Michela Cameletti · Francesco Finazzi
Editors

Quantitative Methods in Environmental and Climate Research

Springer

Quantitative Methods in Environmental and Climate Research

Michela Cameletti • Francesco Finazzi
Editors

Quantitative Methods in Environmental and Climate Research

Springer

Editors
Michela Cameletti
Department of Management, Economics and Quantitative Methods
University of Bergamo
Bergamo, Italy

Francesco Finazzi
Department of Management, Information and Production Engineering
University of Bergamo
Dalmine
Bergamo, Italy

ISBN 978-3-030-01583-1 ISBN 978-3-030-01584-8 (eBook)
https://doi.org/10.1007/978-3-030-01584-8

Library of Congress Control Number: 2018968083

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume presents the revised extended version of a selection of contributions submitted at the joint TIES-GRASPA 2017 Conference on Climate and Environment held at the University of Bergamo, Italy, from 24th to 26th of July 2017, as a satellite meeting of the 61st World Statistics Congress—ISI2017, in Marrakesh.

The International Environmetrics Society (TIES, www.environmetrics.org) is a nonprofit organization aimed to foster the development and use of statistical and quantitative methods in the environmental sciences and environmental monitoring and protection. The Italian environmetricians group named GRASPA (www.graspa.org) is active since 1995, and it is a standing group of the Italian Statistical Society (SIS, https://www.sis-statistica.it/) for statistical applications to environmental problems.

The theme for the TIES-GRASPA 2017 Conference was "climate and environment," and the scientific program was a rich mix of methodological and applied topics presented through 59 sessions. The applications presented at the conference included air pollution, ecology, epidemiology, meteorology, atmospheric measurements, climate change, seismology, and remote sensing, while methodological sessions focused on functional and directional data analysis, spatial and spatiotemporal models, computational effective algorithms, and multivariate methods for complex data. In total 221 participants from 29 different countries attended the conference.

The seven contributions of this volume cover the following problems: detection of disease clusters, analysis of harvest data, change point detection in ground-level ozone concentration, modelling of atmospheric aerosol profiles, prediction of wind speed, precipitation prediction, and analysis of spatial cylindrical data. In particular, the paper by Gómez-Rubio et al. proposes a new method for detecting spatial disease clusters by generalizing the spatial scan statistics approach in the context of Bayesian hierarchical models. The problem of cluster detection is thus reformulated as a problem of variable selection of a generalized linear model, which can also include random effects and can be extended to the spatiotemporal and zero-inflated case. The proposed method is computationally efficient thanks to the use of the integrated nested Laplace approximations (INLA). A Bayesian hierarchical

model is also used in the paper by Nadeem et al. for estimating the age distribution of wildlife population such as white-tailed deer in Canada. The proposed method is based on a Leslie matrix and employs a multinomial likelihood depending on Beta-distributed probabilities estimated using age-specific harvest data collected by wildlife agencies.

The detection of sudden changes in spatially and temporally correlated data is the topic of the paper by Sun and Wu. The authors improve the estimation algorithm of the general spatiotemporal autoregressive (GSTAR) model by using a method which is more stable in estimating parameters and more accurate in detecting change-points also in the case of outliers or heavy-tail distributed errors. The paper by Negri et al. deals with uncertainty assessment of vertical profiles of atmospheric aerosol when they are observed using satellite and ground-based instruments. Uncertainty assessment is a key aspect in climate modelling, and the approach proposed in the paper can be extended to other instruments and essential climate variables. Nicolis et al. address the problem of calibrating simulation outputs with observed data, with application to wind speed forecasting. This is achieved by means of spatiotemporal characterized by time-varying basis functions and spatially varying coefficients. Abdelfattah et al. discuss the modelling of the spatiotemporal variability of precipitations over the area of a country. Precipitation dynamics is described using empirical orthogonal function analysis, while dynamic harmonic regression is adopted in order to understand the partial influence of global sea surface temperatures. Modelling space-time phenomena gets complicated when data have a complex support. This is the case of spatial cylindrical series discussed in the paper by Lagona and modelled as a mixture of copula-based bivariate densities whose parameters vary according to a latent Markov random field. The approach is applied to marine currents in the Adriatic Sea.

We wish to thank all the authors and reviewers as well as Eva Hiripi from Springer-Verlag for the excellent cooperation in publishing this volume. We also thank the president of TIES Prof. Alessandro Fassò and the coordinator of GRASPA Prof. Alessio Pollice for organizing a conference that brought together eminent researchers and scientists and young students from all over the world.

Bergamo, Italy Michela Cameletti
Bergamo, Italy Francesco Finazzi

Contents

Fast Bayesian Classification for Disease Mapping and the Detection of Disease Clusters

V. Gómez-Rubio, John Molitor, and Paula Moraga

Keywords Spatial statistics · Disease clusters · Bayesian inference · Integrated nested Laplace approximation

1 Introduction

The detection of disease clusters is an important topic in public health research. Many statistical methods have been proposed (see [22] for a review), and several journals have devoted special issues to this and other related topics (see, e.g. [24, 25]). One popular approach to detecting these clusters of disease is to utilise scan-based methods which are based on a moving window which is placed over the study region at several possible cluster centres and then perform a local test for clustering (see, e.g. [30] for a general description). Scan statistics provide both a test for statistical significance and estimates of, e.g. the relative risk inside versus outside the detected cluster (see, e.g. [31]).

Based on concepts inherent in partition models (see, e.g. [11, 20]), we propose a new approach that uses dummy variables in a regression model to group regions into clusters, similarly as in Gangnon and Clayton [13]. The importance of the clusters is then assessed based on a likelihood calculation that measures the extent to which the clusters capture the variability in the outcome. This is very similar to what [19, 43, 44] have done; however we extend their work to cope with new problems and

V. Gómez-Rubio (✉)
Department of Mathematics, School of Industrial Engineering, Universidad de Castilla-La Mancha, Albacete, Spain
e-mail: Virgilio.Gomez@uclm.es

J. Molitor
College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, USA
e-mail: John.Molitor@oregonstate.edu

P. Moraga
Faculty of Health and Medicine, Lancaster University, Lancaster, UK
e-mail: p.e.moraga-serrano@lancaster.ac.uk

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_1

provide a new way of assessing cluster significance and importance by means of a more general model selection criteria. Anderson et al. [4] also tackle the problem of spatio-temporal clustering by grouping areas and estimating a different risk for each group.

We regard the problem of cluster detection as a problem of variable selection, where covariates include a number of dummy variables that represent all possible clusters. Although the cluster space is quite large, different approaches to search this space can be used (see, e.g. [7] for a discussion on searching large variable spaces).

Our general method is implemented under the general framework of Bayesian hierarchical models, which have been widely used in classification because of their flexibility and ease of fit by means of Markov chain Monte Carlo (MCMC) methods (see, e.g. [15]). However, when the problem is large, there is a huge computational burden that is particularly restrictive when a large number of model parameters must be estimated. Rue et al. [34] have provided a way of approximating the marginal distributions of the parameters in the model. We will use this approximation to fit the model, provide a reasonable estimate of the coefficient of the cluster variables and compute the deviance information criterion (DIC, [35]), which we will use in model selection. In this context, selecting the best set of covariates is equivalent to choosing the best set of clusters in the region. Hence, when fitting individual model to test for different clusters, our approach, based on INLA, will be faster than fitting the same models with MCMC.

Another reason for considering this approach is that most methods for the detection of clusters focus on assessing the presence of a cluster (usually, by testing its significance), but these methods seldom try to relate the cluster to an outcome variable, as is the case when we want to calculate the disease risk associated with the cluster in question. By including the dummy variables as terms in a regression model, we can estimate the increased risk of disease associated with living in an area defined by the cluster. Hence, the methods proposed in this paper will address both detection and risk estimation.

We have decided to use Bayesian inference for several reasons. First of all, hierarchical models allow us to extend the former model to include other sources of variability, such as random effects. Secondly, credible intervals for the cluster coefficients will be obtained, so that the significancy of the clusters can be computed. Finally, Bayesian inference provides several methods for model assessment that will be helpful to define the location and number of clusters, as described in the next section.

Other authors have already tackled the problem of disease cluster detection using similar approaches. Anderson et al. [3] and Bilancia and Demarinis [6] have recently proposed a similar approach, but we believe that our description on the methods covers a wider range of situations. Wakefield and Kim [39] consider the problem of cluster detection using partition models, and they use the Bayes factor to select areas of clusterings. However, they use a full MCMC approach for model fitting that can be very time-consuming. Knorr-Held and Rasser [20], Gangnon and Clayton [14], Gangnon [12] have proposed a similar approach based on reversible jump MCMC and partition models, but our approach is different. We are trying to find

a small number of clusters of high risk as a spatial scan statistic does. In addition, our method is also able to estimate the number of relevant clusters at a most lower computational cost compared to previous approaches.

Finally, we believe as Rothman [33] that it is more important to assess exposure and the detection of risk factors than simply detecting clusters. However, by highlighting areas of unexplained high risk, we believe that new risk factors can be identified and new hypothesis about the aetiology can be proposed.

This paper is organised as follows. Firstly, we introduce the spatial scan statistic and classification methods in Sect. 2. We then describe the basic generalisation of the spatial scan statistic using generalised fixed-effects models in Sect. 3. Section 4 describes our model fitting using Bayesian hierarchical models and INLA. Another extension using generalised mixed-effects models is described in Sect. 5. Zero-inflated models for the detection of disease clusters are discussed in Sect. 6. Section 7 describes how to extend these ideas to detect clusters in space and time. A simulation study is summarised in Sect. 8. Finally, some examples are discussed in Sect. 9, and a discussion is provided in Sect. 10.

2 Scan Methods for the Detection of Disease Clusters

2.1 Spatial Scan Methods

The spatial scan statistic [21] scans the study region with a huge number of overlapping windows and determine the windows which group together an unusual number of cases, possibly adjusting for multiple testing. The collection of windows depends on the application. Typically, the spatial version uses circular windows constructed by consecutively aggregating nearest-neighbouring areas until a proportion of the total study population is included.

Conditioning on the observed total number of cases, the scan test statistic S is defined as the maximum likelihood ratio over all possible windows Z: $S = max_Z \frac{L(Z)}{L_0}$, where $L(Z)$ is the likelihood for window Z and L_0 is the likelihood function under the null hypothesis which states that the probability of being a case inside Z is equal to the probability of being a case outside Z. The mathematical formulation of S depends on the probability model used. For example, a Poisson model is used for data where the number of events are Poisson distributed, and a Bernoulli model for case-control type data.

The window with the maximum likelihood constitutes the most likely cluster, the cluster that is least likely to have occurred by chance. Its statistical significance is obtained through Monte Carlo hypothesis testing. Thus, the previous procedure is repeated for a large number of replicas of data generated under the null hypothesis, say R, and their respective test statistics are calculated. The test statistic of the observed data is combined with these, and the set of the $R + 1$ values are ordered. If M is the rank of the observed test statistic, a p-value equal to $M/(R + 1)$ would

be obtained. Apart from the most likely cluster, secondary clusters can also be identified, ordered according to the value of S.

There will always be a secondary cluster which is almost identical to the most likely one and with almost the same likelihood, expanding or reducing the size of the initial cluster, but clusters of this type provide little additional information. Normally the option chosen is to show the secondary clusters which do not overlap with the most probable cluster, as they can be of greater interest.

In the end, some areas will be assigned to a small number of clusters, where the number of cases is higher than expected. The remaining areas can be regarded as assigned to another group with no significant increase in the number of cases.

2.2 Classification of Disease

As seen in Sect. 2.1, detecting clusters of disease can be regarded as a classification problem where n areas are assigned to $K + 1$ groups. K is the number of spatial clusters with increased risk and the areas not in any of these clusters are assigned to another group. Usually, we face a double estimation problem here: finding the number of clusters K and allocation of areas to each cluster. We note that a partition, or a set of allocations of areas to clusters, can be represented by dummy variables $c^{(j)} = (c_1^{(j)}, \ldots, c_n^{(j)})$ $(j = 0, \ldots, K)$, so that they are defined as follows:

$$c_i^{(j)} = \begin{cases} 1 \text{ if element } i \text{ belongs to group } j \\ 0 \text{ otherwise} \end{cases}$$

where $c^{(0)}$ denotes a null cluster and represents areas not included in any other cluster. This group will not contribute to the computation of the model likelihood and will effectively be ignored in the spatial analysis.

In our context of disease mapping each cluster is made of several contiguous areas which are neighbours and that have a significant increased risk when considered together. Hence, some constraints on $c^{(j)}$ will be imposed to show this spatial aggregation. In Sect. 4.1 we describe how to explore the set of possible cluster covariates to find the significant ones, similarly as the spatial scan statistic.

3 Classification of Disease: Generalised Linear Models

Generalised linear models (GLM, [27]) provide a suitable way of modelling public health data, which are often modelled using a Poisson or binomial distribution.

In the Poisson case, the observed number of cases O_i in area i is distributed as

$$O_i | \mu_i \sim Po(\mu_i);\ i = 1, \ldots, n$$

where the mean is taken usually as $\mu_i = e_i\theta_i$. Here, e_i represents the expected number of cases and θ_i the relative risk. The expected number of cases e_i is often computed using internal age-sex standardisation (see, e.g. [41]). If other covariates need to be accounted for to compute e_i, they can be used in the standardisation [10, 21] or e_i can be computed from a set of covariates x_i using a log-linear regression model [41].

Furthermore, the mean can be modelled on a vector of covariates $\mathbf{x_i}$:

$$\log(\mu_i) = \log(e_i) + \log(\theta_i) = \log(e_i) + \alpha + s(\mathbf{x_i})$$

Here $s(\cdot)$ represents a generic function on the covariates as the relationship between the covariates and the disease risk can be non-linear [10]. Usually, it will be a linear function, but other non-linear functions (e.g. splines) could be used. If a cluster covariate $c^{(j)}$ is added, then we have the following:

$$\log(\theta_i) = \alpha + s(\mathbf{x_i}) + \gamma_j c_i^{(j)}$$

Coefficient γ_j indicates the importance and significance of the cluster defined by dummy variable $c^{(j)}$.

Instead of a Poisson, a binomial distribution may be preferred (e.g. for common diseases). Hence, we will denote that observed cases come from a binomial distribution as

$$O_i|\pi_i \sim \text{Binom}(\pi_i, N_i);\ i = 1, \ldots, n$$

where π_i is the probability of being a case and N_i the population at risk.

As in the Poisson case, covariates can be considered by using, for example, a logit link function on π_i in the usual manner [29]:

$$\text{logit}(\pi_i) = \text{logit}(\hat{p}) + \log(\theta_i)$$

where θ_i depends on some covariates and cluster variables, and $\hat{p}$ is the observed proportion of cases:

$$\hat{p} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} N_i}$$

Note that $\text{logit}(\hat{p})$ is an offset in the linear predictor and that it can also be written as

$$\text{logit}(\hat{p}) = \log\left(\frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n}(N_i - O_i)}\right)$$

This is the observed odds ratio, and it can be regarded as a reference odds ratio as it provides a measure of the overall incidence of the disease.

Similarly as in the Poisson case, modelling θ_i on the covariates and the cluster covariate $c^{(j)}$ can be done as follows:

$$\log(\theta_i) = \alpha + s(\mathbf{x_i}) + \gamma_j c_i^{(j)}$$

3.1 Adjustment for Relevant Covariates

As we are interested in detecting clusters of disease which cannot be explained by known or possible risk factors (i.e. observed covariates), it is important to remove the effect of the covariates beforehand. Further, risk factors obtained through the use of stratified covariates can be used to compute reference rates, as discussed, for example, by Ferrándiz et al. [10]. Given that our new approach also provides a more general framework to incorporate this information and the cluster specification into a single model, a model without any cluster covariate can be fitted and use the fitted values as an offset in a model with cluster covariates.

