} = \frac{a^3}{432 b^3\, k^6} A^4 \tag{17.31}$$

At $O(\varepsilon^3)$, one obtains the usual NLS equation:

$$A_\tau + i3\,b\,k A_{\zeta\zeta} - i\frac{a^2}{6\,b\,k} A\,|A|^2 = 0 \tag{17.32}$$

At higher order, one obtains variations in A on longer timescales. The most convenient procedure is to simply combine these slow and ultra-slow time tendencies with the NLS equation to obtain the Generalized-NLS (GNLS) equation

$$\begin{aligned} A_\tau + i3\,b\,k A_{\zeta\zeta} - i\frac{a^2}{6\,b\,k} A\,|A|^2 \qquad & \text{[Generalized-NLS]} \\ + b A_{\zeta\zeta\zeta} - \frac{a^2}{2\,bk^2}\, A^2 A^*_\zeta - \frac{a^2}{3\,bk^2}\,|A|^2 A_\zeta & \\ + i\frac{a^4}{864 b^3 k^5}|A|^4 A + i\frac{2a^2}{3bk^3}\, A^2\, A^*_{\zeta\zeta} & \\ - i\frac{a^2}{6bk^3}\,\left(A_\zeta\right)^2\, A^* + i\frac{a^2}{3bk^3}\, A_\zeta\, A^*_\zeta\, A = 0 & \end{aligned} \tag{17.33}$$

where the second line is the correction at $O(\varepsilon^4)$ and the last two lines are that at $O(\varepsilon^5)$, respectively. (Our approximation, (17.26–17.32) above removes the long wave and fundamental parts of the residual at $O(\varepsilon^5)$ to eliminate these secular terms, but not the higher harmonics.)

The perturbative solutions at various orders will be compared with exact KdV solutions in the next-to-last section below.

17.3.1 NLS Dilation Group Invariance

The NLS equation with arbitrary coefficients can always be reduced to a convenient standard form through the following group invariance theorem.

Theorem 17.1 (Reduction of NLS Equation to Canonical Form) *The NLS equation with arbitrary coefficients, but written in a frame of reference moving at the linear long wave speed* c_0,

$$i\mathscr{A}_\tau + F\mathscr{A}_{\xi\xi} + G|\mathscr{A}|^2\mathscr{A} = 0, \tag{17.34}$$

is related to the solution u of the canonical form

$$iu_t + u_{xx} + 2|u|^2 u = 0 \tag{17.35}$$

by

$$t = \frac{G}{2}\tau \tag{17.36}$$

$$x = \sqrt{\frac{G}{2F}}\xi \tag{17.37}$$

$$u = \mathscr{A} \tag{17.38}$$

Proof Substitution of the assumed form into the partial differential equation followed by cancellation of the parameters. ■

17.3.2 Defocusing

The NLS equation has two qualitatively different cases which cannot be converted into one another by any rescaling of space, time or amplitude. When the signs of the dispersive ($A_{\zeta\zeta}$) and nonlinear terms are the *same*, nonlinearity and dispersion can balance so as to create an envelope solitary wave. When the signs of the dispersive and nonlinear terms are *opposite*, nonlinearity *accelerate* dispersion, a sort of "superlinear" dispersion.

In the KdV regime, the signs of the dispersive coefficient, $-3bk$, and the nonlinear coefficient, $a^2/(6bk)$, are *always opposite* independent of the signs and magnitude of the coefficients a and b of the KdV equation. Envelope solitary waves are impossible even though the KdV equation has solitary waves consisting of a single crest. Dispersion and nonlinearity both act to widen the envelope, at least in the asymptotic limit as $t \to \infty$. (The envelope may compress for a finite time interval if the initial conditions are appropriately contrived, but will always widen and shrink in amplitude for sufficiently large time.)

Kevrikis, Frantzeskakis and Carretero have recently published a tome of four hundred pages devoted entirely to the physics of the NLS defocusing case [10]. Their book, Yang [11] and Osborne's [12] treatise supply many details omitted here. Osborne puts much more emphasis on water waves. His first great contribution was the discovery of large amplitude internal gravity solitons in the Andaman Sea [13].

17.3.3 Focusing, Envelope Solitons and Resonance

The veins and arteries of the internet and landline telephony are fiber optical cables which transmit information as pulses of light. Unfortunately, dispersion is inexorable in linear physics. Repeater stations are needed every ten to twenty kilometers to amplify and narrow the light wavepackets lest dispersive spreading overlap Mrs. Jones' gossip with Dr. Paciorek's order to sell ten thousand dollars worth of cocoa futures to the ruination of both the social and financial life of the republic.

Lasers, though, can generate *nonlinear* wavepackets. Although the scale and speed of planetary equatorial ocean waves is vastly different from the tiny streaks of brightness whizzing down a glass highway, both are governed by the NLS equation.

Envelope solitary waves are localized wave packets that do not disperse. Tremendous effort has been expended on understanding the nonlinear dynamics of wavepackets in the hope that networks with vast numbers of repeaters may soon be replaced by repeater-free fiber optics carrying nothing but envelope solitons. Although nonlinear wavepacket dynamics has been greatly generalized for water waves and especially fiber optics, NLS theory still lies at its heart.

The NLS equation is solvable by the inverse scattering transform. The solitary waves as in the KdV case are each associated with a discrete eigenvalue of the eigenproblem whose potential energy function is the initial condition for the time-dependen PDE.

A full description of NLS phenomenology must be left to the books by Jianke Yang [11] and Alfred Osborne [12]. However, we can and should briefly review the following major types of NLS solutions.

1. uniform envelope (nonlinear plane wave)
2. soliton generation threshold
3. cnoidal and dnoidal waves
4. double soliton
5. elastic collisions
6. breathers
7. polycnoidal waves
8. sideband instability and FPU recurrence

By elementary rescalings, the NLS equation with general coefficients can be reduced to the canonical form used by Yang [11], which we shall employ to catalog NLS phenomena:

$$iu_t + u_{xx} + 2|u|^2 u = 0 \tag{17.39}$$

17.3.4 Nonlinear Plane Wave

It must be always remembered that the solution to the NLS equation is merely the *envelope* of the wave packet, and may therefore be *complex-valued*. The nonlinear plane wave is

$$u = A\exp(i2A^2k^2t) \tag{17.40}$$

where the envelope A is a constant. When multiplied by the carrier wave, the product of envelope and carrier wave is still sinusoidal in space and time. The uniform envelope represents merely an adjustment to the wavenumber and frequency of the carrier wave. For most purposes, the uniform envelope is of no interest whatsoever. However, slightly altered carrier waves play a fundamental role in the theory of sideband instability as described below.

17.3.5 Envelope Solitary Wave

NLS solitons must be described by the phrase "envelope solitary wave" because these are not a single peak, but rather are many peaks and valleys traveling together as a coherent structure.

The discrete eigenvalues of a scattering problem determine each solitary wave emerging from a given initial condition, just as for the KdV equation, but the NLS scattering problem has eigenvalues which in general are *complex-valued*. Let ξ and η be the real and imaginary parts of the eigenvalue. The envelope soliton is then

$$u(x,t) = 2\eta\text{sech}(2\eta(x + 4\xi t - x_0))\,\exp\left\{-2i\xi x - 4i(\xi^2 - \eta^2)t + i\sigma_0\right\} \tag{17.41}$$

where the parameters x_0 and σ_0 are the initial position and phase of the soliton. The peak amplitude of the envelope is 2η while its velocity is -4ξ.

The factor

$$u(x,t) = 2\eta\exp\left\{-2i\xi x - 4i(\xi^2 - \eta^2)t + i\sigma_0\right\} \tag{17.42}$$

is a sine wave, identical in form to the carrier wave. Indeed, it has also the form of the nonlinear plane wave and like it, is merely a slight adjustment to the wavenumber and frequency. For every soliton with carrier wave of wavenumber k that is traveling at speed -4ξ plus the linear group velocity $c_g(k)$, there is a stationary solitary wave ($\xi = 0$) associated with wavenumber $k + 2\xi$. The degree of freedom represented by ξ does not increase the span of the single-soliton solutions and as such should largely be ignored.

17.3.6 NLS Cnoidal and Dnoidal Waves

The spatially periodic solutions to the NLS equation solve the boundary value problem

$$u_{xx} - w\,u + 2\,u^3 = 0. \tag{17.43}$$

where w is a constant. This has two classes of solutions, dubbed the "dnoidal" and "NLS cnoidal" waves. For both classes, $c = (u_{xx} + 2u^3)/u$ independent of x, which is sometimes a convenient alternative to the elliptic integral formulas below.

DNOIDAL BRANCH

The parameters are the spatial period P and the shape/elliptic modulus parameter $S \in (0, \infty]$. The "dnoidal wave" is so-called because its envelope is proportional to the elliptic function dn. For weak waves, this branch asymptotes to a constant. Define $\varepsilon \equiv (\pi/P)S$. Then

$$u(x) = \pm\varepsilon \sum_{j=-\infty}^{\infty} \operatorname{sech}(\varepsilon(x - Pj)) \tag{17.44}$$

$$= \pm\frac{\pi}{P}\left\{1 + 2\sum_{n=1}^{\infty} \operatorname{sech}\left(\frac{n\pi^2}{\varepsilon P}\right) \cos\left(\left[\frac{2\pi}{P}\right] nx\right)\right\} \tag{17.45}$$

$$= \pm\frac{2K}{P} \operatorname{dn}\left(\frac{2K}{P}x; m_{ell}\right) \tag{17.46}$$

$$k_{ell}(S) = \frac{\sum_{j=-\infty}^{\infty} \operatorname{sech}((j + 1/2)\pi/S)}{\sum_{j=-\infty}^{\infty} \operatorname{sech}(j\pi/S)} \qquad [S < 1] \tag{17.47}$$

$$= \frac{\sum_{j=-\infty}^{\infty} (-1)^j \operatorname{sech}(j\pi S)}{\sum_{j=-\infty}^{\infty} \operatorname{sech}(j\pi S)} \qquad [S > 1] \tag{17.48}$$

$$K(S) = \frac{\pi}{2} + \pi \sum_{j=1}^{\infty} \operatorname{sech}\left(\frac{j\pi}{S}\right) \qquad [S < 1] \tag{17.49}$$

$$= S\frac{\pi}{2} + S\pi \sum_{j=1}^{\infty} \operatorname{sech}(j\pi S) \qquad [S > 1] \tag{17.50}$$

$$c = \varepsilon^2 \left\{1 + 12\sum_{n=1}^{\infty} \frac{\cosh(n\varepsilon P)}{\sinh^2(n\varepsilon P)}\right\}$$

$$= 4\frac{K^2(S)}{P^2}(2 - k_{ell}^2) \tag{17.51}$$

Note that these series converge for all S; the ranges shown merely indicate where each series is most efficient.

CNOIDAL BRANCH

The cnoidal branch is approximated by a cosine function for small amplitude. The parameter $\varepsilon \equiv (2\,\pi/P)S$ which differs by a factor of two from the definition that is convenient for the dnoidal wave.

$$u(x) = \varepsilon \sum_{j=-\infty}^{\infty} (-1)^j \,\mathrm{sech}\,(\varepsilon(x - j\,P/2)) \tag{17.52}$$

$$= \frac{4\pi}{P}\left\{ \sum_{n=1}^{\infty} \mathrm{sech}\left(\frac{(2n-1)\pi^2}{\varepsilon\, P}\right) \cos\left(\left[\frac{2\pi}{P}\right](2n-1)\,x\right)\right\} \tag{17.53}$$

$$= \frac{4\,k_{ell}\,K}{P}\,\mathrm{cn}\left(\frac{4K}{P}\,x; m_{ell}\right) \tag{17.54}$$

where $m_{ell} = k_{ell}^2$.

Note that c becomes negative for small amplitude waves.

$$c = 32\,\frac{K^2(\varepsilon)}{P^2}\,\left\{k_{ell}^2(\varepsilon) - \frac{1}{2}\right\} \tag{17.55}$$

For both branches of periodic traveling waves, the Nonlinear Superposition Principle applies: The pattern function for the imbricate series of the dnoidal and NLS-cnoidal waves is the hyperbolic secant function, the shape of envelope solitary waves.

17.3.7 N-Soliton Solutions

There is an analytical formula for an exact N-soliton solution, too messy to be repeated here (but see [11]). If the real parts for all the discrete eigenvalues are distinct, then the envelope solitons travel at different speeds. Like KdV solitary waves, their collisions are elastic as illustrated in the left panel of Fig. 2.1 on page 35 of Yang [11]. Just as for the collisions of KdV solitary waves, NLS collisions do phase-shift and position-shift the solitary waves; the post-collison trajectory of each is parallel to the pre-collision trajectory but they do not coincide as illustrated in Fig. 17.2.

17.3.8 Breathers

Boyd's Law of Parametric Probability: A Set of Measure Zero Occurs About Half the Time.

When the NLS inverse scattering eigenproblem has two eigenvalues and the *real parts* of the eigenvalues are *equal*, the two solitary waves do not separate, but instead

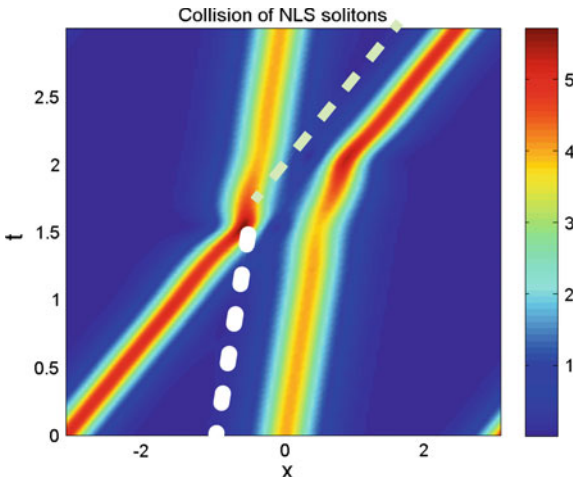

Fig. 17.2 The *white dashed* and *dotted curves* extend the trajectories of the two solitary waves as they would move in the absence of the collision; these curves make it easier to see the position shifts created by the soliton–soliton collision

propagate as a single coherent structure. This traveling wave is not steady state except for translation, but instead oscillates in time with a period of $\pi/(\eta_1^2-\eta_2^2)$ where η_1 and η_2 are the imaginary parts of the scattering eigenvalues. Because the periodic heaving up and down of this bound state of solitons resembles, at least to the imaginative, the rising and falling of the chest during respiration, such nonlinear waves are known as "breathers".

One might suppose that breathers would be very rare since it is obviously a special constraint that the real parts of the scattering eigenvalues should be equal. However, as expressed by the humorous quote at the beginning of this subsection, special species of events that occupy only a tiny part of parameter space often are common in applications. It is in fact very easy to generate breathers. Figure 17.3 shows a breather generated by an initial condition which is the imbrication of 2sech(x).

More complex breathers are possible. Merely increasing the initial condition by a factor of 3/2 gives a bound state of three solitary waves. The breather is non-translating and is centered at $x = 0$ for all t. The plot of $u(0, t)$ shows that this large amplitude breather is periodic in time, but has two distinct minima in each period. (Fig. 17.4)

17.3.9 Modulational ("Sideband") Instability, Self-focusing and FPU Recurrence

In the focusing case, $\text{sign}(nonlinearity) = \text{sign}(dispersiveterm)$, the sine wave, which is the uniform envelope $u(x, t) \equiv constant$, is always unstable. This instability is variously called "modulational instability", "sideband instability", or the "Benjamin–Feir" instability [14].

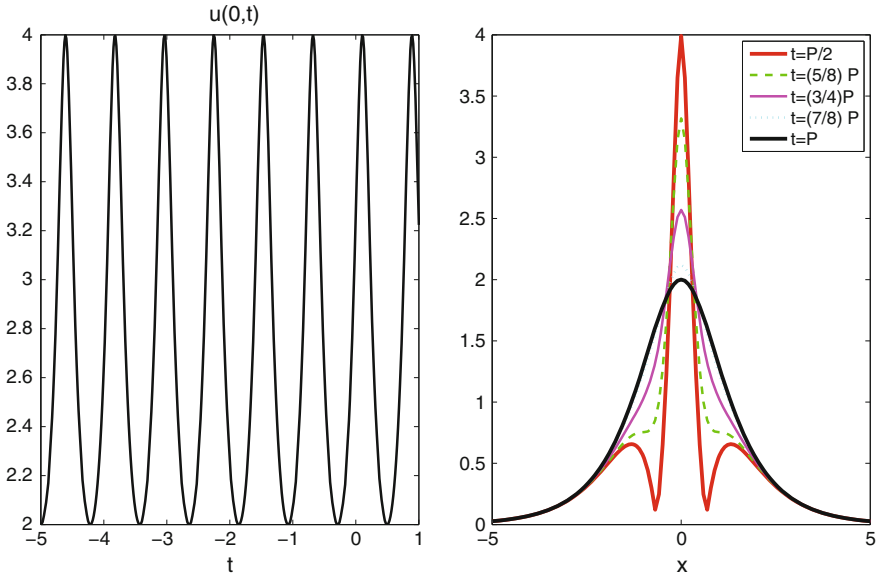

Fig. 17.3 The periodized initial condition $u(0,t) = \sum_{m=-2}^{2} 2\text{sech}(x - 8\pi m)$ gives rise to a breather which is a nearly sinusoidal function of time as illustrated by the time series of u at the center of the domain [*left plot*]. The shapes of $u(x,t)$ over a half period are illustrated on the right

The reason for the name "sideband instability" is that this instability begins as an interaction of a sinusoidal wave with two perturbations that are waves whose wavenumbers are only slightly different from those of the primary (large amplitude). Waves with wavenumbers very similar to a carrier wave of wavenumber k are said to be "sidebands" of the primary wave.

Most of the hard work of understanding sideband instability from first principles has already been accomplished by deriving the NLS equation.

What happens when resonant triad instability does not occur? It turns out that there is another possibility: Sideband instability. The key idea is that waves with wave numbers close to that of the primary wave will necessarily have frequencies close the primary wave, too. Thus these near-neighbors or "sidebands" are always in a sense weakly resonant with the primary wave.

Sideband instability has a growth rate proportional to the *square* of the amplitude of the primary wave instead of the amplitude itself. The reason is that the quadratic nonlinear interaction between two waves of $O(a)$ amplitude, both with $k \approx k_0$, will generate waves of $O(a^2)$ amplitude a sum of $k \approx 0$ [long wave component] and $k \approx 2k_0$ [second harmonic]. The quadratic interaction of the second harmonic or long wave with the primary wave will then generate a resonant wave with wavenumber k_0 and $O(a^3)$ amplitude. The growth rate of this resonant wave at the expense of the $O(a)$ primary wave is $O(a^2)$. In ordinary resonant triad interactions, the resonant wave is $O(a^2)$ and the growth rate is *linear* rather than quadratic in the amplitude of

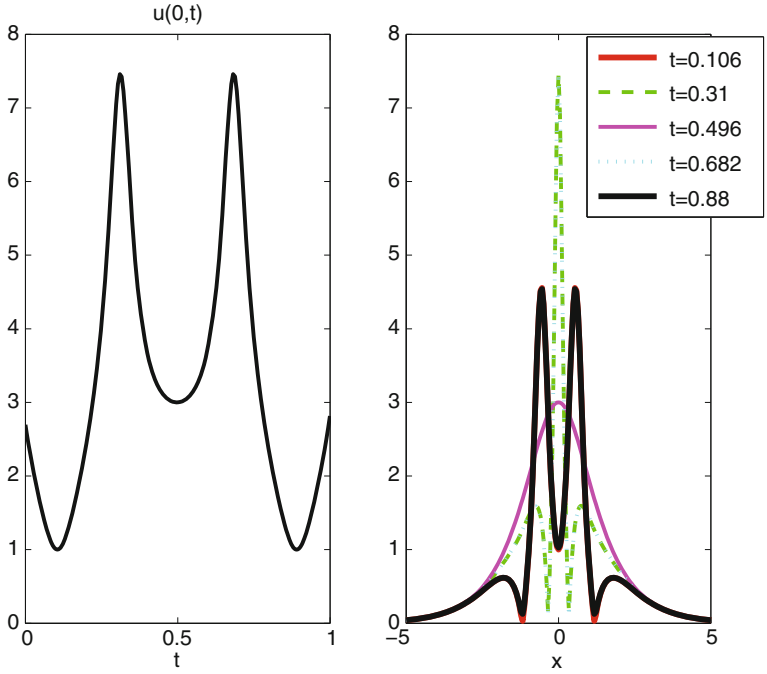

Fig. 17.4 Same as previous figure except that the periodized initial condition was $u(0,t) = \sum_{m=-2}^{2} 3\text{sech}(x - 8\pi m)$, a factor of 3/2 larger. The breather is time periodic as shown by the graph of $u(0,t)$ [*left plot*], but there are two distinct minima on each period. The *right plot* shows $u(x,t)$ for five different times that are inflection points in the time series

the primary wave. Sideband instability will inevitably be swamped by resonant triad instability if the latter occurs. When there is no resonant triad instability, however, the sideband instability mechanism can be extremely important (Fig. 17.5).

The pioneering study was a joint experimental-theoretical work by T. Brooke Benjamin and J. Feir [14]. Their analysis was restricted to plane waves, using a single wave as the basic state and then examining the effects of small sinusoidal perturbations with almost the same wave number. We can greatly streamline the labor, however, by using the Nonlinear Schroedinger equation instead as our starting point where the NLS equation is

$$iA_\tau + rA_{xx} + \nu|A|^2 A = 0 \tag{17.56}$$

Since we are interested in the stability of a plane wave, we shall look for solutions in the form

$$u = A(x,t)\exp(ik[x - c_p(k)t]) + c.c. \tag{17.57}$$

The basic state, equivalent to that of Benjamin and Feir, is almost equivalent to taking A_0 equals constant. We must say "almost" because the phase speed $c_p(k)$ that appears in the phase factor is that computed from the *linear* dispersion relation. In reality, nonlinearity will change the frequency of the wave, so we must allow our basic state A, call it A_0, to be a function of time. The $\exp(ikx)$ factor is already exactly what we want in the wave, so we can take A_0 to be independent of space. The NLS equation reduces to

$$iA_\tau + \nu|A|^2 A = 0 \tag{17.58}$$

This equation, defining

$$a \equiv |A_0| : \text{a constant} \tag{17.59}$$

has the plane wave solution

$$A_0 = a\exp(i\delta t), \qquad \delta \equiv \nu a^2 \tag{17.60}$$

Let us now perturb this plane wave by setting

$$A = A_0 + \Delta, \qquad |\Delta| \ll 1 \tag{17.61}$$

Substituting this into the NLS equation (17.56) and linearizing about A_0 gives

$$i\Delta_t + \Delta_{xx} + \nu A_0^2 \Delta^* + 2\nu a^2 \Delta = 0 \tag{17.62}$$

Assume that Δ has the specific form

$$\Delta = A_+ \exp(i\mu x) + A_- \exp(-i\mu x) \tag{17.63}$$

Contrary to what one might reflexively imagine, it is *not* necessary that A_+ and A_- be complex conjugates because the solution of the NLS equation itself is usually *complex-valued*. When we synthesize the solution of the original problem, we automatically include A^* so that u will be real even if A is not. Thus, A_+ and A_- are completely independent quantities. Physically, the assumed form for A consists (of three wavenumbers $k, k+\mu, k-\mu$) plus their complex conjugates. The wave numbers $k \pm \mu$are said to be the "sidebands" of wavenumber k, provided that μ is small compared to k. We shall see that in fact μ must be small for instability.

If we substitute (17.63) into (17.62) and match factors of $\exp(i\mu x)$, we get

$$i\frac{dA_0}{dt} + \delta A_0 = 0 \tag{17.64}$$

$$i\frac{A_+}{dt} - r\mu^2 A_+ + 2\delta A_+ + \delta \exp(2i\delta t) A_-^* = 0 \tag{17.65}$$

$$i\frac{dA_-}{dt} - r\mu^2 A_- + 2\delta A_- + \delta \exp(2i\delta t) A_+^* = 0 \tag{17.66}$$

The mutual interaction of A_+ and A_- will actually excite small forced components with wave numbers $(k \pm 2\mu)$ and so on, but these will be an order of magnitude smaller than the principal sidebands, and thus it suffices to solve the trio of equations above. The first of these, of course, is already been solved by $A_0 = a\exp(i\nu a^2 t)$ but is repeated for completeness.

To solve this coupled set, it is convenient to define

$$\varepsilon = r\mu^2 \tag{17.67}$$

The coupled set becomes

$$i\frac{dA_0}{dt} + \delta A_0 = 0 \tag{17.68}$$

$$i\frac{A_+}{dt} - \varepsilon A_+ + 2\delta A_+ + \delta \exp(2i\delta t) A_-^* = 0$$

$$i\frac{dA_-}{dt} - \varepsilon A_- + 2\delta A_- + \delta \exp(2i\delta t) A_+^* = 0$$

Further simplification follows by taking the complex conjugate of the last equation in (17.68) so that are unknowns are consistently A_+ and A_-^*,

$$-i\frac{dA_-^*}{dt} - \varepsilon A_-^* + 2\delta A_-^* + \delta \exp(-2i\delta t) A_+ = 0 \tag{17.69}$$

To eliminate the pesky factors of $\exp(2i\delta t)$, define

$$A_+ \equiv \exp(2i\delta t)\,\overline{A}_+ \tag{17.70}$$

This reduces (17.68) to a pair of coupled constant coefficient differential equations in A_-^* and $\overline{A}_+$, which can be solved by the usual device [for attacking constant coefficient ODEs] of assuming that both unknowns are proportional to $\exp(i\lambda t)$. The resulting pair of algebraic equations has a nontrivial solution if and only if the determinant of the matrix vanish, i.e.,

$$\det\left(\begin{vmatrix} (-\lambda - \varepsilon) & \delta \\ \delta & (\lambda + 2\delta - \varepsilon) \end{vmatrix}\right) = 0 \tag{17.71}$$

which happens only at zeros of the "characteristic polynomial" of the matrix,

$$\lambda^2 + 2\delta\lambda + \delta^2 + 2\delta\varepsilon - \varepsilon^2 = 0 \tag{17.72}$$

with the solution

$$\lambda = -\delta \pm \sqrt{\varepsilon(\varepsilon - 2\delta)} \tag{17.73}$$

Thus, there is instability whenever (i) $|\varepsilon| < 2|\delta|$ and (ii) the NLS coefficients are of the same sign. Recall that $\varepsilon \equiv r\mu^2$ and $\delta \equiv \nu a^2$ and note that $\text{sign}(r) = \text{sign}(\nu)$ for the focusing case. It follows that ε and δ are of the same sign. (The radical is unchanged if the signs of both ε and δ are simultaneously switched, so it is only their *relative* signs that counts.) Differentiating the radical with respect to ε shows that the maximum growth rate occurs when

$$\varepsilon = \delta \tag{17.74}$$

$$\max(\Im(\lambda)) = \delta \tag{17.75}$$

Translating these results into r, ν, a and μ, we find that

1. The necessary and sufficient condition for instability is that the NLS dispersive and nonlinear coefficients r and ν have the *same sign*.
2. Wavenumber of maximum instability is

$$\mu_{max} = a\sqrt{\frac{\nu}{r}} \tag{17.76}$$

3. Maximum growth rate is

$$\max(\Im(\lambda)) = \nu a^2 \tag{17.77}$$

where a is the amplitude of the fundamental, ν is the Landau constant, r is the coefficient of A_{xx} in the NLS, $\sigma \equiv \nu a^2$, $\varepsilon = r\mu^2$ and μ is the sideband wavenumber.

The most striking fact about the necessary and sufficient condition for instability is its simplicity. For the KdV equation, the NLS equation has ν and r with *opposite* signs, so the KdV wave train is always *stable*. However, the Landau constant has frequent sign changes for finite wavenumber k, so there are wide ranges in parameter space where Rossby waves are modulationally unstable.

However, the result of sideband instability is not chaos. Rather, after the transient appearance and disappearance of localized peaks, the solution returns to a very close approximation of the initial condition. This is called "Fermi-Pasta-Ulam" [FPU] recurrence after those who discovered it in discrete lattice crystal dynamics. Yuen and Ferguson showed that for the NLS equation with periodic boundary conditions, a small amplitude plane wave generates only a single peak whose rise and fall is simply periodic as illustrated in the left panel of Fig. 17.5 [15]. A larger amplitude plane wave destabilizes a larger number of sidebands. As the growth of some sidebands weakens the primary wave, the wavenumber of the most unstable sideband changes,

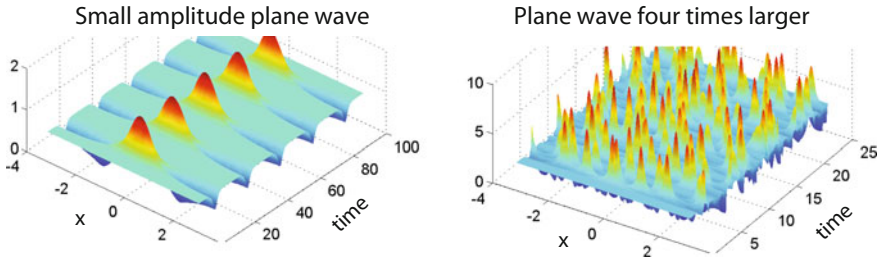

Fig. 17.5 Sideband instability of the uniform plane wave [*green*]. *Left* plane wave amplitude was $1/\sqrt{2}$. *Right* same except that the initial condition was multiplied by a factor of four. For both, the instability was triggered by the perturbation $0.0001\cos(x)$ and the spatial period was 2π

and other sidebands rapidly amplify, leading to a complex, but not chaotic evolution as shown on the right in Fig. 17.5.

The general solution to the NLS equation with periodic boundary conditions can be approximated arbitrarily well by a multivariate hyperelliptic function with a sufficiently large number of periods. It is not necessary to be struck dumb by the apparent impenetrability of hyperelliptic functions and the Nonlinear Fourier Transform. (Although Osborne's book is very readable given its length and complex subject, it is difficult to imagine hyperelliptic functions and integrals being widely taught in American physical oceanography programs.) Classic concepts like linearized instability and soliton dynamics can identify the coherency within the sea of complexity.