In the Poisson case, this is equivalent to considering the expected number of cases as

$$E_i = e_i \exp(\hat{\alpha} + \hat{s}(\mathbf{x_i})) \tag{1}$$

Here we can see how the standardised expected cases e_i are modulated by the covariates. $\hat{\alpha}$ and $\hat{s}(\cdot)$ represent the fitted values for α and $s(\cdot)$. The mean of the Poisson distribution for observation i is now modelled as

$$\log(\mu_i) = \log(E_i) + \log(\theta_i), \tag{2}$$

with $\log(E_i)$ an offset in the model.

In the binomial case, a similar model can be used for the covariates if instead of a common log odds ratio $\text{logit}(\hat{p})$ we consider a *modulated* log odds ratio $\text{logit}(\hat{p}_i)$ for each observation:

$$\text{logit}(\hat{p}_i) = \log\left(\frac{\sum_{i=1}^n O_i}{\sum_{i=1}^n (N_i - O_i)}\right) + \hat{\alpha} + \hat{s}(\mathbf{x_i})$$

The resulting model for the probability of the binomial distribution is

$$\text{logit}(\pi_i) = \text{logit}(\hat{p}_i) + \log(\theta_i) \tag{3}$$

with $\text{logit}(\hat{p}_i)$ an offset in the linear predictor.

Once we have filtered for the covariates, cluster covariates can be included by modelling the relative risk θ_i using cluster covariates or other effects to perform

the detection of clusters as described in Sect. 4.1. For example, if a single cluster covariate is included in the model we will have the following:

$$\log(\theta_i) = \gamma_1 c_i^{(1)}$$

4 Bayesian Hierarchical Models for the Detection of Disease Clusters

4.1 Detection of Clusters

The spatial scan statistic [21] described in Sect. 2.1 performs the detection of clusters by proposing a large number of putative cluster candidates that are later examined for significance. Here, we proceed similarly using our Bayesian linear model framework with cluster indicator variables. We do this by expressing a set of putative clusters as binary covariates in a linear model and then assess the significance (in terms of Bayesian posterior probabilities) as one would using any other set of binary covariates. This allows us to exploit the computational advantages of the Laplace approximation estimation approach to analysing linear models (via the R-INLA software) whilst still framing the model in a fully Bayesian manner.

We first obtain our set of candidate clusters by creating a vector $c^{(1)} = (1, 0, \ldots, 0)$ which contains a '1' corresponding to an area that is in to be a member of the putative cluster to be examined. We start with a single area and then create larger and larger contiguous clusters by simply adding other areas in turn (nearest areas to the cluster centre first). Each cluster covariate will have components $c_i^{(j)}$, which will be 1 if area i is in cluster j and 0 otherwise. We will stop adding new areas to the cluster when a certain percentage of the total population has been reached. By repeating this procedure using all possible areas (or cases) as cluster centres, we will obtain C number of putative clusters and associated cluster covariates $\{c^{(j)}\}_{j=1}^{C}$.

In order to assess the relationship these clusters have with our outcome of interest, we fit a generalised linear model with a single covariate. The linear predictor η_i (for the Poisson and binomial cases) will be

$$\eta_i = \text{offset}_i + \gamma_j c_i^{(j)} \tag{4}$$

where offset_i is the standard offset denoting, say, population size of each area. γ_j is the coefficient of the cluster variable. Note that we have deliberately not included an intercept term in this model, and because of this, the risk baseline is set to one. Furthermore, the adjustment for covariates is done such as it is included in the offset to filter for their effects, as explained in Sect. 3.

In our formulation of the problem of the detection of disease clusters, we propose fitting a hierarchical Bayesian model to account for the different sources

of uncertainty. In particular, we fit

$$\begin{aligned} O_i|\ldots &\sim f(O_i;\ldots) \\ \eta_i &= \text{offset}_i + \gamma_j c_i^{(j)} \\ \gamma_j &\sim N(0, \tau_\gamma) \end{aligned} \tag{5}$$

Here $f(O_i;\ldots)$ represents the likelihood of the data; η_i is a linear predictor on the covariates that is conveniently related to the mean of the distribution of the data by a an appropriate link function. Cluster coefficients are assigned a vague normal prior with zero mean and precision $\tau_\gamma = 0.01$.

4.2 Cluster Selection

We choose our clusters by fitting many models, each corresponding to a particular cluster configuration, and then we choose the model with the lowest DIC value. Significance of each cluster can be assessed by computing the posterior probability of the coefficient of being higher than zero, e.g. $P(\gamma_j > 0|y)$. Note that this measure of 'significance' is Bayesian in nature, as it encapsulates both point estimate and error into one quantity. Further, one can compute the probability that $\gamma_j > \gamma_t$, where γ_t is some threshold chosen to have substantive, subject-area significance.

Hence, the DIC will give us a ranking of the cluster configurations according to how the cluster variables model important areas of extreme risk. Furthermore, $P(\gamma_j > 0|y)$ will actually tell us whether the cluster has a 'significantly' high risk, where significance denotes $P(\gamma_j > 0|y) > 0.95$ or, equivalently, $P(\gamma_j < 0|y) < 0.05$. Our aim is to build a final model with all significant non-overlapping clusters so that model fitting cannot be improved (measured through the DIC) and all cluster covariates are significant (measured through $P(\gamma_j < 0|y)$).

In summary, we will fit different models considering one cluster variable at a time, and we will rank them according to the DIC (in ascending order). We will also compute the posterior probability of the cluster coefficient being lower than zero to consider only clusters with a significant high risk, i.e. those with $P(\gamma_j < 0|y) < 0.05$. The ranked putative clusters will then be processed to obtain a set of significant and non-overlapping clusters in the study region, as described in Sect. 4.3. This will define the number of cluster and the definite partition of the areas in the study region in clusters.

4.3 Number of Clusters

In order to obtain the set of clusters C_d in our final model, the cluster with the lowest DIC will be considered first, and then other non-overlapping clusters will be added

in turn so that significant clusters are in the final model, or they are removed because they overlapped with a cluster with a lower DIC that had been added previously.

This is similar to a forward stepwise variable selection. As more and more clusters are added to the final model, the value of the DIC should continue to decrease, and then at a certain point, it will increase. We could stop adding clusters when $P(\gamma_j < 0|y)$ is large (i.e. higher than 0.05) or when the difference in the DIC is small. However, it is not clear what a small difference in the DIC is. Several authors propose different solutions. Spiegelhalter et al. [35] suggest that differences higher than 5 are substantial. We will take this criterion with caution, and we will discuss this in Sect. 9, where we analyse some examples with real data.

The resulting significant cluster variables will produce a partition of the areas into multiple clusters, similarly as the partition models proposed by other authors Ferreira et al. [11], Knorr-Held and Rasser [20]. Zhang et al. [45] propose a modified test for the spatial scan statistic for multiple clusters, but we believe that our approach to include multiple clusters in the model is more general (as it is based on variable selection procedures and regression models). Furthermore, Zhang et al. [45] approach is based on removing some areas in the most likely and subsequent clusters which is very ad hoc to the model proposed in their paper, whilst our approach can handle multiple clusters regardless of the model used.

Once we have the list of significant and non-overlapping clusters, they can put together into a final model as follows:

$$\eta_i = \text{offset}_i + \sum_{j \in C_d} \gamma_j c_i^{(j)} \tag{6}$$

where C_d is the subset of the cluster covariates that we have selected. Note that this model is effectively adjusting for several clusters.

Hence, our method will produce a partition of the observations into $|C_d|$ cluster groups plus another group of areas with no increased risk. The idea of a cluster as a well-defined area with a discontinuity in risk along its boundary is an oversimplistic notion because risk is a continuous spatial process. However, this does not stop it from being a useful device to find areas of high risk, but findings should be interpreted with caution because spatial risk variation is a continuous process. Also, note that Eq. (6) can be regarded as a low-rank approximation to a spatially continuous Gaussian process.

The fact that the model also includes covariates, which can potentially have a smooth spatial variation, means that risk estimates may also show smooth spatial variation. In the next section, we introduce an extension to this model that includes random effects that can be used to better account for risk variation. In particular, this will allow for within-cluster risk variation whilst still providing a classification of the areas into groups via cluster dummy variables.

5 Classification of Disease: Generalised Mixed-Effects Models

5.1 Motivation

Overdispersion often occurs when working with count data if relevant covariates are not taken into account in the model. Gómez-Rubio et al. [17] propose using a Monte Carlo test for spatial scan statistics where the observed number of cases are drawn from a negative binomial distribution. Loh and Zhou [26] propose a similar Monte Carlo test, but the observed cases are now simulated from a Poisson GLM with spatially correlated random effects.

A more general approach is the use of GLMs with random effects in the cluster selection procedure that we have described in the previous section. This not only allows for a more flexible modelling of the data, but also by including area-level random effects, it is also possible to model risk variation within clusters.

5.2 GLM with Random Effects

Our proposal to deal with overdispersion is to extend the previous model to include random effects under the framework of generalised mixed-effects models (GLMM) [28]. In this case the resulting model will be

$$
\begin{aligned}
O_i|\ldots,u_i &\sim f(O_i;\ldots,u_i)\\
\\
\eta_i &= \text{offset}_i + \gamma_j c_i^{(j)} + u_i\\
u_i|\sigma_u^2 &\sim N(0,\sigma_u^2)\\
\\
\gamma_j &\sim N(0,\tau_\gamma)\\
(\sigma_u^2)^{-1} &\sim Ga(1,0.00005)
\end{aligned}
\tag{7}
$$

where $f(O_i;\theta_i,u_i)$ is the likelihood of the data and u_i represents the random effects, which are assumed to be independent and normally distributed. Cluster coefficients are again assigned a vague Gaussian prior, and variance of the random effects σ_u^2 is assigned a vague inverted Gamma prior.

Zhang and Lin [43], Anderson et al. [3], Bilancia and Demarinis [6] propose the use of spatially correlated random effects when looking for disease clusters. We believe that this may be problematic because we are trying to model the unexplained spatial variation by means of cluster variables and there may be some clash between the cluster variables and the spatial random effects.

5.3 Selection and Number of Clusters Using GLMM

We will follow the same approach as described in the previous section to find the disease clusters in the region by conducting repeated evaluations of the model for different cluster covariates and computing the DIC. Note that the DIC is also a feasible tool to compare mixed-effects models because it accounts for the complexity of the random effects.

This is another advantage of using a Bayesian approach. In a frequentist framework, we will need to define a model selection criterion which accounts for the complexity of the random effects. See, for example, the cAIC proposed by Vaida and Blanchard [38]. However, the cAIC is only developed for normal responses, and it would need to be extended to a more general case to accommodate the use of non-Gaussian likelihoods.

Note that now the linear predictor of the model will look like the following:

$$\eta_i = \text{offset}_i + \sum_{j \in C_d} \gamma_j c_i^{(j)} + u_i$$

This is very similar to other popular models for spatial risk variation, such as the one proposed by Besag et al. [5]. In this model, the linear predictor is made of the sum of some linear effect on the covariates (that we have included in the offset) plus independent random effects (same as u_i above) plus spatially correlated random effects. In our model we have replaced the spatially correlated random effects by cluster components with the aim of identifying groups in the data, but that have a spatial nature (by the way they have been built). Hence, there is an evident connection between the model proposed above and standard methods for measuring spatial variation of disease risk, with the cluster components being a low-rank approximation to the underlying spatial process.

Cluster detection based on models for spatial risk variation (such as the one in [5]) is based on declaring a cluster as a region within which we are confident risk exceeds some threshold level. However, we think there are some advantages of our method over this procedure. First, to detect clusters using a risk surface, it is required to choose a threshold level. We think it is not always clear what level should be chosen and different clusters can be detected using different levels. Second, our method allows looking for clusters with a maximum population. Using a varying risk surface, it would be necessary to examine each of the clusters detected to see if they fulfil these constrains, and this can be a difficult task especially when there are a large number of regions and periods of time.

6 Classification of Disease: Zero-Inflated Models

Another common problem, particularly with very rare diseases, is that the number of zeros observed tends to be very high, and this can affect the results of the detection of clusters [9, 16]. Ugarte et al. [36] propose the use of zero-inflated models to accommodate the high number of zeros observed.

Zero-inflated models are a mixture model with two components: a probability mass function with all its mass at zero, which occurs with probability π, and another distribution that generates the non-zero values, which occurs with probability $1-\pi$. In our case this distribution may be a Poisson or binomial.

The probability of observing n_i cases is given as follows:

$$\Pr(O_i = n_i) = \begin{cases} \pi + (1-\pi) f(0;\theta_i) & n_i = 0 \\ (1-\pi) f(n_i;\theta_i) & n_i = 1, 2, \ldots \end{cases} \tag{8}$$

Hence, we can follow a similar approach for the detection of clusters of disease by including cluster covariates in the second term of the mixture, as part of the $f(n_i;\theta_i)$ term [16]. In this way, we will be looking for clusters after adjusting for the effect of the zero inflation.

For model fitting, π must be assigned a prior distribution. In particular, $\text{logit}(\pi)$ is assigned a Gaussian distribution with mean -1 and precision 0.2. This is the default in R-INLA and π has its prior mode around 0.27. As we shall see in one of our examples, this prior assumption does not seem to have a strong impact (in this particular example). However, this prior can be set in R-INLA using a wide range of distributions and parameters for cases when this prior is too informative.

Note that π could vary among regions, so that it can be replaced by π_i in Eq. (8). π_i can also be further modelled to depend on a number of covariates (see, e.g. [40]). However, it may be difficult to disentangle the effects between the two terms in the mixture model.

6.1 Cluster Selection

When we deal with mixture models the problem of estimating the complexity of the model becomes more difficult, and we should be cautious. Many times it is not clear how many parameters we are trying to estimate. Burnham and Anderson [8, pp. 342–344] discuss this issue in detail. Note that in simple cases, such as when comparing similar models (as in our case), we know that the complexity of the different models is similar. For all these reasons, the DIC should be used with care, and the marginal likelihood of the model could also be used as an indicator to choose the best clusters. In any case, $P(\gamma_j < 0|y)$ should be a good indicator to find the clusters in the study region.

7 Space-Time Clusters

When there are different measures for each area in time, we may be interested in detecting clusters in space and/or time. Jung [19] points out, in the discussion of her paper, how to use the spatial scan statistic to look for prospective or retrospective clusters. For the case of Poisson data, this new model can be formulated by modelling $\mu_{i,t}$, the mean in area i and time t, as follows:

$$\log(\mu_{i,t}) = \log(E_{i,t}) + \gamma_j c_{i,t}^{(j)} \tag{9}$$

$c_{i,t}^{(j)}$ is now the cluster variable in space and time.

In order to define space-time clusters, a temporal window is defined together with a spatial one. The easiest approach is to consider homogeneous blocks in space and time, i.e. for each time period that is considered the same, areas are included in the cluster. Regarding the size of the temporal window, [23] consider a time frame of up to half the total number of time slots.

When dealing with space-time models, we need to account for the temporal trend, so that the clusters detected are not the result of the natural variation of the disease. This could be done by considering time when the disease rates and expected cases are computed. Alternatively, a smooth term can be considered in the linear predictor in Eq. (9). Splines or a different random intercept term for each time period are convenient ways of modelling temporal variation. This approach has the advantage of adjusting for non-linear changes in time which may distort the cluster detection when a linear term is used.