17.4 KdV from NLS

The NLS equation is usually derived by perturbation theory from a system of equations far more complicated than the KdV equation. However, if that more complicated system can be approximated in the longwave limit by the KdV equation, then the KdV equation must be consistent with the limit $k \to 0$ of the NLS equation. It follows that if we denote the nonlinear coefficient of the NLS equation by $\nu(k)$, then

$$\nu(k) \sim \frac{a^2}{6bk} \qquad \leftrightarrow a = \sqrt{\lim_{k\to 0}\{6bk\nu(k)\}} \tag{17.78}$$

where a is the coefficient of the nonlinear term of the KdV equation. Thus, the "Landau constant" $\nu(k)$ always has a *simple pole* at $k = 0$ when the longwave ($k \to 0$) limit is described also by the KdV equation.

The dispersive coefficient of the KdV equation must be consistent with the linear dispersion relation for both the underlying system and the KdV equation itself and is therefore always

$$b = -(1/6)\omega_{kkk}(k = 0) \tag{17.79}$$

where ω_{kkk} denotes the third derivative of the frequency with respect to wavenumber. The expression for the nonlinear coefficient of the KdV equation is thus completely determined by the longwave limits of the linear dispersion relation and of $\nu(k)$ and can be rewritten

$$a = \sqrt{-\omega_{kkk}(0) \lim_{k\to 0}\{k\nu(k)\}} \tag{17.80}$$

Thus, if one has a computer program to evaluate $\nu(k)$, then it is not necessary to write separate software or redo the usual KdV derivation to obtain the coefficients of the KdV model. The dispersive and nonlinear coefficients of the KdV equation are always given by (17.79) and (17.80).

Marshall and Boyd [2], who illustrate this principle for equatorially-trapped waves in the ocean, show that something even stronger is true. Wave packets exhibit other kinds of resonances including long wave/short wave resonances and second harmonic resonances. These, too, generate poles in the Landau constant as a function of k, and the coefficients of the "long wave/short wave" system and of the "second harmonic" system fall out from the linear dispersive theory plus the residue of $\nu(k)$ at the resonance pole.

The mathematics of the pole of the Landau constant at $k = 0$ is clear, but what is the physics? The answer is that as $k \to 0$, the phase speeds and group velocities of the second harmonic (wavenumber $= 2k$) and the longwave (wavenumber zero) differ from those of the fundamental (wavenumber $= k$) only by $O(k^2)$. Thus, as $k \to 0$, these three components are increasingly close to resonance. This resonance in turn forces the Landau constant to grow as $k \to 0$.

There is one mildly misleading aspect to the pole: The amplitudes of the higher order terms in the multiple scales series increase inversely with k, too. Thus, the timescale for nonlinear interaction does not decrease to zero as at a true singularity, but rather the perturbation expansion collapses, with higher terms as large as the fundamental, for sufficiently small k.

17.4.1 The Landau Constant: Poles and Resonances

The nonlinear coefficient of the NLS equation is called the "Landau constant" and denoted by $\nu(k; m, n)$. It depends on the longitudinal wavenumber k of the carrier wave plus the latitudinal and vertical mode numbers. In weakly nonlinear perturbation theory, each field is the product of the NLS solution (for the envelope of the wavepacket) multiplied by a sinusoidal function of x and t (the "carrier wave") multiplied by a latitudinal factor and a depth-dependent function which are the same north-south and baroclinic modes as in linear wave theory.

As explained in Marshall and Boyd [2], the Landau constant is the sum of messy integrals that are the products of the usual linear latitudinal and baroclinic modes. The underlying principles are simple and the integrals are nonsingular and therefore are easy numerical quadratures, but the bookkeeping is tedious and best left to their

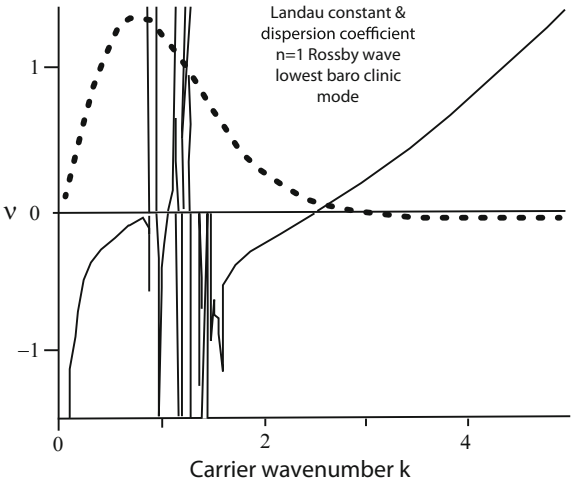

Fig. 17.6 Landau constant versus zonal wavenumber k for the lowest symmetric latitudinal Rossby mode in the first baroclinic mode. Calculated by the methods of Marshall and Boyd [2]. The dispersive coefficient of the NLS equation (*dashed*) is everywhere positive for $k \in [0, 2.7]$ with simple zeros at both endpoints of this interval. The wavepackets change from focusing to defocusing and back again over small intervals in the middle of the wavenumber range illustrated

article. The important point is that $\nu(k)$ has both smooth, nonsingular sign reversals plus many poles, especially for Rossby waves as illustrated by the particular but representative case of the lowest latitudinal, lowest baroclinc Rossby mode in Fig. 17.6.

Each of the Landau poles is associated with a resonance. In the long wave limit, $k \to 0$, for Rossby waves, the KdV equation applies as shown above and there is always a pole. The reason is that Rossby waves become nondispersive as $k \to 0$ and thus every Rossby longitudinal wavenumber is in resonance with every other for a given latitudinal and vertical mode number. Each long/wave short wave resonance, second harmonic resonance, and triad resonance is also a pole.

It is unfortunate that the parameter space is so chopped up by resonances. The behavior of Rossby wavepackets will not be simple. It also behooves us to thoroughly understand each species of resonance, and we shall now discuss each in turn.

17.5 Weakly Dispersive Waves

The failure of the NLS perturbation theory as $k \to 0$ is rather disconcerting because the KdV equation is usually derived as a longwave approximation. Thus, there is only a narrow range of small k where the wavenumber of the carrier wave of the wavepacket is sufficiently small that the KdV equation is a good model, and simul-

taneously so large that the NLS model is accurate. The NLS/KdV connection would seem to be much ado about a very narrow parameter range.

However, the equatorial Kelvin wave in the ocean, which was the target of our investigation [16], is quite different. In the absence of a mean flow, the Kelvin wave is *nondispersive*. Latitudinal variations in the mean east-west ocean currents induce a weak dispersion in the Kelvin mode, but to second order in the strength of the shear, the dispersion relation is quadratic in the east-west wavenumber. It follows that the third derivative term in the KdV equation captures the correct dispersive behavior for the Kelvin wave even when the wavenumber is *not* small.

One consequence is that KdV-type Kelvin solitary waves are not restricted to long zonal scales (highly anisotropic solitons, much wider east-west than north-south); the KdV equation is a good approximation in describing *"round"* solitons of *equal longitudinal and latitudinal width*. A second consequence is that oceanic equatorial Kelvin wavepackets are captured for the entire interesting range of parameters by deriving the NLS equation from the corresponding KdV approximation.

Similar comments apply to other species of waves. When the waves are nearly nondispersive, the KdV approximation will apply to moderate as well as long length scales. The NLS approximation derived from the KdV equation will similarly have a wide range of applicability.

17.6 Numerical Experiments

Chen and Boyd [17] carried out interesting experiments which were repeated and expanded by the second author in recent months. Although many variations in amplitude and width of the envelope were tried, the discussion will concentrate on one representative case which evolved from the initial condition

$$u(x,0) = 6\mu\,\mathrm{sech}(\lambda x)\cos(x+\pi/3) \tag{17.81}$$

where the nonlinearity and dispersion parameters are $\mu = 1/10$ and $\lambda = 1/10$. The domain has periodic boundary conditions where the period $P = 128\,\pi$. The Generalized-NLS with quintic nonlinearity (17.33) and also the KdV equation were solved by a Fourier pseudospectral method with 1024 grid points, combined with a fourth order Runge–Kutta time marching method with a time step of 1/400 on the interval $t \in [0, 200]$. The high spatial and temporal resolution for the Generalized-NLS equation is, except for ease of programming, quite unnecessary; the whole point is that the envelope $A(\zeta, \tau)$ varies *slowly* with both space and time.

Figure 17.7 compares the initial and final wavepackets, showing that the amplitude has decreased to almost exactly one-third of its initial value. The width has increased by roughly the square of the reciprocal of the amplitude decrease, approximately a factor of nine. The packet has become so wide that, thanks to the periodic boundary conditions, packets centered on neighboring periods are now overlapping. The scalloping or wiggliness of the envelopes near the edges of the periodicity interval

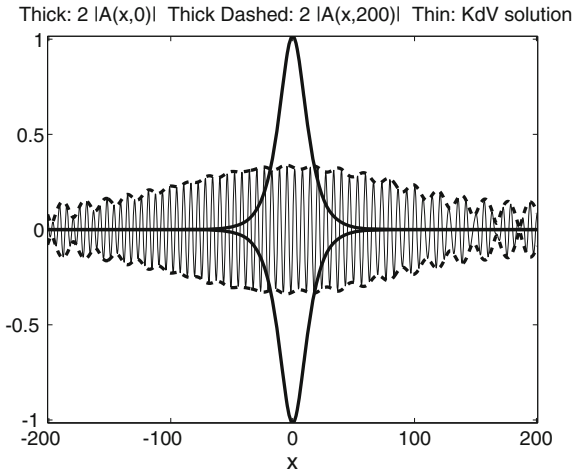

Fig. 17.7 The *tall solid curve* is the initial envelope. (Because the solution is $A\exp(ikx) + A^*\exp(-ikx)$, we show twice $|A(x,0)|$.) The *thick dashed curve* is the envelope at $t = 200$. The *thin curve* is the KdV solution at $t = 200$. The GNLS-predicted solution was also plotted, but is graphically indistinguishable from the exact solution. All plots were scaled by dividing by the maximum initial value of the KdV solution

at $t = 200$ is the spoor of this overlap-of-packets; we have done many experiments with both the GNLS and standard NLS equations, and this wiggliness-at-the-edges always seems to develop when the packet widens beyond the edges of the periodicity interval.

Figure 17.8 shows the evolution in the errors of the Fourier coefficients with time. Tracking the errors in the envelope cannot be done by comparing it to the numerical KdV solution because the latter is the product of the envelope of the carrier wave plus all the higher harmonics. Instead, the solution of the quintic Generalized NLS equation was used as the "exact" solution to approximate the errors in the perturbative solutions at each order. Zeroth order included only $A(\zeta, \tau)$, first order added A^{10} and A^{12}, and so on. Clearly, the method of multiple scales has been successful with good uniformity in time.

Figure 17.9 compares the initial and final Fourier coefficients with the predicted Fourier coefficients. The predicted fourth harmonic (rightmost dashed bump) is an order of magnitude too small, but all the higher harmonics have decreased considerably over time both in the perturbative prediction and in the true KdV solution. The very dramatic narrowing of the longwave component, which is the leftmost bump, is also predicted with quantitative accuracy by perturbation theory.

Figure 17.10 compares the absolute values of the final coefficients of the KdV solution with the errors in absolute value of the third order perturbative prediction. Another intriguing surprise is that the *absolute* errors in Fourier coefficients are *largest* for the *fundamental* (that is, in the wavenumber range $k/2$ to $(3/2)k$ where here $k = 1$). Because of the large size of the fundamental, this is also the range where the *relative* error is smallest. The *relative* errors are about 1 part in 10,000. For the fourth harmonic and above, the relative error is very large. Overall, though, the perturbation theory does well for this case where the perturbation parameters are 1/10.

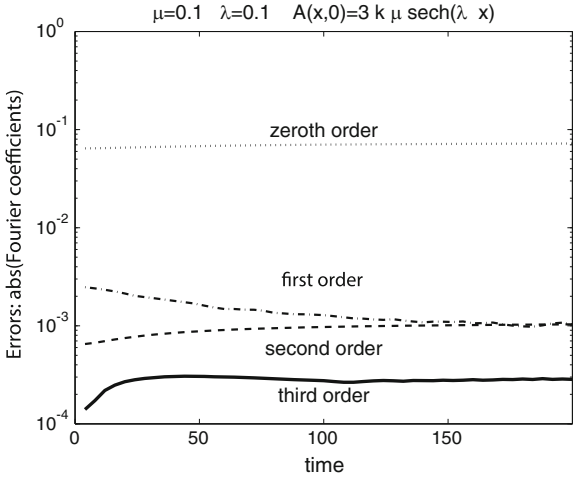

Fig. 17.8 The curves are the maximum differences in the absolute values of the Fourier coefficients of u_{KdV} and $u^{(j)}$ where j is the perturbative order. The errors have been scaled by dividing by the maximum Fourier coefficient at each time

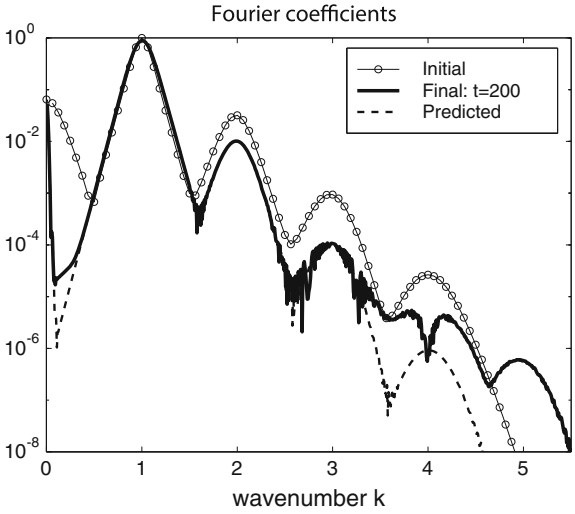

Fig. 17.9 The absolute values of the initial and final Fourier coefficients are the thin line with circles and the thick solid line, respectively. The GNLS-predicted coefficients are also shown as the *dashed curve*, but are difficult to distinguish from the exact coefficients except under the fourth and fifth harmonics (wavenumber > 3.5)

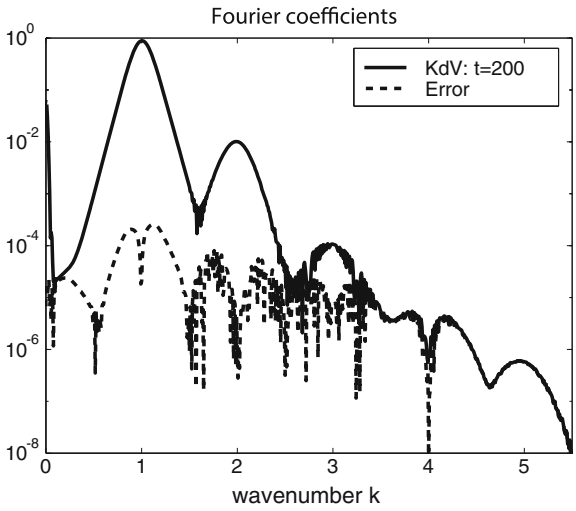

Fig. 17.10 The *thick solid line* is the absolute value of the Fourier coefficients of the KdV solution at the final time, just as in the previous figure. *The lower curve* (*dashed*) is the difference between the absolute values of the coefficients of the KdV solution and the prediction of third order perturbation theory. Note that for the fourth harmonic, $k \in [-3.5, 4.5]$, and larger wavenumber, the error merges with the coefficient curve so that the relative error for the fourth and higher harmonics is close to 100%

Although the dramatic narrowing (in wavenumber space) of the longwave component is well-captured by the perturbation theory, we have not been able to arrive at a simple heuristic explanation. Perhaps the reader will be more ingenious!

Similarly, we do not understand why the fourth harmonic is an order of magnitude larger than predicted by perturbation theory, nor why it dips to near-zero amplitude at a wavenumber of $4k$, right in the center of the harmonic.

We performed the usual experiments, halving the timestep and also independently doubling the width of the spatial domain, but these puzzling features were unaltered.

17.7 Nonlinear Schroedinger Equation (NLS) Summary

We have shown the following:

- The NLS equation derived from the KdV equation is always the *defocusing* case; envelope solitary waves cannot form in the KdV solution.
- NLS perturbation theory is here explicitly extended to two orders beyond the NLS equation itself [95].
- The Generalized-NLS (GNLS) equation, which describes the evolution of the envelope, is quintically nonlinear and third order in space.

- A special quirk of the KdV equation is that the Generalized-NLS, when carried to one order beyond the usual NLS, captures the *linear* dispersion of the KdV equation exactly. For packets of finite amplitude, however, there is some small error because the nonlinear terms cannot exactly represent the nonlinearity of the KdV equation.
- The nonlinear coefficient of the NLS equation has a pole in k as the wavenumber k of the carrier wave tends to zero.
- The residue of this pole gives an explicit expression for the nonlinear coefficient of the KdV equation and vice-versa.
- When the underlying waves are only weakly dispersive, as true of equatorial Kelvin waves in the ocean, the KdV equation is an accurate model even for waves of moderate or short length scale. The NLS-from-KdV theory then has a similarly large region of validity.
- Our numerical experiments show good agreement between the NLS/GNLS perturbation theory and the exact KdV wavepacket for small and moderate times.

17.8 Resonances: Triad, Second Harmonic and Long-Wave Short Wave

Although the triad interaction is a basic paradigm of modern fluid dynamics, multiple rationales have appeared in the literature, and none is entirely satisfactory. One useful perspective begins with a Hough–Hermite discretization of the nonlinear shallow water equations, that is, a Galerkin spectral method using the normal modes (Hough functions) of the linear shallow water equations as the basis functions. Ripa [18] and Tribbia [19] show that a normal mode spatial discretization of a quadratically nonlinear system of partial differential equations will yield a big system of quadratically nonlinear ordinary differential equations in time [18, 19]. If the zonal factors of the basis function are of the form $a_k(t)\exp(ikx)$ plus its complex cojugate $a^*_{-k}\exp(-ix)$, the product of a_1 with a_2 will appear in the equation for da_3/dt if and only if $k_1 + k_2 = \pm k_3$ where the sign ambiguity arises because each basis function is accompanied by its complex conjugate. However, the dependence on depth and latitude implies that each pair of interacting longitudinal wavenumbers is a double infinity of three-dimensional basis functions, and the wavenumber k may be any real number.

Thus the dynamics can always be described as an infinite number of triad interactions in the sense that the time derivative of the coefficient $a_n(t)$ of the n-th normal mode (here, a Hough–Hermite vector of functions) is driven by a sum of terms each the product of a numerical "interaction coefficient" I_{njk} with the product of two other coefficients, $a_j a_k$.

Thus, a trio of wavepackets will exchange energy between themselves. In the absence of resonance, constructive interference will turn to destructive interference

Table 17.1 Resonances: selected references

Reference	Remark
Boyd [20]	Triads involving long Rossby waves
Boyd [8]	Derivation of NLS equation, which admits envelope solitons and second harmonic and long wave/short wave resonances
Boyd [21]	Long wave/short wave resonance
Boyd [17]	Second harmonic resonance
Boyd [22]	Radiative decay of nonlocal Rossby solitons
Boyd [23]	Review of nonlinear equatorial waves
Boyd [24]	4-mode model
Boyd [25]	Numerical experiments: tilted modons and nonlocal soliton radiation
Greatbatch [26]	Kelvin wave fronts and nonlinear spin-up
LeSommer, Reznik and Zeitlin [27]	Numerical study of response to localized perturbation
Domaracki and Loesch [28]	Triad resonance; triad wave packet PDEs
Loesch and Deininger [29]	Systems of interlocking, interacting triads
Marshall and Boyd [2]	Nonlinear equatorial waves in 3D continuously-stratified ocean
Madjaa, Rosales Tabak and Turner [30]	Topographic resonance coupling Kelvin and other equatorial waves
Raupp, Silva Dias [31]	Atmospheric waves
Raupp, Silva Dias, Tabak and Milewski [32]	Atmospheric triads
Bates and Grimshaw [33]	Resonances between barotropic and baroclinc modes

in a cyclic dance. The changes of one half-period are reversed on the next half-period. If the wave amplitude are small, the exchanges remain small, bounded and boring.

If the triad of waves satisfy resonance conditions, this simply means that the wavepackets always remain in phase as they propagate. Even tiny waves can be drastically reshaped by nonlinearity through these nonperiodic exchanges. Astronomers call such steadily accumulating nonperiodic changes and exchanges "secular" effects. Articles on equatorial resonances are collected in Table 17.1.

It is impossible to analyze such a plethora of terms; how then can one justify truncating this PDE to a trio of ordinary differential equations in time to three equations in the three unknowns?

Lorenz' answer was his Principle of "Maximum Simplification", which is that much may be learned from studying the most drastic truncation of a spectral model that has nontrivial physics, even if not qualitatively realistic. His 1960 article solved the triad equations [34] because a trio of mutually interacting waves is the simplest Low Order Spectral Truncation (LOST) that cycles energy between modes.

Bretherton [35] and Domaracki and Loesch [28] use multiple scales perturbation theory to justify the retention of one triplet of waves. In this framework, interactions

of the resonant waves remain perpetually in phase whereas nonresonant interactions are periodic in time. For very tiny amplitudes, only the resonant, secular interactions are significant. The logic is consistent, but not well-attuned to reality. Energy may leak from mode to mode, even without resonances, like water seeping from compartment to compartment of a torpedoed ship, whose shockwave-sprung bulkheads are no longer watertight. In practical terms, Lorenz' type thinking must underlie the relevance of the triad model.

Multiple scales perturbation theory yields an integrable trio of partial differential equations that describes spatially localized wave packets traveling at the linear group velocity (altered slightly by nonlinear corrections) of each packet [28]. Bers, Kaup and Reiman [36], Kaup [37] and Martin and Segur [38] provide a comprehensive introduction. Here we shall follow the usual practice and discuss only disturbances with spatially uniform envelopes.

What falls out of the theory is a six degree of freedom system that contains one wavenumber as a parameter plus a "detuning parameter" δ that measures closeness to resonance. Ripa [18] solves the general case, but the complex-valued solutions of his study are not enlightening. Following standard practice, we shall restrict further attention to the system of three real-valued equations in three unknowns in which the phases of the waves are perfectly in phase so as to maximmize the strength of the interaction. Little generality is lost because both "imperfect resonance" ($w_1 + w_2 = \pm w3 + \delta, \delta \neq 0$), and imperfect phase matching weaken the wave interactions without changing the qualitative behavior. We shall refer to the resulting canonical system as "Triply-Simplified" because, to summarize, we impose three specializations:

1. Plane waves [rather than wave packets]
2. Resonance is perfect [detuning parameter/s $\delta = 0$]
3. The phases of the (ever-oscillating) triad components are such that the unknowns are all-real-valued and the rate of energy transfer is maximized.

Recall that, with variables nondimensionalized so as to eliminate Lamb's parameter ε, the dispersion relation for equatorial waves of zonal wavenumber k and latitudinal mode number n, linearized abut a state of rest, is

$$w^3 - (k^2 + 2n + 1)w - k = 0 \tag{17.82}$$

Including both positive and negative k and w, the resonance conditions are

$$w_1 + w_2 + w_3 = 0 \tag{17.83}$$

$$k_1 + k_2 + k_3 = 0 \tag{17.84}$$

The approximate dispersion relations from Chap. 3 can be invoked to simplify a rather complicated algebraic problem:

$$w = \frac{-k}{k^2 + 2n + 1}\{1 + \Delta\}, \qquad |\Delta| \leq \frac{0.039}{(n + 1/2)^2}, \qquad n \geq 1, \text{ [Rossby]}$$

$$w = \sqrt{k^2 + 2n + 1}\,\{1 + \Delta\}\,, \qquad |\Delta| \le \frac{0.141}{n + 1/2} \quad \text{[gravity waves]} \tag{17.85}$$

or the improved approximation

$$w = \pm\sqrt{k^2 + 2n + 1}\left\{1 \pm \frac{k}{(2\,(k^2 + 2n + 1)^{3/2}}\right\}\{1 + \Delta\} \tag{17.86}$$

$$|\Delta| \le \frac{0.021}{(n + 1/2)^2} \qquad \text{[improved gravity waves]} \tag{17.87}$$

The system consists of three copies of the dispersion relation plus the two resonance conditions. We can specify one wavenumber and all three latitudinal mode numbers arbitrarily and then have five equations in five unknowns. A detailed analysis of the resonance conditions for general triads is given in Sect. 17.13.

Second harmonic resonance and long wave/short wave resonance are special cases in which two of the three waves are identical. The nonlinear coefficient ("Landau constant") of the NLS equation is singular at both types of resonances.

17.9 Second Harmonic Resonance

Definition 17.1 (*Second Harmonic Resonance*) Second Harmonic Resonance is a special case of triad resonance in which two of the three waves are identical ("fundamental" mode of wavenumber k) while the third is the "second harmonic" of the "fundamental", that is, the wavenumber of the third component is $2k$. For equatorial waves of mode numbers m and n, the condition for second harmonic resonance is

$$w(2k, m) = 2w(k, n) \tag{17.88}$$

where n is the latitudinal mode number of the fundamental and m is the mode number of the second harmonic.

Second harmonic resonance has a two-fold significance. First, this resonance always transfers energy monotonically (in time) and irreversibly from the fundamental to the second harmonic, and is thus a mechanism of long-to-short-wave cascade. Second, the nonlinear coupling coefficient ("Landau constant") in the Nonlinear Schroedinger equation has a pole at each second harmonic resonance.

Equatorial second harmonic resonance for the nonlinear shallow water beta-plane model has been studied by Loesch, Domaracki and Deininger [28, 29], Ripa [39–42], Boyd [8, 17] and Marshall and Boyd [2]. Denoting the frequency and wavenumber of the fundamental by w and k, the resonance conditions are

$$w^3 - (k_2^2 n + 1)w - k = 0 \tag{17.89}$$

$$8w^3 - (8k^2 + 4m + 2)w - 2k = 0 \tag{17.90}$$

This system can be solved exactly as explained in [17]. The sole non-trivial real root is

$$k = \sqrt{\left[\frac{2}{3}(n-m)\right]\left[\frac{1}{1-9/q^2}\right]} \tag{17.91}$$

where q is the parameter

$$q \equiv 4(2n+1) - (2m+1) \tag{17.92}$$

Detailed analysis shows the following:

1. The latitudinal mode number m of the resonant second harmonic is *always odd*.
2. Resonance is possible only between (a) two Rossby waves or
   (b) two westward-traveling gravity waves
3. Rossby–Rossby resonance is

$$k = \sqrt{\left[\frac{2}{3}(n-m)\right]\left[\frac{1}{1-9/\{4(2n+1)-(2m+1)\}^2}\right]} \tag{17.93}$$

$$\approx \sqrt{\frac{2}{3}(n-m)} \tag{17.94}$$

4. Westward Gravity–Westward Gravity resonance is

$$m = 4n + 1 \tag{17.95}$$

$$k = \frac{1}{2}\sqrt{n + 1/3} \tag{17.96}$$

5. The gravity waves resonance formulas for $n = 0$ generate the mixed Rossby-gravity wave results for $n = 0$.

When the resonant waves are sine waves or wavepackets with very broad envelopes, the triad ODEs collapse to the pair

$$\frac{dA}{dt} = -I\,A\,H, \qquad \& \qquad \frac{dH}{dt} = I\,A^2 \tag{17.97}$$

where $A(t)$ and $H(t)$ are the amplitudes of the fundamental and second harmonic, respectively. The *general* solution is

$$A(t) = a\,\mathrm{sech}(aI(t-t_0)), \qquad \& \qquad H(t) = -a\tanh(aI(t-t_0)) \tag{17.98}$$

where a and t_0 are arbitrary constants, the constants of integration needed to fit the two initial conditions.

Irrespective of these, the *second harmonic always drains all the energy monotonically from the fundamental*.

17.9.1 Landau Constant Poles

If the n-th latitudinal mode Rossby wave or westward-propagating gravity wave becomes resonant with its own second harmonic at wavenumber $k = k_{res}$, then the Nonlinear Schroedinger model that approximates the evolution of the envelope of a wavepacket becomes a busted, inaccurate model. This breakdown manifests itself as a simple pole in the nonlinear coefficient, the "Landau constant", at the resonant wavenumber:

$$\nu(k) \approx \frac{R}{k - k_{res}}, \qquad R = \frac{I^2}{2\left\{c_g(2k, m) - c_g(k, n)\right\}} \tag{17.99}$$

17.9.2 Barotropic/Baroclinic Triads

On the equatorial beta-plane, the barotropic mode is not trapped latitudinally. However, the usual one-and-a-layer model has no coupling to the barotropic mode. When the stratification is continuous and the nonlinear terms are projected onto the vertical eigenmodes, energy injected into the barotropic mode is not confined to the equator, but instead is irreversibly lost to extra-tropical latitudes.

Boyd showed that on a spherical earth, equatorial trapping is proportional to $\varepsilon + s^2$ where $s = 0, 1, 2 \ldots$ is the integral zonal wavenumber and not merely ε alone as on the standard equatorial beta-plane; it is possible for barotropic ($\varepsilon = 0$) waves to be equatorially trapped if the zonal wavenumber is large, an effect completely missed by the classic equatorial beta-plane (see Sect. 3.12).

Bates and Grimshaw devised an "extended equatorial beta-plane" which retains this equatorial trapping due to the zonal wavenumber [33]. This allows them to analyze resonant coupling between baroclinic and barotropic modes that are both equatorial.

Although motivated by atmospheric applications and therefore slightly outside the scope of this work, it would be remiss not to note similar barotropic/baroclinic equatorial wave studies by Biello and Majda [43–45].

17.10 Long Wave/Short Wave Resonance

Definition 17.2 (*Long Wave/Short Wave Resonance*) A packet of ultralong Rossby waves is resonant with a packet of short westward-traveling Rossby waves or short westward-traveling gravity waves when the resonance condition

$$c_p(0, m) = c_g(k, n) \tag{17.100}$$

is satisfied where the ultralong Rossby wave is of latitudinal mode number m and $c_p(0, m)$ is its phase velocity in the zero wavenumber limit and the short wavepacket has carrier wavenumber k and latitudinal mode number n. "Long" and "short" refer to zonal wavenumber; the ultralong Rossby packet has a carrier wavenumber of zero, that is, its carrier wave is the trivial function $\mathfrak{C} \equiv 1$.

Recall that $c_p(0, m)$ is also the group velocity since Rossby waves are nondispersive in the $k = 0$ limit, and the group and phase velocities are identical for nondispersive waves; the resonance condition could thus be equivalently stated as an equality of group velocities. The resonance is then obvious: it demands that in linear dynamics, the short and long wavepackets must travel westward at the same (group velocity) speed.

Their nonlinear dynamics is described by

$$i\, A_t + (1/2)\omega_{kk} A_{xx} = \nu_L A\, B \tag{17.101}$$

$$B_t = -\mu\, (|A|^2)_{xx} \tag{17.102}$$

where A, which is usually complex-valued, is the amplitude of the short wave packet and B, which is real-valued, is the longitude-and-time dependence of the long Rossby wave group. The second wavenumber derivative of $\omega(k)$ is denoted by $\omega_{kk}(k)$. Note that this system is written in the coordinate system moving at the common speed of the two wavepackets.

Boyd [46] provides a full derivation of these equations on the equatorial beta-plane. The Long Wave/Short Wave equations are an integrable system that can be solved by the inverse scattering method (Benney [47], Ma [48] and Ma and Redekopp [49].)

Boyd's study noted that

1. The latitudinal mode number m of the long Rossby waves is always odd, or equivalently, the long wave mode must always be symmetric with respect to the equator.
2. From the usual Rossby wave dispersion relation, it follows that $c_p(0; n) = -1/(2n + 1)$
3. For a given pair of latitudinal mode numbers (m, n), the carrier wave number of the resonant short waves, if Rossby waves, is

$$k^2 = -(2n + 1) - \left(m + \frac{1}{2}\right)\left(-1 + \sqrt{2m + 16n + 9}\right) \quad \text{[Rossby short wavepacket]}$$

where it is necessary for a Rossby–Rossby resonance that $m > n$ and m is odd.

4. If the short wave packets are westward-traveling gravity waves, the resonant carrier wave number is

$$k = \frac{\sqrt{2n+1}}{2m+1} + \frac{1}{2\sqrt{2n+1}} \tag{17.103}$$

where again the long Rossby wave must be symmetric with respect to the equator (m odd.)

Ma and Redekopp [49] and earlier authors they cite show that the LSR equations have the following properties:

1. Spatially localized solitons which are the union of an envelope soliton traveling in concert with a long Rossby bell soliton.
2. Coherent structures with shortwave envelopes that are uniform except for a spatially localized "hole" at the location of a KdV-type long wave soliton.
3. Phase-jump solitons where, for the short wave packet, only the transition in phase is spatially localized; this transition always travels in resonance with a KdV-type solitary wave.
4. The spatially uniform short wave envelope is *always unstable*; it breaks up into spatially localized solitary waves.
5. On the spatially periodic domain, Fermi-Pasta-Ulam (FPU) recurrence periodically cycles the disturbance from uniform envelope to a train of solitary waves and back again, endlessly.
6. An initial condition which is purely a packet of short waves is always unstable with respect to the long waves which will grow linearly with time; the reverse is not true.
7. A pulse of ultralong Rossby waves will not trigger unstable growth of the short wavepacket, but will merely produce oscillations in the phase of the short wave wavepacket.

The one-sided character of instability — the short wave is unstable while the long wave disturbance is not — is consistent with general resonant triad theory. Domaracki and Loesch [28] observed that it is always the wave of *highest frequency* which is *unstable*. Recalling that the long Rossby waves of mode number have a phase speed $c = -1/(2m+1)$ and that for all wave species $\omega = kc$, it follows that the limit $k \to 0$ is also a zero frequency limit. Thus, the long Rossby disturbance will never be the highest frequency part of a long wave/short wave resonance.

Boyd [21] gives extensive tables of the resonances.

There has been no identification of long wave/short wave resonance in either observations or numerical models. However, this type of resonance has been observed in many non-geophysical systems.

17.10.1 Landau Constant Poles

Due to long wave/short wave resonance, the NLS Landau constant of the NLS equation for latitudinal mode n will have a pole of the form:

$$\nu^{NLS} = -\frac{\nu_L}{\omega_{kk}}\mu\frac{1}{k - k_{res}} \tag{17.104}$$

17.11 Triad Resonances: The General Case Continued

17.11.1 A Brief Catalog of Triad Concepts

To return to the general triads, it is useful to summarize some important properties.

1. The resonance conditions are a system of polynomial equations that can be reduced to a single equation in a single unknown of at most sixth degree.
2. Each combination of three latitudinal mode numbers plus the single continuous parameter, which by convention we take as one of the wavenumbers, generates as many as eighteen solutions.
3. All species and all mode numbers can participate in one or more triads.
4. Modes can only interact as a triad if the sum of their latitudinal mode numbers is ODD.
5. The interaction coefficients of a triad are the coupling constants of a Hough–Hermite Galerkin discretization. These are given explicitly in the appendix of [28]; the analytical formulas are extremely lengthy.
6. Ripa [18, 39] analyzes various "non-local" resonances, that is, resonances that couple long waves (small zonal wavenumber) to triad modes of large zonal wavenumber.
7. One type is "Bragg scattering" in which a short gravity wave is scattered into a short gravity wave traveling in the opposite direction through the catalysis of a long Rossby wave.
8. Many triads overlap in the sense that modes participate in more than triad. It is therefore necessary to analyze chains of ten to one hundred modes as illustrated in Fig. 1 of [29].
9. Most of the energy in these chains remains in the modes of lowest latitudinal wave number such as the Kelvin wave, Yanai mode and $n = 1$ Rossby wave.
10. In midlatitude resonances, the unstable mode whose interaction coefficient is different from the other two (and therefore the unstable mode) is the mode of intermediate total wavenumber; Ripa claims for equatorial triads that the unstable mode is always the mode of intermediate "slowness" [the reciprocal of the phase speed].
11. Multiple scales perturbation theory yields a trio of partial differential equations that are a completely integrable system (solvable by the inverse scattering

method) for the wavepackets traveling at the nonlinear-modified group velocity appropriate to each latitudinal mode, provided that the wavenumbers and frequencies of the carrier waves satisfy the resonance conditions.

12. When the triad components are plane waves and also the resonances are exact (as opposed to *near*-resonances with non-zero detuning parameters) and lastly the phases of the three modes are in phase for maximum energy exchange, the amplitudes of the waves satisfy a trio of three quadratically-nonlinear ODEs in time that (i) can be solved explicitly for all initial and parameter values in terms of Jacobian elliptic functions and (ii) are identical with Euler's equations for the three angles of rotation about the orthogonal principal axes of a freely-falling body.

We amplify on some of these topics in what follows.

17.11.2 Rescalings

Theorem 17.2 (Triply-Restricted Triad Equations) *In their most general, phase-locked perfectly resonant form, the triad equations are*

$$\begin{aligned} \frac{dA_1}{dt} &= -\, I_1\, a_2\, a_3 \\ \frac{dA_2}{dt} &= I_2\, a_1\, a_3 \\ \frac{dA_3}{dt} &= -I_3\, a_1\, a_2 \end{aligned} \tag{17.105}$$

where the constant I_j are all the same sign. Without loss of generality, we assign the mode whose interaction coefficient is different in sign from the other two to be "mode 2".

Through the change of unknowns

$$A_1 = \frac{1}{\sqrt{-\, I_2\, I_3}} a_1 \tag{17.106}$$

$$A_2 = \frac{1}{\sqrt{I_1\, I_3}} a_2 \tag{17.107}$$

$$A_3 = \frac{1}{\sqrt{-\, I_1\, I_2}} a_3, \tag{17.108}$$

the triad equations are reduced to the canonical form

$$\frac{da_1}{dt} = -\,a_2\,a_3$$
$$\frac{da_2}{dt} = a_1\,a_3$$
$$\frac{da_3}{dt} = -\,a_1\,a_2 \tag{17.109}$$

Theorem 17.3 (Triad Invariants) *1. If* $g(t)(= [a_1(t), a_2(t), a_3(t)])$ *is a solution to the canonical triad equations, then so is*

$$\lambda g(\lambda[t + \varphi]) \tag{17.110}$$

for arbitrary translation [phase shift] φ *and dilation* λ*, both constants independent of time. In other words, the triad solutions are invariant with respect to translation and dilation.*

2. The triad conserves both energy and enstrophy. However, it is more convenient to work with new invariants that are linear combinations of the energy and enstrophy. These convenient invariants are

$$F_1 \equiv a_1^2(t) + a_2^2(t) \tag{17.111}$$

$$F_2 \equiv a_1^2(t) + a_3^2(t) \tag{17.112}$$

Proof For proposition one, the basic idea is to substitute (17.110) into the triad equations and cancel out all factors of λ and φ see [23]. The invariance of F_1 follows from multiplying the first triad equation by a_1, the second by a_2 and adding. The left-hand side is

$$a_1\frac{da_1}{dt} + a_2\frac{da_2}{dt} = \frac{1}{2}\left(\frac{d(a_1^2 + a_2^2)}{dt}\right) \tag{17.113}$$
$$= \frac{1}{2}\frac{dF_1}{dt} \tag{17.114}$$

The right-hand side is

$$-\,a_1a_2a_3 - (-a_1a_2a_3) = 0 \tag{17.115}$$

It follows that $dF_1/dt = 0$ as claimed, and similarly for the time-independence of F_2. ■

The triad equations conserve both energy E and enstrophy F. Surfaces of each are triaxial ellipsoids in the three-dimensional space spanned by the triad unknowns (modal amplitudes) $\{a_1, a_2, a_3\}$ where "triaxial" means that all three axes are in general unequal. Linear combinations of E and F give new invariants F_1 and F_2 which

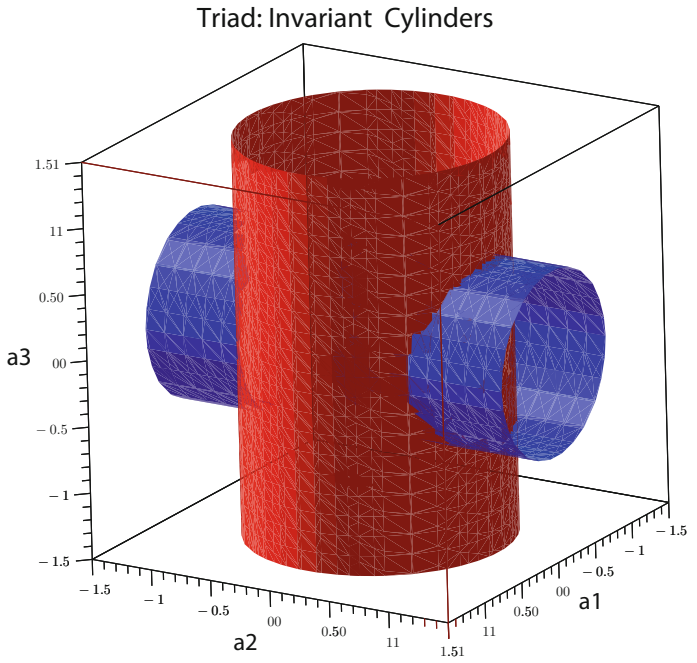

Fig. 17.11 The initial conditions determine two infinitely long cylinders which are invariant surfaces in the sense that the trajectory must always lie on both surfaces so that energy and enstrophy are conserved. This implies that all trajectories are periodic since the trajectory must lie on the closed curve where the invariant cylinders intersect

depend on only two coordinates and are therefore easier to visualize. As illustrated in Fig. 17.11, the isosurfaces of the new invariants F_1 and F_2 are infinitely long cylinders parallel to one of the coordinate axes; these can be regarded as ellipsoids in which one axis is infinitely long.

Once the initial conditions have placed the trajectory on a particular cylinder, it must remain on that cylinder forever. To be on both cylinders simultaneously, the trajectory must lie wholly on the closed curve which is the intersection of the two cylinders. This curve may always be parameterized in principle by time. The existence of the two invariants is sufficient to show that the solution is always periodic. Note, however, that trajectories with infinite periods are allowed.

17.11.3 The General Explicit Solutions

Theorem 17.4 (Explicit Solutions to the Triply-Simplified Triad Equations) *First, adopt the convention that mode two is the one whose interaction coefficient is opposite in sign to the other two where the "interaction coefficient" for mode* j *is* $da_j/dt/\prod_{k=1,k\neq j}^{3} a_k$. *Assign mode one to be the larger of the two other modes at*

$t = 0$, that is, adopt the convention that $|a_1(0)| \geq |a_3(0)|$. The general solution to

$$\begin{aligned} \frac{da_1}{dt} &= -\, a_2\, a_3 \\ \frac{da_2}{dt} &= a_1\, a_3 \\ \frac{da_3}{dt} &= -\, a_1\, a_2 \end{aligned} \tag{17.116}$$

can be expressed in terms of elliptic functions with elliptic modulus

$$m_{ell} = \frac{a_2(0)^2 + a_3(0)^2}{a_2(0)^2 + a_1(0)^2} \tag{17.117}$$

as

$$\begin{aligned} a_1 &= \Lambda\, \frac{2K}{\pi} dn(\Lambda\, \frac{2K}{\pi}\, (t+\phi); m_{ell}) \\ &= \Lambda\, \left\{ 1 + 4 \sum_{n=1}^{\infty} \frac{q^n}{1+q^{2n}}\, \cos\left(2n\Lambda(t+\phi)\right) \right\} \end{aligned} \tag{17.118}$$

$$\begin{aligned} a_2 &= \sqrt{m_{ell}}\, \Lambda\, \frac{2K}{\pi}\, sn(\Lambda\, \frac{2K}{\pi}\, (t+\phi); m_{ell}) \\ &= \Lambda\, 4 \sum_{n=1}^{\infty} \frac{q^{n-1/2}}{1-q^{2n-1}}\, \sin\left((2n-1)\Lambda(t+\phi)\right) \end{aligned} \tag{17.119}$$

$$a_3 = \sqrt{m_{ell}}\, \Lambda\, \frac{2K}{\pi}\, cn(\Lambda\, \frac{2K}{\pi}(t+\phi); m_{ell}) \tag{17.120}$$

$$= \Lambda\, 4 \sum_{n=1}^{\infty} \frac{q^{n-1/2}}{1+q^{2n-1}}\, \cos\left((2n-1)\Lambda(t+\phi)\right) \tag{17.121}$$

where

$$\Lambda = \frac{\pi}{2\, K(m_{ell})} \sqrt{\frac{a_2(0)^2 + a_3(0)^3}{m_{ell}}} \tag{17.122}$$

$$\phi = \frac{1}{\Lambda}\, \frac{\pi}{2\, K(m_{ell})}\, dn^{-1}\left(\frac{a_1(0)}{\Lambda}\, \frac{\pi}{2K(m_{ell})}\, ; m_{ell} \right) \tag{17.123}$$

The solution is periodic *in time with the period*

$$P = \frac{2\pi}{\Lambda} \tag{17.124}$$

where $K(m_{ell})$ is the complete elliptic integral of the first kind.

Note: $sn(u, m_{ell})$ and $cn(u, m_{ell})$ are periodic with period $4K$ in u and 2π in y where $u = 2K(m_{ell})/\pi$. The function dn is periodic with period $2K$ in u and π in y.

Proof Application of the standard elliptic function derivative identities. The identity $\mathrm{sn}^2 + \mathrm{cn}^2 = 1$, which is true for all t and all m_{ell} is used to obtain the formula for Λ. The phase ϕ follows from evaluating $a_1(0)$ and then solving for ϕ. ■

To impose general initial conditions, a three-step procedure is necessary. First, determine the elliptic modulus from $m_{ell} = (a_2(0)^2 + a_3(0)^2)/(a_2(0)^2 + a_1(0)^2)$.

Second, define $\alpha_j(t) \equiv a_j(t; m_{ell}, \Lambda = 1)$. Replace t by $t + \phi$, and increase the phase ϕ in small steps until $\alpha_1(\pi)/\alpha_2(\phi)$ in the formulas given above match the ratio in the desired initial conditions, $a_1(0)/a_2(0)$. Third, compute $\Lambda = a_1(0)/\alpha_1(\phi)$.

17.12 Linearized Stability Theory

One interesting question is: Which components are stable to small perturbations? In the context of linearized instability theory, this is equivalent to the question: If most of the initial energy is in a single component and the initial amplitude of the other two is smaller by $O(\varepsilon)$ where $\varepsilon \ll 1$, what happens?

Because of our convention that a_1 is always initially the larger of the two modes with negative interaction coefficients, there are only two cases: most of the energy in a_1 or alternatively most of the energy in a_2. Let the large amplitude component have initial amplitude B. The equation for the time derivative of the big component has the two smaller components on the right-hand side and is therefore $O(\varepsilon^2)$. It is thus consistent to take the big component to be a constant to first order. The other two equations, neglecting terms that are quadratic in the perturbation, are

Case I: Stability of the dn component

$$\begin{aligned} \frac{da_2}{dt} &= B\, a_3 \\ \frac{da_3}{dt} &= -B\, a_2 \end{aligned} \tag{17.125}$$

These reduce to the single equation

$$\frac{d^2 a_3}{dt^2} + B^2 a_3 = 0 \tag{17.126}$$

This has *oscillatory* solutions: components a_2 and a_3 execute a dance with ever-small amplitudes while a_1 oscillates slightly in response, but never varying by more than $O(\varepsilon^2)$ from its initial value B. Case II: Instability of the sn component

$$\frac{da_1}{dt} = -B\ a_3$$
$$\frac{da_3}{dt} = -B\ a_1 \tag{17.127}$$

These reduce to the single equation

$$\frac{d^2a_3}{dt^2} - B^2a_3 = 0 \tag{17.128}$$

This has one solution which *grows exponentially* with time:

$$a_3(t) \sim \text{constant} \quad \exp(Bt) \tag{17.129}$$

as well as a stable solution that decays at the same rate. The linearized instability theory is not uniformly valid, but fails when the perturbations are sufficiently large. However, when two modes are initially small compared to the third, unstable (a_2) component, the elliptic functions greatly simplify. For further simplicity, shift the origin of the time coordinate, $t = 0$, to coincide with a node of one of the perturbations.[2] The initial conditions are then

$$a_1(0) = \varepsilon\ B, \qquad a_2(0) = B, \qquad a_3(0) = 0 \tag{17.130}$$

$$m_{ell} = 1 - \varepsilon^2 + O(\varepsilon^4) \tag{17.131}$$

$$\Lambda = B\ \ \{1 + O(\varepsilon^2)\} \tag{17.132}$$

$$\begin{aligned} \phi &= -\frac{K(1-\varepsilon^2)}{B} \\ &\approx \frac{\log(\varepsilon/4)}{B} \end{aligned} \tag{17.133}$$

When the elliptic modulus $m_{ell} \approx 1$, the peaks of the hyperbolic functions in the imbricate series are very narrow compared to the temporal period. The amplitudes of the growing perturbations are smallest in between the peaks of the sech functions. It is a good approximation, roughly halfway between each peak, to replace each infinite series of hyperbolic functions by just the two nearest neighbors (three nearest neighbors for a_2) so that

[2] Recall that the triad system is translation-invariant.

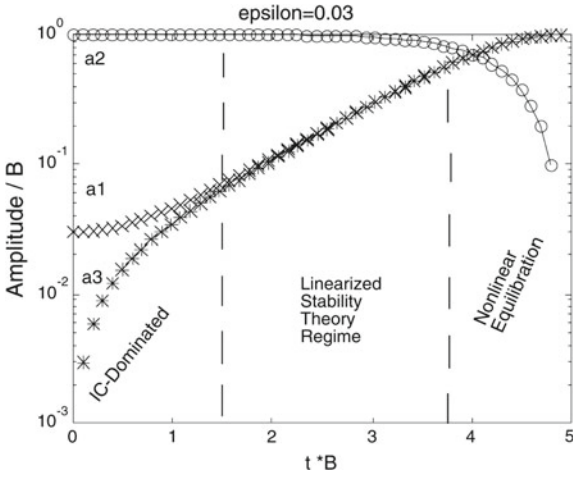

Fig. 17.12 Instability. The a_2 component, initially at unit amplitude, is unstable to small perturbations in the other two components. There are three regimes, separated by the vertical dashed dividing lines. The leftmost-regime is controlled by the initial conditions, abbreviated "IC" on the graph, more than by the dynamics of the differential equation. On this log/linear plot, exponential growth is a straight line on the graph as occurs in the middle regime where linearized stability theory is a good description of the flow. *Solid-with-circles* a_2. x's: a_1. *Asterisks* a_3

$$
\begin{aligned}
a_1(t) &\approx B\,\mathrm{sech}(Bt + \log(\varepsilon/4)) + B\,\mathrm{sech}(Bt - \log(\varepsilon/4)) \\
a_2(t) &\approx -B\tanh(Bt + \log(\varepsilon/4)) + (B/2)\tanh(Bt - \log(\varepsilon/4)) \\
&\quad + (B/2)\tanh(Bt - 3\log(\varepsilon/4)) \\
a_3(t) &\approx -B\,\mathrm{sech}(Bt + \log(\varepsilon/4)) + B\,\mathrm{sech}(Bt - \log(\varepsilon/4))
\end{aligned}
\tag{17.134}
$$

Figure 17.12 graphs the solution for a typical case. For small t, the initial conditions ("IC") dominate. The amplitude of the third mode a_3 is small compared to a_1 because $a_3(0) = 0$. For intermediate time, both perturbations grow exponentially in time; on the log-linear plot, exponential growth is a straight line with positive slope. For larger times, nonlinear equilibration slows the growth of a_1 and a_3 and the unstable component begins to fall rapidly. Finally, the perturbations a_1 and a_3 reach finite (but no longer small) maxima.

17.12.1 Vacillation and Index Cycles

It was observed by the Bergen School in the 1920's that atmospheric storm intensity (for so-called "synoptic scale" cyclones and anticyclones) increased and decreased with an irregular period whose order-of-magnitude is about two weeks. Such oscil-

lations in magnitude became known as the "index cycle" because synoptic meteorologists developed an index to quantify these oscillations.

In laboratory experiments, such as those of Fultz, Pfeffer and coworkers, and theoretical studies, such as those of Pedlosky, Loesch, Nathan and others, related oscillations were noted and dubbed "vacillation".

The triad equation exhibits vacillation or index cycle behavior in the sense that its instabilities do not grow without bound nor asymptote to a steady magnitude. Rather, after the perturbations have reached a maximum, they begin to decay while the unstable component begins to recover. The elliptic function solutions show that the amplitudes of the three Fourier components oscillate forever.

The analogy with real index cycles or vacillation is not perfect. Triad vacillation is always periodic. In more complex models, vacillation and index cycles are *chaotic* with an *irregular* separation in time between neighboring maxima of any of the components. Still, the fact that amplitude oscillations occur at all in the simplest and most regular model, the triad equations, tells us that we should not be shocked or even surprised to see similar oscillations in the atmosphere and ocean.

17.12.2 Euler Equations and Football

Euler showed that the angles of a freely-falling body evolved in time according to a system of three equations *identical* in form to the triad equations:

$$\begin{aligned}
\frac{d\omega_1}{dt} &= \frac{(I_2 - I_3)}{I_1}\omega_2\ \omega_3 \\
\frac{d\omega_2}{dt} &= \frac{(I_3 - I_1)}{I_2}\omega_3\ \omega_1 \\
\frac{d\omega_3}{dt} &= \frac{(I_1 - I_2)}{I_3}\omega_1\ \omega_2
\end{aligned} \tag{17.135}$$

where the ω's are rotational angles with respect to rotation about each of the three principal axes of rotation and the I_j are the corresponding moments of inertia. Euler performed a linearized stabilized analysis identical to that presented above. Rotation about the axis of intermediate moment of inertia is unstable.

An airborne (American) football is an amusing illustration (Fig. 17.13). Two of the moments of inertia are equal; both correspond to end-over-end rotation. By a slight modification of the linearized analysis given earlier, one can show that motion about the axes of equal moments is unstable whereas rotation with respect to the long axis is stable.

The stability of rotation about the axis which goes from one "point" of the football to the other is the quarterback's friend. It would be quite impossible to throw a "spiral", as football players refer to a pass which rotates only about this long axis as it flies through the air, were it not for the stability of this motion. (As the quarterback

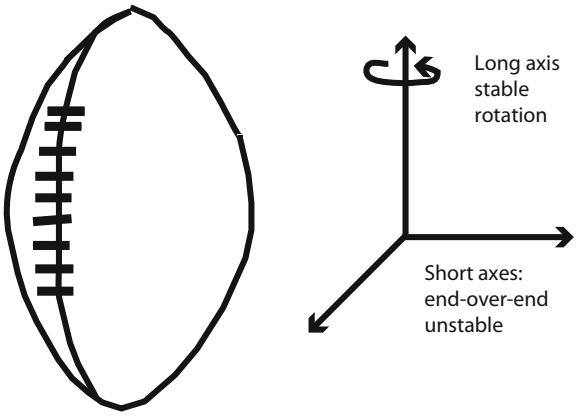

Fig. 17.13 An American football can rotate about three orthogonal axes of rotation. When the football is airborne, the rotation rates about each axis are coupled by nonlinear ordinary differential equations identical to the resonant triad equations. Rotation about the long axis that connects the pointed ends of the ball ("spiral") is stable. Rotation about the other two axes ("end-over-end") is unstable

snaps his arm forward, his wrist and fingers almost invariably apply a little end-over-end motion to the ball, too.) The end-over-end wobble tends to damp out, rapidly, not through friction or viscosity but rather through nonlinear energy transfer to the spiral rotation.

The instability of end-over-end motion is the reason that so many field goal tries that look good when the ball leaves the kicker's foot hook badly to the left or the right and miss the goalposts. A kicking motion gives the ball an initial rotation which is almost purely end-over-end and therefore unstable. As the ball sails towards the goalpost, more and more of this motion is converted into a spin about the long axis of the football. This spiralling motion makes the ball hook like a Lionel Messi soccer strike, initially far left of the goal, deliberately hit off-center to spin and thus curve back to land just inside the goal. Sometimes when a TV camera is mounted high above the goalposts, one can actually pick up the rotation of the football as it approaches the goalposts and observe that it usually has a noticeable "spiral" motion superimposed on the end-over-end tumble by the time it nears the goalposts.

17.12.3 Lemniscate Case

The Fourier series and imbricate series converge at the same rate when $m_{ell} = 1/2$, the "elliptic lemniscate". The triad solutions for this case with a temporal period of 2π are illustrated in Fig. 17.14. The leading terms of the Fourier series show very rapid convergence:

$$a_1(t) = 1 + 0.172\,\cos(2\,t) + 0.007469\,\cos(4\,t) + 0.000322\,\cos(6\,t) \quad (17.136)$$

$$a_2(t) = 0.8691\,\sin(t) + 0.03594\,\sin(3\,t) + 0.001553\,\sin(5\,t) \quad (17.137)$$

$$a_3(t) = 0.7971\,\cos(t) + 0.03593\,\cos(3\,t) + 0.001553\,\cos(5\,t) \quad (17.138)$$

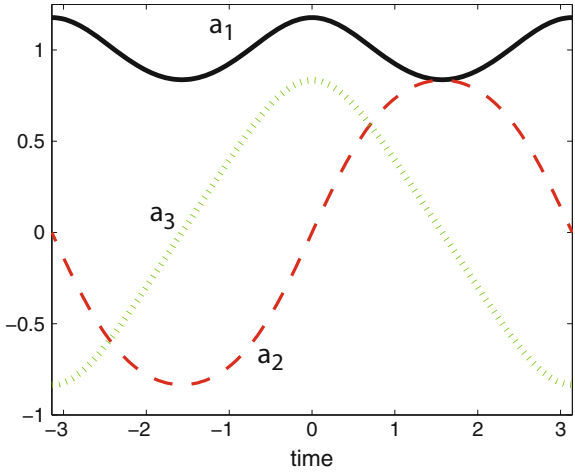

Fig. 17.14 A resonant triad for the elliptic modulus $m_{ell} = 1/2$, the lemniscate case. Note that $a_1(t)$ [*thick solid black curve*], the component proportional to the elliptic function"dn", is equal to the constant one modulated by fluctuations whose amplitude is only 0.18. $a_2(t)$ [*red dashed*] and $a_3(t)$ [*green dotted*] look a lot like $\sin(t)$ and $\cos(t)$ because to a good approximation, they are $\sin(t)$ and $\cos(t)$

Even for this weakly nonlinear case, there is considerable cycling of energy between two modes. Strongly nonlinear, $m_{ell} \rightarrow 1$ triads are (ironically) best analyzed by *linear* dynamics as illustrated in Sect. 18.12 and as pointed out by Leonard Euler a quarter of a millennium ago.

Note this is a stable oscillation as predicted by the linearized instability theory of the previous section. Linearizing about $a_1(t) = 1$ predicts that a_1 will remain constant to lowest order while $a_2(t)$ and $a_3(t)$ are both sinusoidal with a phase difference of $\pi/2$. The amplitudes are undetermined by the linearized analysis except that $a_2(t)$ and $a_3(t)$ should have equal amplitudes. The Fourier series for the lemniscate case shows that the amplitudes of $\sin(t)$ in $a_2(t)$ and of $\cos(t)$ in $a_3(t)$ are 0.869 and 0.797, respectively, which are not equal, but differ by only 9%.

17.12.4 Instability and the Lemniscate Case

The lemniscate case is also the boundary between stability and instability. When the elliptic modulus $m_{ell} = 1/2$, the minimum of a_1 [dn function] just touches the maximum of a_2, the sn function. For $m_{ell} < 1/2$, the dn component is always larger than the other two throughout the entire temporal period. For such small elliptic moduli, it only makes sense to linearize about a_1 because it is always the largest, but we saw in the preceding section that linearizing about a_1 [dn] yields only stability.

When $m_{ell} > 1/2$, a_2 is larger than a_1 for at least a small interval around the minima of a_1, and linearizing about a_2 is now sensible at least in this time neighborhood. But the linearized analysis of earlier sections yields only instability.