Anderson et al. [4] address the problem of spatio-temporal clustering by providing a separate partition of areas and times so that a different risk is estimated for each group.

8 Simulation Study

In this section we assess the performance of the methods proposed above by means of a simulation study. It is based on the cases of brain cancer in basic health zones in the province of Navarre (Spain), which we have used as a case study in Sect. 9.2. Altogether, there are 40 different areas and 129 cases.

The datasets have been simulated assuming different cluster sizes (5 or 15 areas), relative risks (1.5 or 3) and two covariates to produce mild or strong overdispersion if they are not taken into account in the model. The covariate that produces strong overdispersion has been simulated by taking the y coordinate of each region, re-scaling the values (by subtracting the mean and dividing by its standard deviation) and then randomly assigning the values to the areas. This is a simple way of creating a covariate from the real dataset that has no spatial pattern but that induces a strong overdispersion as the values are between -2.01 and 1.77. By dividing these values

by 3, we have created another covariate that induces mild overdispersion in the data when not taking into account in the model because its values are in the range of the exceed risk in the simulated clusters. Hence, this will also provide a way of measuring how overdispersion affects the cluster detection procedure. Altogether, we have eight combinations according to the covariate, relative risk and cluster size.

In this setting, the mean μ_i of each area is

$$\log(\mu_i) = 1 \cdot x_i + \log(RR) \cdot I_i$$

where x_i is the value of the simulated covariate in area i, RR the cluster relative risk and I_i a binary variable that indicates whether area i is in the cluster.

In order to obtain the simulated value of the number of cases, we have conditioned on the total number of cases of the real dataset (129) and distributed them at random using a multinomial distribution with probabilities $\mu_i / \sum_{i=1}^{40} \mu_i$, $i = 1, \ldots, 40$.

To simulate the zero-inflated data, we have considered that six areas have a structural zero. This follows from the estimated value of π obtained after fitting a zero-inflated model to the actual dataset (see Sect. 9.2). First, we have sampled the six areas with structural zeros completely at random, and, for the remaining areas, we have assigned the number of cases similarly as in the previous case. Note that now the cases are distributed over 34 regions only.

This setting is very similar to the simulation study in Bilancia and Demarinis [6], which considered a region with 60 different areas and used a Poisson model with covariates and cluster effects to create the simulated datasets. However, our simulation study also includes GLMs, GLMMs and ZIP models to analyse datasets based on Poisson and zero-inflated Poisson distributions. Hence, our study is more general.

Finally, in order to assess the performance of the method, we have followed the procedure described before to remove overlapping clusters and keep non-overlapping cluster only. Then, this classification of the areas has been used to compute sensitivity and specificity, by comparing how our method has classified an area and whether it actually belongs to a cluster.

Table 1 summarises the average cluster size, sensitivity and specificity of cluster detection using a particular model (GLM, GLMM or ZIP) over the 100 simulated datasets for each combination of relative risk, cluster size, covariate and data model (Poisson and ZIP).

In general, specificity is quite high, which means that the proportion of areas with no exceeded risk included in a cluster is quite low. Sensitivity behaves similarly as in other simulation studies (see, e.g. [6]), and it primarily increases with size and risk inside the cluster. Small clusters (i.e. size 5) with high risk (i.e. 3) seem to be the best detected clusters. The average size of the detected clusters also depends on the actual size and risk, but seems to be dominated by whether adjustment for the

Table 1 Results of simulation study to assess the performance of our method for the detection of disease clusters

	Simulation parameters			Poisson data			ZIP data		
Model	Cov. adj.	RR	Size	cl. size	Sens	Spec	cl. size	Sens	Spec
Covariate with mild overdispersion									
GLM	No	1.50	5.00	11.20	0.62	0.77	10.80	0.54	0.77
GLM	No	3.00	5.00	7.10	0.96	0.93	7.99	0.83	0.89
GLM	No	1.50	15.00	14.50	0.67	0.82	11.70	0.47	0.82
GLM	No	3.00	15.00	13.10	0.81	0.96	12.23	0.73	0.95
GLM	Yes	1.50	5.00	6.80	0.44	0.87	8.84	0.48	0.82
GLM	Yes	3.00	5.00	7.20	0.98	0.93	6.97	0.83	0.92
GLM	Yes	1.50	15.00	6.40	0.38	0.97	8.72	0.40	0.89
GLM	Yes	3.00	15.00	12.30	0.77	0.97	11.89	0.72	0.96
GLMM	No	1.50	5.00	7.80	0.40	0.83	5.02	0.27	0.90
GLMM	No	3.00	5.00	4.82	0.86	0.98	1.64	0.21	0.98
GLMM	No	1.50	15.00	10.07	0.48	0.89	6.20	0.27	0.91
GLMM	No	3.00	15.00	9.63	0.63	0.99	1.71	0.11	1.00
GLMM	Yes	1.50	5.00	5.89	0.38	0.89	5.37	0.29	0.89
GLMM	Yes	3.00	5.00	5.91	0.92	0.96	2.61	0.37	0.98
GLMM	Yes	1.50	15.00	8.39	0.44	0.93	6.27	0.30	0.93
GLMM	Yes	3.00	15.00	10.87	0.70	0.99	3.12	0.20	0.99
ZIP	No	1.50	5.00	8.93	0.47	0.81	9.05	0.52	0.82
ZIP	No	3.00	5.00	6.69	0.93	0.94	6.38	0.85	0.94
ZIP	No	1.50	15.00	11.98	0.56	0.86	10.11	0.44	0.86
ZIP	No	3.00	15.00	13.00	0.82	0.97	12.55	0.79	0.97
ZIP	Yes	1.50	5.00	5.58	0.37	0.89	5.54	0.38	0.90
ZIP	Yes	3.00	5.00	6.12	0.93	0.96	5.38	0.84	0.97
ZIP	Yes	1.50	15.00	8.09	0.42	0.93	6.30	0.33	0.94
ZIP	Yes	3.00	15.00	12.53	0.80	0.98	12.28	0.76	0.97
GLM	No	1.50	5.00	12.90	0.64	0.72	9.15	0.36	0.79
GLM	No	3.00	5.00	12.00	0.82	0.77	11.04	0.73	0.79
GLM	No	1.50	15.00	10.60	0.39	0.81	10.31	0.32	0.78
GLM	No	3.00	15.00	11.50	0.59	0.90	11.10	0.56	0.89
GLM	Yes	1.50	5.00	6.00	0.46	0.89	8.17	0.45	0.83
GLM	Yes	3.00	5.00	6.00	0.82	0.95	7.21	0.82	0.91
GLM	Yes	1.50	15.00	7.30	0.44	0.97	9.32	0.42	0.88
GLM	Yes	3.00	15.00	13.00	0.83	0.98	11.71	0.70	0.95
GLMM	No	1.50	5.00	0.00	0.00	1.00	0.00	0.00	1.00
GLMM	No	3.00	5.00	0.00	0.00	1.00	0.00	0.00	1.00
GLMM	No	1.50	15.00	0.00	0.00	1.00	0.00	0.00	1.00
GLMM	No	3.00	15.00	0.00	0.00	1.00	0.00	0.00	1.00
GLMM	Yes	1.50	5.00	5.98	0.46	0.89	4.97	0.28	0.90
GLMM	Yes	3.00	5.00	5.67	0.85	0.96	2.47	0.29	0.97
GLMM	Yes	1.50	15.00	8.09	0.40	0.91	4.62	0.22	0.95
GLMM	Yes	3.00	15.00	9.99	0.64	0.98	1.91	0.11	0.99

(continued)

Table 1 Continued

	Simulation parameters			Poisson data			ZIP data		
Model	Cov. adj.	RR	Size	cl. size	Sens	Spec	cl. size	Sens	Spec
Covariate with strong overdispersion									
ZIP	No	1.50	5.00	11.84	0.55	0.74	11.07	0.48	0.75
ZIP	No	3.00	5.00	14.38	0.90	0.72	12.36	0.77	0.76
ZIP	No	1.50	15.00	13.79	0.49	0.74	12.29	0.40	0.75
ZIP	No	3.00	15.00	13.20	0.64	0.85	10.92	0.52	0.87
ZIP	Yes	1.50	5.00	5.90	0.46	0.90	4.93	0.36	0.91
ZIP	Yes	3.00	5.00	6.04	0.88	0.95	5.32	0.80	0.96
ZIP	Yes	1.50	15.00	7.91	0.39	0.92	6.94	0.35	0.93
ZIP	Yes	3.00	15.00	12.99	0.82	0.97	11.36	0.71	0.97

The summary is split according to the two covariates used in the simulation study that will induce mild (top) or strong (bottom) overdispersion when not accounted for in the model

covariate is done. Adjusting for covariates seems to have a positive effect on the cluster detection as it increases sensitivity and specificity, as well as provides better estimates of the cluster size. This supports the use of the methods described in this paper for the detection of disease clusters including covariates and cluster variables.

Regarding the use of different models for cluster detection for Poisson data, GLMs provide good detection, which increases with cluster size and risk. Cluster detection is also better when the appropriate covariates have been adjusted for in the model. GLMMs increase specificity but are not able to detect clusters under strong overdispersion as the two patterns cannot be disentangled. Under mild overdispersion, they detect smaller clusters and have a high specificity. This is probably due to the fact that random effects pick up some of the extra variation within the cluster areas and these produces smaller cluster being detected. ZIP models provide similar results as GLMs, probably because the estimates of π are very close to zero and it becomes a Poisson model. Under strong overdispersion, they seem to have lower sensitivity and specificity.

Regarding the performance for ZIP datasets, Poisson models have lower sensitivity and specificity than for Poisson datasets. In particular, this happens when there is no adjustment for significant covariates. GLMMs also have a lower sensitivity but a higher specificity, which probably means that zero inflation and clustering are picked up by the random effects in the model. ZIP models perform worse for ZIP data than for Poisson data. However, they are still better than using a Poisson model in terms of sensitivity when there is strong overdispersion. With no overdispersion, they have higher specificity than Poisson models and similar sensitivity.

Finally, all analyses have been run on a high performance computer using six nodes so that computations are run in parallel. Total computing time for each dataset was between 20 and 140 s for this region. For each dataset, we have fit about $40 \times (40 \times 0.5) = 800$ models because we have considered clusters that are up to a 50% of the total population. This means fitting about six models (i.e. investigating six

putative clusters) per second in the worst-case scenario and about 40 models per second in the best-case scenario. Fitting any of these models using MCMC using a reasonable number of iterations will require, at least, 1 s, which means that we could fit about six models per second using six nodes on the computer in the best-case scenario.

9 Examples

The methods presented in this paper have been applied to a number of case studies. They have been implemented using the R programming language [32]. Packages `DCluster` [17] and `R-INLA` have been used to embed fitting Bayesian hierarchical models with INLA into the general framework of spatial scan statistics. All datasets described here can be obtained by downloading R package `DClusterm` [18] from CRAN.

9.1 Cancer in Upstate New York

In the first example, we revisit the dataset on leukaemia incidence in New York [42]. This dataset comprises cases of leukaemia in upstate New York and its possible relationship to TCE-contaminated waste sites. Ahrens et al. [2] highlight the importance of accounting for relevant factors to avoid drawing misleading conclusions on the causes of leukaemia. This example is particularly interesting because covariates measure not only socio-economic variables but also proximity to putative pollution centres.

Data are available at the census tract level, for which number of cases, population and other risk factors are available. Raw expected cases e_i were computed using the population in each census tract. Covariate standardised expected number of cases E_i were computed fitting a Poisson regression with offset $\log(e_i)$ on three covariates: the percentage of the population aged 65 or more, percentage of population who own their home and a measure of exposure based on the inverse distance to the nearest TCE site. Then, the fitted values from this model were used to compute the expected number of cases using Eq. (1).

Our analyses will use raw data with no covariate adjustment (e.g. using e_i) and analysis after covariate adjustment (e.g. with E_i) for the spatial scan statistic, GLMs and GLMMs. When summarising the clusters found for a particular model and adjustment for covariates, we have only shown non-overlapping clusters.

9.1.1 Spatial Scan Statistic

Firstly, we have run the spatial scan statistic as described in [1]. This computes the p-value using a Gumbel distribution (often used to model the maximum of

Table 2 Summary of non-overlapping clusters detected using the spatial scan statistic

ID	Centre	Statistic	Size	Observed	Expected	SMR	p-value
No covariates							
CL1	36007014300	287,500.94	29	103.63	62.97	1.65	0.00e+00
CL2	36023990600	4912.04	9	44.50	22.78	1.95	0.00e+00
CL3	36067000400	475.74	16	44.69	25.56	1.75	7.77e−16
CL4	36011991300	207.43	4	27.30	13.75	1.99	3.02e−06
Adjusting for covariates							
CL2	36023990700	1212.02	9	43.43	22.17	1.96	0.00e+00
CL5	36067005700	185.81	3	11.13	4.95	2.25	1.85e−02
CL3	36067000400	84.90	16	44.69	25.56	1.75	9.86e−03

ID is a label for the cluster, centre is the census tract where the centre of the cluster is, size is the number of regions in the cluster, p-value is the one associated to the test statistic, observed is the number of cases in the cluster, expected the number of expected cases in the cluster and SMR is the standardised mortality ratio in the cluster

a number of samples from a distribution) instead of Monte Carlo methods. In all cases, we have considered clusters containing up to a 15% of the total number of expected cases and a p-value lower than 0.05. Now clusters are ordered according to increasing p-values, so that the cluster with the lowest p-value is ranked first. The number of clusters have been reduced by removing overlapping clusters as explained in Sect. 4.1, and the remaining clusters are outlined in Table 2 and shown in Fig. 1.

Clusters detected have been labelled according to their location. In order to compare the clusters among the different methods, we have assigned the same label to two clusters that overlap, even if they do not share all their areas completely.

9.1.2 Cluster Selection Using GLMs

The results obtained with our method based on cluster covariates and GLMs are shown in Table 3 and displayed in Fig. 1. As it can be seen, the results are very similar to those found with the spatial scan statistic. However, our method is also able to provide an estimate of the cluster risk, a 95% credible interval and the posterior probability of the risk being lower than 1, i.e. $P(\gamma_j < 0|y)$. In addition, we have displayed the DIC of the model including the cluster alone and all the previous clusters detected (under column *DICadj*). This is the DIC for a model that adjusts for several clusters and will give us information about how several clusters perform together.