Figure 17.15 is identical to Fig. 17.14 except that the elliptic modulus has been increased to $m_{ell} = 0.99994420407$, or equivalently the parameter S in the analytical

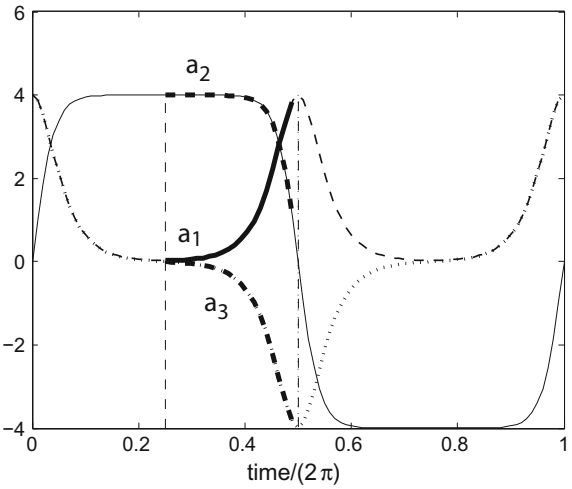

Fig. 17.15 A resonant triad for the elliptic modulus $m_{ell} = 0.9999442$. The instability occurs on the interval enclosed by the vertical *dashed lines*, $t/P \in [1/4, 1/2]$; the curves have been thickened in this interval

solutions has been increased from one to four. The instability analyzed earlier unfolds over only one-fourth of the complete period, marked on the diagram as the interval between the dashed vertical lines and also by thickening the curves for each unknown. When t is equal to one quarter of the temporal period, $a_3(t)$ is zero, $a_1(t)$, the dn function, is at its very small minimum, and $a_2(t)$ is large and almost steady. Over the next quarter-period, $a_1(t)$ and $a_3(t)$ grow exponentially until $a_2(t)$ begins to change. Over the remaining three-quarters of the period, the triad cycles back to its starting point in endless vacillation.

17.13 Resonance Conditions: A Problem in Algebraic Geometry

Definition 17.3 (*Triad Resonance Conditions*) Three equatorial waves are said to be in triad resonance when both the frequencies and the wavenumbers satisfy the following five constraints.

$$w_1^3 - (k^2 + 2n + 1)w_1 - k = 0 \tag{17.139}$$
$$w_2^3 - (k^2 + 2m + 1)w_2 - k = 0 \tag{17.140}$$
$$w_3^3 - (k^2 + 2p + 1)w_3 - k = 0 \tag{17.141}$$

$$w_1 + w_2 \pm w_3 = 0 \tag{17.142}$$

$$k + k_2 \pm k_3 = 0 \tag{17.143}$$

where (n, m, p) are latitudinal mode numbers (always integers) and where the continuous parameter is k_1, which is written as k [unsubscripted] to emphasize its role as parameter.

The system consists of three copies of the dispersion relation plus the two resonance conditions. We can specify one wavenumber and all three latitudinal mode numbers arbitrarily and then have five equations in five unknowns. By using the two linear equations, the resonance conditions can be reduced to a polynomial system of three polynomials in three unknowns. Bezout's Theorem states that the three cubics have at most a total of 27 solutions. The multiple branches incorporate all the wave species except for the Kelvin wave and Yanai waves, which require special treatment.

To proceed, arbitrarily choose the signs in the linear equations to be positive, and and choose the unknowns to be (w_1, w_2, k_3) in the reduced system. The trio of three equations in three unknowns is then

$$w_1^3 - (k^2 + 2n + 1)w_1 - k = 0 \tag{17.144}$$

$$w_2^3 - (\{k + k_3\}^2 + 2m + 1)w_2 + k + k_3 = 0 \tag{17.145}$$

$$-(w_1 + w_2)^3 + (k_3^2 + 2p + 1)(w_1 + w_2) - k = 0 \tag{17.146}$$

plus the back-solve constraints

$$w_1 + w_2 = -w_3 \tag{17.147}$$

$$k + k_3 = -k_2 \tag{17.148}$$

This system is simpler than the general case of three cubic polyomials in three unknowns because the first cubic is not a function of all three unknowns but only one. We can therefore solve this cubic for $w_1(n; k)$ independently of the rest of the system, either numerically or by using the perturbation series from [50]:

$$w_1^{Rossby} = -\frac{k}{2n+1+k^2}\left\{1 + \frac{k^2}{(n+1+k^2)^3} + 3\,\frac{k^4}{(2n+1+k^2)^6} + 12\,\frac{k^6}{(2n+1+k^2)^9} + \cdots\right\}$$

$$w_1^{GW} = \sqrt{2n+1+k^2}\left\{\pm 1 + \frac{1}{2}\,\frac{k}{(2n+1+k^2)^{3/2}} \mp \frac{3}{8}\,\frac{k^2}{(2n+1+k^2)^3}\right\}$$

where the expansion applies to both eastward-traveling and westward-traveling with different choices of sign.

The resultant[3] of the two remaining copies of the dispersion relation with respect to k_3 is a sixth degree polynomial in w_2:

$$R = p_6 w_2^6 + p_5 w_2^5 + p_4 w_2^4 + p_3 w_2^3 p_2 w_2^2 + p_1 w_2 + p_0 = 0 \qquad (17.149)$$

where

$$\begin{aligned}
p_6 &= 4\,{w_1}^2 - 4\,k^2 \\
p_5 &= -12\,w_1\,k^2 + 8\,m w_1 + 12\,{w_1}^3 - 8\,p w_1 \\
p_4 &= 4\,k^2 + 4\,k^2 m - 14\,k^2 {w_1}^2 + 4\,m^2 + 20\,{w_1}^2 m - 20\,p {w_1}^2 \\
&\quad - 8\,m p + 4\,k w_1 + k^4 + 13\,{w_1}^4 + 4\,k^2 p + 4\,p^2 \\
p_3 &= -16\,{w_1}^3 p + 16\,{w_1}^3 m + 8\,m w_1\,k^2 + 8\,m^2 w_1 - 16\,p w_1\,m - 8\,{w_1}^3 k^2 + 8\,{w_1}^2 k \\
&\quad + 8\,p w_1\,k^2 + 6\,{w_1}^5 + 8\,p^2 w_1 + 2\,k^4 w_1 + 8\,w_1\,k^2 \\
p_2 &= -4\,p w_1\,k + 4\,m {w_1}^2 k^2 + 4\,p {w_1}^2 k^2 + 6\,{w_1}^3 k - 4\,m w_1\,k - 3\,{w_1}^2 + k^4 {w_1}^2 \\
&\quad - 2\,{w_1}^4 k^2 - 4\,k w_1 + 4\,{w_1}^4 m - 4\,{w_1}^4 p \\
&\quad + {w_1}^6 + 4\,m^2 {w_1}^2 - 2\,k^3 w_1 - k^2 + 4\,k^2 {w_1}^2 - 8\,p {w_1}^2 m + 4\,p^2 {w_1}^2 \\
p_1 &= 2\,p w_1 - 3\,{w_1}^3 + 2\,{w_1}^4 k - 4\,{w_1}^2 k - 4\,p {w_1}^2 k - w_1\,k^2 - 2\,m w_1 - 2\,k^3 {w_1}^2 - 4\,m {w_1}^2 k \\
p_0 &= {w_1}^2 + 2\,p {w_1}^2 + k w_1 - {w_1}^4 + k^2 {w_1}^2
\end{aligned} \qquad (17.150)$$

Ripa previously showed that the resonance definitions could be reduced to a polynomial of sixth degree.

17.13.1 Selection Rules and Qualitative Properties

Theorem 17.5 (Triad Mode Number Selection) *When the normal modes are linearized about a state of rest, the interaction coefficients of a trio of waves is* zero *unless the sum of the latitudinal mode numbers is* odd. *To put it another way, the modes of a resonant triad must fall into one of the following symmetry classes:*

1. *Sym/Sym/Sym*
2. *Sym/Antisym/Antisym*

Proof This theorem was known to Domaracki and Loesch and to Ripa, who state it less formally. The crucial point is that if two modes have both the same latitudinal symmetry, then quadratically nonlinear terms involving these will necessarily be symmetric with respect to the equator. (When both modes are antisymmetric, the negative signs that arise in both modes *individually* when reflected about the equator *cancel* in the quadratically nonlinear product.) If we then take the inner product of

[3]The resultant of two bivariate polynomials in the unknowns (x, y) is the determinant of the so called Sylvester matrix which is formed from the coefficients. The resultant with respect to x is a function of only the single variable y; its zeros are all possible y-values of the zeros of the system of two polynomial equations in two unknowns [50].

such pairs of nonlinear terms with a mode which is antisymmetric with respect to the equator, the integral will be zero because it is the integral over all y of a triple product which changes sign when reflected about the equator. All integrals over $y \in [-\infty, \infty]$ of antisymmetric functions with an inner product weight which is symmetric with respect to the equator will be zero because the integration over the hemisphere where the integrand is positive will be canceled by integration over the other half-interval where the integrand is equal in magnitude but opposite in sign.

Similarly, the nonlinear terms that are the interaction between two modes of opposite symmetry are antisymmetric with respect to the equator and therefore yield a nonzero interaction coefficient only when multiplied by an antisymmetric mode. ■

Ripa introduces the "slowness",

$$s \equiv \frac{k}{w} = \frac{1}{c} \tag{17.151}$$

and asserts that the mode of intermediate slowness is always the unstable mode.

Ripa has a good discussion of "nonlocal" triads, by which he means triads composed of modes with frequencies of different orders of magnitude, such as strong interactions among short gravity waves catalyzed by a long Rossby wave.

17.13.2 Limitations of Triad Theory

Resonant triad theory has multiple severe limitations. The first is that there are so many resonances and near-resonances that focusing on a single triad is as useless at focusing on a single bubble in turbulent convection. Worse still, a single wave may be resonant in more than one triad, allowing energy to leak from one triad to another. Such chains of mutually resonant triads are a major theme of Loesch and Deininger [29]. Their Fig. 1 is a closed system of forty-eight triads. Richard Deininger, who drew the diagrams as an undergraduate, regarded the task as his own personal Twelve Labors of Hercules (private communication 1980). Contemporary dynamicists, though armed with much faster floating point arithmetic and vastly superior graphics, will likely be equally exhausted.

The second limitation is that resonance is a small amplitude concept. If we define a significant energy exchange to be, say, the halving of the energy initially in a given mode, then in the limit of infinitesimal amplitude, the rate of energy transfer tends to zero and perfect resonance is required. If the resonance is imperfect in the sense that

$$w_1 + w_2 + w_3 = \delta^{detuning} \tag{17.152}$$

the direction of energy will reverse sign after a time interval of $P_{exchange} = O(1/\delta^{detuning})$. The total amount of energy exchanged is the product of the amplitude

ε of the components of the triad multiplied by $P_{exchange}$. As the amplitude increases, therefore, more and more triads join the dance of large energy exchanges. For sufficiently large amplitude, every trio of modes is strongly fluctuating, and "selective resonant transfers" must be replaced by an adjective like "wave turbulence" [51].

17.14 Solitary Waves in Numerical Models

Kindle [52], Greatbatch [26], Williams and Wilson [53] and Boyd [25, 26, 52, 53] have all shown that it is easy to generate Rossby solitary waves in numerical models. We will discuss some sample results from each in turn.

Kindle [52] tested the analytical theory of Boyd [20] by using the lowest order perturbative soliton as the initial condition for a 1-1/2 layer model. To the extent that the perturbative solution is accurate, the twin troughs[4] should propagate westward without change in shape. The soliton is of moderately large amplitude; the maximum thermocline displacement (off the equator) is 1/4 the undisturbed layer thickness and $B = 0.48$, which is 90% of the threshold for recirculation. Nonetheless, the prediction is highly accurate. Only the very small ripples left behind as the soliton propagates westward indicate error as shown in his Fig. 2a. By contrast, Fig. 2b shows that the corresponding *linear* solution has dispersed into a shallow pair of troughs followed by two prominent crests and a long train of gradually diminishing oscillations. Again as predicted, the nonlinear solution has traveled as far westward in 120 days as the linear solution in 160 days.

KdV theory also predicts that an initial condition which reduces the thickness of the upper layer, that is, one of opposite sign from the soliton, will exhibit "superlinear" dispersion: the nonlinearity will defocus the peak so that wave energy spreads more rapidly in x, faster than for the linear solution. Kindle confirmed this prediction by making another run with an initial condition of the same structure but opposite sign (as in Fig. 16.20).

Kindle's other experiments generated Rossby solitons by the reflection of a Kelvin wave from the eastern boundary of the ocean after the latter had been excited by a transient wind stress which was applied to a zonal patch of finite width in the western part of the sea. His results are too extensive to fully describe, but in a typical experiment, a wind stress of 0.5 dynes/cm^2 applied for a month in a wind patch 6,000 km across generated a soliton with $B = 0.46$; the half-maximum-to-half-maximum zonal width of the solitary wave is about 1200 km. The soliton amplitude is determined by the *product* of the strength of the wind stress and the zonal width of the region where the stress is applied.

Long duration events (two months or more) generate multiple solitary waves. In Kindle's idealized El Niño, the width of the wind patch was only 2,000 km, but in contrast to the experiment above, the eastward wind stress was left on for the

[4] In the 1-1/2 layer formalism, a *crest* in ϕ is equivalent to a thickening of the upper layer, and therefore a *trough* in the thermocline. Kindle has chosen to illustrate the thermocline trough.

duration after being switched on at $t = 0$. The result was two solitons with $B = 0.43$ and $B = 0.33$, respectively, with the larger solitary wave leading its little brother westward as predicted by KdV theory.

One interesting conclusion is that wind stress events which are very broad in x or t or both can nevertheless create solitary waves which have much smaller space and time scales. (For example, the experiment with the wind stress 6,000 km across generated a reflected Rossby soliton only one-fifth as wide as the patch of wind.) Part of the explanation is that the linear group velocity of the Kelvin wave is three times that of the Rossby wave. A Kelvin pulse of width W will be reflected in a time interval of length W since the Kelvin wave has unit speed. However, the reflected Rossby wave will have width $W/3$ because this is as far as the leading edge of the Rossby packet can travel before the trailing edge leaves the coast. In addition to this linear contraction of the pulse, the nonlinearity will further focus the crest – it is precisely this nonlinear focusing that balances the dispersion to make the soliton possible.

Nonetheless, this threefold compression-by-reflection is very important. Kindle [52] notes that "it appears Rossby solitary waves may be most easily generated by the reflection of Kelvin waves from the eastern boundary". His experiments showed that it is possible to generate Rossby solitons directly by an eastward wind stress, but only when the stress is narrow and of short duration. The compression-by-reflection mechanism is very important in converting very large scale wind stress events into intense, small-scale solitons.

In the numerical work of Greatbatch [26], the wind is turned on impulsively over the whole basin and left on. As in Kindle's calculations, the reflection of a Kelvin wave from the eastern boundary will generate solitons as shown by his Fig. 3.[5] However, if the wind stress changes extend all the way to the eastern boundary, this coast will act as a soliton- wavemaker beginning the instant the (westward) wind stress is turned on. He shows that the result is a pair of solitons which completely dominate the flow field. The corresponding linear solution is missing the two vortex pairs – the solitons – that are the most striking and intense features of the nonlinear flow.

The conclusion is amusing. The inverse scattering theory of the Korteweg-deVries equation shows that an initial condition will normally generate at least one soliton unless the flow at $t = 0$ is everywhere of the wrong sign. For the bounded ocean, not even this escape clause exists: Greatbatch obtains solitons for *either* sign of the wind stress as long as the transient wind extends as far as the eastern coast. In this sense, solitons are more readily found in equatorial waves than in non-rotating water waves even though the latter is where solitary crests were first observed.

Greatbatch also generated a strong equatorial mean flow in his experiments, and generalized the theory of Boyd [20, 54] to include the doppler-shifting and other effects of a mean latitudinally-varying current. He found good agreement between

[5] Note that source of the incident downwelling Kelvin wave is different. A wind stress event which is confined to a patch will generate such a wave from the *eastern edge* of the patch if the wind is *eastward*. A global wind stress will excite a thermocline-deepening Kelvin wave at the *western coast* of the ocean if the wind stress is *westward*.

the predicted and observed soliton phase speeds after correcting for the mean flow. His numerical experiments also show that the solitons are very stable and persistent even when the mean current is very strong. He shows a case with a mean flow so that intense that it is actually unstable, breaking down into meanders. Nonetheless, a train of solitary waves forms, and they propagate westward with little change in form.

The bottom panel also shows that there is a broad patch of dead water between the two solitons. The exact solution for the KdV solitary wave shows that the amplitude of the soliton decreases *exponentially* fast as one moves away from the peak, so two solitons can be very close together in a geographical sense, and yet have only an exponentially small overlap so that they are truly "solitary".

Westward-propagating vortex pairs ("modons") are highly unstable on the mid-latitude beta plane, but the equatorial wave guide stabilizes vortex pairs that straddle the equator, even when given a strong initial tilt. As discovered by Williams and Wilson and explained theoretically by Boyd, higher latitudinal mode solitary waves are weakly nonlocal through radiation of sinusoidal Rossby waves of lower latitudinal mode number. The amplitude and wavelength of the radiation are in good agreement with nonlocal soliton theory.

The interaction of free Rossby waves with the semi-transparent equatorial waveguide and the ensuing wave-mean flow interaction are described in Reznik and Zeitlin [55].

Andrew Madja has written many articles on nonlinear dynamics with Biello, Dutrifroy, Rosales and numerous other collaborators [30, 43–45]. Most have an atmospheric motivation, and therefore are properly excluded here, but fortunately Majda has been a prolific writer of reviews [56] and books with [57, 58] as the most relevant. Majda is a Grand Master of applied mathematics and it is fortunate that we can read his perspective in his own voice, unfiltered.

17.15 Gerstner Trochoidal Waves and Lagrangian Coordinate Descriptions of Nonlinear Waves

In 1802, Frantiŝek Josef Gerstner (1756–1832) discovered a family of exact solutions for nonlinear water waves [59, 60]. His solution using Lagrangian rather than Eulerian coordinates was independently discovered by Rankine in 1863 [61]. Marie-Louise Dubreil-Jacotin (1905–1972) extended Gerstner's solution to stratified fluids [62]. Chia-Shun Yih (1918–1997) showed that Gerstner's solution also described edge waves, that is, waves propagating parallel to a straight coastline on water of depth varying linearly in the direction normal to the coast [63].

However, in a review of the history of water waves, Craik wrote "The Gerstner wave solution was long overlooked; even today it is usually regarded more as a curiosity than a result of practical importance because the wave is not irrotational." [p. 3 of [64]]

In 2001, two events revived interest in the long-dormant theory of Gerstner waves. First, a highly influential computer graphics paper showed that multiple copies of Gerstner's analytical formulas were a superb way to generate realistic water waves. This method is now very widely used, motivated entirely by plausible appearance rather than any pretense of scientific accuracy. The other event was strictly mathematical when, to quote [65]: "A mathematical analysis of Gerstner's solution was performed by [Adrian] Constantin in [66] which investigated rigorously, for the first time, the evolution of the fluid domain under the propagation of Gerstner's wave. It was shown, using a mixture of analytical and topological methods, that it is indeed possible to have a motion of the whole fluid body where all the particles describe circles with a depth-dependent radius — the fluid domain as a whole evolves in a manner which is consistent with the full governing equations."

This touched off a flurry of work in the mathematics community on Gerstner waves which has still not subsided.

Constanin himself gave this field a second wind when he found Gerstner solutions on the equatorial beta-plane [67], which in turn triggered further works by himself [68–71] and and a dozen others [72–75]. Sastre-Gomez' paper is a good review that also proves that Constantin's (2012) solution actually exists [73].

Unfortunately, although a couple of papers have appeared in the *Journal of Physical Oceanogrphy*, the authors have been entirely mathematics professors publishing mostly in mathematics journals.

Jan Erik Weber's paper that sought Gerstner waves in the laboratory wrote in the abstract, "A comparison with experimental data for the drift of thin plastic sheets in wave tanks is made, showing that the presence of viscosity-modified Gerstner waves cannot be ruled out on the basis of these observations." Cannot be ruled out? And the last sentence of the article is "Like Gerstner [4], we have assumed that the wave field is present initially. The eventual generation of Gerstner waves in a viscous fluid from a no-wave state has not been discussed here. This intriguing matter should be further investigated by combining laboratory experiments and theoretical analyses."

It is astounding that such basic questions as how to generate the waves is still wide open. The initialization question has in fact always discouraged studies of Gerstner waves because the waves are strongly vortical whereas conventional water wave models are of irrotational waves, which is what wind etc., normally generate.

There is a need for an oceanographic Freeman Dyson to reconcile two wildly different formalisms. The British physicist famously showed that the renormalization schemes of Feynman and Schwinger, enormously different in appearance and philosophy, would in fact always give the same answer.[6] So far, equatorial waves in Lagrangian coordinates and nonlinear equatorial waves in Eulerian coordinates have lacked a Great Reconciliation.

[6]Luck as well as Dyson's brilliance played a role in the Great Quantum Electrodynamics Renormalization Reconciliation. Dyson thoroughly learned Schwinger's approach during a lengthy summer school at the University of Michigan where Schwinger was a prominent lecturer; he learned Feynman's formalism while carpooling with the older physicist on a three day cross-country trip where there was little to do, in those days before auto entertainment systems, except discuss quantum physics.

17.16 Potential Vorticity Inversion

The nonlinear shallow water wave equations have a conserved potential vorticity

$$q \equiv (\zeta + y)/(1 + \phi) \qquad \text{[Potential Vorticity]}$$

where ζ is the usual relative vorticity. Middle atmosphere dynamicists have recently shown that the potential vorticity plays a central role in flow in that region. In particular, it is possible to "invert the potential vorticity" so as to deduce the winds, pressure, density, and so on from this single scalar quantity.

Unfortunately, this concept doesn't work very well on the equatorial beta-plane even in the simple case of linear dynamics. The linearized potential vorticity is

$$q_{linear} \equiv \zeta - y\,\phi \qquad \text{when } |\zeta|, |\phi| \ll 1 \tag{17.153}$$

Since the Rossby modes are strongly rotational whereas the gravity waves are strongly divergent, but have comparatively little vorticity, one might legitimately hope to expand an observed distribution of linearized q in Hermite functions and then deduce the amplitude of each Rossby mode from this under the assumption that the contributions of the gravity waves are negligible.

The bad news is that the Kelvin wave, which is probably *the* single most important mode in equatorial dynamics (because it is the gravest, and therefore the most easily excited) has *no* linearized potential vorticity as proved in the next section. In consequence, the observed q-field gives us no information whatsoever about the Kelvin wave.

17.16.1 A Proof that the Linearized Kelvin Wave Has Zero Potential Vorticity

The potential vorticity of the nonlinear shallow water equations is

$$q = \frac{\zeta + y}{1 + \phi} \tag{17.154}$$

The relative vorticity $\zeta = v_x - u_y$ simplifies to $\zeta = -u_y$ because the north-south velocity is identically zero for the Kelvin wave. When the wave is small, $1/(1+\phi) \approx 1-\phi+O(\phi^2)$. Neglecting terms that are quadratic in the wave amplitude, the potential vorticity of an infinitesimal amplitude Kelvin wave is

$$q_{Kelvin} \approx -u_y - y\phi \tag{17.155}$$

Because $u = \phi \exp(-y^2/2)$, $u_y = -yu = -y\phi$, the terms cancel identically giving

$$q_{Kelvin} = 0 \tag{17.156}$$

as asserted in the previous subsection.

17.17 Coupled Systems of KdV or RLW Equations

Replacement of the KdV equation by the RLW equation corrects only the dispersion relation to $O(k^2)$; substituting the RLW equation for KdV does nothing to correct the proportionality between S, v, and D. However, the effects of finite zonal length scale on the latitudinal structure of S, v and D are non-secular; they do not increase with time or with distance away form the boundaries; they are always small corrections if the zonal scale is large. However, the difference between a dispersive wave packet and a nondispersive wave packet will become greater and greater as the packet propagates along because the dispersion, even if weak, will eventually cause the packet to break up and spread whereas the nondispersive packet will merely travel on, its shape in x forever unchanging.

The nonlinear terms can be incorporated into the factors F_1 and F_3 of the linear shallow water equations in a similar heuristic manner: the procedure is to simply evaluate them using the linear, nondispersive solutions to specify the latitudinal structure. Since the eigenfunctions of the linear shallow water wave equations are a complete set, it has been pointed out independently by Kasahara [76] and Ripa [39] that one can use them as Galerkin basis functions in exactly the same way as one would use trigonometric functions, Hermite functions, or spherical harmonics [39, 76]. The complications are that (i) each eigenfunction has three components, i.e., the eigenfunctions are 3-vectors, and that (ii) the eigenfunctions are different for each different zonal wavenumber. The advantage is that one can jettison gravity wave functions and keep only the Kelvin and Rossby wave basis functions which is basically just what we're doing with the long wave approximation. Thus, this *ad hoc* philosophy for treating the nonlinear terms can be given a rigorous justification, too.

Unfortunately, the latitudinal structure of Rossby waves is also a function of k. The denominators of $(1 + c)$ and $(1 - c)$ in (16.40) tend to 1 as $k \to \infty$ in contrast to the their long wave limits of 2/3 and 4/3 (Fig. 17.16). This moderate error does not alter the fact that the RLW model is a better model than the KdV equation for equatorial Rossby waves. In particular, if we solve an initial value problem, the RLW equation will correctly propagate energy both east and west of the initial pulse, which is impossible with the KdV equation. The structure of the wavefront with eastward group velocity will not be quite right: the u and ϕ fields will have a component of $\psi_2(y)$ that is a little too small while the contribution of $\psi_0(y)$ is exaggerated. The nonlinear modifications to the ripples will be wrong, too, but if the initial condition has a long zonal scale, the eastward-propagating part of the disturbance will be of small to moderate amplitude, and the poor treatment of nonlinearity may not matter either.

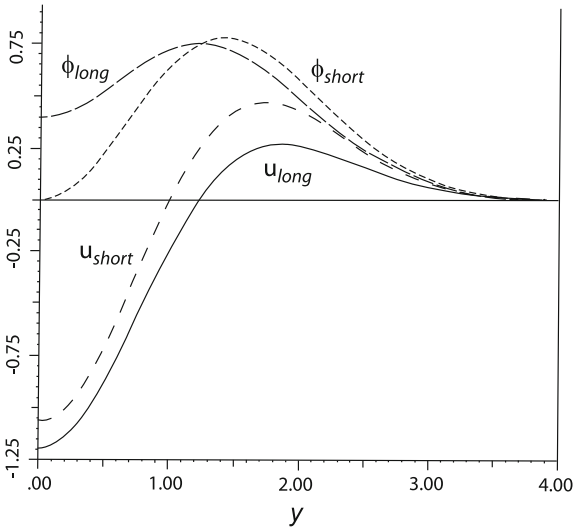

Fig. 17.16 The latitudinal structure functions for the east-west velocity u and height field ϕ for the $n = 1$ [lowest symmetric] Rossby mode in the limits of zero wavenumber ($k = 0$) and infinite wavenumber ($k \to \infty$), the "short" wave limit. The long wave structure functions are $\phi^{long} = 3/8 \frac{\sqrt{2}e^{-1/2\,y^2}(2\,y^2-3)}{\sqrt[4]{\pi}}$ and $u^{long} = 3/8 \frac{\sqrt{2}e^{-1/2\,y^2}(2\,y^2+1)}{\sqrt[4]{\pi}}$

A strength of the long wave approximation is that in the limit of very large zonal scale (small k), the vector eigenfunctions of the shallow water wave equations become independent of k. Instead of having to take a discrete sum or Fourier integral over zonal wavenumbers, we can lump the x and t dependence of a given latitudinal mode into an amplitude function $A_{n+1}(x, t)$ which is multiplied by the appropriate function of y to give the components of S, v, and D (or u, v and ϕ as you please). By doing so, we are making an $O(k^2)$ error in the nonlinear terms, but we greatly simplify the problem. It is certainly better to include the nonlinear terms in an approximate way than to neglect them entirely as done in the linear long wave model.

Thus, we conclude that it is possible to generalize the linear, long wave model to a "pseudo-long wave" set that includes both dispersion and nonlinearity without sacrificing the advantages of (i) filtering out gravity waves and (ii) lumping the $x - t$ dependence of a given Rossby mode into a single amplitude function. Analytical solutions of this extended long wave model are probably not possible, but it is very easy to integrate numerically (at least in the RLW form).

Why bother? The answer lies in the foci and caustics of the linear, long wave model for normal modes in a basin, discussed in Chap. 6 above. Even if the amplitude is small and dispersive effects weak over most of the ocean basin, both nonlinearity and dispersion must inevitably become important at caustics and foci because the wave amplitude is blowing up along them, and the length scale of wave variation is tending to 0. The "pseudo-long wave" model would be a cheap tool for understanding how dispersion and nonlinearity modify the linear results of Cane and Sarachik [101]. It must be noted, however, that the full nonlinear shallow water wave equations are themselves relatively inexpensive to integrate with today's computers. The benefits of normal modes can be easily retrieved by post-processing whose first step is to expand the fields in Hermite functions.

References

1. Bender CM, Orszag SA (1978) Advanced mathematical methods for scientists and engineers. McGraw-Hill, New York
2. Marshall HG, Boyd JP (1987) Solitons in a continuously stratified equatorial ocean. J Phys Oceanogr 17:1016–1031
3. Johnson RS (1976) Nonlinear, strongly dispersive water waves in arbitrary shear. Proc R Soc Lond A 338:101–114
4. Johnson RS (1997) A modern introduction to the mathematical theory of water waves, vol 18. Cambridge texts in applied mathematics. Cambridge University Press, Cambridge
5. Tracy ER, Osborne AR, Bergamasco L (1988) On the Nonlinear Schrodinger limit of the Korteweg-deVries equation. Physica D 32:83–106
6. Tracy ER, Larson JW, Osborne AR, Bergamasco L (1991) The relationship between the spectral theories for the periodic Korteweg-deVries and Nonlinear-Schrodinger equations. In: Osborne AR (ed) Nonlinear topics of ocean physics: Fermi Summer School. Course LIX, North-Holland, pp 769–825
7. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
8. Boyd JP (1983) Equatorial solitary waves, Part II: envelope solitons. J Phys Oceanogr 13:428–449
9. Kevorkian J, Cole JD (1996) Multiple scale and singular perturbation methods. Springer, New York
10. Kevrikis P, Frantzeskakis J, Carretero R (2015) The defocusing nonlinear schoedinger equation: from dark solitons to vortices and vortex rings. SIAM, Philadelphia
11. Yang J (2010) Nonlinear waves in integrable and nonintegrable systems, mathematical modeling and computation, vol 16. SIAM, Philadelphia
12. Osborne A (2010) Nonlinear ocean waves and the inverse scattering transform, international geophysics, vol 67. Academic, New York
13. Osborne AR, Burch TL (1980) Internal solitons in the Andaman sea. Science 208(4443):451–460
14. Benjamin TB, Feir JE (1967) The disintegration of wave trains on deep water. Part 1. Theory. J Fluid Mech 27(3):417–430
15. Yuen HC, Ferguson WE Jr (1978) Relationship between Benjamin-Feir instability and recurrence in the nonlinear Schoedinger equation. Phys Fluids 21(8):1275–1278
16. Chen GY, Boyd JP (2002) Nonlinear wave packets of equatorial Kelvin waves. Geophys Astrophys Fluid Dyn 96(5):357–379
17. Boyd JP (1983) Second harmonic resonance for equatorial waves. J Phys Oceanogr 13:459–466
18. Ripa P (1981) On the theory of nonlinear interactions among geophysical waves. J Fluid Mech 103:87–115
19. Tribbia JJ (1984) A simple scheme for high-order nonlinear normal initialization. Monthly Weather Rev 112:278–284
20. Boyd JP (1980) Equatorial solitary waves, Part I: Rossby solitons. J Phys Oceanogr 10:1699–1718
21. Boyd JP (1983) Long wave/short wave resonance in equatorial waves. J Phys Oceanogr 13:450–458
22. Boyd JP (1989) Non-local equatorial solitary waves. In: Nihoul JCJ, Jamart BM (eds) Mesoscale/synoptic coherent structures in geophysical turbulence: proceedings 20th Liege Coll. on hydrodynamics. Elsevier, Amsterdam, pp 103–112
23. Boyd JP (1991) Nonlinear equatorial waves. In: Osborne AR (ed) Nonlinear topics of ocean physics: Fermi Summer School. Course LIX, North-Holland, pp 51–97
24. Boyd JP (2006) Fourier pseudospectral method with Kepler mapping for traveling waves with discontinuous slope: application to corner waves of the Ostrovsky-Hunter equation and equatorial Kelvin waves in the four-mode approximation. Appl Math Comput 177(1):289–299

25. Boyd JP (2002) Equatorial solitary waves. Part 5. Initial value experiments, co-existing branches and tilted-pair instability. J Phys Oceangr 32(9):2589–2602
26. Greatbatch RJ (1985) Kelvin wave fronts, Rossby solitary waves and nonlinear spinup of the equatorial oceans. J Geophys Res 90:9097–9107
27. Le Sommer J, Reznik GM, Zeitlin V (2004) Nonlinear geostrophic adjustment of long-wave disturbances in the shallow-water model on the equatorial beta-plane. J Fluid Mech 515:135–170
28. Domaracki A, Loesch A (1977) Dynamics of closed systems of resonantly interacting equatorial waves. J Atmos Sci 34:486–498
29. Loesch A, Deininger RE (1979) Dynamics of closed systems of resonantly interacting equatorial waves. J Atmos Sci 36:1490–1497
30. Majda AJ, Rosales RR, Tabak EG, Turner CV (1999) Interaction of large-scale equatorial waves and dispersion of Kelvin waves through topographic resonances. J Atmos Sci 56(24):4118–4133
31. Raupp CFM, Dias PLS (2005) Excitation mechanism of mixed Rossby-gravity waves in the equatorial atmosphere: role of the nonlinear interactions among equatorial waves. J Atmos Sci 62(3):1446–1462
32. Raupp CFM, Dias PLS, Tabak EG, Milewski P (2008) Resonant wave interactions in the equatorial waveguide. J Atmos Sci 65(11):3398–3418
33. Bates ML, Grimshaw RHJ (2014) An extended equatorial plane: linear spectrum and resonant triads. Geophys Astrophys Fluid Dyn 108(1):1–19
34. Lorenz EN (1960) Maximum simplification of the dynamic equations. Tellus 12:243–254
35. Bretherton FP (1964) Low frequency oscillations trapped near the equator. Tellus 16:181
36. Bers A, Kaup DJ, Reiman A (1976) Space-time evolution of nonlinear three-wave interactions. I. Interaction in a homogeneous medium. Phys Rev Lett 37(4):182–185
37. Kaup DJ (1976) 3-wave interaction nondispersive phenomenon. Stud Appl Math 55(1):9–44
38. Martin RA, Segur H (2016) Toward a general solution of the three-wave partial differential equation. Stud Math Appl 137(1):70–92
39. Ripa P (1982) Nonlinear wave-wave interactions in a one-layer reduced-gravity model on the equatorial beta-plane. J Phys Oceangr 12:97–111
40. Ripa P (1983) Resonant triads of equatorial waves in a one-layer model. Part I. Non-local triads and triads of waves with the same speed. J Phys Oceangr 13:1208–1226
41. Ripa P (1983) Resonant triads of equatorial waves in a one-layer model. Part II. General classification of resonant triads and double triads. J Phys Oceangr 13:1227–1240
42. Ripa P (1984) Harmonics resonance and chaos in the equatorial waveguide. AIP conference proceedings. American Institute of Physics/Woodbury, Ensenada, pp 537–545
43. Biello JA, Majda AJ (2004) The effect of meridional and vertical shear on the interaction of equatorial baroclinic and barotropic Rossby waves. Stud Appl Math 112(4):341–390
44. Majda AJ, Biello JA (2004) Boundary layer dissipation and the nonlinear interaction of equatorial baroclinic and barotropic Rossby waves. Geophys Astrophys Fluid Dyn 98:85–127
45. Majda AJ, Biello JA (2003) The nonlinear interaction of barotropic and equatorial baroclinic Rossby waves. J Atmos Sci 60(15):1809–18212
46. Boyd JP, Deng D, Chen QC, Gao S (2013) Applications of bivariate Fourier series for solving the Poisson equation in limited-area modeling of the atmosphere: higher accuracy with a boundary buffer strip discarded and an improved order-raising procedure. Monthly Weather Rev 141(11):4154–4164
47. Benney DJ (1977) General theory for interactions between short and long waves. Stud Appl Math 56(1):81–94
48. Ma YC (1978) Complete solution of long-wave-short-wave resonance equations. Stud Appl Math 59(3):201–221
49. Ma YC, Redekopp LG (1979) Some solutions pertaining to the resonant interaction of long and short waves. Phys Fluid 22:1872–1876

50. Boyd JP (2014) Solving transcendental equations: the Chebyshev polynomial proxy and other numerical rootfinders. Perturbation Series and Oracles. SIAM, Philadelphia
51. Medvedev SB, Zeitlin V (2005) Weak turbulence of short equatorial waves. Phys Lett A 342:217–227
52. Kindle J (1983) On the generation of Rossby solitons during El Nino. In: Nihoul JCJ (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 353–368
53. Williams GP, Wilson RJ (1988) The stability and genesis of Rossby vortices. J Atmos Sci 45:207–249
54. Boyd JP (1980) The nonlinear equatorial Kelvin wave. J Phys Oceangr 10:1–11
55. Reznik GM, Zeitlin V (2009) The interaction of free Rossby waves with semi-transparent equatorial waveguide - wave-mean flow interaction. Nonlinear processes in geophysics 16(3):381–392
56. Khoulde B, Majda AJ, Stechmann SN (2013) Climate science in the tropics: waves, vortices and PDEs. Nonlinearity 26(1):R1R68
57. Majda A (2003) Introduction to PDEs and waves for the atmosphere and ocean, vol 9. Courant lecture notes. American Mathematical Society, Rhode Island
58. Majda A, Wang X (2006) Nonlinear dynamics and statistical theories for basic geophysical flows. Cambridge University Press, Cambridge
59. Gerstner F (1802) Theorie der Wellen. Abhandlunger der Koniglichen Bohmischen Geselschaft der Wissenschaften (Prague) Reprinted in 1809
60. Gerstner F (1809) Theorie der Wellen samt einer daraus abgeleiteten Theorie der Deichprofile. Ann Phys 2:412–445
61. Rankine WJM (1863) On the exact form of waves near the surface of deep water. Phil Trans R Soc Lond 153:127–138
62. Dubreil-Jacotin ML (1932) Sur les ondes de type permanent dans les liquides heterogenes. Atti Acad Naz Lincei, Rend Classe Sci Fis Mat Nat 15(6):814–919
63. Yih CS (1966) Note on edge waves in a stratified fluid. J Fluid Mech 24(NC11):765–767
64. Craik ADD (2004) The origins of water wave theory. Ann Rev Fluid Mech 36:1–24
65. Henry D (2008) On Gerstners water wave. J Nonlinear Math Phys 15(Supplement 2):87–95
66. Constantin A (2001) On the deep water wave motion. J Phys A 34(7):1405–1417
67. Constantin A (2012) An exact solution for equatorially trapped waves. J Geophys Res 117(C05):029
68. Constantin A (2013) Some three-dimensional nonlinear equatorial flows. J Phys Oceangr 43(1):165–175
69. Constantin A (2014) Some nonlinear, equatorially trapped, nonhydrostatic internal geophysical waves. J Phys Oceangr 44(2):781–789
70. Constanin A, Germain P (2013) Instability of some equatorially trapped waves. J Geophy Res-Oceans 118(6):2802–2810
71. Constantin A, Johnson RS (2015) The dynamics of waves interacting with the Equatorial Undercurrent. Geophys Astrophys Fluid Dyn 109(4):311–358
72. Henry D (2015) Internal equatorial water waves in the f-plane. J Nonlinear Math Phys 22(4):499–506
73. Sastre-Gomez S (2015) Global diffeomorphism of the Lagrangian flow-map defining equatorially trapped water waves. Nonlinear Anal-Theory Methods Appl 125:725–731
74. Ionescu-Kruse D (2015) On Pollard's wave solution at the Equator. J Nonlinear Math Phys 22(4):523–530
75. Compelli A, Ivanov R (2015) On the dynamics of internal waves interacting with the equatorial undercurrent. J Nonlinear Math Phys 22(4):531–539
76. Kasahara A (1977) Numerical integration of the global barotropic primitive equations with Hough harmonic expansions. J Atmos Sci 34:687–701
77. Stiassnie M, Peregine DH (1980) Shoaling of finite-amplitude surfaces waves on wateer of slowly-varying depth. J Fluid Mech 97:783–805
78. Munk W (1949) The solitary wave theory and its application to surf problems. Ann New York Acad Sci 51(3):376–424

79. Keller JB (1948) The solitary wave and periodic waves in shallow water. Ann New York Acad Sci 51(3):345–350
80. Peregine DH (1983) Breaking of waves on beaches. Ann Rev Fluid Mech 15(1):149–178
81. Zhao Q, Fu ZT, Liu SK (2001) Equatorial envelope Rossby solitons in a shear flow. Adv Atmos Sci 18(3):418–428
82. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, Part I: theory and methods. J Atmos Sci 35:2236–2258
83. Boyd JP (1978) The effects of latitudinal shear on equatorial waves, Part II: applications to the atmosphere. J Atmos Sci 35:2259–2267
84. Boyd JP, Christidis ZD (1983) Instability on the equatorial beta-plane. In: Nihoul J (ed) Hydrodynamics of the equatorial ocean. Elsevier, Amsterdam, pp 339–351
85. Boyd JP (1984) Equatorial solitary waves, Part IV: Kelvin solitons in a shear flow. Dyn Atmos Oceans 8:173–184
86. Boyd JP (1985) Equatorial solitary waves, Part 3: modons. J Phys Oceanogr 15:46–54
87. Boyd JP, Moore DW (1986) Summability methods for Hermite functions. Dyn Atmos Oceans 10:51–62
88. Boyd JP, Christidis ZD (1987) The continuous spectrum of equatorial Rossby waves in a shear flow. Dyn Atmos Oceass 11:139–151
89. Boyd JP (1985) Barotropic equatorial waves: the non-uniformity of the equatorial beta-plane. J Atmos Sci 42:1965–1967
90. Boyd JP (1991) Monopolar and dipolar vortex solitons in two space dimensions. Wave Motion 57:223–243
91. Boyd JP (1998) High order models for the nonlinear shallow water wave equations on the equatorial beta-plane with application to Kelvin wave frontogenesis. Dyn Atmos Oceas 28(2):69–91
92. Boyd JP, Natarov A (1998) A Sturm-Liouville eigenproblem of the Fourth Kind: a critical latitude with equatorial trapping. Stud Appl Math 101:433–455
93. Natarov A, Boyd JP (2001) Beyond-all-orders instability in the equatorial Kelvin wave. Dyn Atmos Oceans 33(3):181–200
94. Natarov A, Boyd JP (2002) Shafer (Hermite-Padé) approximants for functions with exponentially small imaginary part with application to equatorial waves with critical latitude. Appl Math Comput 125:109–117
95. Boyd JP, Chen GY (2001) Weakly nonlinear wavepackets in the Korteweg-deVries equation: the KdV/NLS connection. Math Comput Simul 55(4–6):317–328
96. Boyd JP (2002) Deleted residuals, the QR-factored Newton iteration, and other methods for formally overdetermined determinate discretizations of nonlinear eigenproblems for solitary, cnoidal, and shock waves. J Comput Phys 179(1):216–237
97. Boyd JP (2005) The short-wave limit of linear equatorial Kelvin waves in a shear flow. J Phys Oceangr 35(6):1138–1142
98. Boyd JP, Zhou C (2008) Uniform asymptotics for the linear Kelvin wave in spherical geometry. J Atmos Sci 65(2):655–660
99. Boyd JP, Zhou C (2008) Kelvin waves in the nonlinear shallow water equations on the sphere: nonlinear traveling waves and the corner wave bifurcation. J Fluid Mech 617:185–205
100. Zhou C, Boyd JP (2013) Cross-equatorial structures of equatorially trapped nonlinear Rossby waves. Dyn Atmos Oceans 64(1):53–61
101. Cane MA, Sarachik ES (1981) The response of a linear baroclinic equatorial ocean to periodic forcing. J Marine Res 39:652–693

Appendix A
Hermite Functions

Abstract Hermite functions play such a central role in equatorial dynamics that it is useful to collect information about them from a variety of sources. Hille-Watson-Boyd convergence and rate-of-convergence theorems, a table of explicit formulas for the Hermite coefficients of elementary functions, Hermite quadrature and integral representations for the Hermite functions and so on are included. Another table lists numerical models employing Hermite functions for oceanographic and meteorological applications. Recurrence formulas are provided not only for the normalized Hermmite functions themselves, but also for computing derivatives and products of powers of y with Hermite functions. The Moore-Hutton series acceleration and Euler acceleration for slowly-converging Hermite series are also explained.

Just because there's an exact formula doesn't mean it's necessarily a good idea to use it.

— Lloyd N. "Nick" Trefethen, FRS

A.1 Normalized Hermite Functions: Definitions and Recursion

The normalized Hermite functions ψ_n have the orthogonality property that

$$\int_{-\infty}^{\infty} \psi_n(y)\psi_m(y)\,dy = \delta_{mn} \tag{A.1}$$

where δ_{mn} is the usual Kronecker δ, which is one when $m = n$ and zero otherwise. The Hermite functions can be efficiently computed using a three-term recursion relation from two starting values:

$$\psi_0(y) \equiv \pi^{-1/4}\exp(-(1/2)y^2); \qquad \psi_1(y) \equiv \pi^{-1/4}\sqrt{2}\;y\;\exp(-(1/2)y^2) \tag{A.2}$$

$$\psi_{n+1}(y) = \sqrt{\frac{2}{n+1}}\;y\,\psi_n(y) - \sqrt{\frac{n}{n+1}}\,\psi_{n-1}(y) \tag{A.3}$$

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0

Table A.1 Normalization factors for the Hermite functions
The normalization factor is $d_n = \pi^{1/4}\, 2^{n/2}\, \sqrt{n!}$ where the orthonormal Hermite functions are related to the unnormalized Hermite functions a $\psi_n(y) = \exp(-[1/2]y^2)\, H_n(y)/d_n$ where

n	$d_n = \pi^{1/4}\, 2^{n/2}\, \sqrt{n!}$	$1/d_n$
0	1.331335	0.751126
1	1.882792	0.531126
2	3.765585	0.265563
3	9.223761	0.108416
4	26.088736	0.038330

The unnormalized Hermite polynomials H_n are ill-conditioned in numerical applications, but have coefficients which are integers (Table A.1).

The relationship between normalized and unnormalized Hermite functions is given in Table A.1.

$$
\begin{aligned}
\psi_0 &= \pi^{-1/4} \exp(-(1/2)y^2) \\
\psi_1 &= \frac{\sqrt{2}}{\pi^{1/4}}\, y\, \exp(-(1/2)y^2) \\
\psi_2 &= \left\{ \sqrt{2}\, y^2 - \frac{1}{\sqrt{2}} \right\} \pi^{-1/4} \exp(-(1/2)y^2) \\
\psi_3 &= \left\{ \sqrt{4/3}\, y^3 - \sqrt{3} y \right\} \pi^{-1/4} \exp(-(1/2)y^2) \\
\psi_4 &= \left\{ \sqrt{2/3}\, y^4 - \sqrt{6} y^2 + \frac{\sqrt{6}}{4} \right\} \pi^{-1/4} \exp(-(1/2)y^2) \\
\psi_5 &= \left\{ \frac{2}{15}\sqrt{15}\, y^{15} - \frac{2}{3}\sqrt{15} y^3 + \frac{\sqrt{15}}{2} y \right\} \pi^{-1/4} \exp(-(1/2)y^2)
\end{aligned}
\tag{A.4}
$$

The lowest few $\psi_n(y)$ are graphed in Fig. A.1.

The Hermite functions satisfy recursion relations which are very useful in applications such as

$$
y\, \psi_n = \sqrt{(n+1)/2}\, \psi_{n+1} + \sqrt{n/2}\, \psi_{n-1} \tag{A.5}
$$

The first derivative can be computed by either of the recurrences

$$
\frac{d\psi_n}{dy} = -\sqrt{(n+1)/2}\, \psi_{n+1} + \sqrt{n/2}\, \psi_{n-1}, \qquad n \geq 0 \tag{A.6}
$$

$$
\frac{d\psi_n}{dy} = -y\, \psi_n + \sqrt{2n}\, \psi_{n-1} \qquad n \geq 0 \tag{A.7}
$$

These must be initialized by the derivatives of ψ_0 and ψ_1 which are

$$\frac{d\psi_0}{dy} = -\frac{1}{\sqrt{2}}\,\psi_1; \tag{A.8}$$

$$= -y\psi_0 \tag{A.9}$$

$$= -y\,\pi^{-1/4}\exp(-(1/2)y^2) \tag{A.10}$$

For the derivative of ψ_1, an explicit formula is easier than the recurrences

$$\frac{d\psi_1}{dy} = \sqrt{2}\,\left\{1 - y^2\right\}\,\psi_0 \tag{A.11}$$

$$= \sqrt{2}\pi^{-1/4}\exp(-(1/2)y^2)\,\left\{1 - y^2\right\} \tag{A.12}$$

A.2 Raising and Lowering Operators

The similarity of the results of the operations of differentiation and of multiplication by y make it possible to define so-called "raising" and "lowering" operators. The motive for these names is self-explanatory in that the raising operator $\mathfrak{R}$, when applied to a Hermite function of degree n, gives a result which is proportional to the Hermite function of the next highest degree. Similarly, the lowering operator $\mathfrak{L}$ reduces the degree of a Hermite function by one.

$$\mathfrak{R} \equiv (d/dy - y) \qquad \text{["raising operator"]} \tag{A.13}$$

$$\mathfrak{R}\,\psi_n = -\sqrt{2(n+1)}\,\psi_{n+1} \tag{A.14}$$

$$\mathfrak{L} \equiv (d/dy + y) \qquad \text{["lowering operator"]} \tag{A.15}$$

$$\mathfrak{L}\,\psi_n = \sqrt{2n}\,\psi_{n-1} \tag{A.16}$$

The Hermite eigenoperator

$$\mathfrak{H} \equiv \frac{d^2}{dy^2} - y^2; \qquad \mathfrak{H}\psi_n = -(2n+1)\psi_n \tag{A.17}$$

can be written in terms of the raising and lower operators as

$$\mathfrak{H} \equiv \frac{1}{2}\,\{\mathfrak{R}\,\mathfrak{L} + \mathfrak{L}\,\mathfrak{R}\} \tag{A.18}$$

The operations of differentiation and multiplication by y can also be expressed in terms of the raising and lowering operators:

$$y\psi_n = \frac{1}{2}\left(\mathfrak{L} - \mathfrak{R}\right)\psi_n \tag{A.19}$$

$$= \sqrt{n/2}\,\psi_{n-1} + \sqrt{(n+1)/2}\,\psi_{n+1} \tag{A.20}$$

$$\frac{d}{dy}\psi_n = \frac{1}{2}\left(\mathfrak{L} + \mathfrak{R}\right) \tag{A.21}$$

$$= \sqrt{n/2}\,\psi_{n-1} - \sqrt{(n+1)/2}\,\psi_{n+1} \tag{A.22}$$

A.3 Integrals of Hermite Polynomials and Functions

The identity for derivatives can be interpreted as a recurrence for the integrals of the Hermite functions

$$\int^y dy'\psi_0 = \sqrt{\sqrt{\pi}}\,\frac{1}{\sqrt{2}}\mathrm{erf}\left(\frac{1}{\sqrt{2}}\,y\right) \tag{A.23}$$

$$\int^y dy'\psi_1 = -\frac{\sqrt{2}}{\sqrt{\sqrt{\pi}}}\exp\left(-\frac{1}{2}\,y^2\right) \tag{A.24}$$

$$\int^y dy'\psi_2 = -\psi_1 + \sqrt{\frac{1}{2}}\int^y dy'\psi_0(y') \tag{A.25}$$

$$= -\psi_1 + \frac{\pi^{1/4}}{2}\mathrm{erf}(y/\sqrt{2}) \tag{A.26}$$

$$\int^y dy'\psi_3 = -\sqrt{2/3}\psi_2 + \sqrt{2/3}\int^y dy'\psi_1(y') \tag{A.27}$$

$$= -\sqrt{2/3}\psi_2 - \frac{2}{\sqrt{3}}\psi_0 \tag{A.28}$$

$$\int^y dy'\psi_{n+1} = -\sqrt{\frac{2}{n+1}}\,\psi_n + \sqrt{\frac{n}{n+1}}\int^y dy'\psi_{n-1}(y') - \tag{A.29}$$

The unnormalized Hermite polynomials H_n obey simpler identities

$$\int_0^x H_n(y)dy = \frac{1}{2(n+1)}\left(H_{n+1}(x) - H_{n+1}(0)\right) \tag{A.30}$$

$$\int_0^x \exp(-y^2)H_n(y)dy = H_{n-1}(0) - \exp(-x^2)\,H_{n+1}(x) \tag{A.31}$$

A.4 Integrals of Products of Hermite Functions

Integrals of products of three or more Hermite functions are ubiquitous in the weakly nonlinear theory of equatorial waves, and sometimes to calculate projections in linear theory, too. Generalizing the work of Busbridge [1] and Azor, Gillis and Victor [2], Ripa [3] gave asymptotic formulas both for the Hermite functions themselves and for the integral of the product of J Hermite functions with a Gaussian of arbitrary width [3]. Because of their complexity, these formulas are not repeated here.

A.5 Higher Order and Symmetry-Preserving Recurrences

From these two fundamental recursions for ψ_n and its first derivative, a number of higher order identities can be obtained. Examples are

$$\psi_{n,yy} = \left\{ -(2n+1) + y^2 \right\} \psi_n \tag{A.32}$$

$$\psi_{n,yyy} = \left\{ -(2n+1) + y^2 \right\} \psi_{n,y} + 2\, y\, \psi_n \tag{A.33}$$

The Hermite functions of a given symmetry can be generated by a recurrence that connects only functions of that symmetry:

$$\sqrt{2}\psi_2 = \left\{ y^2 - 1 \right\} \psi_0 \tag{A.34}$$

$$\sqrt{6}\psi_3 = \left\{ y^2 - 3 \right\} \psi_1 \tag{A.35}$$

$$\sqrt{(n+1)(n+2)}\psi_{n+2} = \left\{ y^2 - (2n+1) \right\} \psi_n - \sqrt{n(n-1)}\, \psi_{n-2}, \qquad n \geq 2 \tag{A.36}$$

where the starting values $\psi_0(y)$ or $\psi_1(y)$ are given by Eq. A.2.

This same recurrence can also be re-written as an expression for $y^2\, \psi_n$:

$$y^2\, \psi_n = \frac{1}{2}\sqrt{(n+1)(n+2)}\psi_{n+2} + \left(n + \frac{1}{2} \right) \psi_n + \frac{1}{2}\sqrt{n(n-1)}\, \psi_{n-2}, \qquad n \geq 2 \tag{A.37}$$

This is useful in deriving Hermite–Galerkin discretizations of differential equations.

Similarly

$$\begin{aligned} y^4 \psi_n &= 4\sqrt{(n+1)(n+2)(n+3)(n+4)}\psi_{n+4} + 4(n+3/2)\sqrt{(n+1)(n+2)}\psi_{n+2} \\ &+ 3(1+2n+2n^2)\psi_n + 2(2n-1)\sqrt{(n-1)n}\, \psi_{n-2} + \sqrt{(n-3)(n-2)(n-1)n}\psi_{n-4}, \qquad n \geq 4 \end{aligned} \tag{A.38}$$

The special cases are

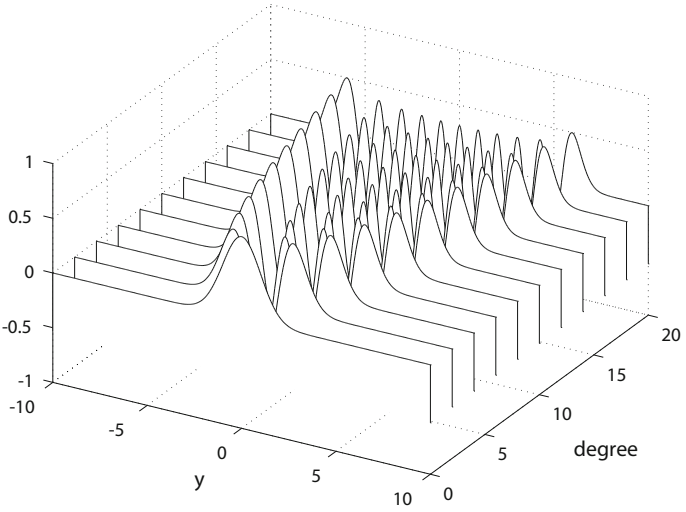

Fig. A.1 Hermite functions $\psi_n(y)$ versus y and degree n. The functions have been scaled so that $\max(\psi_n)(y) = 1$

$$y^4\psi_n = 4\sqrt{(n+1)(n+2)(n+3)(n+4)}\psi_{n+4} + 4(n+3/2)\sqrt{(n+1)(n+2)}\psi_{n+2} + 3(1+2n+2n^2)\psi_n, \qquad n = 0, 1 \tag{A.39}$$

$$y^4\psi_n = 4\sqrt{(n+1)(n+2)(n+3)(n+4)}\psi_{n+4} + 4(n+3/2)\sqrt{(n+1)(n+2)}\psi_{n+2} + 3(1+2n+2n^2)\psi_n + 2(2n-1)\sqrt{(n-1)n}\,\psi_{n-2}, \qquad n = 2, 3 \tag{A.40}$$

Note that the special cases are the general case if we define all Hermite functions of negative degree to be identically equal to zero (Fig. A.1).

A.6 Unnormalized Hermite Polynomials

These are defined by the starting values

$$H_0 = 1; \qquad H_1 = 2\,y \tag{A.41}$$

and the recursion

$$H_{n+1} = 2\,y\,H_n \;-\; 2\,n\,H_{n-1} \tag{A.42}$$

These satisfy the differentiation law

$$H_{n,y} = 2n\, H_{n-1} \tag{A.43}$$

and the orthogonality integral

$$\int_{-\infty}^{\infty} \exp(-y^2)\, H_m\, H_n\, dy = \begin{cases} \pi^{1/2}\, 2^n\, n! & [m = n] \\ 0 & [m \neq n] \end{cases} \tag{A.44}$$

$$\begin{aligned} H_0 &= 1 \\ H_1 &= 2y \\ H_2 &= 4y^2 - 2 \\ H_3 &= 8y^3 - 12y \\ H_4 &= 16y^4 - 48y^2 + 12 \\ H_5 &= 32y^5 - 160y^3 + 120y \end{aligned} \tag{A.45}$$

Because many formulas are simpler when expressed in terms of unnormalized Hermite polynomials — note the absence of square roots and π from (A.41–A.43) — they have been listed here. However, the unnormalized functions increase very rapidly with N as shown by (A.44) and Table A.3, so we shall use only the normalized Hermite functions ψ_n in the rest of this work (Table A.2).

$$\psi_n(y) \equiv \exp(-[1/2]y^2) H_n(y) \frac{1}{\pi^{1/4}\, 2^{n/2} \sqrt{n!}} \tag{A.46}$$

Equivalently,

$$\exp(-[1/2]y^2)\, H_n(y) = \pi^{1/4}\, 2^{n/2}\, \sqrt{n!} \psi_n(y) \tag{A.47}$$

A.7 Zeros of Hermite Series

Theorem A.1 (Hermite Companion Matrix) *Let $f_N(x)$ denote a polynomial of degree N written in "Hermite form" as*

$$f_N(x) = \sum_{j=0}^{N} a_j\, H_j(x) \tag{A.48}$$

or a function that is a truncated series of Hermite functions,

$$f_N(x) = \sum_{j=0}^{N} a_j \, \psi_j(x) \tag{A.49}$$

where $\psi_j(x) = \exp(-[1/2]x^2)\, H_j(x)$. *(Note that the Gaussian factor that is part of the definition of the Hermite functions has no effect on the zeros of the sum of the series.) Then all roots of* $f_N(x)$*, both on and off the canonical expansion interval,* $x \in [-\infty, \infty]$*, are eigenvalues of the* $N \times N$ *matrix* H *whose elements are*

$$H_{jk} = \begin{cases} (1/2)\delta_{2,k}, & j = 1, \quad k = 1, 2, \ldots, N \\ (j-1)\delta_{j,k+1} + \frac{1}{2}\delta_{j,k-1}, & j = 2, \ldots, (N-1), \\ (-1)\frac{1}{2}\frac{a_{j-1}}{a_N} + (N-1)\delta_{k,N-1}, & j = N \end{cases} \tag{A.50}$$

where δ_{jk} *is the usual Kronecker delta function.*

For a quintic polynomial,

$$\begin{vmatrix} 0 & (1/2) & 0 & 0 & 0 \\ 1 & 0 & (1/2) & 0 & 0 \\ 0 & 2 & 0 & (1/2) & 0 \\ 0 & 0 & 3 & 0 & (1/2) \\ (-1)\frac{a_0}{2a_5} & (-1)\frac{a_1}{2a_5} & (-1)\frac{a_2}{2a_5} & (-1)\frac{a_3}{2a_5} + 4 & (-1)\frac{a_4}{2a_5} \end{vmatrix} \tag{A.51}$$

[4, 5].