The DIC of the model with no cluster covariates can be used as a threshold to compare all the other models with cluster covariates. For the model with no covariates, it is 957.62, whilst when we adjust for the covariates, it becomes 882.72. As seen in Table 3, only a few clusters have an important (i.e. higher than 5, as suggested by Spiegelhalter et al. [35]) difference with the DIC of the null model (i.e. a model with no cluster covariates). We shall consider these clusters as definitive clusters, whilst the others may have appeared due to the random variation of the

Fig. 1 Clusters detected by the different methods discussed in this paper (top) and zoom around Syracuse city in Onondaga County (bottom). Clusters have been labelled according to their location, and it is possible that two overlapping clusters obtained with different models have the same label even if they do not have all their areas in common

Table 3 Summary of non-overlapping clusters detected with the Bayesian GLM

ID	Centre	Size	Obs.	Exp.	SMR	DIC	DICadj	γ_j	$P(\gamma_j < 0 \mid y)$	95% C.I. γ_j
No covariates										
CL1	36007014000	33	107.81	67.18	1.60	938.90	938.90	0.47	6.20e−06	(0.28, 0.66)
CL2	36023990700	9	43.43	22.17	1.96	943.73	925.01	0.67	1.11e−05	(0.36, 0.96)
CL4	36011991100	4	27.31	13.75	1.99	949.27	916.65	0.69	3.05e−04	(0.29, 1.04)
CL3	36067001500	4	14.17	6.29	2.25	952.36	911.38	0.81	2.95e−03	(0.25, 1.30)
CL6	36067002200	6	10.11	3.91	2.58	952.81	906.57	0.95	3.21e−03	(0.28, 1.52)
CL5	36067005700	3	11.13	4.95	2.25	953.92	902.87	0.81	6.67e−03	(0.18, 1.36)
CL7	36017990400	2	9.98	4.52	2.21	954.73	899.97	0.79	1.28e−02	(0.12, 1.37)
CL8	36053030502	2	7.54	3.29	2.29	955.60	897.95	0.83	2.26e−02	(0.05, 1.49)
Adjusting for covariates										
CL2	36023990700	9	43.43	22.17	1.96	871.23	871.23	0.61	6.07e−05	(0.30, 0.90)
CL5	36067005700	3	11.13	4.95	2.25	874.36	862.86	1.15	4.42e−04	(0.52, 1.69)
CL3,CL6	36067001500	37	70.37	50.95	1.38	877.64	857.77	0.33	4.73e−03	(0.09, 0.56)
CL8	36053030502	2	7.54	3.29	2.29	879.06	854.11	1.01	7.18e−03	(0.24, 1.67)

ID is a label for the cluster, centre is the census tract where the centre of the cluster is, size is the number of regions in the cluster, observed is the number of cases in the cluster, expected the number of expected cases in the cluster, SMR is the standardised mortality ratio in the cluster, DIC is the model with a single cluster covariate, DICadj is the DIC of a model with the cluster covariates form the previous rows in the model and γ_j is the associated cluster coefficient

data. It should be noted that all these 'spurious' clusters have, in general, a very small size.

This might happen because the DIC does not penalise enough. Broman and Speed [7] also acknowledge this problem and suggest multiplying the penalty term (in our case, the effective number of parameters) by a constant. We prefer to stop adding clusters to the list when the decrease in the DIC is not important (e.g. lower than 5), but we have reported all clusters with significant cluster coefficients. As we will see below, this is less of a problem when random effects are considered.

9.1.3 Cluster Selection Using GLMMs

Next, we have performed cluster detection when i.i.d. random effects are included in the model. This is the model shown in Eq. (7), in Sect. 5.2. The results are shown in Table 4 and displayed in Fig. 1. The DIC of the models without cluster variables are 923.87 (no adjustment for covariates) and 882.70 (adjusting for covariates).

In this case the comparison of the DIC of the univariate models and that of the null model is better seen when adjusting for several clusters. In addition, we have seen that some of the detected clusters (not included in the table) have a non-significant increased risk because the 95% credible intervals contains the zero.

It should be noted that when the expected counts are not adjusted for the covariates Dean's test for overdispersion gives a p-value of 4.70e−08, whilst the p-value when adjusting for covariates is 4.10e−02 (which means very weak overdispersion). This is the reason why the differences in the DIC between GLMs and GLMMs are so small, and they provide very similar values when covariates are taken into account.

9.2 Analysis of Zero-Inflated Data: Brain Cancer in Navarre, Spain

In this example we consider the number of cases of brain cancer among the male population in Navarre, Spain, in the period 1988–1994 (see [36, 37] for details). The aggregation level is the basic health zone (BHZ), and the expected number of cases has been computed using standardisation by age group. This dataset has been positively tested for zero inflation. Hence, we will use a zero-inflated Poisson (ZIP) model in this case. A similar analysis can be found in [16] using maximum likelihood estimation.

We have conducted an analysis using the spatial scan statistic, our method using both Poisson and zero-inflated Poisson models described in Eq. (8), in Sect. 6. The only cluster detected using a Poisson and ZIP model is shown in Table 5. The DIC of the null model (without any cluster covariate) is 142.02. Figure 2 shows the SMR of brain cancer and the clusters detected with the different methods.

Table 4 Summary of non-overlapping clusters detected with the Bayesian GLMM

ID	Centre	Size	Obs.	Exp.	SMR	DIC	DICadj	γ_j	$P(\gamma_j < 0 \mid y)$	95% C.I. γ_j
No covariates										
CL1	36007014000	18	69.70	61.90	1.13	921.30	921.30	0.51	6.55e−04	(0.21, 0.80)
CL6	36067002200	6	10.11	4.25	2.38	922.01	919.13	0.94	1.13e−02	(0.18, 1.63)
CL2	36023990400	6	33.71	17.76	1.90	922.24	918.64	0.74	7.34e−04	(0.28, 1.17)
CL3	36067001500	4	14.17	8.71	1.63	923.28	916.17	0.77	1.05e−02	(0.09, 1.39)
CL4	36011991100	4	27.31	24.81	1.10	923.75	909.42	0.65	9.37e−03	(0.11, 1.17)
Adjusting for covariates										
CL2	36023990700	9	43.43	23.50	1.85	870.94	870.94	0.61	5.97e−05	(0.30, 0.90)
CL5	36067005700	3	11.13	3.53	3.15	873.48	862.75	1.14	4.41e−04	(0.50, 1.70)
CL3,CL6	36067001500	37	70.37	50.34	1.40	876.80	857.80	0.33	4.74e−03	(0.09, 0.56)
CL8	36053030502	2	7.54	2.75	2.75	878.08	854.19	1.01	7.11e−03	(0.22, 1.68)

Centre is the census tract that is the centre of the cluster, size is the ID is a label for the cluster, centre is the census tract where the centre of the cluster is, size is the number of regions in the cluster, observed is the number of cases in the cluster, expected the number of expected cases in the cluster, SMR is the standardised mortality ratio in the cluster, DIC is the model with a single cluster covariate, DICadj is the DIC of a model with the cluster covariates form the previous rows in the model, and γ_j is the associated cluster coefficient

Table 5 Summary of non-overlapping clusters of brain cancer in Navarre (Spain), 1988–1994

ID	Model		Centre			Statistic	Size	Observed	Expected	SMR	p-value
CL1	Spat. Sac Stat.		Puente la Reina			127.13	9	78	60.366	1.29	5.1e−03
ID	Model	Centre	Size	Obs.	Exp.	SMR	DIC	γ_j	$P(\gamma_j < 0\|y)$	95% C.I. γ_j	
CL1	Poisson	Artajona	4	17	8.51	2.00	138.45	0.69	6.14e−03	(0.186, 1.14)	
CL1	ZIP	Artajona	4	17	8.51	2.00	137.46	0.69	6.14e−03	(0.186, 1.14)	

ID is a label for the cluster, centre is the basic health zone (BHZ) where the centre of the cluster is, size is the number of regions in the cluster, observed is the number of cases in the cluster, expected the number of expected cases in the cluster, SMR is the standardised mortality ratio in the cluster, DIC is the model with a single cluster covariate and γ_j is the associated cluster coefficient

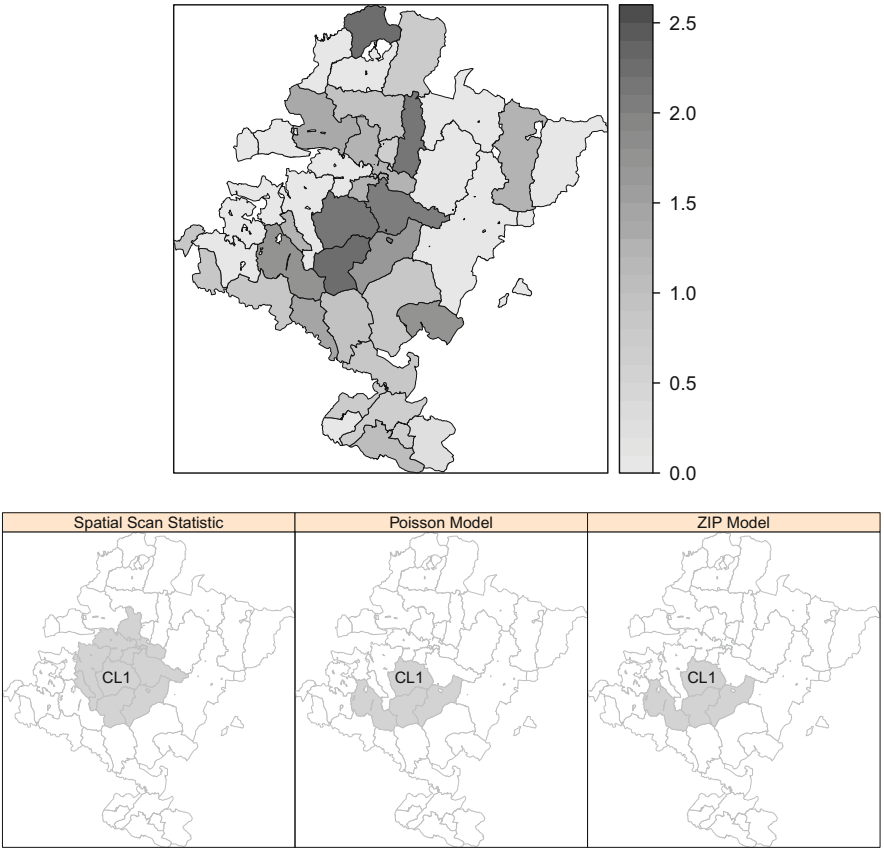

Fig. 2 Standardised Mortality Ratios (top) and clusters detected (bottom) of brain cancer in Navarre, Spain, 1988–1994

Regarding the clusters detected, all methods detected only one cluster. In the case of the spatial scan statistic, the size of the cluster detected is 9, which has two areas with zero cases. On the other hand, the methods based on the Poisson and ZIP model detected an overlapping cluster of size 4, with no areas with zero cases. Hence, accounting for zero-inflation provides a better detection of the clusters as areas with no cases are not included in the cluster. Furthermore, the posterior mean of π is 0.161, with a 95% credible interval of $(0.038, 0.397)$, which clearly indicates zero-inflation and is very similar to the value obtained in [36]. Hence, we believe that the default prior on π is not too informative in this case.

10 Discussion

In this paper we have proposed a new methodology for the detection of disease clusters based on the use of partition models and dummy variables in Bayesian hierarchical models to assign each area to a cluster. Although the main ideas are very similar to those found in [3, 6, 19, 39, 43, 44], our approach is based on methods to detect clusters in space, and that can be easily extended to space-time clusters, using different configurations of the dummy variables to mimic the spatial scan statistic. This new approach has several advantages. First of all, we have used INLA [34] to fit the Bayesian models, and the detection of clusters is done by means of the DIC [35]. Hence, we are able to detect the most significant clusters and avoid the use of simulation techniques, which are very time-consuming and computer-intensive. Second, the same approach is used to detect clusters regardless of whether fixed effects or mixed effects are included in the model. Detection of disease clusters with mixture models (such as zero-inflated Poisson) can also be tackled with our approach. Furthermore, the code used to develop the examples in this paper will eventually be included in the DClusterm R package [18], which is currently available on CRAN and implements similar (i.e. non-Bayesian) methods.

In order to show the potential of these ideas, we have conducted a simulation study that confirms that ours is a valid approach to detect disease clusters in a wide number of situations. Furthermore, we have considered several datasets that show different problems of cluster detection. In some ways, our results are similar to those found by other authors. However, we are able to expand on these results with our approach by quantifying the increase in risk within a disease cluster, and we are even able to model different risks within the clusters by means of random effects.

As a general advice on how to use the models presented in this paper, the GLM presented in Sects. 3 and 4 should be used when only the effect of the clusters and covariates is thought to affect the data. The model with random effects presented in Sect. 5 is useful to accommodate overdispersion in the data or the effect of unmeasured covariates (through the random effects). This a very general model that can be used with different types of likelihoods as well. Finally, the model based on the zero-inflated GLM, described in Sect. 6, is particularly addressed for situations where we find a large number of zero counts such as, for example, the analysis of rare diseases. Although the DIC will make a selection of the clusters, it is also important to consider the posterior probability of its associated coefficient of being higher than zero, as this is the actual indicator of increased risk.

Finally, the approach presented in this paper could be improved in a number of ways. First of all, another model selection criteria could be considered in addition to the DIC. Bilancia and Demarinis [6] use the DIC and CPO, but do not consider zero-inflated models. Although the presented method can be implemented in parallel, this aspect of the method could be improved by including a better algorithm to ensure that the same cluster is not tested twice, which may happen when two cluster centres correspond to regions which overlap. The lack of parallelization is not a problem with our current approach, but it could be used to reduce the computational burden

related to the analysis of larger datasets. In general, other Bayesian approaches could be considered in order to tackle the problem of cluster selection and risk assessment under different models for the observed number of cases. This would potentially reduce the problem of multiple testing inherent to the current approach.

References

1. Abrams AM, Kulldorff M, Kleinman K (2006). Empirical/asymptotic p-values for monte carlo-based hypothesis testing: an application to cluster detection using the scan statistic. Adv Dis Surveill 1(1):1
2. Ahrens C, Altman N, Casella G, Eaton M, Hwang JTG, Staudenmayer J, Stefanescu C (1999) Leukemia clusters in upstate New York: how adding covariates changes the story. Environmetrics 12(7):659–672
3. Anderson C, Lee D, Dean N (2014) Identifying clusters in Bayesian disease mapping. Biostatistics 15(3):457–469
4. Anderson C, Lee D, Dean N (2017) Spatial clustering of average risks and risk trends in Bayesian disease mapping. Biometrical J 59(1):41–56
5. Besag J, York J, Mollie A (1991) Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43(1):1–59
6. Bilancia M, Demarinis G (2014) Bayesian scanning of spatial disease rates with integrated nested laplace approximation (INLA). Stat Methods Appl 23(1):71–94
7. Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc Ser B 64(4):641–656
8. Burnham KP, Anderson DR (2002) Model selection and multimodel inference. A practical Information-theoretic approach, 2nd edn. Springer, New York
9. Cançado A, da Silva C, da Silva M (2014) A spatial scan statistic for zero-inflated poisson process. Environ Ecol Stat 21:627–650
10. Ferrándiz J, Abellán JJ, Gómez-Rubio V, López-Quílez A, Sanmartín P, Abellán C, Martínez-Beneito MA, Melchor I, Vanaclocha H, Zurriaga O, Ballester F, Gil JM, Pérez-Hoyos S, Ocaña R (2004) Spatial analysis of the relationship between cardiovascular mortality and drinking water hardness. Environ Health Perspect 112(9):1037–1044
11. Ferreira J, Denison DGT, Holmes CC (2002) Partition modelling. In: Lawson AB, Denison DGT (eds) Spatial cluster modelling, Chap 7. Chapman & Hall/CRC, Boca Raton, pp 125–145
12. Gangnon RE (2006) Impact of prior choice on local bayes factors for cluster detection. Stat Med 25:883–895
13. Gangnon RE, Clayton MK (2000) Bayesian detection and modelling of spatial disease clustering. Biometrics 56:922–935
14. Gangnon RE, Clayton MK (2003) A hierarchical model for spatially clustered disease rates. Stat Med 22:3213–3228
15. Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice. Chapman & Hall, Boca Raton, FL
16. Gómez-Rubio V, López-Quílez A (2010) Statistical methods for the geographical analysis of rare diseases. Adv Exp Med Biol 686:151–171
17. Gómez-Rubio V, Ferrándiz-Ferragud J, López-Quílez A (2005) Detecting clusters of disease with R. J Geogr Syst 7(2):189–206
18. Gomez-Rubio V, Serrano PEM, Rowlingson B (2018) DClusterm: model-based detection of disease clusters. R package version 0.2
19. Jung I (2009) A generalized linear models approach to spatial scan statistics for covariate adjustment. Stat Med 28(7):1131–1143

20. Knorr-Held L, Rasser G (2000) Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56:13–21
21. Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26(6):1481–1496
22. Kulldorff M (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework. J Am Stat Assoc 101(475):1289–1305
23. Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health 88:1377–1380
24. Lawson A (ed) (2005). Statistical methods in medical research special issue on disease mapping, vol 14(1). SAGE Publications, Thousand Oaks
25. Lawson AB, Gangnon RE, Wartenberg D (eds) (2006). Statistics in medicine. Special issue: developments in disease cluster detection, vol 25(5). Wiley, New York
26. Loh JM, Zhou Z (2007) Accounting for spatial correlation in the scan statistic. Ann Appl Stat 1:560–584
27. McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
28. McCullogh CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley, New York
29. Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A (General) 135(3):370–384
30. Openshaw S, Charlton M, Wymer C, Craft AW (1987) A Mark I geographical analysis machine for the automated analysis of point datasets. Int J Geogr Inf Syst 1:335–358
31. Prates MO, Kulldorff M, Assunção RM (2014) Relative risk estimates from spatial and space-time statistics: are they biased? Stat Med 33:2634–2644
32. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
33. Rothman KJ (1990) A sobering start for the cluster busters' conference. Am J Epidemiol Suppl. No. 1(132):S6–S13
34. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximation (with discussion). J R Stat Soc Ser B 71(2):319–392
35. Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J R Stat Soc Ser B 64(4):583–616
36. Ugarte MD, Ibáñez B, Militino AF (2004) Testing for poisson zero inflation in disease mapping. Biom J 46(5):526–539
37. Ugarte MD, Ibáñez B, Militino AF (2006) Modelling risks in disease mapping. Stat Methods Med Res 15:21–35
38. Vaida F, Blanchard S (2005) Conditional Akaike information for mixed-effects models. Biometrika 92(2):351–370
39. Wakefield J, Kim A (2013) A Bayesian model for cluster detection. Biostatistics 14:752–765
40. Walker SF, Bosch J, Gomez V, Garner TWJ, Cunningham AA, Schmeller DS, Ninyerola M, Henk DA, Ginestet C, Arthur C-P, Fisher MC (2010) Factors driving pathogenicity vs. prevalence of amphibian panzootic chytridiomycosis in iberia. Ecol Lett 13:372–382
41. Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley, Hoboken, NJ
42. Waller L, Turnbull B, Clark L, Nasca P (1992) Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence in TCE-contaminated dumpsites in upstate New York. Environmetrics 3:281–300
43. Zhang T, Lin G (2009) Cluster detection based on spatial associations and iterated residuals in generalized linear mixed models. Biometrics 65:353–360
44. Zhang T, Lin G (2009) Spatial scan statistics in loglinear models. Comput Stat Data Anal 53:2851–2858
45. Zhang Z, Assunção R, Kulldorff M (2010) Spatial scan statistics adjusted for multiple clusters. J Probab Stat 2010:1–11

A Novel Hierarchical Multinomial Approach to Modeling Age-Specific Harvest Data

Khurram Nadeem, Entao Chen, and Ying Zhang

Keywords Leslie matrix · White-tailed deer · Harvest data · Beta distribution · Population reconstruction

1 Introduction

Population reconstruction methods, originally developed in the context of quantitative stock assessment in fisheries [16], provide a useful framework for estimating population demographic trends in harvested wildlife population [6, 7, 17]. The traditional reconstruction techniques, such as the Downing method [6] and the sex-age-kill (SAK) method [15], are giving way to recently developed more powerful likelihood-based statistical population reconstruction (SPR) methods [5, 9, 18]. A key attraction of the SPR methodology is that it relies on age-specific or age-at-harvest data that are routinely collected by wildlife agencies over a large geographic scale. The analysis is based on a product multinomial likelihood of observed age-specific cohorts' harvest counts as a function of initial recruitment, survival, and harvesting mortality processes. However, age-at-harvest data alone are insufficient to parse these demographic processes as auxiliary data are needed to ensure identifiability of the associated demographic parameters even in the simplest of SPR models [9]. The auxiliary data sources normally include a separate radiotelemetry study, independent abundance estimates or hunter catch-effort indices. Strong assumptions on the form of natural survival, such as constant survival for all age classes and years, are further required to fit these models [8, 9].

We develop a new modeling approach to estimate the age distribution, i.e., proportion of animals in various age classes, in a short time range from age-at-harvest data without requiring auxiliary data source. Our approach is based on stable age distribution properties of a Leslie age-classified matrix projection model

K. Nadeem · E. Chen · Y. Zhang (✉)
Acadia University, Wolfville, NS, Canada
e-mail: entaochen18@163.com; ying.zhang@acadiau.ca

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_2

through a Beta distribution approximation to the natural survivorship curve [13]. We estimate the yearly age distribution using a hierarchical multinomial model for the yearly age-specific harvest counts. Unlike the existing SPR methodology, which requires auxiliary reporting data to account for underreporting of harvest counts, our modeling approach automatically adjusts for reporting errors by assuming that the reported data constitute a random sample from the overall harvest count.

The rest of this paper will be presented in the following manner. After introducing a motivational example in Sect. 2, we propose a harvest Leslie matrix model and our estimation approach in Sect. 3. In Sect. 4, we investigate the estimation performance of our approach by generating age-at-harvest data using the stochastic harvest Leslie matrix model. In Sect. 5, we perform the analysis of a motivational example regarding the age distribution for a white-tailed deer population from year 2009 to 2013 in Nova Scotia, Canada.

2 Age-Specific Harvest Data

The Department of Natural Resources in Nova Scotia (NSDNR), Canada, is responsible for regulating white-tailed deer (Odocoileus virginianus) harvest across the whole province. Apart from a keystone species of the local fauna, white-tailed deer provides a valuable aesthetic and recreational resource for human residents. Harvesting also provides a management tool to maintain the deer herd size at levels tolerable for farmers and other factions of the society. The data considered are the jawbone age measures collected from harvesting of white-tailed deer from the year 2009 to the year 2013. Figure 1 shows the age frequency distributions of the jawbone samples in these 5 years. In each year, a sample of white-tailed deer jawbones collected from hunters, as well as through various programs within the NSDNR, was collected. The sampled jawbones were aged by advanced technologies to obtain precise estimates. Population age distribution information is important for wildlife population monitoring programs. Harvest data are routinely collected by wildlife management systems in Canada. Our goal is to develop a novel modeling approach with this type of harvest data in order to provide with the reliable population age profile.

3 Model Development

3.1 Harvest Leslie Matrix Model

The Leslie matrix, in principle, is an age-classified matrix projection model for modeling and predicting population growth based on survival and fertility information [1, 11, 12, 19]. The Leslie matrix divides the continuous variable *age*

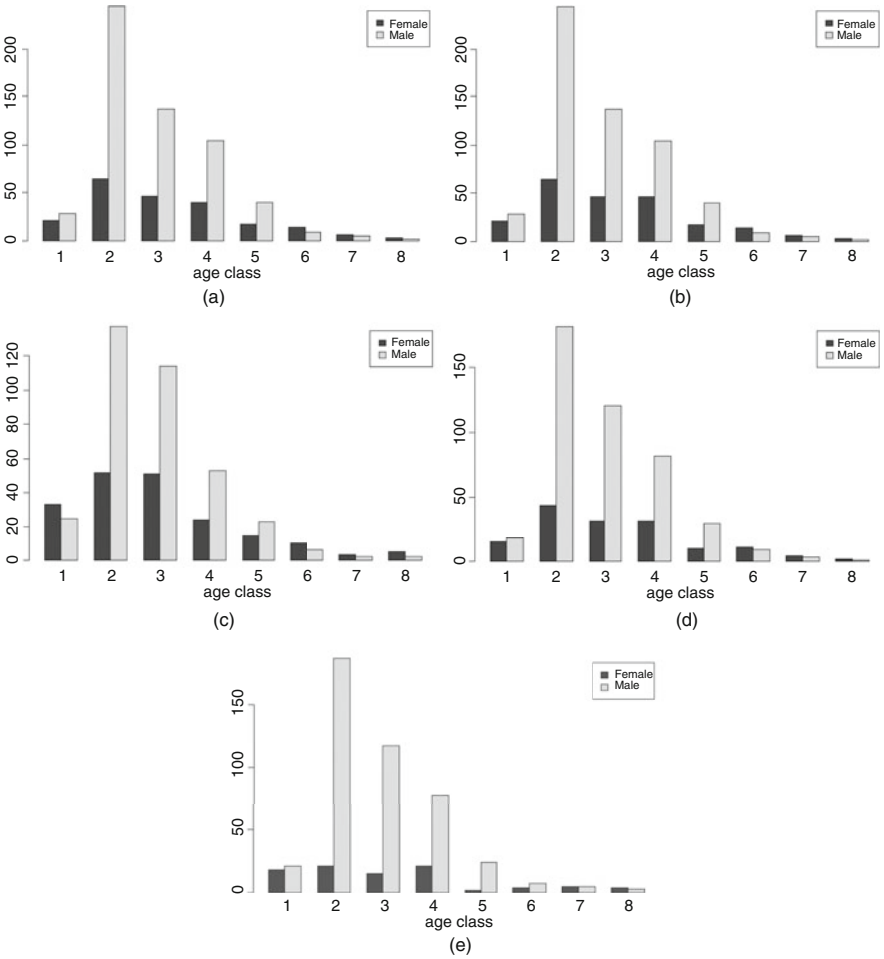

Fig. 1 Age class frequency distributions of jawbone samples from year 2009 to year 2013. (**a**) Year 2009 jawbone sample. (**b**) Year 2010 jawbone sample. (**c**) Year 2011 jawbone sample. (**d**) Year 2012 jawbone sample. (**e**) Year 2013 jawbone sample

into discrete age classes. Starting from 1, the age class i includes individuals in ages $(i-1) < x \leq i$. An individual of age i means that it belongs to age class i instead of natural age i. Suppose the animal population can be divided into A age classes. We use n_{it} to denote the number of individuals in age class i at time t such that a population vector $\mathbf{n}_t = (n_{1t}, n_{2t}, \ldots, n_{At})$ describes the population age distribution at time t. A simple Leslie matrix contains two components—a vector of fertilities, $\mathbf{F} = (F_1, F_2, \ldots, F_A)$, on the first row and a vector of survivals

$\mathbf{S} = (S_1, S_2, \ldots, S_{A-1})$, on the sub-diagonal:

$$\mathbf{M} = \begin{bmatrix} F_1 & F_2 & F_3 & \cdots\cdots & \cdots & F_A \\ S_1 & 0 & \cdots\cdots\cdots & \cdots & 0 \\ 0 & S_2 & \ddots & & & \vdots \\ \vdots & \ddots & S_3 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots\cdots\cdots & 0 & S_{A-1} & 0 \end{bmatrix}, \tag{1}$$

where F_i denotes the number of female animals born to a female of age i, and S_i denotes the probability of a female of age i surviving to age $i+1$. This form of Leslie matrix is also known as the *female-only* Leslie matrix [19], and thus $\mathbf{n}_t$ corresponds to the age distribution of females in the population. If sex components need to be considered, fertilities F_i and survivals S_i should be scaled by the sex ratio, or a two-sex Leslie matrix [19] should be employed. To simplify our demonstration, we will employ the one-sex Leslie matrix (1) and assume that the fertilities and survivals have been scaled by the sex ratio.

Starting from the initial population vector $\mathbf{n}_0$, the vector $\mathbf{n}_{t+1}$ can be projected by right multiplying the Leslie matrix $\mathbf{M}$ by $\mathbf{n}_t$:

$$\mathbf{n}_{t+1} = \mathbf{M} \cdot \mathbf{n}_t. \tag{2}$$

Noting that in the Leslie matrix (1), S_i denotes the overall survival from age class i to age $i+1$, it is the complement of the overall mortality. When we equip the Leslie matrix with harvest parameters, it is necessary to distinguish the *natural mortality* and the *harvest mortality*. Deaths caused by harvesting belong to the harvest mortality, while deaths caused by sources other than harvesting belong to the natural mortality. We use H_i to denote the probability that an individual in age class i survives the harvest, and now S_i represents the probability of surviving the natural death from age i to age $i+1$. In wildlife management, harvesting of mammal game species is regulated by setting up harvesting seasons, which are generally some continuous time segments in a year. The harvesting season in Nova Scotia usually lasts 1 month starting in late October. Although harvesting and natural death can happen simultaneously during the harvest season, we assume that harvesting plays a dominant role and thus natural death is negligible in the harvest season. Therefore, harvesting and natural death can be considered as independent events, and by probability theory, we have the following:

$$\begin{aligned} &P\{\text{survive from class } i \text{ to class } i+1\} \\ &\quad = P\{\text{survive natural death } \cap \text{ survive harvest}\} \\ &\quad = S_i H_{i+1}. \end{aligned} \tag{3}$$

We also assume that harvesting and reproduction are independent because the harvest season generally does not overlap with the breeding season of animals. We define a harvest matrix $\mathbf{H}$ as

$$\mathbf{H} = \begin{bmatrix} H_1 & 0 & \cdots & \cdots & 0 \\ 0 & H_2 & \ddots & & \vdots \\ \vdots & \ddots & H_3 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & H_{A-1} & 0 \\ 0 & \cdots & \cdots & 0 & H_A \end{bmatrix}, \tag{4}$$

to project the harvest yields. Suppose $\mathbf{n}_t$ is counted right after the harvest season of year t, then the population of next year $\mathbf{n}_{t+1}$, can be projected by

$$\mathbf{n}_{t+1} = \mathbf{HMn}_t. \tag{5}$$

That is, the population lives through a full-year life cycle described by the Leslie matrix (1) and then suffers harvesting before it enters the next time to be counted in year $t+1$. Thus we use $\mathbf{P} = \mathbf{HM}$ to denote the harvest Leslie matrix as

$$\mathbf{P} = \begin{bmatrix} F_1H_1 & F_2H_1 & \cdots & \cdots & F_AH_1 \\ S_1H_2 & 0 & \cdots & \cdots & 0 \\ 0 & S_2H_3 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & S_{A-1}H_A & 0 \end{bmatrix}, \tag{6}$$

where $F_i, i = 1, \ldots A$, are fertilities; $S_i, i = 1, \ldots, (A-1)$, are natural survival probabilities; and $H_i, i = 1, \ldots, A$, are harvest survival probabilities. Since $\mathbf{H}$ is a diagonal matrix, $\mathbf{P}$ has the same structure as $\mathbf{M}$. By the Perron–Frobenius theorem, $\mathbf{P}$ has a dominant eigenvalue and the associated eigenvector that determine the population growth rate and the stable age distribution. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_A)^T$ be the vector of the stable age distribution determined by (6), it then satisfies

$$\mathbf{P}\boldsymbol{\pi} = \lambda\boldsymbol{\pi}, \tag{7}$$

where λ is the right eigenvalue of (6). Writing (7) explicitly, we have

$$S_1H_2\pi_1 = \lambda\pi_2$$
$$S_2H_3\pi_2 = \lambda\pi_3$$

$$S_3 H_4 \pi_3 = \lambda \pi_4 \tag{8}$$

$$\vdots$$

$$S_{A-1} H_A \pi_{A-1} = \lambda \pi_A$$

Assuming the population is a birth-pulse population and the Leslie matrix is built right after the breading season, the survival probabilities are approximated by

$$S_i = \frac{l(i)}{l(i-1)}, \tag{9}$$

where $l(i)$ is the survivorship function of the underlying population, that is, the probability that an individual survives up to time i. Note that the approximation of S_i is based on the birth type (birth-pulse or birth flow) and time when the matrix is built during the life circle. Different approximations can be found in [3]. Lynch and Fagan [2] modeled the survivorship function as

$$l(i) = 1 - F_{\text{Beta}}(i/A), \tag{10}$$

where $F_{\text{Beta}}(\cdot)$ denotes the cumulative density function of a Beta distribution with nonnegative shape parameters (α,β) and A is the maximum lifespan of the species. Here age a is rescaled by A to fall in the interval [0,1], the support of the Beta probability distribution. The shape parameters (α,β) and *longevity* A describe the shape and scale of the survivorship function, respectively [13]. The Beta survivorship function (10) has been found to provide excellent fits to survivorship schedules of a wide range of mammal species [13, 14]. Notice that the estimation of $l(i)$ using (10) is based on annual survival counts of a cohort starting from birth to death of the last cohort members. If the population is stationary ($\lambda = 1$), solution to the age distribution may be written as

$$\pi_i = \left(1 - F_{\text{Beta}}\left(\frac{i-1}{A}, \alpha, \beta\right)\right) H_i \ldots H_2 \pi_1, \tag{11}$$

where $i = 2, \ldots, A$. Thus the age distribution can be obtained by estimating the beta distribution shape parameters (α,β) given the information of harvest probabilities.