A.8 Zeros of Hermite Functions

The roots of the n-th Hermite function are important because they are the quadrature points for the Hermite–Gauss numerical integration scheme (next section) and also are the collocation points for Hermite function pseudospectral schemes for solving differential equations. $\psi_n(y)$ has n real zeros on the interval spanned by the turning points, $y \in [-\sqrt{2n+1}, \sqrt{2n+1}]$.

The *average* separation between zeros on this interval is asymptotically (as $n \to \infty$)

$$\text{average}_j(y_{j+1}^{root} - y_j^{root}) \sim 2^{3/2}/\sqrt{n} \tag{A.52}$$

$$\min_j(y_{j+1}^{root} - y_j^{root}) \sim \frac{\pi}{\sqrt{2}}\frac{1}{\sqrt{n}} \tag{A.53}$$

The average is thus $(4/\pi)$ greater than the minimum.

The zeros may be numerically calculated by specializing the Hermite companion matrix to $a_N = 1$ while all other $a_j = 0$. Then all roots of $\psi_N(x)$ are eigenvalues of the $N \times N$ tridiagonal matrix H whose elements are

$$H_{jk} = \begin{cases} (1/2)\delta_{2,k}, & j = 1, \quad k = 1, 2, \ldots, N \\ (j-1)\delta_{j,k+1} + \frac{1}{2}\delta_{j,k-1}, & j = 2, \ldots, N \end{cases} \tag{A.54}$$

where δ_{jk} is the usual Kronecker delta function.

For a quintic polynomial,

$$\begin{vmatrix} 0 & (1/2) & 0 & 0 & 0 \\ 1 & 0 & (1/2) & 0 & 0 \\ 0 & 2 & 0 & (1/2) & 0 \\ 0 & 0 & 3 & 0 & (1/2) \\ 0 & 0 & 0 & 4 & 0 \end{vmatrix} \tag{A.55}$$

[4, 5].

A.9 Gaussian Quadrature

A.9.1 Gaussian Weighted

$$\int_{-\infty}^{\infty} \exp(-y^2) g(y) \approx \sum_{j=1}^{N} \mathfrak{w}_j \, g(y_j) \tag{A.56}$$

$$\mathfrak{w}_j = \frac{2^{N-1} N! \sqrt{\pi}}{N^2 [H_{N-1}(y_j)]^2} \tag{A.57}$$

$$= \frac{1}{N \exp(y_j^2) [\psi_{N-1}(y_j)]^2} \tag{A.58}$$

A.9.2 Unweighted Integrand

$$\int_{-\infty}^{\infty} f(y) \approx \sum_{j=1}^{N} w_j \, f(y_j) \tag{A.59}$$

$$w_j = \exp(y_j^2)\varpi_j \tag{A.60}$$

$$= \exp(y_j^2)\,\frac{2^{N-1}N!\sqrt{\pi}}{N^2[H_{N-1}(y_j)]^2} \tag{A.61}$$

$$= \frac{1}{N[\psi_{N-1}(y_j)]^2} \tag{A.62}$$

A.10 Pointwise Bound on Normalized Hermite Functions

Theorem A.2 (Bound on Hermite Functions)

$$|\,\psi_n(y)\,| \;\le\; 0.816 \tag{A.63}$$

(From Abramowitz and Stegun [6], p. 787.)

Equation (5.2) of [7] shows that $|\psi_n(y)|$ is asymptotically no smaller than $0.63n^{-1/12}$, so the bound (A.63) is a tight one in the sense that it overestimates the maximum of $\psi_n(y)$ by at most a factor of 2 for all $n \le 200$. The bound also proves the following:

Theorem A.3 (Hermite Truncation Error)

$$|\,f(y) \;-\; \sum_{n=0}^{N} a_n\,\psi_n(y)\,| \;\le\; 0.816 \sum_{n=N+1}^{\infty} |a_n| \tag{A.64}$$

A.11 Asymptotic Approximations

A.11.1 Interior Approximations

"Cosine Approximations": $n >>,\ |y| < \sqrt{2n+1}$

$$\psi_{2n}(y) \sim (-1)^n\,\frac{\sqrt{(2n)!}}{\pi^{1/4}2^n\,n!}\,\cos\left\{\sqrt{4n+1}\,y\right\}\,\frac{1}{\left\{1-y^2/(4n)\right\}^{1/4}} \tag{A.65}$$

$$\sim (-1)^n 0.56419\,n^{-1/4}\,\cos\left(2\sqrt{n}\,y\right)\,\frac{1}{\left\{1-y^2/(4n)\right\}^{1/4}} \tag{A.66}$$

$$\psi_{2n+1}(y) \sim (-1)^n \frac{\sqrt{(2n)!}}{\pi^{1/4} 2^n n!} \sin\left\{\sqrt{4n+3}\, y\right\} \frac{1}{\left\{1 - y^2/(4n)\right\}^{1/4}} \quad \text{(A.67)}$$

$$\sim (-1)^n 0.56419\, n^{-1/4} \sin\left(2\sqrt{n}\, y\right) \frac{1}{\left\{1 - y^2/(4n)\right\}^{1/4}} \quad \text{(A.68)}$$

Phase errors accumulate with increasing $|y|$ because the local wavenumber is not $\sqrt{2n+1}$ but $\sqrt{2n+1-y^2}$ as incorporated into the WKB method>

"WKB Approximations": $n >>,\ |y| < \sqrt{2n+1}$

$$\psi_{2n}(y) \sim (-1)^n 0.56419\, n^{-1/4} \cos\left(\sqrt{2n+1}\, \mathfrak{P}(y/\sqrt{2n+1})\right) \frac{1}{\left\{1 - y^2/(4n)\right\}^{1/4}}$$

$$\psi_{2n+1}(y) \sim (-1)^n 0.56419\, n^{-1/4} \sin\left(\sqrt{2n+1}\, \mathfrak{P}(y/\sqrt{2n+1})\right) \frac{1}{\left\{1 - y^2/(4n)\right\}^{1/4}}$$

where

$$\mathfrak{P}(Z) \equiv \int_0^Z dx \sqrt{1-x^2} \quad \text{(A.69)}$$

$$= (1/2) Z\sqrt{1-Z^2} + (1/2)\arcsin(Z) \quad \text{(A.70)}$$

$$= Z - (1/6)Z^3 - (1/40)Z^5 - (1/112)Z^7 - (5/1152)Z^9 - (7/2816)Z^1 1 + \cdots,\ |Z| \le 1$$

$$= \pi/4 - (\sqrt{2}2/3)(1-Z)^{3/2} + (\sqrt{2}/10)(1-Z)^{5/2} + (\sqrt{2}/112)(1-Z)^{7/2} + (\sqrt{2}/576)(1-Z)^{9/2} + (5\sqrt{2}/11264)(1-Z)^{11/2},\ |1-Z| \le 2 \quad \text{(A.71)}$$

Note that these interior approximations are most accurate for small $|y|$, and fail very close to and beyond the turning points, $|y_t| \approx \sqrt{2n+1}$. The coefficients of the cosine and sine factors are the exact values of $\psi_{2n}(0)$.

A.11.2 Airy Approximation Near the Turning Points

The Hermite functions change from oscillatory behavior, as described in the previous subsection, to monotonic, exponential decay at the "turning points",

$$y_t \approx \sqrt{2n+1} \quad \text{(A.72)}$$

In the vicinity of the turning points, the Hermite functions are approximated in terms of the Airy function $\mathrm{Ai}(z)$ as

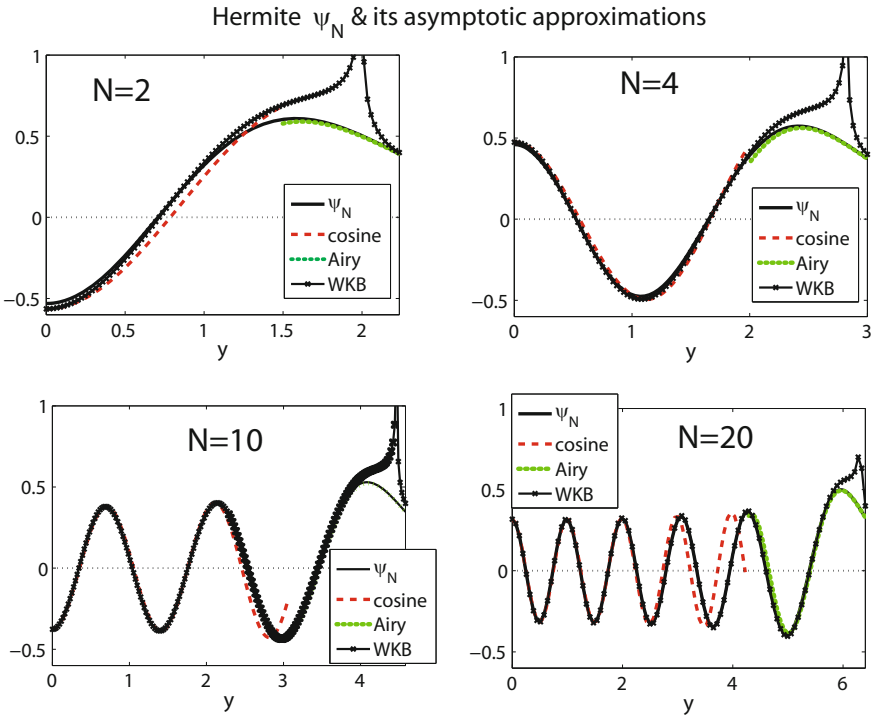

Fig. A.2 Asymptotic approximations to Hermite functions of various degrees N. The horizontal scale extends to the turning point for each mode. Only the positive interval $y \in [0, \sqrt{2N+1}]$ is graphed because Hermite functions are always either symmetric or antisymmetric with respect to the origin, and thus the graph for negative y would be the mirror image of that for positive latitude. (Mirror image with a sign flip when N is odd.) The cosine approximation is exactly what the name implies except for a slow amplitude modulation by a factor of $\left\{1 - y^2/(4n)\right\}^{-1/4}$. The approximation is very accurate for small y, but develops a phase error as y increases. The WKB approximation almost perfectly captures this phase variation, and breaks down only very close to the maxima of ψ_N closest to the turning point. The Airy approximation is indistinguishable from the Hermite function near the turning point, but becomes inaccurate for small y. In the upper left hand panel, $N = 2$, which is hardly a "large" value of N, but the asymptotic approximations are very good anyway

$$\psi_n(y) \sim 1.260\,(2n+1)^{-1/12}\,\mathrm{Ai}\left\{\mathrm{sign}(y)\,1.260\,(2n+1)^{1/6}\,|y - y_t|\right\} \qquad \text{(A.73)}$$

Uniformly valid approximations, valid not only near the turning points but in wider domains, can be found in the handbooks [6, 8]. These are of much greater complexity than those presented here, alas. However, Fig. A.2 shows that the simpler non-uniform approximations are extremely accurate. This in turn implies that the conceptual model of Hermite-as-sine-wave between the turning points is sound.

A.12 Convergence Theory

Boyd's book [9] contains a thorough discussion of numerical use of Hermite functions. Here, we give a brief summary of how the rate of convergence of a Hermite series depends upon the analytical properties of the solution. In general, both the rate at which the function decays along the real y axis and also the location of singularities of $f(y)$ in the complex y-plane affect the asymptotic behavior of the Hermite coefficients. Some preliminary definitions are helpful. Note that the limits in the convergence theorems are usually supremum limits.

Definition A.1 (*Exponential order of real axis decay*) The "exponential order of real axis decay" k is the least upper bound of j for which

$$f(y) = O\left(e^{-p|y|^j}\right) \tag{A.74}$$

for some constant p as $|y| \to \infty$ along the real axis. The function is said to be "Sub-Gaussian" or "Super-Gaussian" respectively if

$$k < 2 \qquad \text{["Sub-Gaussian"]} \tag{A.75}$$

$$k > 2 \qquad \text{["Super-Gaussian"]} \tag{A.76}$$

Definition A.2 (*Algebraic order of real axis decay*) The algebraic order of real axis decay j is the least upper bound of j for which

$$f(y) = O\left(1/|y|^j\right) \tag{A.77}$$

as $|y| \to \infty$ along the real axis. Note that $j = \infty$ if the exponential order of decay is greater than zero.

Definition A.3 (*Algebraic order of convergence*) The Algebraic order of convergence q is the largest number for which

$$\lim_{n\to\infty} \mid a_n \mid n^q < \infty, \qquad n \gg 1 \tag{A.78}$$

where the a_n are the coefficients of the series. (For a Fourier series, the limit must be finite for both the cosine coefficients a_n and the sine coefficients b_n.)

Alternative definition: if the coefficients of a series are a_n and if

$$a_n \sim O[1/n^q], \qquad n \gg 1 \tag{A.79}$$

then q is the algebraic order of convergence.

Definition A.4 (*Infinite order convergence*) If the algebraic order of convergence k is unbounded – in other words, if the coefficients a_n decrease faster than $1/n^k$ for

ANY finite power of k – then the series is said to have the property of "infinite order", "exponential", or "spectral" convergence.

Alternative definition: If

$$a_n \sim O[\exp(-qn^r)], \qquad n \gg 1 \tag{A.80}$$

with q a constant for some $r > 0$, then the series has infinite order or "exponential" convergence.

The equivalence of the second definition to the first is shown by the identity

$$\lim_{n\to\infty} n^k \exp(-qn^r) = 0, \qquad \text{all k, all r} > 0 \tag{A.81}$$

([6], p. 68). The reason for giving two definitions is that (A.80), which is more obvious and easier to understand, does not cover all possible cases. The terms "exponential" and "infinite order" are synonyms and may be used interchangeably.

Definition A.5 (*Exponential index of convergence*) The "exponential index of convergence" r is given by

$$r \equiv \lim_{n\to\infty} \frac{\log \mid \log(\mid a_n \mid) \mid}{\log(n)} \tag{A.82}$$

An equivalent definition is that if s and $q > 0$ are constants and

$$a_n \sim O(s \exp[-qn^r]), \qquad n \gg 1, \tag{A.83}$$

then the "exponential index of convergence" is the exponent r.

If the algebraic order of decay j is finite, then the Hermite coefficients will decay algebraically with n also, and have a finite algebraic order of convergence. In general, each increase of j by one will increase the algebraic order of convergence q by 1/2. A simple integration-by-parts arguments shows that

$$q \le (j/2) - 1 \tag{A.84}$$

if $f(y)$ has no singularities on the real y-axis, but this bound may be too conservative; the functions that give various fields in the Yoshida jet satisfy $q = j/2 + 1/4$ as explained in [7]. References to convergence theory are cataloged in Chap. 17, Sect. 4, of [9]. The most important point about algebraically-converging Hermite series is that for small q they are almost useless without a summability method as explained in the next section.

Theorem A.4 (Hille's Hermite Width-of-Convergence Strip)

(i) The domain of convergence of a Hermite series is the infinite strip *about the real y-axis bounded by*

$$|Im(y)| = w \tag{A.85}$$

(ii) For an entire *function, that is, $f(y)$ which has no singularities except at infinity, the width is*

$$w = \begin{cases} \infty & \text{if } k > 1 \\ 0 & \text{if } k < 1 \end{cases} \tag{A.86}$$

(iii) For singular *functions, that is, functions with poles and/or branch points at a finite distance off the real axis, let τ denote the absolute value of the imaginary part of the location of that singularity which is closest to the real y-axis, and let p denote the constant in the definition of real axis decay. Then*

$$w = \begin{cases} \tau & \text{if } k > 1 \\ \text{smaller of } [p, \tau] & \text{if } k = 1 \\ 0 & \text{if } k < 1 \end{cases} \tag{A.87}$$

Theorem A.5 (Hermite Rate-of-Convergence)
(i) If the strip of convergence has a finite width w, then

$$a_n \sim O\left(e^{-w\sqrt{2n+1}}\right), \tag{A.88}$$

that is, subgeometric convergence with exponential convergence index $r = 0.5$.
(ii) For entire functions with order of real axis decay p [i. e., functions decaying as fast as $\exp(-constant|y|^p)$ *as $|y| \to \infty$], the exponential convergence index r is*

$$\begin{aligned} r &= p/(2[p-1]) && \text{(Super-Gaussian, that is, } p > 2) \\ r &= p/2 && \text{(Sub-Gaussian, that is, } p < 2) \end{aligned} \tag{A.89}$$

(iii) Geometric convergence is possible only for entire functions and only for these two cases:

1. *$f(y)$ has $p = 2$ [Gaussian decay], that is, decay along the real axis proportional to* $\exp(-Qy^2)$ *where Q need not equal 1/2]*
2. *$f(y)$ is Super-Gaussian ($p > 2$) and the scaling factor α is varied with the truncation N as described in Boyd [7] where the scaled basis is the set of $\psi_n(\alpha y)$ where $\alpha > 0$ is independent of y.*

(iv) If the function $f(y)$ decays ALGEBRAICALLY *with y so that*

$$f(y) \sim O\left(1/|y|^{\delta}\right) \qquad \text{as} \qquad |y| \to \infty \tag{A.90}$$

and if the function has p continuous derivatives on $[-\infty, \infty]$ (with the (p+1)-st derivative continuous except at a finite number of points), then

$$a_n \sim O\left(\frac{1}{n^{q/2}}\right) \tag{A.91}$$

where

$$q = \textit{smaller of } [p + 3/2,\ \delta - 5/6]. \tag{A.92}$$

Equation (A.91) only gives a lower bound on the convergence rate; it is known that the Hermite coefficients for $f(y) \equiv 1$ converge as $O(n^{-1/4})$ even though (A.91) can guarantee only that the coefficients will diverge no faster than $n^{5/12}$.

Part (i) was proved by Hille [10–13], building on earlier work by G.N. Watson, (ii) and (iii) by Boyd [7], and (iv) by Bain [14]). Hille gives additional theorems not quoted here in [15, 16].

Several conclusions emerge from this convergence theory. When a spectral series — an ordinary Fourier or Chebyshev expansion — is used on a finite interval, the coefficients decrease at least as fast as a geometric series ($r \geq 1$) unless the function is singular or discontinuous on the expansion interval itself. This is not true for Hermite series except for the special case that (i) $f(y)$ is entire *and* (ii) the function decays as a Gaussian. Thus, the popular choice of $\exp(-by^2)$ as a factor in wind stresses and models of mean currents is eminently sensible on at least two accounts: (i) the Gaussian is one of the handful of functions for which the Hermite coefficients are known analytically and (ii) it belongs to the very special class of functions whose Hermite expansions converge geometrically fast.

The second conclusion is that Hermite functions without series acceleration and summability methods are likely to be a very poor numerical tool except for problems on the open ocean with forcings which decay exponentially as $|y| \to \infty$. If the forcing decays only algebraically with y, the coefficients of the Hermite expansion will decay only algebraically. Even if the forcing is localized, the Hermite coefficients will still have (small!) finite algebraic order of convergence in the presence of eastern or western boundaries since the coastal Kelvin waves that run up and down the coast, generated by reflection of equatorially trapped waves, are not themselves equatorially trapped and decay only algebraically away from the equator.

This has not prevented successful use of Hermite series to solve many problems, but it is usually necessary to employ summability methods as discussed in the next section.

A.13 Abel–Euler Summability, Moore's Trick, and Tapering

When the function $f(y)$ decays only algebraically with $|y|$, the convergence of the Hermite series is extremely slow, and thousands of terms may be necessary to obtain an accurate answer. Unfortunately, coastal Kelvin waves always decay as $O(1/|y|^{1/2})$, so any problem with boundaries will require special tricks.

Boyd and Moore discuss a family of such tricks. Define the n-th partial sum formally as

$$S_N \equiv \sum_{n=0}^{N} a_n \psi_n(y), \tag{A.93}$$

The goal is to compute the limit

$$S \equiv \lim_{n\to\infty} S_N \tag{A.94}$$

During his doctoral studies, Dennis Moore observed that such slowly converging series typically have *alternating partial sums*. If S_{N-1} undershoots the sum, then $S_N > S$. It follows that the *average* of these two partial sums,

$$T_N \equiv (S_{N-1} + S_N)/2 \qquad \text{[Once-averaged Partial Sums]} \tag{A.95}$$

is a better approximation to S than either S_{N-1} or S_N. (In the literature of sequence acceleration schemes, this averaging is called "Hutton's" Method.)

The once-averaged partial sums T_N also alternated in Moore's equatorial application. This suggested a second round of averaging:

$$V_N \equiv (T_{N-1} + T_N)/2 \qquad \text{[Twice-averaged Partial Sums]} \tag{A.96}$$

$$= \sum_{n=0}^{N-2} a_n\,\psi_n(y) + (3/4)\,a_{N-1}\,\psi_{N-1}(y) + (1/4)\,a_N\,\psi_N(y) \tag{A.97}$$

Applying the averaging four times gives

$$S_N \equiv \sum_{n=0}^{N-4} a_n\,\psi_n(y) + (15/16)\,a_{N-3}\,\psi_{N-3}(y) + (11/16)\,a_{N-2}\,\psi_{N-2}(y)$$
$$+ (5/16)\,a_{N-1}\,\psi_{N-1}(y) + (1/16)\,a_N\,\psi_N(y) \qquad \text{[Moore's Averaging Method]} \tag{A.98}$$

Morse and Feshbach (Chap. 4 of [17]) give a very good discussion of repeated averaging in the context of summing the asymptotic expansion for the exponential integral.

Boyd later observed that repeating the averaging process until all available terms are exhausted is "Abel–Euler Summability", also known as "Euler acceleration of series".

Abel's inspiration is to observe that the series

$$f(y) = \sum_{n=0}^{\infty} a_n\,\psi_n(y) \tag{A.99}$$

is a special case ($\zeta = 1$) of the more general function

$$f(y;\zeta) = \sum_{n=0}^{\infty} \zeta^n\,a_n\,\psi_n(y) \tag{A.100}$$

If the a_n are decreasing very slowly (algebraically) with n, this implies that, regarded as a power series in ζ, $f(y;\zeta)$ has a radius of convergence of unity. That is, for $|\zeta| < 1$, the series converges exponentially (not algebraically) fast whereas for $|\zeta| > 1$ the sum diverges to infinity exponentially fast. The original sum, $f(y;\zeta = 1)$, thus corresponds to evaluating the series at a point on its radius of convergence. This in turn implies that $f(y;\zeta)$ must have a singularity — that is, a branch point or pole such that all sufficiently high order ζ-derivatives of $f(y;\zeta)$ are infinite at the singularity — on the circle $|\zeta| = 1$ in the complex ζ-plane. However — and this is very, very important — the fact that the original series $f(y)$ is *alternating* implies that the singularity must be at $\zeta = -1$, that is, on the side of the circle of convergence directly *opposite* the point where we want to evaluate $f(y;\zeta)$, $\zeta = 1$.

The reader is entitled to ask: So what? The answer is that because $f(y;\zeta)$ is a power series in ζ, one may attack the sum (A.100) using every tool in the arsenal of complex variable theory for "analytically-continuing" a power series beyond its radius of convergence. In particular, one may invoke a second idea: Euler's transformation of a power series. Euler introduced a new variable (conformal mapping of ζ to z) via

$$z \equiv \frac{2\zeta}{\zeta + 1} \qquad \Leftrightarrow \qquad \zeta = \frac{z}{2 - z} \tag{A.101}$$

and then showed that the first N terms of the ζ-series could be rearranged into the first N terms of a power series in z merely through linear recursion. Since the singularity at $\zeta = -1$ is mapped to $z = \infty$, the point where we must sum $f(y;\zeta(z))$ to recover the original series (A.99), $\zeta = 1$, is well within the radius of convergence of the transformed series in z. Conformal mapping in general and its great value in physics and engineering are reviewed in [18, 19].

Boyd and Moore [20] explain how this transformation of power series is equivalent to repeated averaging of partial sums. They also present numerical examples to illustrate the great effectiveness both of Moore's four-fold averaging and the exponentially-convergent Euler acceleration method.

Table A.2 collects the weights for averagings of various orders. The fourth row contains the weights that Moore applied in his thesis to the last four terms of each series.

The n-th Hermite function is exponentially small for $|y| > \sqrt{2n+1}$. Consequently, any finite truncation of the Hermite sum of a function like $f(y) = 1$ is necessarily non-uniform in space. The N-term truncated (but unaccelerated) sum will be corrupted by oscillations for small y. Figure A.3 shows that Moore–Hutton weighting can purge the spurious oscillations for small y. The error is non-uniform in latitude: acceleration fails near the "cliff" where the highest terms in the sum reach their turning latitudes and then fall exponentially for larger $|y|$. This "cliff" is located roughly at $\sqrt{2N_{highest}+1}$ where $N_{highest}$ is the degree of the Hermite function of highest degree retained in the truncation. (Most Hermite sums in the various theories presented earlier are sums over even degree Hermite only or over odd degree only, in which case $N_{highest}$ is roughly double the number of terms N in the series.) One can

Table A.2 Hutton/Euler coefficient weights w_{Mj}; the M-th row gives the weights for an M-fold Hutton averaging. A series with term $\sum_{n=0}^{N} a_n$ is accelerated by being evaluated as $\sum_{n=0}^{N-M-1} a_n + \sum_{j=0}^{M} w_{M,j} a_{N-M+j}$. The user may choose any $M \le N$

M	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
0	1	–	–	–	–	–	–	–	–
1	1	1/2	–	–	–	–	–	–	–
2	1	3/4	1/4	–	–	–	–	–	–
3	1	7/8	1/2	1/8	–	–	–	–	–
4	1	15/16	11/16	5/16	1/16	–	–	–	–
5	1	31/32	13/16	1/2	3/16	1/32	–	–	–
6	1	63/64	57/64	21/32	11/32	7/64	1/64	–	–
7	1	127/128	15/16	99/128	1/2	29/128	1/16	1/128	–
8	1	255/256	247/256	219/256	163/256	93/256	37/256	9/256	1/256

obtain a good approximation for any finite range in latitude by taking a sufficiently large number of terms in the series and applying the acceleration.

A.14 Alternative Implementation of Euler Acceleration

The phase speed for Rossby waves can be expanded in a power series in k^2 derived by substituting

$$c = \sum_{j=0}^{\infty} c_{2j} k^{2j} \tag{A.102}$$

into the cubic equation

$$c^3 - (\nu + k^2)c - 1 = 0 \tag{A.103}$$

where

$$\nu \equiv 2n + 1 \tag{A.104}$$

The cubic polynomial has a triple point when $k^2 = -\nu$, so the series converges only for $k \le \sqrt{2n+1}$.

The radius of convergence can be greatly extended, probably to the entire positive real k-axis, by replacing k^2 by the new small parameter ð[1] and expanding in powers of ð.

[1] This symbol is the old Anglo-Saxon letter "eth" and is part of the standard American Mathematical Society symbol set.

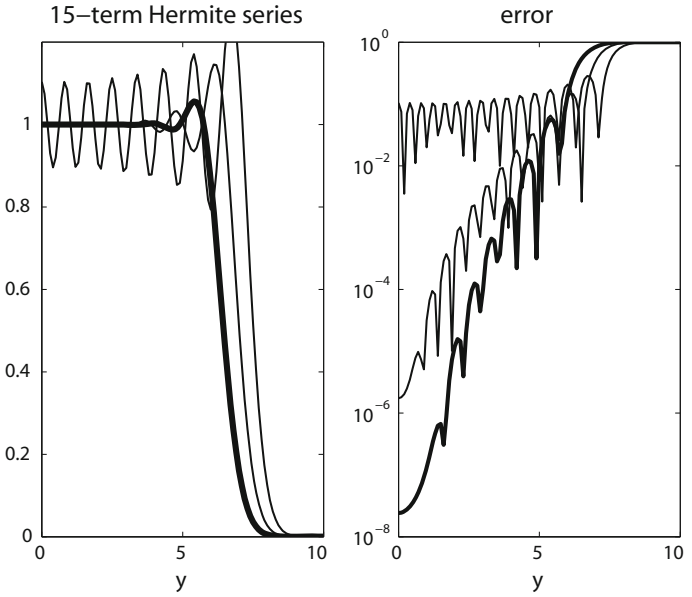

Fig. A.3 The errors in a fifteen term Hermite approximations to the "trivial function" $f(y) \equiv 1$, both with and without series acceleration. All odd coefficients are zero, so the sums include terms in $\{\psi_0, \psi_2, \ldots \psi_{28}\}$ only. All Hermite functions in a sum of the first N even functions (up to and including $\psi_{2N-2}(y)$) are exponentially small for $y > \sqrt{4N-3}$. The *top error curve* is that of the unaccelerated series. The *middle curve* applies Moore's thesis trick, which is weighting the last four terms of the series by multiplying the terms by 15/16, 11/16, 5/16 and 1/16. The *thick curve* (*bottommost error curve*) applies eight weights to the last eight terms of the series

$$k^2 = \eth \frac{\nu}{1-\eth} \qquad \Leftrightarrow \qquad \eth = \frac{k^2}{k^2+2n+1} \tag{A.105}$$

This parameter replacement is often simpler than weighting the original power series as described in the preceding section.

A.15 Tapering

A second, alternative trick is what we shall call "Gaussian tapering". For the Yoshida jet, the slow convergence arises because the wind stress is a constant at all latitudes (and therefore non-decaying in $|y|$). "Tapering" means multiplying the wind stress by $\exp(-\Delta y^2)$ to force exponential decay in $|y|$. If $\Delta \ll 1$, the wind stress will differ little from one except at high latitudes, but the Hermite series will convergence exponentially fast.

Boundaries generate coastal Kelvin waves, alas. However, Cane and Moore (1981) have shown in the latitudinal geostrophy (a. k. a. "long wave" or "low frequency") approximation, the sum of the Kelvin mode and its boundary reflection can be calculated in analytical closed form, so Hermite expansions are needed only to calculate the "particular solution", which does not involve the boundaries. Consequently, the series of papers by Cane and Sarachik have employed this artifice of the "Gaussian-tapered" wind stress extensively: for the particular solution, any stress that decays exponentially fast for large $|y|$ will yield exponentially convergent Hermite series. Although other forms of tapering could be used, the Gaussian is simple and particularly well-suited to Hermite expansions as discussed in the section on convergence theory, Sect. A.12.