3.2 Beta Distribution-Based Hierarchical Multinomial Model

Given the age-at-harvest data, with the assumptions that:

(i) The underlying harvested population follows a stochastic matrix process model with a mean matrix **P** that may be characterized by (6)
(ii) The process is stationary

(iii) If the data were reported sample values, it should be a simple random sample of the total harvest data

then the procedure proposed below is to estimate the stable age distribution $\boldsymbol{\pi}$ based on Eq. (11).

We summarize some key notations in Table 1. Specifically, we denote the age-at-harvest data matrix by $\mathbf{X}_{Y\times A}$, over Y years and A age classes. Depending on the context, entries of $\mathbf{X}_{Y\times A}$ represent the harvest counts of a reported sample thereof. Furthermore, we denote random variables and their realized values by uppercase and lowercase letters, respectively. The italic, boldfaced letters represent vector-valued random variables. The age classes are denoted by a in this section. Assumptions (i) and (ii) indicate that the mean level of the age distribution is converged to a constant vector of age distribution and the mean level of the total size of the population is converged to a constant over time. Thus we can use the approximation (11) to model the age proportions. Assumption (iii) guarantees that the model likelihood using reported age-at-harvest data is identical to the one using total age-at-harvest data. We will further discuss the model likelihood after we specify the model.

We model the annual age-specific harvest, X_{ya}, by the following binomial distribution:

$$P(X_{ya} = x_{ya}|N_{ya}) = \text{Binom}(N_{ya}, p_{ya}), \tag{12}$$

Table 1 Notation and definitions of various quantities

Notation	Description
N_{ya}	Number of animals alive of age a in year y; $a = 1, 2, \ldots, A$; $y = 1, 2, \ldots, Y$
$N_{y\cdot}$	Annual population size in year y, i.e., $N_{y\cdot} = \sum_{a=1}^{A} N_{ya}$
x_{ya}	Number of animals harvested and reported in age class a in year y. The corresponding harvest vector is denoted as $\boldsymbol{x}_y = (x_{y1}, x_{y2}, \ldots, x_{yA})^T$
$x_{y\cdot}$	Total reported harvest size in year y, i.e., $x_{y\cdot} = \sum_{a=1}^{A} x_{ya}$. The corresponding total harvest vector is denoted by $\boldsymbol{x}_{\cdot} = (x_{1\cdot}, x_{2\cdot}, \ldots, x_{Y\cdot})^T$
$\mathbf{X}_{(Y\times A)}$	Reported age-at-harvest data matrix: $\mathbf{X}_{(Y\times A)} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_Y^T \end{bmatrix}$
$h_{y\cdot}$	Total, reported plus unreported, harvest count in year y
π_a	The underlying mean proportion of the population in age class a
π_{ya}	True latent proportion of the population of age a in year y
p_{ya}	Binomial probability of harvesting an animal of age a in year y from age-specific population of size N_{ya}

where p_{ya} is the probability that an animal belonging to age class a is harvested in year y. Further conditioning on the total annual harvest $X_{y\cdot}$, we propose the following multinomial model for yearly harvest counts, $\mathbf{X}_y = (X_{y1}, X_{y2}, \ldots, X_{yA})^T$:

$$P(\mathbf{X}_y = \mathbf{x}_y | X_{y\cdot}) = \text{Multinom}(X_{y\cdot}, \boldsymbol{\pi}_y^{(x)}), \tag{13}$$

where we parameterize the probability vector $\boldsymbol{\pi}_y^{(x)} = (\pi_{y1}^{(x)}, \pi_{y2}^{(x)}, \ldots, \pi_{yA}^{(x)})^T$ as follows:

$$\pi_{ya}^{(x)} = \frac{E(X_{ya}|N_{ya})}{\sum_{a=1}^{A} E(X_{ya}|N_{ya})}. \tag{14}$$

That is, $\pi_{ya}^{(x)}$ is the proportion of total expected harvest that falls in age a in year y. Noticing that $E(X_{ya}|N_{ya}) = N_{ya}\, p_{ya}$ and denoting $\pi_{ya} = N_{ya}/N_{y\cdot}$, (14) can be written as

$$\begin{aligned} \pi_{ya}^{(x)} &= \frac{N_{ya}\, p_{ya}}{\sum_{a=1}^{A} N_{ya}\, p_{ya}} \\ &= \frac{(N_{ya}/N_{y\cdot})\, p_{ya}}{(\sum_{a=1}^{A} N_{ya}\, p_{ya})/N_{y\cdot}} \\ &= \frac{\pi_{ya}\, p_{ya}}{\sum_{a=1}^{A} \pi_{ya}\, p_{ya}} \end{aligned} \tag{15}$$

Here, we model the *latent* age-specific abundance proportions, π_{ya}, as

$$\pi_{ya} = g^{-1}(g(\pi_a) + \epsilon_{ya}), \tag{16}$$

where $g(\cdot)$ is a smooth invertible link function, ϵ_{ya} is the age-specific random effect that is independently distributed as $N(0, \sigma_a^2)$, and π_a is the age distribution given by (11). Throughout this paper, we use the logit link function: $g(\pi) = \ln(\pi/(1-\pi))$.

The likelihood function for the observed age-at-harvest data may be written as

$$L(\boldsymbol{\theta}; \mathbf{X}|\mathbf{x}_{\cdot}) = \prod_{y=1}^{Y} \int P(\mathbf{X}_y; \alpha, \beta, \mathbf{p}_y | x_{y\cdot}, \boldsymbol{\epsilon}_y) g_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}_y; \sigma_{\boldsymbol{\epsilon}}^2) d\boldsymbol{\epsilon}_y, \tag{17}$$

where $\sigma_{\boldsymbol{\epsilon}}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_A^2)^T$ is the variance vector of the age-specific random effects introduced in (16), $\mathbf{p}_y = (p_{y1}, p_{y2}, \ldots, p_{yA})$ is the age-specific harvest probability vector in year y, $\mathbf{X}$ is the $Y \times A$ age-at-harvest data matrix, and $g_{\boldsymbol{\epsilon}_y}(\cdot)$ denotes joint density function of the random effects vector $\boldsymbol{\epsilon}_y = (\epsilon_{y_1}, \ldots, \epsilon_{y_A})$.

In the above derivations, we have implicitly assumed that $\mathbf{X}$ is the full age-at-harvest data matrix, i.e., entries of $\mathbf{X}$ are the true harvest counts without any underreporting. The following theorem, whose proof is relegated to Appendix,

shows that the likelihood (17) remains unchanged even when only a fraction of the harvested animals are aged and reported as a simple random sample.

Theorem 1 *Let* $\mathbf{h}_y = (h_{y1}, h_{y2}, \ldots, h_{yA})^T$ *now represent the full (reported plus unreported) age-specific harvest vector, which is modeled by the likelihood function given by*

$$L(\boldsymbol{\theta}; \mathbf{H}|\mathbf{h}_{\cdot}) = \prod_{y=1}^{Y} \int P(\mathbf{H}_y; \alpha, \beta, \mathbf{p}_y | h_{y\cdot}, \boldsymbol{\epsilon}_y) g_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}_y; \sigma_\epsilon^2) d\boldsymbol{\epsilon}_y, \tag{18}$$

where $h_{y\cdot} = \sum_{a=1}^{A} h_{ya}$, $\mathbf{h}_{\cdot} = (h_{1\cdot}, h_{2\cdot}, \ldots, h_{Y\cdot})^T$, *and* $\mathbf{H}$ *is the full age-at-harvest data matrix. Then, if* $x_{y\cdot} \leq h_{y\cdot}$ *and the animals are reported as a simple random sample with replacement (SRSWR) from the total* $h_{y\cdot}$ *harvested animals, the likelihood function corresponding to the reported age-at-harvest data matrix* $\mathbf{X}$ *is given as* $L(\boldsymbol{\theta}; \mathbf{X}|\mathbf{x}_{\cdot})$ *as defined in (17).*

The above theorem emphasizes that with the guarantee of assumption (iii), we can directly use the reported harvest sample as our input data without knowing the reporting rate, as long as it is a simple random sample of the full age-at-harvest data. Technically, $x_{y\cdot}$ cannot be a SRSWR from $h_{y\cdot}$ as we assume that none of the harvested animals are duplicated in the total age-reported harvest count, $x_{y\cdot}$. Rather, simple random sampling without replacement (SRSWoR) is a more reasonable sampling model. However, as we remark in Appendix, the SRSWR is a good approximation to SRSWoR as the annual harvest count, $h_{y\cdot}$, is normally sufficiently large for most of the harvested populations.

In model (17), we generally denote the yearly age-specific harvest probability vector as $\mathbf{p}_y = (p_{y1}, p_{y2}, \ldots, p_{yA})$. However, in the hunting process, hunters seldom hunt animals based on age of animals precisely. Besides, the action of hunters tends to be stable in a short time unless new management policies on hunting are introduced during the period. Based on this fact, we loosen the assumption on age-specific harvests and consider the harvest is time-invariant and uniform to all ages. However, there actually exists a selection between juveniles and adults. It is difficult to identify the age in hunting, but it is possible to tell apart juveniles and adults. For example, juvenile white-tailed deers are mostly raised by their mothers and have small antlers, while adult white-tailed deers usually have larger and sharper antlers. Hunters may tend to hunt adult deers for the reason such as trophy winning and avoid hunting juvenile deers for the sustainability of the population. Based on these assumptions, we introduce two time-invariant harvest probabilities in place of the yearly age-specific harvest, with the additional assumption that:

(iv) The harvest probabilities for juveniles and adults are constants over time.

We denote the harvest to juveniles and adults as p_{ju} and p_{ad}, respectively. In the white-tailed deer example described in Sect. 2, it is reasonable to consider all deer of at least 2 years of age as adults. Thus the overall model in this case may be

summarized as follows:

$$\begin{aligned}
\pi_a &= \left(1 - F_{\text{Beta}}\left(\frac{a-1}{A}, \alpha, \beta\right)\right) p_{ad}^{a-1} \pi_1 \\
\pi_{ya} &= g^{-1}(g(\pi(a)) + \epsilon_{ya}),\ \epsilon_{ya} \sim N(0, \sigma_a^2) \\
\pi_{ya}^{(x)} &= \frac{\pi_{ya} p_a}{\sum_{a=1}^{A} \pi_{ya} p_a} \\
p_1 &= p_{ju};\ p_2 = p_{ad}, \ldots, p_A = p_{ad} \\
\mathbf{X_y} | x_{y\cdot}, \boldsymbol{\epsilon}_y &\sim \text{Multinom}(x_{y\cdot}, \boldsymbol{\pi}_y^{(x)})
\end{aligned} \tag{19}$$

where $\boldsymbol{\pi}_y^{(x)} = (\pi_{y1}^{(x)}, \ldots, \pi_{yA}^{(x)})$. We will focus on this particular Beta distribution based hierarchical multinomial model (*BetaHM*) in the rest of the paper. We argue that the *BetaHM* model is also applicable to other species which reflect a hunting selection between juveniles and adults.

Since the likelihood functions (17) or (18) contain integrals over the random effects on age proportions, theoretical derivation of parameter estimators is challenging. For this reason, we attempt to perform an extensive simulation to evaluate the model estimation performance. A simulation trial includes generating an age-at-harvest data table and the trial of fitting the *BetaHM* model to the simulated age-at-harvest data table. In a single simulation trial, we generate a 5×8 age-at-harvest data table as the primary input of the *BetaHM* model. We perform 200 simulation trials in each scenario in order to collect the point estimates and the associated standard deviation of the model parameters. We employ non-informative Bayesian approach for model estimation in our simulations through the JAGS sampler. JAGS is a program for analysis of Bayesian hierarchical models using Markov chain Monte Carlo (MCMC) simulation. The Beta distribution parameters (α, β) are given log-Norm$(0, 1)$ prior distributions; the random effect variance (precision) parameters σ_i^{-2} are all given Gamma$(0.01, 0.01)$ prior distributions; and the harvest probabilities (p_{ju}, p_{ad}) are given Unif$(0, 1)$ distributions. All computations are performed via the *rjags* package under the R 3.2.2 computing environment.

4 Simulation Study

In each simulation trial, we obtain a set of point estimates of the parameters. With 200 simulation trials, we calculate the mean and standard error of the point estimates of each parameter.

4.1 Generating Data

We employ the stochastic matrix simulation framework proposed in Chen [4] to simulate population data and the corresponding age-at-harvest data. The objective is to generate the population and the age-at-harvest data which are similar to those generated from the exact *BetaHM* model setting but with life circle fluctuations that can resemble the real animal world in order to evaluate the robustness of the *BetaHM* parameterization procedure. The models assume stability and stationarity of the animal population, which can be ensured by setting up appropriate fertility, survival and harvest parameters of the mean Leslie matrix $\mathbf{P}$ to generate a mean growth rate $\lambda_{\mathbf{P}} = 1$. The stochastic Leslie matrix simulation model should naturally produce randomness in the age proportions by introducing probability distributions on the Leslie matrix parameters with their underlying mean values. The harvest data are generated from a binomial process according to age-specific harvest probabilities.

4.1.1 Matrix Parameter Settings

We consider a population with eight age classes and a 5-year successive age-specific harvest table as data. We use two harvest probabilities, p_{ju} for juveniles (age class = 1) and p_{ad} for adults (age classes ≥ 2), which are set to 0.05 and 0.20, respectively. We employ a Beta$(1, 3)$ distribution to approximate the *natural survivorship function* and further generate the postbreeding survival probability parameters (9). These natural survival parameters are approximated to 0.67, 0.63, 0.58, 0.51, 0.42, 0.30, and 0.13. We adjust the fertility parameters $\mathbf{f}$ so that $\lambda_{\mathbf{P}} = 1$. Without loss of generality, we adopt $\mathbf{f} = (0, 1.05, 1.05, 1.05, 1.05, 1.05, 1.05, 1.05)$, even though there are other possible choices of $\mathbf{f}$ which produce $\lambda_{\mathbf{P}} = 1$.