A.16 Hermite Functions on a Finite Interval

Remarkably, Hermite functions can also be applied successfully on a finite interval. Since this is a very active research frontier, but not directly relevant to equatorial dynamics, we leave the topic to [21–26].

A.17 Hermite–Galerkin Numerical Models

Many articles in tropical oceanography and meteorology employ Hermite series. In Table A.4, we have collected a selection of articles that are primarily numerical, or pay a lot of attention to numerical issues in building geophysical models.

Hermite functions have not conquered the world because the Fast Fourier Transform (FFT) cannot be used to manipulate them with blazing speed as true of Chebyshev polynomials and Fourier. In domains that are not the tensor-product of one-dimensional intervals, global spectral methods can only be applied with difficulty. For multiply-connected oceans, Chebyshev single-domain methods have been replaced by Legendre polynomial spectral elements. There are no Hermite spectral elements.

Even so, the simplicity of Hermite spectral methods and the close connections between basis functions and physics will likely ensure that Hermite number-crunchers never entirely disappear. As reviewed in [9], Hermite functions are very popular in plasma physics, especially for the Vlasov equation, and in quantum mechanics, where Hermite functions are the eigenfunctions of the quantum harmonic oscillator.

Table A.3 Hermite Function Series
For completeness, some expansions are given in more than one way. Equations marked with CS use the functions α_{2n+1} introduced by Cane and Sarachik (1976). Equations marked with A denote the asymptotic form of the coefficients as $n \to \infty$. Note that most expansions involve only even degree Hermite functions (symmetric about $y = 0$) or odd degree Hermite functions; often the form of coefficient is for a_{2n} rather than a_n, so this is clearly marked. $f(y) = \sum_{n=0}^{\infty} a_n \psi_n(y)$ $D_n(y)$ denotes the usual parabolic cylinder function. The numbers $\alpha_{2n+1} \equiv \sqrt{2n+1}\sqrt{(2n+1)!}/(2^n n!)$ are auxiliary coefficients used extensively by Cane and Sarachik

$f(y)$	a_n
$\exp\left\{-[1/2]y^2 + i\sqrt{2}by\right\}$	$a_n = \pi^{1/4} \exp(-[1/2]b^2) \frac{i^n b^n}{\sqrt{n!}}$
1	$a_{2n} = \pi^{1/4}\sqrt{2}\,\frac{\sqrt{(2n)!}}{2^n n!}$ $= \pi^{1/4}\sqrt{2}\,\alpha_{2n+1}/\sqrt{2n+1}$ CS $\sim 1.682\,(2n)^{-1/4}$ $[n \to \infty]$ A
y	$a_{2n+1} = \pi^{1/4}\,\frac{\sqrt{(2n+1)!}}{2^{n-1}\,n!}$ $= 2\,\pi^{1/4}\,\alpha_{2n+1}/\sqrt{2n+1}$ CS $= 2\pi^{1/4}\Lambda_n$ $\sim 2.378\,(2n)^{1/4}$ $[n \to \infty]$ A
$\exp(-b^2\,y^2)$	$a_{2n} = \pi^{1/4}\,\frac{\sqrt{2}}{\sqrt{1+2b^2}}\,\frac{\sqrt{(2n)!}}{2^n\,n!}\left\lvert\frac{1-2b^2}{1+2b^2}\right\rvert^n$ $\sim 1.682\frac{\sqrt{2}}{\sqrt{1+2b^2}}\,(2n)^{-1/4}\left\lvert\frac{1-2b^2}{1+2b^2}\right\rvert^n$ $[n \to \infty]$ A
$\exp\left(-[1/2][(1-c^2)/(1+c^2)]y^2\right)$	$a_{2n} = \pi^{1/4}\sqrt{1+c^2}\,\frac{\sqrt{(2n)!}}{2^n\,n!}\,c^{2n}$ $= \pi^{1/4}\sqrt{1+c^2}\,\alpha_{2n+1}\,(2n+1)^{-1/2}\,c^{2n}$ CS $\sim \sqrt{1+c^2}\,n^{-1/4}\,c^{2n}$ $[n \to \infty]$ A
$y\,\exp\left(-[1/2][(1-c^2)/(1+c^2)]y^2\right)$	$a_{2n+1} = \pi^{1/4}(1+c^2)^{3/2}\,\frac{\sqrt{(2n+1)!}}{2^{n+1/2}\,n!}\,c^{2n}$ $= \pi^{1/4}\,2^{-1/2}\,(1+c^2)^{3/2}\,\alpha_{2n+1}\,c^{2n}$ CS $\sim 0.8409\,(1+c^2)^{3/2}\,(2n)^{1/4}\,c^{2n}$ $[n \to \infty]$ A
$\exp(-[1/2]y^2)/(b^2+y^2)$	$a_{2n} =$ $(-1)^n\pi^{1/4}\,\frac{\sqrt{(2n)!}}{b\exp(-[1/2]b^2)2^{-[n+1/2]}}\,D_{-(2n+1)}\left\{b\sqrt{2}\right\}$ $\sim$ $(-1)^n\,\frac{1.7724}{b\exp(-[1/2]b^2)\,n^{1/4}}\,\exp\left\{-b\sqrt{4n+1}\right\}$ $[n \to \infty]$ A

A.18 Fourier Transform

The Hermite functions are (to within a factor of i^n) their own Fourier transform. Consequently, one can sometimes easily obtain the Hermite series of a function $f(y)$ by calculating the Hermite expansion of its Fourier transform $F(k)$ and then multiplying the coefficients by i^n. If we define the forward and inverse transforms via

Table A.4 Hermite function numerical models of the ocean and atmosphere

References	Comments
Anderson [27]	Time-dependent model of the tropical ocean
Tribbia [28, 29]	Hough–Hermite Galerkin method
Boyd and Moore [20]	Euler sequence acceleration of algebraically-converging Hermite series with applications to oceanography
Marshall and Boyd [30]	Equatorial ocean waves
Smith [31]	Hermite–Galerkin model of tropical ocean
Holvorcem [32], Holvorcem and Vianna [33, 34], Vianna and Holvorcem [34]	Hermite Summation methods including bivariate Green's functions; integral equations
Tse and Chasnov [35]	Hermite for vertical coordinate Convection in unstable layer surrounded by unbounded stable layers; Fourier in x, y
Meiron, Moore and Pullin [36]	Fourier-Hermite basis for two-dimensional Stuart vortex arrays
Majda and Khoulder [37]	Atmospheric applications
Shamir and Paldor [38]	Global nonlinear model exploiting that spherical harmonics have Hermite asymptotic approximation

$$F(k) \equiv (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(iky)\, f(y) dy, \tag{A.106}$$

$$f(y) \equiv (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(-iky)\, F(k)\, dk \tag{A.107}$$

then

$$i^n\, \psi_n(k) \equiv (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(iky)\, \psi_n(y) dy, \tag{A.108}$$

Thus, the Hermite functions are, to within i^n, their own Fourier transforms.

For the Hermite functions of degree $4n$, $i^{4n} = 1$ so that $\psi_{4n}(y)$ is "self-reciprocal" with respect to the Fourier Transform. Conversely, a function which is Fourier Transform self-reciprocal has a Hermite series expansion involving only the Hermite functions whose degree is an integer multiple of four.

When the argument of the Hermite polynomials is changed from y to $y/\sqrt{2}$, the transforms are modified as follows

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(ky) \exp\left(-\frac{y^2}{2}\right) H_{2n}\left(\frac{y}{\sqrt{2}}\right) = (-1)^n 2^n k^{2n} \exp\left(-\frac{k^2}{2}\right) \tag{A.109}$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sin(ky) \exp\left(-\frac{y^2}{2}\right) H_{2n+1}\left(\frac{y}{\sqrt{2}}\right) = (-1)^n 2^{n+1/2} k^{2n+1} \exp\left(-\frac{k^2}{2}\right) \tag{A.110}$$

These transforms are essential to the Hermite Distributed Filter (HDAF) for accelerating slowly convergent Fourier series and Chebyshev polynomial expansions invented by Jared Tanner [39].

A.19 Integral Representations

$$\psi_n(y) = \frac{2^{1+n/2}}{\sqrt{n!}\,\pi^{3/4}} \exp([1/2]y^2) \int_{-\infty}^{\infty} \exp(-t^2)\, t^n \cos\{2yt - (n/2)\pi\}\, dt \tag{A.111}$$

Source: Abramowitz and Stegun [6], p. 785.

$$\psi_n(y) = \frac{2^{n/2}}{\sqrt{n!}\,\pi^{3/4}} \exp(-[1/2]y^2) \int_{-\infty}^{\infty} \exp(-t^2)\, (y + it)^n\, dt \tag{A.112}$$

Source: Gradshteyn and Rhyzik (1965), p. 1033.

$$\psi_n(y) = \frac{\sqrt{n!}}{i\, 2^{1+n/2}\ \pi^{5/4}} \exp(-[1/2]y^2) \oint_C \frac{\exp(-z^2 + 2zy)}{z^{n+1}}\, dz \tag{A.113}$$

where C is a circle of radius a in the complex z-plane where a is positive but otherwise arbitrary. Source: Abramowitz and Stegun [6].

Appendix B
Expansion of the Wind-Driven Flow in Vertical Modes

Abstract The wind-driven ocean circulation has a convergent expansion in vertical normal modes even though all modes individually have zero wind stress at the surface. This appendix explains that the mismatch between homogeneous Neumann boundary conditions on the eigenmodes and inhomogeneous boundary conditions on the series sum does have observable consequences (the sum over the barotropic mode and the baroclinic modes converges slowly) but not fatal consequences, which would be non-convergence.

> Why do our ideas about the ocean circulation have such a peculiarly dreamlike quality?
>
> — Henry Stommel, title of a pamphlet published by Stommel on his own printing-press in the 1950s, said to be highly influential among ocean theorists although never published in another form, as noted on pp. xvi–xvii of [40].

The eigenfunctions of any Sturm–Liouville eigenproblem must satisfy *homogenous* boundary conditions; a differential equation boundary value problem ceases to be an eigenproblem whenever the boundary conditions and/or the differential equation itself are inhomogeneous.

Consistent with this, the eigenfunctions of the vertical structure equation satisfy homogeneous Neumann boundary conditions. Each baroclinic mode is stress-free. However, the wind-driven ocean circulation is a huge and essential part of ocean flow. Nonzero wind stress implies nonzero vertical derivative at the sea surface, directly contradicting the property of all baroclinic modes of zero vertical derivative and therefor zero stress at the surface. A more precise statement is the following:

Definition B.1 (*Wind Stress Boundary Condition Paradox*) Denote the components of the wind stress vector at the sea surface by (τ^x, τ^y) and let ν denote the viscosity. The upper boundary condition is then

$$u_z = \tau^x/\nu; \qquad v_z = \tau^y/\nu \tag{B.1}$$

However, the eigenfunctions (baroclinic modes) of the vertical structure equation individually satisfy no-surface-stress boundary conditions, that is, homogeneous Neumann boundary conditions at the ocean surface:

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0

$$\mathfrak{z}_{m,z} = 0 \qquad \text{at } z = 0 \tag{B.2}$$

The paradox is: How can an infinite series of Sturm–Liouville eigenfunctions sum to non-zero surface wind stress when the vertical derivative at the sea surface is individually zero for each term in the series?

Obviously, it is impossible for the normal mode expansion to accurately represent the derivatives of u and v right at the surface. However, the theory of spectral expansions shows that a mis-match between the $u(z)$ and the expansion functions at a single point need not prevent the expansion from converging everywhere else.

The best way to show this is through a specific example. When the static stability is constant, the baroclinic eigenfunctions are $\cos(n\pi z/D)$ where n is the vertical mode number and D the ocean depth. (The depth-independent constant in the eigenfunction series is the barotropic mode.) When the flow is steady and variations in the Coriolis parameter are ignored, we obtain the classical Ekman solution

$$v(z) = b\ \exp(-\alpha z)\ \sin(\alpha z - \pi/4) \tag{B.3}$$

where α and b are constants. Since our interest in this appendix is solely in the mathematical properties of the expansion of $v(z)$ as a Fourier cosine series, we shall normalize by setting $b = 1$, $D = \pi$. Then

$$v(z) = (1/2)a_0 + \sum_{n=1}^{\infty} a_n\ \cos(nz) \tag{B.4}$$

where

$$a_n = (2/\pi) \int_0^{\pi} v(z) \cos(nz) dz \tag{B.5}$$

$$= \frac{1}{\sqrt{2}\,\pi}\, n \left\{ \frac{1}{\alpha^2 + (\alpha + n)^2} - \frac{1}{\alpha^2 + (\alpha - n)^2} \right\} \tag{B.6}$$

$$\sim 0.900\alpha\, n^{-2}, \qquad n \to \infty \tag{B.7}$$

and where we have neglected terms from the bottom ($z = -\pi$).

The $O(n^{-2})$ convergence rate shows the heavy price that must be paid for the mis-match between the vertical modes and the wind-driven flow: the Fourier series converges very slowly. In a numerical model, one would switch to Chebyshev polynomials to obtain an expansion which converges exponentially fast in n. When solving problems via separation-of-variables in any branch of physics, however, this algebraic convergence with n is a common burden that must be accepted philosophically. The reward is that we can analyze the ocean on a mode-by-mode basis instead of having to immediately resort to three-dimensional numerical models.

Figure B.1 with α arbitrarily chosen as 3.5 show that the vertical modes expansion most emphatically does work. In the left panel of Fig. B.1, the sum of the first 50

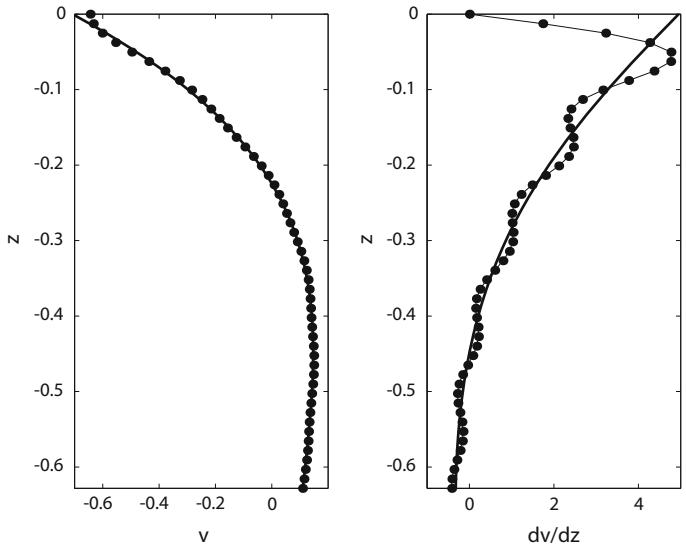

Fig. B.1 *Left panel* Comparison of the exact solution [*solid line*] for $v(z)$, the component of Ekman flow parallel to the wind stress, with the sum of the first 50 terms in its Fourier cosine series, which is marked by black disks. Note that only the upper quarter of the ocean is shown for clarity even though the expansion is accurate over the whole depth of the sea. *Right panel* same as the left except that the first derivative is illustrated. Note the very steep slope of the Fourier series (oscillating *solid line*) near $z = 0$, which corresponds to the surface of the ocean, and note also that the first peak overshoots the exact derivative of $v(z)$ (*smooth solid line*) by about 9%. These are the characteristic spoor of Gibbs' Phenomenon

vertical modes is almost indistinguishable from that of the exact Ekman flow in spite of the mis-match of the first derivatives at the surface. The right panel shows that the Fourier series for the first derivative also converges below the surface even though the sum of the series is zero to all orders at $z = 0$ where the exact first derivative has its largest value.

The derivative series is composed only of sine functions, and these are antisymmetric with respect to $z = 0$. Consequently, the Fourier sine series for the derivative converges to a function whose value jumps from -5 for $z < 0$ to $+5$ for $z > 0$. This jump discontinuity in the first derivative, although it has only a mathematical existence — the flow fields are always well-behaved and smooth — causes the coefficients of the derivative series to decrease only as $O(1/n)$. Near the surface $z = 0$, the first peak of the derivative series overshoots the exact derivative by about 9% — the famous Gibbs' Phenomenon [9, 41]. As more terms are added to the series, this 9% overshoot remains, but the first peak below the surface moves upward as N increases, closer and closer to the surface — and the series converges more and more accurately to $u(z)$ at all lower depths.

The assumption of constant N^2 is poor for the actual ocean, but this is *mathematically* irrelevant: if we used the actual eigenfunctions of a realistic stratification, the

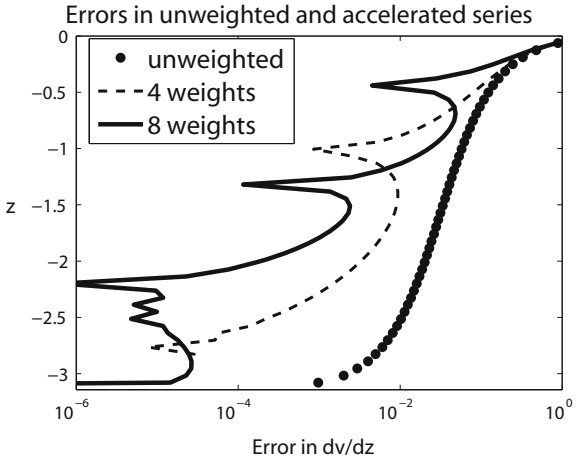

Fig. B.2 Errors in the approximation of dv/dz for the same series as shown in the *right panel* in the previous figure. Note that the vertical scale has been increased by a factor of five to show the entire depth of the ocean. The *dots* show the errors in the unaccelerated series. The *dashed curve* is the fourfold Hutton acceleration used by Moore. The *thick solid curve* is the eightfold Hutton acceleration; the last eight terms of the series are weighted by $\{255/256, 247/256, 219/256, 163/256, 93/256, 37/256, 9/256, 1/256\}$, respectively

general theory of Sturm–Liouville eigenfunctions shows that we would obtain the same $O(1/n^2)$ convergence for the coefficients and the same Gibbs' Phenomenon for the first derivative near $z = 0$.

Series acceleration methods have been very useful for speeding up the convergence of slowly-converging series as discussed in Appendix A. Acceleration schemes weight the last k terms for some user-chosen k. Moore showed the usefulness of the k-fold Hutton acceleration in his thesis.

In the limit that all N terms in the N-th partial sum are multiplied by acceleration weights, the iterated Hutton scheme becomes the Euler acceleration [20]. The good news is that very rapid, even exponential convergence can be recovered, but the improvement is spatially nonuniform. This is not a flaw for the Hermite series illustrated in Appendix A; the Hermite sums are only unaccelerated for large latitude y where the equatorial beta-plane is invalid anyway. For Hermite series, the Moore/Hutton/Euler acceleration is an unqualified success.

Unfortunately, the same is not true for eigenfunction series in depth. Figure B.2 shows that series acceleration can greatly reduce error in the depths, but there is *no improvement* near the *surface*.

Even so, the conclusion is that in spite of the paradox of mis-matched first derivatives at the surface, the expansion into vertical normal modes is a useful and successful tool even for flows with non-vanishing wind stresses provided only that we are careful to take enough terms in the expansion, and defeat the rather slow $O(1/n^2)$ convergence through a series acceleration method.

Vertical normal modes in the atmosphere are very different because the atmosphere is semi-infinite, not bounded, and the impedance mismatch at the tropopause reflects most wavelengths only weakly. Even the purely numerical use of vertical modes is not straightforward [42–44].

Appendix C
Potential Vorticity, Streamfunction and Nonlinear Conservation Laws

Abstract The nonlinear shallow water equations on the equatorial betaplane are blessed with valuable conservation laws. Potential vorticity conservation, among its other merits, justifies potential vorticity (PV) inversion but one must be careful: the Kelvin wave has zero potential vorticity and therefore is vorticity-invisible. The quasi-geostrophic streamfunction is replaced by an exact conservation law for the "mass streamfunction". Bernoulli's Principle is useful in the theory of equatorial modons and coherent structures. Pseudomomentum conservation has not yet seen extensive applications.

Proofs are to mathematics what spelling (or even calligraphy) is to poetry.

— Vladimir I. Arnold in *Mathematics: Frontiers and Perspectives*, edited by V.I. Arnold, M. Atiyah, P.D. Lax and B. Mazur. American Mathematical Society (2000).

C.1 Potential Vorticity

The nonlinear shallow water equations have a number of conservation laws (Ripa [48]).

Theorem C.1 (Potential Vorticity) *The potential vorticity for the equatorial shallow water equations is*

$$q \equiv \frac{v_x - u_y + y}{h} \tag{C.1}$$

Denote the dissipation and/or wind stress terms in the momentum equations by $\mathbf{F} = (F_1, F_2)$.

$$q_t + uq_x + vq_y = \frac{1}{h}\,\hat{z} \cdot curl\mathbf{F} \tag{C.2}$$

In the absence of dissipation and momentum forcing, the potential vorticity is conserved following the motion:

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0

$$Dq/Dt = 0 \tag{C.3}$$

where $D/Dt = \partial_t + u\partial_x + v\partial_y$ is the usual "total" or "particle-following" derivative. In other words, in the absence of dissipation and forcing, each blob of fluid preserves its value of q as it moves.

For small amplitude waves, the potential vorticity deviation from its mean value $q = y$ is given by

$$q' = v_x - u_y - y\phi \tag{C.4}$$

Although the relative vorticity

$$\zeta = v_x - u_y \tag{C.5}$$

is a sum of *three* Hermite functions for each mode, the potential vorticity for the n-th mode is proportional to a *single* Hermite function, $\psi_n(y)$.

C.2 Potential Vorticity Inversion

For middle latitude dynamics, potential vorticity q is a kind of alchemical elixir. "Potential vorticity inversion" is the jargon for deducing horizontal and vertical velocities, pressure, temperature and density from this single enchanted field. As shown by Hoskins, McIntyre and Robertson [45] and many sequels, the errors are small away from the equator even for three-dimensional flow.

Unfortunately, potential vorticity inversion has a serious flaw on the equatorial beta-plane even in the simple case of linear dynamics: the infinitesimal amplitude Kelvin wave has *zero* potential vorticity. Because Rossby modes are strongly rotational whereas the gravity waves are strongly divergent, but have comparatively little vorticity, one might legitimately hope to expand an observed distribution of linearized q in Hermite functions and then deduce the amplitude of each Rossby mode from this q-series under the assumption that the contributions of the gravity waves are negligible.

The bad news is that the Kelvin wave is *invisible* to potential vorticity. Because it is the lowest and simplest equatorial mode, the Kelvin wave probably is more strongly excited and more energetic than any other equatorial mode. Thus, a purely-q inversion scheme is doomed to failure.

However, because of its simple structure, the Kelvin wave can be estimated by taking the inner product of $\psi_0(y)$ with the nondimensional zonal velocity and height fields. The rest of the low-frequency motion can be inverted from the potential vorticity.

Unfortunately, the poor quality of oceanic data compared to meteorological observations has discouraged Kelvin-plus-q inversion. The potential vorticity requires *derivatives* of the horizontal velocity, and these are hard to obtain accurately from noisy, sparse data.

C.3 Mass-Weighted Streamfunction

C.3.1 General Time-Varying Flows

Two-dimensional flow is incompressible if and only if $u_x + v_y = 0$. One can then replace the two velocities by a single quantity, the streamfunction $\tilde{\psi}$, such that $u = -\tilde{\psi}_y$ and $v = \tilde{\psi}_x$. Unfortunately, the equation of continuity in the shallow water set is

$$h_t + (uh)_x + (vh)_y = 0 \tag{C.6}$$

where h is the total depth of a column of fluid. Because the free surface (or in the one-and-a-layer model interpretation, the thermocline) is free to move up and down, the divergence of the horizontal velocity is usually non-zero. However, it is possible to generalize the concept of a streamfunction as follows.

Definition C.1 (*Mass-Weighted Streamfunction*) The mass-weighted streamfunction for the shallow water equations is a function $\psi(x, y)$ such that

$$uh = -\psi_y - \int^x h_t(x', y)\, dx', \qquad vh = \psi_x \tag{C.7}$$

The reason for the name is that the derivatives of the mass-weighted streamfunction do not give the horizontal velocities, but rather quantities uh and vh which give the total mass transport in a fluid column.

Although h complicates the formulas, ψ *always* makes it possible to reduce the nonlinear shallow water equations to two equations (the momentum equations) in two unknowns (ψ, h). Unfortunately, the equatorial two-equation set is rather messy and has not been employed in any investigations to date.

C.3.2 Streamfunction for Steadily-Traveling Waves

For the special case of waves steadily-translating at a phase speed c, however, the mass-weighted streamfunction equations simplify to

$$uh = -\psi_y + c\,\phi, \qquad vh = \psi_x \qquad \text{[Steady-Translation Only]} \tag{C.8}$$

(Note that equatorially-trapped motions cannot be steadily-translating in any direction except longitude; a wave that propagates at an angle to either meridians or latitude circles will be *refracted* by variations in the Coriolis parameter and therefore cannot merely translate; a steady north-south velocity would eventually make the (un-refracted) disturbance a *non*-equatorial wave.)

C.4 Streakfunction

For steadily-traveling waves, it is convenient to define a modified streamfunction:

Definition C.2 (*Streakfunction*) The streakfunction $\Psi(x, y)$ is the streamfunction for flow in a coordinate system *moving* at the phase speed c so that the traveling wave is stationary in this reference frame. The streakfunction is defined by

$$\Psi \equiv \psi + cy \tag{C.9}$$

or equivalently by

$$(u - c)h = -\Psi_x\,, \qquad vh = \Psi_x \qquad \text{[Steady-Translation Only]} \tag{C.10}$$

The difference $(u - c)$ is the east-west velocity in the moving reference frame. Thus, the streakfunction allows the simple statement: the mass-weighted velocities in the moving reference frame are simply the y and x derivatives of the mass-weighted streakfunction.

One important distinction between ψ and Ψ is that for equatorially-trapped motions, ψ decays to zero as $y \to \infty$ whereas Ψ asymptotes to cy, and thus is not equatorially-confined.

The contours or isolines of the streamfunction and streakfunction are known as "streamlines" and "streaklines", respectively. In the moving reference frame, the velocity vector is tangent to a streakline at every point on the streakline.

Theorem C.2 (Streakfunction Properties)

1. *For a* steadily-translating *disturbance moving east-west at a constant phase speed c, q-conservation is equivalent to:*

   (a). The Jacobian of q and Ψ is zero:

   $$J(q, \Psi) = 0 \qquad \Leftrightarrow \qquad -\Psi_y\, q_x + \Psi_x\, q_y = 0 \tag{C.11}$$

   (b). The potential vorticity is a function of the streakfunction only, that is, for some function F,

   $$q = F(\Psi). \tag{C.12}$$

2. *If the disturbance is a "localized" disturbance which tends to zero at infinity, then on "open" streaklines, that is, isolines of Ψ that extend to spatial infinity where $u, v, \phi \to 0$,*

   $$q = \frac{\Psi}{c} \equiv F(\Psi) \qquad \textit{[Open Streaklines Only]} \tag{C.13}$$

Boyd [46].

As discussed in Chap. 17, at least one branch of Rossby solitary waves has $q = \Psi/c$ everywhere, and not merely on the exterior of the largest closed streakline. However, there is also at least one branch of solitons where the relationship between q and Ψ is nonlinear in the interior of the outermost closed streakline. The break in functional form for $q(\Psi)$ from $q = \Psi/c$ on the exterior to a nonlinear function of Ψ in the interior implies that the potential vorticity is not smooth, probably with a discontinuity in gradient on the outermost closed streakline.

The quantity $\mathscr{B}$ defined below is a close cousin of the usual Bernoulli function $h + (1/2)u^2 + (1/2)v^2$. The modified functional has some properties similar to those of the streakfunction, though it is a different quantity.

Theorem C.3 (Modified Bernoulli Function)

1. *The modified Bernoulli function*

$$\mathscr{B} \equiv \frac{1}{2}cy^2 + \phi + \frac{1}{2}\left([u-c]^2 - c^2 + v^2\right) \qquad \textit{[Modified Bernoulli Function]} \tag{C.14}$$

is exactly *conserved following the motion for a* steadily-translating *disturbance moving east-west at a constant phase speed c. $\mathscr{B}$-conservation is equivalent to:*

(a). The Jacobian of $\mathscr{B}$ and Ψ is zero:

$$J(\mathscr{B}, \Psi) = 0 \qquad \Leftrightarrow \qquad -\Psi_y \mathscr{B}_x + \Psi_x \mathscr{B}_y = 0 \tag{C.15}$$

(b). The modified Bernoulli function is a function of the streakfunction only, that is,

$$\mathscr{B} = G(\Psi) \tag{C.16}$$

for some function G. (Moore and Niler [47], note that $G(\Psi) = dF/d\Psi$ where $F(\Psi)$ is the streakfunction-vorticity functional of the previous theorem [47]).

2. *If the disturbance is a "localized" disturbance which tends to zero at infinity, then on "open" streaklines, that is, isolines of Ψ that extend to spatial infinity where $u, v, \phi \to 0$, $G(\Psi)$ must be quadratic*

$$\mathscr{B} = \frac{1}{2c}\Psi^2 \equiv G(\Psi) \qquad \textit{[Open Streaklines Only]} \tag{C.17}$$

Proofs and applications can be found in [46, 47].