4.1.2 Matrix Randomness Settings

Without loss of generality, we employ uniform distributions to generate randomness to all of the Leslie matrix parameters. We employ Unif(1.05 ± 0.2) uniformly on $(f_2, \ldots, f_7)$, leaving $f_1 = 0$ constant. We employ Unif(0.67 ± 0.20), Unif(0.63 ± 0.20), Unif(0.58 ± 0.20), Unif(0.51 ± 0.20), Unif(0.42 ± 0.20), Unif(0.30 ± 0.15), and Unif(0.13 ± 0.05) on the natural survival parameter vector. We do not generate any randomness on the harvest parameters, as the *BetaHM* models model the harvest probabilities as fixed parameters.

We set the size of each age class uniformly equal to 10,000 which forms our initial population vector with the total size equal to 80,000. We project the population vector 50 times with the stochastic harvested Leslie matrix framework defined above. We retain the population- and age-specific harvest data from the last five projections to be fitted by the *BetaHM* models.

The assumptions remain the same as specified in Sect. 3. For simplicity, we use one random effect variance σ^2 (instead of the age-specific random effect variance) to fit the data generated.

4.2 Results

Table 2 shows that the harvest probabilities are estimated as $\hat{p}_{ju} = 0.0625$ and $\hat{p}_{ad} = 0.287$. The ratio between two harvest estimates is $0.0625/0.287 = 0.2177$, which is close to the true ratio $0.05/0.20 = 0.25$, preserving the order of two true harvest probabilities ($p_{ju} < p_{ad}$). Figure 2 reveals that the boxplots of the estimated age proportion capture the true mean population age distribution.

Table 2 Mean (standard error) of estimates of the parameters of the *BetaHM* model, fitting to data generated from the stochastic Leslie matrix setting using the type III survivorship curve Beta(1, 3)

Parameter	Value	Estimate
α	1	1.6312(0.6101)
β	3	2.888(0.4132)
p_{ju}	0.05	0.0625(0.0191)
p_{ad}	0.2	0.2877(0.0956)
σ^2	×	0.1312(0.0528)

× means the underlying true value of the parameter is unobserved in generating data

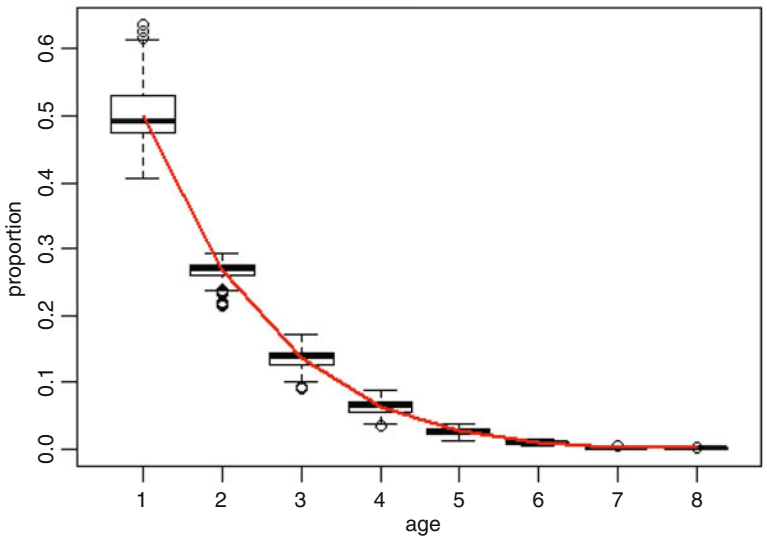

Fig. 2 Mean true age distribution curve and the boxplots of the estimates obtained from the *BetaHM* model fitting to data generated from the stochastic Leslie matrix setting using the type III survivorship curve Beta(1, 3)

As shown in Table 2, α and β are not precisely estimated. This implies the true parametric form of the Beta survivorship curve may not be recovered from the estimates $\hat{\alpha}$ and $\hat{\beta}$. As Eq. (19) shows that the age distribution is determined jointly by $\hat{\alpha}$, $\hat{\beta}$ and $\hat{p}_{ad}$, this is likely inducing confounding and bias in the corresponding estimates. However, our simulation study shows that estimation of the underlying age distribution is robust against this parameter confounding as the true form of the age distribution is accurately estimable in almost all cases. See Figs. 2, 3, 4.

The above Beta(1, 3) survivorship curve is referred to the type III survivorship. We also study the cases with the type I and type II survivorship curves. The type I survivorship is a convex decreasing curve, and we use Beta(5, 1) for example. The type II survivorship is an approximate linear decreasing curve, and we use Beta(0.5, 0.5) for example. With the change of survivorship curve, the constructed survival probabilities used in the harvested Leslie matrix model are also changed. We adjust the fertility parameters in the harvested Leslie matrix model to ensure it produces a mean-stationary population. The mean (standard error) of estimates are summarized in Tables 3 and 4. The estimated age distributions are displayed in Figs. 3 and 4.

Tables 3 and 4 show that neither the type I nor the type II survivorship can be recovered when α and β are poorly estimated. The harvest ratio $\hat{p}_{ju}/\hat{p}_{ad}$ is calculated as $0.033/0.164 = 0.201$ from Table 3 and as $0.057/0.240 = 0.237$ from Table 4. Similarly to the estimation results in Table 2, the estimated harvest ratio is preserving the order of two harvest rates but underestimating that true ratio 0.25. Figures 3 and 4 show that the age distribution is generally estimated well, except for

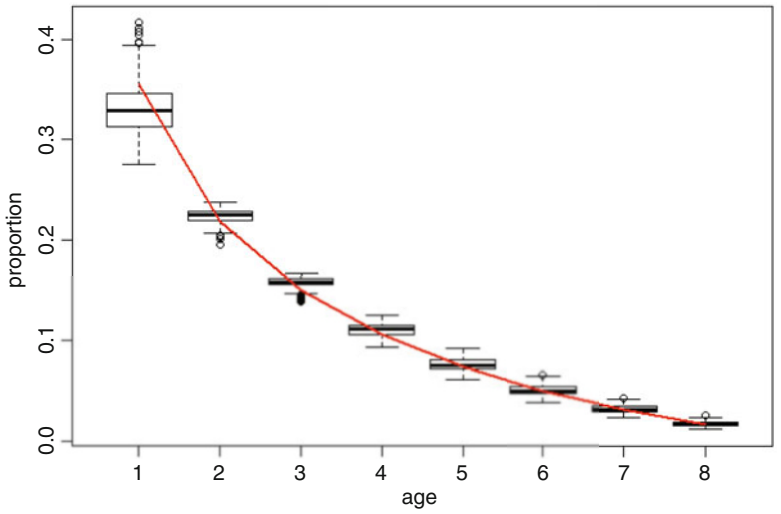

Fig. 3 Mean true age distribution curve and the boxplots of the estimates obtained from the *BetaHM* model fitting to data generated from the stochastic Leslie matrix setting using the type I survivorship curve Beta(5, 1)

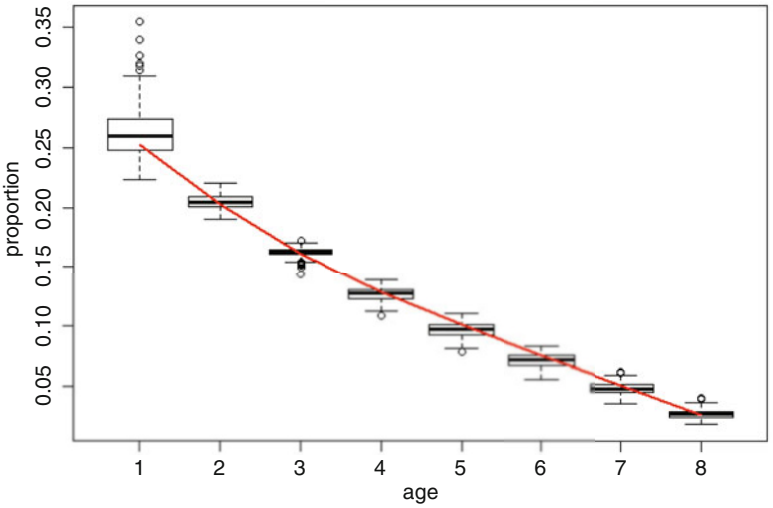

Fig. 4 Mean true age distribution curve and the boxplots of the estimates obtained from the *BetaHM* model fitting to data generated from the stochastic Leslie matrix setting using the type II survivorship curve Beta(0.5, 0.5)

Table 3 Mean (standard error) of estimates of the parameters of the *BetaHM* model, fitting to data generated from the stochastic Leslie matrix setting using the type I survivorship curve Beta(5, 1)

Parameter	Value	Estimate
α	5	2.4430(0.9445)
β	1	0.6967(0.1636)
p_{ju}	0.05	0.0333(0.0066)
p_{ad}	0.2	0.1645(0.0371)
σ^2	×	0.1558(0.0599)

Table 4 Mean (standard error) of estimates of the parameters of the *BetaHM* model, fitting to data generated from the stochastic Leslie matrix setting using the type II survivorship curve Beta(0.5, 0.5)

Parameter	Value	Estimate
α	0.5	1.8430(0.9420)
β	0.5	0.5352(0.1450)
p_{ju}	0.05	0.0572(0.0104)
p_{ad}	0.2	0.2400(0.0456)
σ^2	×	0.1069(0.0417)

the slight underestimation of age class 1 in Fig. 3 and the slight overestimation of age class 1 in Fig. 4.

We conclude that the *BetaHM* model can reasonably estimate the age distributions of the models described above. The beta shape parameters α, β and the harvest parameters p_{ju}, p_{ad} should be regarded as "assisting" parameters. Even though they are not correctly estimated, they can jointly ensure the age distribution to be

estimated well. The estimated ratio between two harvest probabilities maintains the correct order of the true ratio, despite of the shape of survivorship curve.

5 Motivating Example

In this section, we demonstrate the application of the *BetaHM* model by fitting the model to data described in Sect. 2. Although the jawbones are obtained from hunters on a voluntary basis, there is considerable variation in age and sex of jawbones returned by a specific hunter across multiple years. This ensures that collected sample is arising from a random catch, thereby making it sufficiently representative of the overall harvested deer population in a given year. We first visualize the age distributions of the jawbone samples in these 5 years, which are plotted in Fig. 1. From Fig. 1 we discover that the age distributions of the male deer jawbone samples are similar across 5 years, while those of the female jawbone samples are different across 5 years. We also notice that the male jawbones take a large proportion of the total jawbone sample in each year, suggesting that hunters should have sex selection in harvesting. In the distributions of male jawbone samples, there exists an age selection between juveniles and adults. The deer population size in the juvenile class is generally higher than in other classes, but on the contrary, we receive only a small proportion in the juvenile class in the reported jawbone sample (age class 1 in Fig. 1). This is the reason why we propose two harvest probabilities p_{ju} and p_{ad}. As the sex selection is a potential issue affecting the model estimation result, we only fit the *BetaHM* model to the male jawbone samples and estimate the age distribution in the male-only population. We assume that the male-only population is mean stable in its age distribution and mean stationary in the total size.

The harvest of juvenile male white-tailed deer is estimated as 0.0237, and the harvest of adult male white-tailed deer is estimated as 0.323. Notice that the *BetaHM* model cannot correctly estimate the harvest probabilities, we only interpret the ratio between two harvest probabilities. The ratio p_{ad}/p_{ju} is $0.323/0.0237 \approx 14$, which means the harvest of adults is 14 times higher than of the juveniles. See Table 5.

Figure 5 shows that the male white-tailed deer age proportion decreases as age increases. The juveniles take about 40% of the population, dominating the whole population. Since we do not have the reference age distribution of the male white-tailed deer for comparison, the interpretation of the age distribution estimation should be considered with caution.

Table 5 Estimation result of one trial fitting the *BetaHM* model to the male jawbone data

Parameter	Mean estimates(Stds)
α	1.988(0.6837)
β	2.007(0.4211)
p_{ju}	0.0237(0.0042)
p_{ad}	0.323(0.0553)
σ^2	0.0299(0.0406)

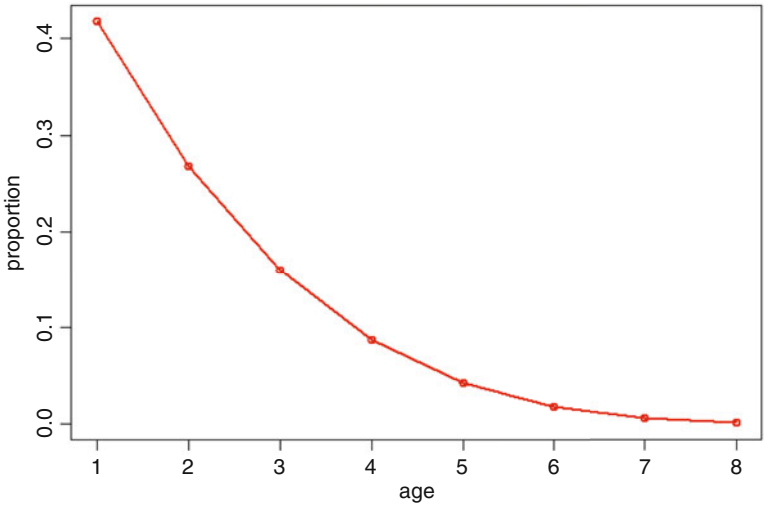

Fig. 5 Age distribution estimated by one trial of the BetaHM model fitting to the male jawbone data

6 Conclusion

The harvest data are routinely collected. Based on the stable age distribution properties of a harvest Leslie matrix model, we attempt to propose a population reconstruction model, the BetaHM model to estimate age proportions of the harvested animal population. Our modeling framework makes two basic assumptions concerning the age-structured population dynamics, i.e., (1) the harvested population is stable and stationary and (2) the available aged and reported sample of harvested animals represents a simple random sample from the underlying total annual harvest. Although we incorporate an age-specific structure on the harvest mortality to adjust for juvenile vs adult harvest rates, our modeling approach makes no assumptions about the underlying form of the age-specific natural mortality and fertility rates. This is an important relaxation from wildlife management point of view as direct incorporation of these rates in age-structure population modeling is challenging because auxiliary information about age-specific vital rates is rarely available for game populations. The simulation results from our stochastic Leslie

matrix models with two harvest rates (juvenile vs adult) show that the BetaHM model can reasonably estimate the age proportions and preserve the order between two harvest probabilities.

Appendix

Here we provide the proof of Theorem 1 stated in Sect. 3.2. We omit model parameter notation from the respective probability distribution for simplicity of exposition. We start with the following lemma.

Lemma 1 *Let* $\mathbf{H} = (H_1, H_2, \ldots, H_A)^T$ *and* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_A)^T$ *be random vectors such that*

$$P_H(\mathbf{H} = \mathbf{h}) = Multinom(M, \boldsymbol{\pi}_H) \tag{20}$$

$$P_Z(\mathbf{Z} = \mathbf{z}|\mathbf{h}) = Multinom(1, \boldsymbol{\pi}_Z) \tag{21}$$

where $\boldsymbol{\pi}_H = (\pi_1, \pi_2, \ldots, \pi_A)$, $\boldsymbol{\pi}_Z = (\frac{h_1}{M}, \frac{h_2}{M}, \ldots, \frac{h_A}{M})$ *and* M *is the number of multinomial trials. Then* $P_Z(\mathbf{Z} = \mathbf{z}) = Multinom(1, \boldsymbol{\pi}_H)$.