C.5 The Streamfunction for Small Amplitude Traveling Waves

For small amplitude disturbances such that $uh \approx u$, $vh \approx v$, the streamfunction relationships (C.8) simplify to

$$u = -\psi_y + c\,\phi, \qquad v = \psi_x \qquad \text{[Steady-Translation Only]} \tag{C.18}$$

The definition of v in terms of ψ implies that the streamfunction must be the x-integral of v. It follows that for a traveling wave in the n-th latitudinal mode,

$$\psi(x, y, t) = A(x - ct)\,\psi_n(y) \tag{C.19}$$

for both gravity and Rossby waves.

$$v = \frac{dA(x - ct)}{dx}\,\psi_n(y)$$

By applying the recurrence relations for $y\psi_n(y)$ and $d\psi_n/dy$, one finds — with no approximations except for linearization about a state of rest —

$$u = A(x, t)\left(-\frac{d\psi_n}{dy} + \frac{1}{1 - c^2}\left\{-c^2\frac{d\psi_n}{dy} + c\,y\,\psi_n\right\}\right) \tag{C.20}$$

$$\phi = A(x, t)\left(y\,\psi_n + \frac{1}{1 - c^2}\left\{c^2 y\,\psi_n - c\,\frac{d\psi_n}{dy}\right\}\right) \tag{C.21}$$

This can be rewritten as

$$u = -\frac{d\psi(x, y, t)}{dy} + c\phi = -\frac{1}{1 - c^2}\frac{d\psi(x, y, t)}{dy} + c\frac{1}{1 - c^2}\,y\,\psi(x, y, t) \tag{C.22}$$

$$\phi = y\psi(x, y, t) + cu = \frac{1}{1 - c^2}y\,\psi(x, y, t) - c\frac{1}{1 - c^2}\frac{d\psi(x, y, t)}{dy} \tag{C.23}$$

These relationships apply to all species of equatorial waves, subject only to the approximation that $|A(x - ct)| \ll 1$ so that quadratic terms in A can be ignored. If the magnitude of the phase speed is *small* — which is true only for Rossby modes — then the expressions for u and ϕ can be approximated as

$$u \approx -A(x, t)\frac{d\psi_n}{dy} + O(c) = \frac{d\psi(x, y, t)}{dy} + O(c) \tag{C.24}$$

$$\phi \approx A(x, t)\,y\,\psi_n + O(c) = y\psi(x, y, t) + O(c) \tag{C.25}$$

which are the usual relationships between u and ϕ in quasi-geostrophic theory.

Unfortunately, the phase speed for the gravest ($n = 1$) Rossby mode is as large as $-1/3$ for waves of large zonal scale, and this phase speed is not particularly small. Furthermore, the streamfunction for the linearized Kelvin wave is identically zero (just like the north-south velocity of the Kelvin mode), so the streamfunction is a useless concept for this mode.

Even when c isn't small, and even for gravity waves [but not the Kelvin mode], it is always possible in *linearized* wave theory to deduce u, v and ϕ from ψ. However, only for small $|c|$ are these relationships the same as in quasi-geostrophic theory.

C.6 Other Nonlinear Conservation Laws

Ripa [48] gives three additional conservation laws as follows.

Theorem C.4 (Zonal Momentum Density) *The zonal momentum density per unit mass,*

$$\mathscr{M} \equiv h(u - (1/2)y^2) \tag{C.26}$$

solves

$$\mathscr{M}_t + (u\mathscr{M})_x + (v\mathscr{M})_y + \phi_x + \phi\,\phi_x = hF_1 \tag{C.27}$$

where F_1 denotes the sum of the wind stress and dissipation terms in the x-momentum equation.

Theorem C.5 (Energy Density) *The energy density,*

$$\mathscr{E} \equiv \frac{1}{2}\left\{hu^2 + hv^2 + \phi^2\right\} \tag{C.28}$$

solves

$$\mathscr{E}_t + \left\{u\left(\phi + (1/2)\phi^2 + \mathscr{E}\right)\right\}_x + \left\{v\left(\phi + (1/2)\phi^2 + \mathscr{E}\right)\right\}_y = (u\,F_1 + v\,F_2\,) \tag{C.29}$$

where F_1 and F_2 denotes the sums of the wind stress and dissipation terms in the x-momentum and y-momentum equations.

Theorem C.6 (Zonal Pseudomomentum) *The zonal pseudomomentum,*

$$\mathscr{P} \equiv u\phi - \frac{1}{2}\frac{\left(v_x - u_y - y\phi\right)^2}{h} \tag{C.30}$$

solves

$$\mathscr{P}_t + \{u\,(u + \mathscr{P})\}_x + \{v\,(u + \mathscr{P})\}_y + \frac{1}{2}\left\{\phi^2 - u^2 - v^2\right\}_x = u\,F_1 - \left(v_x - u_y - y\phi\right)\,\hat{z}\cdot curl(\mathbf{F}) \tag{C.31}$$

where $\mathbf{F} = (F_1, F_2)$ is the vector of the sum of the wind stress and dissipation terms in the x-momentum and y-momentum equations.

Conservation laws are essential to the instability criterion analyzed by Fruman and Shepherd [49].

Glossary

Like Adam, naming the birds and the flowers
— J.J. Sylvester [of his invention of many mathematical terms]

aeroarithmomancy "Numerical prediction of weather or climate". (From the Greek $\alpha\varepsilon\rho$o "air", $\alpha\rho\theta\mu o\sigma$, "number", and $\mu\alpha\nu\tau\varepsilon\iota\alpha$, "divination".)

anaster A planet which is not orbiting a star. (From the Greek "α without" and $\alpha\sigma\tau\varepsilon\rho$ "star".) Also known as a "rogue planet", "nomad planet", "free-floating planet", "starless planet", or "orphan planet". Also applied to free-floating objects whose identity (planet versus very dim brown dwarf) is uncertain.

antisymmetric If a function is such that $f(x) = -f(x)$ for all x, it is said to be "antisymmetric with respect to the origin" or to be of "odd parity".

arithmurgy Synonym for "number-crunching". (From the Greek $\alpha\rho\theta\mu o\sigma$, "number", and $-\varepsilon\rho\gamma o\sigma$, "working".)

backscatter (in turbulence) Nonlinear transfer of energy from high wavenumbers to low wavenumbers, also known as "upscaling".

basin resonance See **quasi-normal mode**.

basis set The collection of functions which are used to approximate the solution of a differential equation. The Fourier functions $\{1,\ \cos(nx),\ \sin(nx)$ for $n = 1,\ 2,\ \ldots\}$ and the Chebyshev polynomials $\{T_n(x),\ n = 0,\ 1,\ \ldots\}$ are two examples of basis sets. The elements of a basis set are "basis functions".

bell soliton A solitary wave which has a sech-like or "bell"-like shape such as the soliton of the KdV Equation.

beyond-all-orders Catch-all term for any effect or property whose magnitude is *exponentially* small in $1/\varepsilon$ where ε is a small parameter. Such terms as $\exp(-1/\varepsilon)$ are smaller than ε^n for *any* finite value of n as $\varepsilon \to 0$, and therefore cannot be calculated by a power series expansion in ε. Beyond-all-orders effects are invisible to power series in ε.

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0

beyond-all-orders perturbation theory Perturbative methods that calculate terms which are *exponentially* small in $1/\varepsilon$ where ε is the perturbation parameter [Synonym: "**hyperasymptotic perturbation theory**".]

bifurcation point A point where two branches of a solution $u(\alpha)$ cross. Synonym is "crossing point". For both $\alpha > \alpha_{\text{limit}}$ and $\alpha < \alpha_{\text{limit,}}$ two different solutions exist which meet at the bifurcation point. Newton's method fails at a bifurcation point, but it is possible to "shoot the bifurcation point" or switch branches. (When "bifurcation" is used in a broader sense, a "crossing point" is also called a "trans-critical bifurcation".)

breather A class of solitary wave in which the coherent structure has a standing wave oscillation superimposed upon its translational motion, if any. The function $u(x,t) = \text{sech}(\varepsilon x)\sin(\omega t)$ has the form of a breather.

chirugery Paper-and-pencil calculations, literally "handwork": A term popular in medicine from the Greek $\chi\varepsilon\iota\rho$, "hand", and $-\varepsilon\rho\gamma o\sigma$, "working".

cnoidal wave A spatially-periodic generalization of a solitary wave. The term was coined by Korteweg and deVries, who showed that the KdV equation had exact solutions that could be expressed in terms of the elliptic cosine function, whose abbreviation is "cn". ("-oid" is Greek for "like", so "cnoidal" means "cn-like".) By extension, this term is now applied to similar steadily-translating, spatially periodic solutions, regardless of whether these are described by the "cn" function or not.

collocation points The grid of points where the residual function must vanish. Synonym for "interpolation points".

compatibility conditions A countably infinite set of constraints on the initial conditions of a time-dependent partial differential equation which are necessary for the solution to be analytic everywhere in the space-time domain. Example: an incompressible flow will be strongly singular at $t = 0$ if the initial condition is divergent at some point in the fluid, and will have discontinuities (vortex sheets) at rigid walls if the initial velocity does not satisfy the no-slip boundary condition of vanishing at the walls. If the initial flow does not satisfy additional constraints, the flow will be singular, but the discontinuities are postponed to higher and higher derivatives as more and more of the infinite set of compatibility conditions are satisfied. The constraints are different for each partial differential equation.

completeness A basis set is "complete" for a given class of functions if all functions within the class can be represented to arbitrarily high accuracy as a sum of a sufficiently large number of basis functions.

Couette flow A mean flow which varies linearly between the boundaries. "Plane Couette" flow is the exact steady flow between two infinite plates, one stationary and the other moving parallel to it.

cryptid A conjectured phenomenon which is presently supported only by speculation or by observational evidence insufficient for refereed journal publication. (From zoology.)

cryptospectral An (apparently) non-spectral algorithm which incorporates spectral features, or is a hybrid of spectral methods with others. An example is the low-order polynomial interpolation employed as part of the semi-Lagrangian advection scheme in many spherical harmonic weather models: because the models are dealiased, the polynomial interpolation has near-spectral accuracy and is therefore "cryptospectral".

dipole vortex A pair of contra-rotating vortices, i. e., a cyclone and anticyclone, functioning dynamically as a single coherent structure.

direct matrix method A non-iterative algorithm for solving a matrix equation, as in the phrase "fast direct methods". Examples include Gaussian elimination, Crout reduction, Cholesky factorization, LU factorization, skyline solver, and static condensation.

double cnoidal wave An alternative name for the 2-polycnoidal wave, that is, a nonlinear wave which is independent of time except for steady translation due to two independent phase variables. In other words, $u(x,t) = u(X,Y)$ where $X \equiv k_1(x - c_1 t)$ and $Y = k_1(x - c_1 t)$ where k_1, k_2, c_1 and c_2.

downscaling One-way transfer of information from a global model to a higher resolution limited-area model; also used in a broader sense to describe energy transfer or information transfer from large scales to small. Also called "forward scattering".

envelope solitary wave A nonlinear wave packet which is the product of a spatially localized function ("envelope") that propagates at the group velocity c_g together with a sinusoidal factor ("carrier wave") which travels at the phase velocity c_p. The function $\mathrm{sech}(x - c_g t)\,\sin(k[x - c_p t])$ has the structure of an envelope solitary wave where the sech function is the envelope and the sine function is the carrier wave.

equivalent depth (i) In the one-and-a-layer model, the equivalent depth is the actual mean thickness of the upper layer multiplied by the fractional density difference between the two layers. (ii) With continuous stratification, Lamb's parameter ε, a nondimensional eigenvalue of linearized wave equations, can be converted to an "equivalent depth" H_{eq} via $H_{eq} = (4\Omega^2 a^2/g)\varepsilon^{-1}$ where g is the gravitational constant, $\Omega = (2\pi/84000)s^{-1}$ and a is the earth's radius.

FFT Abbreviation for Fast Fourier Transform.

fold point Synonym for **limit point**.

forward scatter (in turbulence) Nonlinear transfer energy from low wavenumbers to high wavenumbers. Also called **downscaling**. Dominant energy cascade in turbulence; see also **backscatter**.

geometric convergence A series whose coefficients are decreasing as

$$a_n \sim \alpha(n)\,p^n \leftrightarrow \alpha(n)\exp[-n\,|\log(p)|]) \qquad n \to \infty \qquad |p| < 1$$

has the property of "geometric convergence" where $\alpha(n)$ denotes a function that varies algebraically with n, rather than exponentially, such as a power of n. The

reason for the name is that the terms of a geometrically convergent series can always be bounded by those of a geometric series, that is, by the terms of the power series expansion of $a/(b+x)$ for some a and b where these are constants. (All convergent power series have geometric convergence. All Chebyshev series for functions which have no singularities on $x \in [-1,\ 1]$ (including the endpoints) also have geometric convergence).

Gibbs oscillations When a discontinuous function is approximated by a spectral series truncated at N terms, the sum of the truncated series visibly oscillates with a wavelength which is $O(1/N)$. These oscillations are also called **spectral ringing**.

Gibbs Phenomenon When a discontinuous function is approximated by a spectral series truncated at N terms, the sum of the truncated series overshoots the true value of the function at the edge of the discontinuity by a fixed amount (approximately 0.089 times the height of the jump), independent of N. Gibbs Phenomenon includes **Gibbs oscillations**.

Hough functions "Hough functions are the eigenfunctions, linearized about a state of rest, of Laplace's tidal equations, which are also labeled as the shallow water equations on the sphere. The name honors Sydney Samuel Hough, FRS (1870-1923), whose life is described in Dyson (1923) in Hough (1897, 1898). His name is pronounced "Huff".

hydroarithmomancy Synonym for "numerical prediction of ocean flow" or "computational fluid mechanics". (From the Greek $\upsilon\delta\rho$o "water", $\alpha\rho\theta\mu o\sigma$, "number", and $\mu\alpha\nu\tau\varepsilon\iota\alpha$, "divination".)

hyperasymptotics A form of beyond-all-orders perturbation theory in which the error is reduced below the **superasymptotic** error by appending one or more terms of a *second* asymptotic expansion with different scaling assumptions than those of the primary series. Loosely applied to all **beyond-all-orders** methods.

imbricate series A series for a *spatially periodic* function which is the superposition of an infinite number of evenly spaced identical copies of a "pattern" function $A(x)$. It may be shown that all periodic functions have imbricate series in addition to their Fourier expansions, and often the imbricate series converge faster. Imbricate series may be generalized to an arbitrary number of dimensions.

implicit (time-marching) A time-integration scheme in which it is necessary to solve a boundary-value problem to compute the solution at the new time level. The Crank–Nicolson method ("trapezoidal rule time-marching") is an example.

implicitly-implicit A label for a time-dependent partial differential equation in which the time derivative is multiplied by a differential operator L so that it is necessary to solve a boundary value problem at every time step even to apply an explicit time-marching scheme.

infinite order convergence A spectral series possesses the property of "infinite order convergence" if the error decreases faster than any finite inverse power of N as N,

the number of terms in the truncated series, increases. A synonym for "exponential convergence".

interpolant An approximation $P_{N-1}(x)$ whose free parameters or coefficients are chosen by the requirement that

$$f(x_i) = P_{N-1}(x_i) \qquad i = 1, \ldots, N$$

at a set of N grid points. The process of computing such an approximation is interpolation.

isola For a nonlinear equation that depends upon two parameters, α and τ, an "isola" is a one-parameter family of solutions, parameterized by some quantity s, such that $(\alpha(s),\ \tau(s))$ is a simple closed curve in the $\alpha - \tau$ plane. (From Latin "insula", "island", through Italian "isolato", "isolate".) Sometimes restricted to such a closed contour shrunk to a single point, an isolated solution.

KdV Equation The Korteweg-deVries equation, $A_t + aA_x + bAA_x + dA_{xxx} = 0$ where a, b and d are constants that depend on both the latitudinal and baroclinic mode numbers.

Lamb-Chaplygin dipoles. See **modon**.

kymology The study of waves. (From the Greek "$\kappa\upsilon\mu\alpha$", "wave").

limit point A point where a solution $u(\alpha)$ of a nonlinear equation curves back so that there are two solutions for α on one side of $\alpha = \alpha_{\text{limit}}$ and no solutions on the other side of the limit point. As the limit point is approached, $du/d\alpha \to \infty$. Special methods ["pseudoarclength continuation" or "globally convergent homotopy"] are needed to "turn the corner" and march from the lower branch through the limit point onto the upper branch or vice versa. Synonyms are "fold point", "turning point", and "saddle-node bifurcation".

localized, spatially A label applied to a wave or wave disturbance whose amplitude decreases rapidly as $|x| \to \infty$.

long wave approximation See **meridional geostrophy approximation**.

low frequency approximation See **meridional geostrophy approximation**.

meridional Of or pertaining to variations in the north-south direction. Synonym for "latitudinal".

meridional geostropy approximation Neglect of v_t in the linear shallow water equations or neglect of all terms except yu and ϕ_x and perhaps the wind stress in the nonlinear equations of motion. Also known as the "**long wave approximation**" or "**low frequency approximation**".

micropteron A weakly nonlocal solitary wave whose far field radiation is an *algebraic, non-exponential* function of the width of the core. (From the Greek $\mu\iota\kappa\rho o\sigma$, "small", and $\pi\tau\varepsilon\rho o\nu$, "wing".)

MKdV Abbreviation for the Modified Korteweg-deVries equation $u_t + u^2 u_x + u_{xxx} = 0$.

modon A dipole vortex, composed of two contra-rotating vortices of equal strength, which steadily translates as a coherent structure Also known as "Batchelor dipoles" or more accurately, as "Lamb-Chapyglin Vortex Pairs".

monopole vortex A spinning cylindrical column of fluid in which the vorticity [curl of velocity] is one-signed in the vortex core.

nanopteron A weakly nonlocal solitary wave whose far field sinusoidal radiation is an *exponentially* small function of the amplitude of the core. (From the Greek $\nu\alpha\nu o\sigma$, "small", and $\pi\tau\varepsilon\rho o\nu$, "wing".)

NLS Abbreviation for the Nonlinear Schroedinger equation $i\,A_t + (1/2)A_{xx} \pm A|A|^2 = 0$. The *envelope* of a wave packet evolves as a solution to this equation.

optimally-truncated asymptotic series If an asymptotic series is divergent, then for a given ε, the error decreases as more terms are added up to some $N_{opt}(\varepsilon)$ and then increases. The "optimal truncation" is to include only those terms up to and including $O(\varepsilon^{N_{opt}(\varepsilon)})$. An optimally-truncated series is said to be a **superasymptotic** approximation.

parexic analysis Neologism of Cornelius Lanczos for the art of approximation; coequal with pure analysis and numerical analysis as one of three branches of applied mathematics in Lanczos' thinking. (From the Greek "parexic", "nearby", from the roots "para" $\pi\alpha\rho\alpha$ ["almost, quasi"] and "ek" $\varepsilon\xi\omega$ ["out"]. From the preface of *Applied Analysis* (1956) [?], which he says is a book of "parexic analysis".

parity A symmetry of functions such that either $f(x) = f(-x)$ for all x ("symmetric with respect to the origin") or $f(x) = -f(-x)$ for all x ("antisymmetric with respect to the origin"). A symmetric function is said to be of "even parity" while an antisymmetric function is said to be of "odd parity".

periodic A function $f(x)$ is "periodic" with period L if and only if

$$f(x + L) = f(x)$$

for all x.

Poiseuille flow A mean flow which is parabolic between the boundaries. "Circular Poiseuille" flow is the flow in a pipe; the mean velocity is a quadratic function of radius. "Plane Poiseuille" flow is the flow between two infinite, stationary parallel plates, driven by a pressure gradient parallel to the plates.

pseudospectral An algorithm which uses interpolation of the residual function to determine the coefficients of a spectral series that approximates the solution to a differential or integral equation. Synonyms are "orthogonal collocation", "discrete ordinates" and "method of selected points".

quasi-normal mode A solution to the inviscid shallow water wave equations is a true normal mode if it oscillates forever. A solution is a quasi-normal mode if it decays by irreversibly radiating energy out of the equatorial region to higher latitudes so slowly that the decay time scale is large compared to the period of oscillation. Also called "basin resonance".

residual function When an approximate solution u_N is substituted into a differential, integral, or matrix equation, the result is the residual function, usually denoted $R(x;\ a_0,\ a_1,\ \ldots,\ a_N)$. The residual function would be identically zero if the approximate solution were exact.

ronin A perpetual postdoc, a scientist lacking a tenure-track or civil service position. (Japanese for a samurai who lacks a feudal lord, a very deplorable state.)

Rule of Many Names Although it is exasperating that the same thing may have different names in different fields, the importance and breadth of a concept, species or discovery can be measured by the number of times it has been independently discovered and named.

semi-implicit (time-marching) A time-integration method that treats some terms implicitly and others explicitly. Such algorithms are very common in hydrodynamics where diffusion is treated implicitly but the nonlinear advection is explicit; this avoids inverting large systems of *nonlinear* algebraic equations at each time step. Outside of geophysics, usually called implicit-explicit (IMEX).

separable An adjective applied to a partial differential (PDE) equation which denotes that the PDE can be split into a sequence of ordinary differential equations including Sturm–Liouville eigenproblems. The procedure for reducing a PDE to ODEs is the "method of separation of variables".

skew-symmetric A matrix A is skew-symmetric if and only if its tranpose is its negative, that is, $\mathrm{A}^T = -\mathrm{A}$. All eigenvalues of a skew-symmetric matrix are pure imaginary or zero. Similarly, an operator is skew-symmetric if its adjoint is its negative; it, too, has pure imaginary eigenvalues.

silverback A senior professor or scientist. From anthropology where the label was first applied to senior members of a troop of primates, later extended from chimpanzees to anthropologits.

sparse matrix A matrix whose elements are mostly zero. The opposite of a dense or full matrix.

spectral (i) A catch-all term for all methods (including pseudospectral techniques) which expand the unknown as a series of global, infinitely differentiable expansion functions. (ii) In a narrower sense, it denotes those algorithms that use only the expansion functions and their integrals and recurrence relations (as opposed to pseudospectral methods, which need the values of the functions on a grid of interpolation points).

spectral blocking The (spurious) accumulation of energy at or near the smallest resolved scales. Common in nonlinear hydrodynamics; also arises when the CFL

limit is violated only for the shortest resolvable scales. The signature of spectral blocking is that the graph of $\log |a_n|$ versus degree n rises instead of falls with n near $n = N$, the truncation limit. The usual remedy is to add a little dissipation which is highly scale-selective, damping only wavenumbers or coefficients close to the truncation limit.

sponge layer A common strategy for computing flow in an unbounded domain is to impose a boundary at large but finite distance L ("domain truncation"). To absorb waves near the artificial barrier and thus prevent spurious reflections from corrupting the solution everywhere, a large artificial viscosity is often added in the neighborhood of the artificial boundary. The regions of large damping are called "sponge layers" because they absorb the waves.

subcritical bifurcation A bifurcation in which the flow is unstable to finite amplitude perturbations for *smaller* values of a parameter, such as the Reynolds number Re, than the stability boundary of infinitesimal amplitude perturbations.

supercritical bifurcation A bifurcation in which the stability boundary in parameter space is the stability of infinitesimal amplitude perturbations.

superasymptotic error If an asymptotic series is divergent, then for a given ε, the error decreases as more terms are added up to some $N_{opt}(\varepsilon)$ and then increases. If the error of optimally-truncated series, broken off at $N_{opt}(\varepsilon)$ is an *exponential* function of $1/\varepsilon$, such as $\exp(-q/\varepsilon^r)$ where q and r are positive constants – the usual case — the optimally-truncated series is said to be a **superasymptotic** approximation.

weakly nonlocal solitary wave A steadily-translating, finite amplitude wave that decays to a small amplitude oscillation (rather to zero) as one moves away from the core of the disturbance, or that decays by never-ceasing radiation of small amplitude waves whose phase speed matches that of the coherent structure.

zonal Synonym for "longitudinal", adjective pertaining to the east-west direction.

References

1. Busbridge IW (1948) Some integrals involving Hermite polynomials. J Math Soc Lond 23(2):135–141
2. Azor R, Gillis J, Victor JD (1982) Combinatorial applications of Hermite-polynomials. SIAM J Math Anal 13(5):879–890
3. Ripa P (1983) Definite integral of the product of Hermite functions, with applications to the theory of nonlinear interactions among equatorial waves. J Geophys Res Oceans 88(NC14):9741–9744
4. Boyd JP (2013) Finding the zeros of a univariate equation: proxy rootfinders, Chebyshev interpolation and the companion matrix. SIAM Rev 55(2):375–396
5. Boyd JP (2014) Solving transcendental equations: the Chebyshev polynomial proxy and other numerical rootfinders, Perturbation series and Oracles. SIAM, Philadelphia, 460 pp
6. Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
7. Boyd JP (1984) The asymptotic coefficients of Hermite series. J Comput Phys 54:382–410

8. Olver FWJ, Lozier DW, Boisvert RF, Clark CW (eds) (2010) NIST handbook of mathematical functions. Cambridge University Press, New York
9. Boyd JP (2001) Chebyshev and fourier spectral methods. Dover, New York
10. Hille E (1926) A class of reciprocal functions. Ann Math 27(4):427–464
11. Hille E (1939) Contributions to the theory of Hermitian series. Duke Math J 5:875–936
12. Hille E (1940) Contributions to the theory of Hermitian series. II. The representation problem. Trans Am Math Soc 47:80–94
13. Hille E (1940) A class of differential operators of infinite order, I. Duke Math J 7:458–495
14. Bain M (1978) On the uniform convergence of generalized Fourier series. J Inst Math Appl 21(4):379–386
15. Hille E (1961) Sur les fonctions analytiques définies par des séries d'hermite. J Math Pures Appl 40:335–342
16. Hille E (1980) Contributions to the theory of Hermitian series III. Mean values. Int J Math Math Sci 3:407–421
17. Morse PM, Feshbach H (1953) Methods of theoretical physics, 1st edn. McGraw-Hill, New York, 2000 pp, (in two volumes), reissued by Feshbach Publishing (2005)
18. Boyd JP (2011) A proof, based on the Euler sum acceleration, of the recovery of an exponential (geometric) rate of convergence for the Fourier series of a function with Gibbs Phenomenon. In: Hesthaven JS, Ronquist EM (eds) Spectral and higher order methods for partial differential equations: proceedings of ICOSAHOM '09 conference, 22–26 June, Trondheim, Norway, vol 76. Lecture notes in computational science and engineering. Springer, New York, pp 131–140
19. Pearce CJ (1978) Transformation methods in the analysis of series for critical properties. Adv Phys 27(1):89–148
20. Boyd JP, Moore DW (1986) Summability methods for Hermite functions. Dyn Atmos Oceans 10:51–62
21. Boyd JP, Alfaro LF (2013) Hermite function interpolation on a finite uniform grid: defeating the Runge Phenomenon and the partial obsolesence of radial basis functions. Appl Math Lett 26(10):995–997
22. Boyd JP (2015) Hermite function interpolation on a finite interval. Appl Numer Math 87:125–144
23. Martens J (1990) The Hermite transform-theory. IEEE Trans Acoust Speech Signal Process 38(9):1595–1606
24. Martens J (1990) The Hermite transform-theory. IEEE Trans Acoust Speech Signal Process 38(9):1607–1618
25. Fasshauer GE, McCourt MJ (2012) Stable evaluation of Gaussian radial basis function interpolants. SIAM J Sci Comput 34(2):737–762
26. McCourt MJ (2013) Using Gaussian eigenfunctions to solve boundary value problems. Adv Apppl Math Mech 5:569–594
27. Anderson DLT (1973) A low latitude spectral model using Chebyshev-parabolic cylinder functions. Report 7, GARP Programme on Numerical Experimentation W. G. N. E., World Meteorological Organization, Genev
28. Tribbia JJ (1979) Nonlinear initialization on an equatorial beta-plane. Mon Weather Rev 107:704–713
29. Tribbia JJ (1984) A simple scheme for high-order nonlinear normal initialization. Mon Weather Rev 112:278–284
30. Marshall HG, Boyd JP (1987) Solitons in a continuously stratified equatorial ocean. J Physical Oceanogr 17:1016–1031
31. Smith NR (1988) A truncated oceanic spectral model for equatorial thermodynamic studies. Dyn Atmos Oceans 12:313–337
32. Holvorcem PR (1992) Asymptotic summation of Hermite series. J Phys A 25(4):909–924
33. Holvorcem PR, Vianna ML (1992) Integral-equation approach to tropical ocean dynamics. 2. Rossby-wave scattering from the equatorial Atlantic western boundary. J Marine Res 50(1):33–61

34. Vianna ML, Holvorcem PR (1992) Integral-equation approach to tropical ocean dynamics. 1. Theory and computational methods. J Marine Res 50(1):1–31
35. Tse KL, Chasnov JR (1998) A Fourier-Hermite pseudospectral method for penetrative convection. J Comput Phys 142(2):489–505
36. Meiron DI, Moore DW, Pullin DI (2000) On steady compressible flows with compact vorticity; the compressible Stuart vortex. J Fluid Mech 409:29–49
37. Majda AJ, Khoulder B (2001) A numerical strategy for efficient modeling of the equatorial wave guide. Proc Nat Acad Sci 98(4):1341–1346
38. Shamir O, Paldor N (2014) A Hermite-based shallow water solver for a thin ocean over a rotating sphere. J Comput Phys 269:80–97
39. Tanner J (2006) Optimal filter and mollifier for piecewise smooth spectral data. Math Comput 75(254):767–790
40. Arons AB (1981) The scientific work of Henry Stommel. In: Warren BA, Wunsch C (eds) Evolution of physical oceanography. MIT Press, Cambridge, pp i–xx
41. Jerri AJ (ed) (2010) Gibbs phenomenon. Sampling Publishing, Potsdam
42. Callaghan P, Fusco A, Francis G, Salby M (1999) A Hough spectral model for three-dimensional studies of the middle atmosphere. J Atmos Sci 56:1461–1480
43. Mizzi A, Tribbia J, Curry J (1995) Vertical spectral representation in primitive equation models of the atmosphere. Monthly Weather Rev 123(8):2426–2446
44. Kasahara A, Puri K (1981) Spectral representation of three-dimensional global data by expansion in normal modes. Monthly Weather Rev 109:37–51
45. Hoskins BJ, McIntyre ME, Robertson AW (1985) On the use and significance of potential vorticity maps. Q J R Meteor Soc 111(475):877–946
46. Boyd JP (2002) Equatorial solitary waves. Part 5. Initial value experiments, co-existing branches and tilted-pair instability. J Phys Oceangr, 32(9):2589–2602
47. Moore DW, Niiler PP (1974) A two-layer model for the separation of inertial boundary currents. J Mar Res 32:457–484
48. Ripa P (1982) Nonlinear wave-wave interactions in a one-layer reduced-gravity model on the equatorial beta-plane. J Phys Oceangr 12:97–111
49. Fruman MD, Shepherd TG (2008) Symmetric stability of compressible zonal flows on a generalized equatorial β plane. J Atmos Sci 65(6):1927–1940

Index

J.P. Boyd, *Dynamics of the Equatorial Ocean*,
https://doi.org/10.1007/978-3-662-55476-0