Proof Let us evaluate the probability for some fixed index value i^*:

$$P(Z_{i^*} = 1, Z_i = 0 \forall i \neq i^*) = \sum_S P(Z_{i^*} = 1, Z_i = 0 \forall i \neq i^* | \mathbf{H} = \mathbf{h}) P(\mathbf{H} = \mathbf{h}), \tag{22}$$

where S is the sample space containing all possible outcomes under $P(\mathbf{H} = \mathbf{h})$. We can partition S into $M+1$ mutually exclusive and exhaustive subsets $S_0, S_1, \ldots, S_M$, i.e.

$$S = \bigcup_{r=0}^{M} S_r \tag{23}$$

and $S_r \bigcup S_t = \emptyset$. Here, the r-th subset is defined as

$$S_r = \left\{ \mathbf{H} | H_{i^*} = r, \sum_{i \neq i^*} H_i = M - r \right\}. \tag{24}$$

As we have $\mathbf{H} \in S_r \Leftrightarrow H_{i^*} = r$

$$P(\mathbf{H} \in S_r) = P(H_{i^*} = r) = \binom{M}{r} \pi_{i^*}^r (1 - \pi_{i^*})^{M-r} \tag{25}$$

That is, (25) is in fact the marginal probability mass function (pmf) of H_{i^*}. Thus, the sum in (22) can be rearranged as follows:

$$\begin{aligned}
&P_Z(Z_{i^*} = 1, Z_i = 0 \forall i \neq i^*) \\
&= \sum_{r=0}^{M} P(Z_{i^*} = 1, Z_i = 0 \forall i \neq i^* | H_{i^*} = r) P(H_{i^*} = r) \\
&= \sum_{r=0}^{M} \left(\frac{h_1}{M}\right)^0 \left(\frac{h_2}{M}\right)^0 \cdots \left(\frac{r}{M}\right)^1 \cdots \left(\frac{h_M}{M}\right)^0 \binom{M}{r} \pi_{i^*}^r (1 - \pi_{i^*})^{M-r} \\
&= \frac{1}{M} \sum_{r=0}^{M} r \binom{M}{r} \pi_{i^*}^r (1 - \pi^{i^*})^{M-r} \\
&= \frac{1}{M} E(H_{i^*}) = \frac{1}{M} M \pi_{i^*}^r, \qquad (26)
\end{aligned}$$

and

$$P(Z_{i^*} = 0) = 1 - \pi_{i^*}^r, \qquad (27)$$

for some arbitrary index value $i^* = 1, 2, \ldots, A$. This can be represented more compactly as

$$P_Z(\mathbf{Z} = \mathbf{z}) = \pi_1^{Z_1} \pi_2^{Z_2} \cdots \pi_A^{Z_A}, Z_i \geq 0, \sum_{i=1}^{A} Z_i = 1. \qquad (28)$$

Thus, the result follows immediately.

Lemma 2 *Let $\{\mathbf{Z}_j\}_{j=1}^R$ be a collection of independently and identically distributed random vectors, where $\mathbf{Z}_j$ is distributed as*

$$P_Z(\mathbf{Z}_j = \mathbf{z}_j) = Multinom(1, \boldsymbol{\pi}_H), \qquad (29)$$

where $\mathbf{z}_j = (z_{j1}, z_{j2}, \ldots, z_{jA})^T$. Also define a random vector

$$\mathbf{X} = \left(\sum_{j=1}^{R} Z_{j1}, \sum_{j=1}^{R} Z_{j2}, \ldots, \sum_{j=1}^{R} Z_{jA} \right)^T. \qquad (30)$$

Then, the probability mass function of $\mathbf{X}$ is given as

$$P_X(\mathbf{X} = \mathbf{x}) = Multinom(R, \boldsymbol{\pi}_H) \qquad (31)$$

The proof of Lemma 2 can be found in Johnson et al. [10]. Next we show the proof of Theorem 1.

Proof Let $\mathbf{h}_y = (h_{y1}, h_{y2}, \ldots, h_{yA})^T$ denote the complete (reported plus unreported) age-specific harvest distribution of animals. This, in terms of the model (17), has the following conditional pmf:

$$P_H(\mathbf{H}_y = \mathbf{h}_y | h_{y\cdot}, \boldsymbol{\epsilon}_y) = \text{Multinom}(h_{y\cdot}, \boldsymbol{\pi}_y^{(x)}). \tag{32}$$

Also, let $\mathbf{z}_j = (z_{j1}, z_{j2}, \ldots, z_{jA})^T$ be the outcome of the j-th SRSWR draw from the full harvest distribution $\mathbf{h}_y$, $j = 1, 2, \ldots, x_{y\cdot}$. Notice that

$$\mathbf{x}_y = \left(x_{y1} = \sum_{j=1}^{x_{y\cdot}} z_{j1}, x_{y2} = \sum_{j=1}^{x_{y\cdot}} z_{j2}, \ldots, x_{yA} = \sum_{j=1}^{x_{y\cdot}} z_{jA} \right)^T. \tag{33}$$

Thus, it follows from the definition of the multinomial distribution that

$$P_Z(\mathbf{Z}_j = \mathbf{z}_j | \mathbf{h}_y, \boldsymbol{\epsilon}_y) = P_Z(\mathbf{Z}_j = \mathbf{z}_j | \mathbf{h}_y) = \text{Multinom}(1, \boldsymbol{\pi}_y^{(z)}), \tag{34}$$

where $\boldsymbol{\pi}_y^{(z)} = (\frac{h_{y1}}{h_{y\cdot}}, \frac{h_{y2}}{h_{y\cdot}}, \ldots, \frac{h_{yA}}{h_{y\cdot}})$. Then, it also follows from Lemma 1 that

$$P_Z(\mathbf{Z}_j = \mathbf{z}_j | \boldsymbol{\epsilon}_y) = \text{Multinom}(1, \boldsymbol{\pi}_y^{(x)}). \tag{35}$$

Furthermore, setting $\boldsymbol{\pi}_H = \boldsymbol{\pi}_{\boldsymbol{y}}{}^{(x)}$ and $R = x_{y\cdot}$, it also follows from Lemma 2 that

$$P_X(\mathbf{X}_y = \mathbf{x}_y | x_{y\cdot}, \boldsymbol{\epsilon}_y) = \text{Multinom}(\mathbf{x}_y; x_{y\cdot}, \boldsymbol{\pi}_y^{(x)}). \tag{36}$$

This yields

$$P_X(\mathbf{X}_y = \mathbf{x}_y | x_{y\cdot}) = \int P_X(\mathbf{X}_y = \mathbf{x}_y | x_{y\cdot}, \boldsymbol{\epsilon}_y) g_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}_y; \boldsymbol{\sigma}_\epsilon^2) d\boldsymbol{\epsilon}_y. \tag{37}$$

Thus, by the conditional independence of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_Y$, we have

$$L(\boldsymbol{\theta}; \mathbf{X} | \mathbf{x}_\cdot) = \prod_{y=1}^{Y} P_X(\mathbf{X}_y = \mathbf{x}_y | x_{y\cdot}), \tag{38}$$

thereby completing the proof.

Acknowledgements This research was funded by the Nova Scotia Habitat Conservation Fund and Mitacs Accelerate awards. We would like to thank Wildlife Division, Nova Scotia Department of Natural Resources, for providing us the data and support. We would also like to thank Hugh Chipman and Holger Teismann for their support and useful discussions and Anja Haltner for reviewing this paper.

We are also indebted to the editors for their effort of organizing the proceedings of the TIES-GRASPA 2017 conference and anonymous reviewer for the constructive comments which improved the presentation of the paper.

References

1. Bernardelli H (1977) Population waves. In: Mathematical demography. Springer, Berlin, pp 215–219
2. Casella G, Berger RL (2002) Statistical inference, vol 2. Duxbury Pacific Grove, CA
3. Caswell H (2001) Matrix population models. Wiley Online Library
4. Chen E (2017) A Stochastic harvest Leslie matrix simulation model for evaluating wildlife population reconstruction methods using harvest data, Masters Thesis Acadia University
5. Clawson MV, Skalski JR, Millspaugh JJ (2013) The utility of auxiliary data in statistical population reconstruction. Wildl Biol 19(2):147–155
6. Downing RL (1980) Vital statistics of animal populations. Wildlife management techniques manual. Fourth edition. The Wildlife Society, Washington, DC, pp 247–267
7. Fryxell JM, Mercer WE, Gellately RB (1988) Population dynamics of newfoundland moose using cohort analysis. J Wildl Manag 52:14–21
8. Gast C, Skalski JR, Beyer DE (2013) Evaluation of fixed-and random-effects models and multistage estimation procedures in statistical population reconstruction. J Wildl Manag 77(6):1258–1270
9. Gove NE, Skalski JR, Zager P, Townsend RL (2002) Statistical models for population reconstruction using age-at-harvest data. J Wildl Manag 66:310–320
10. Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions, vol 165. Wiley, New York
11. Leslie PH (1945) On the use of matrices in certain population mathematics. Biometrika 33(3):183–212
12. Leslie PH (1948) Some further notes on the use of matrices in population mathematics. Biometrika 35(3/4):213–245
13. Lynch HJ, Fagan WF (2009) Survivorship curves and their impact on the estimation of maximum population growth rates. Ecology 90(4):1116–1124
14. Lynch HJ, Zeigler S, Wells L, Ballou JD, Fagan WF (2010) Survivorship patterns in captive mammalian populations: implications for estimating population growth rates. Ecol Appl 20(8):2334–2345
15. Millspaugh JJ, Skalski JR, Townsend RL, Diefenbach DR, Boyce MS, Hansen LP, Kammermeyer K (2009) An evaluation of sex-age-kill (SAK) model performance. J Wildl Manag 73(3):442–451
16. Quirynen M, Lamoral Y, Dekeyser C, Peene P, van Steenberghe D, Bonte J, Baert AL CT scan standard reconstruction technique for reliable jaw bone volume determination. Int J Oral Maxillofac Implants 5(4):384–389 (1989)
17. Roseberry JL, Woolf A (1991) A comparative evaluation of techniques for analyzing white-tailed deer harvest data. Wildl Monogr 117:3–59
18. Skalski JR, Townsend RL, Gilbert BA (2007) Calibrating statistical population reconstruction models using catch-effort and index data. J Wildl Manag 71(4):1309–1316
19. Skalski JR, Ryding KE, Millspaugh J (2010) Wildlife demography: analysis of sex, age, and count data. Academic, New York

Detection of Change Points in Spatiotemporal Data in the Presence of Outliers and Heavy-Tailed Observations

Bin Sun and Yuehua Wu

Keywords Change-point detection · EM-type algorithm · General spatiotemporal autoregressive model · M-estimation · Outlier

1 Introduction

Spatial-temporal data has been drawing a dramatically increasing attention due to their wide availabilities in many research fields including environmental study, climate change, and biology. They are usually spatially correlated and/or temporally correlated. In the literature, there are many approaches to model the spatial dependence structure as well as the temporal dependence structure in the spatiotemporal data. Research interest also arises on the topic to detect sudden changes occurring in spatiotemporal data over a long time period. These changes could be due to exposure changes, instrument/observer changes, the implementation of government regularities and policies [7], etc.

This paper proposes approaches for the analysis of multiple change- point models when dependency in the data is modeled through a hierarchical Gaussian Markov random field. Integrated nested Laplace approximations are used to approximate data quantities, and an approximate filtering recursions approach is proposed for savings in computational cost when detecting change-points. All of these methods are simulation free. Analysis of real data demonstrates the usefulness of the approach in general. The new models which allow for data dependence are compared with conventional models where data within segments is assumed independent.

Under the framework of Bayesian approaches, [9] presented methods for analyzing multiple change-point models when dependency in the data is modeled through a hierarchical Gaussian Markov random field, and Altieri et al. [1] proposed methods

B. Sun · Y. Wu (✉)
Department of Mathematics and Statistics, York University, Toronto, ON, Canada
e-mail: bsun@mathstat.yorku.ca; wuyh@mathstat.yorku.ca

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_3

for detecting multiple change-points over time in the inhomogeneous intensity of a spatiotemporal point process with spatial and temporal dependence within segments, among others.

On the other hand, under the framework of maximum likelihood methods, [4] and [5] introduced methods for modeling spatiotemporal or spatial data containing changes over time or space. Nappi-Choulet and Maury [4] proposed a hybrid method for incorporating a temporal regime switch into the spatiotemporal autoregressive model to deal with exogenous macroeconomic factors. For spatial data, [5] proposed a test procedure to detect change-points of multidimensional autoregressive processes. Their method works well to find possible structural breaks in the process that can occur at a certain distance from the predefined center. Most recently, [8] proposed a general spatiotemporal autoregressive (GSTAR) model which takes into account the effect of station surroundings, seasonality, temporal correlation among observations at the same spatial location, and spatial correlation among observations from different spatial locations. The model is so multifunctional that it can also be used to detect new influences that largely affected the measurements in the treatment area compared to the control area. However, their method is dependent on the normality assumption.

As the spatial-temporal data is usually observed over a large area and in many years, undetectable outliers can easily occur unexpectedly in any days for any small area because of measurement error or other reasons. The parameter estimation method given in [8] may not be stable or robust. There is a great need to develop a parameter estimation method for the GSTAR model that is resistant to outliers and stable in respect to heavy-tail distributed errors. In the development of such robust methods, M-estimation can play important and complementary roles. Thus we modify the EM-type algorithm given in [8] by replacing the least squares (LS) estimation by M-estimation, which is more stable in estimating parameters in the presence of outliers and/or heavy-tailed observations [2]. We name the modified EM-type algorithm as the $_M$EM-type algorithm. We also modify their change-point detection procedure accordingly, which is more accurate in detecting change-points in the presence of outliers and/or heavy-tailed observations.

The outline of this article is the following. In Sect. 2, a general spatiotemporal autoregressive model is reviewed, and the $_M$EM-type algorithm is presented. Then we describe the procedure for detecting change-points in the treatment area via the GSTAR models. In Sect. 3, a real data application and simulations are given to compare the $_M$EM-type algorithm with the original one and to compare both change-point detection procedures. Section 4 summarizes the results.

2 The GSTAR Model-Based Procedure of Change-Point Detection in the Daily Spatiotemporal Data

In this section, we first introduce the GSTAR model and present a specially designed EM-type algorithm to estimate the model parameters. We then give a change-point detection procedure based on the GSTAR model.

2.1 The GSTAR Model

In the following, we present the GSTAR model given in [8] that takes into account the effect of station surroundings, seasonality, temporal correlation among observations at the same spatial location, and spatial correlation among observations from different spatial locations while allowing the coefficients to vary over time. The GSTAR is defined as the following:

$$y_{i,T(k-1)+t} = \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma} + c_i + \rho \sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} \quad (1)$$
$$-\boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma} - c_l) + \varepsilon_{i,T(k-1)+t},$$

where $\varepsilon_{i,T(k-1)+t}$ are assumed to be independently and identically (iid) normal distributed with mean 0 and variance σ^2; $y_{i,T(k-1)+t}$ is the spatiotemporal variable of interest observed at spatial location i on tth day in the kth year; $t \in \mathcal{S}$ with $\mathcal{S}$ being a set of consecutive days in a year with size T; $W = (w_{il})_{L\times L}$ is a neighborhood matrix to describe the spatial correlation among observations collected from different spatial locations, which satisfies the conditions that $w_{il} \geq 0, w_{ii} = 0$ and $\sum_{l=1}^{L} w_{il} = 1$; $\boldsymbol{x}_{T(k-1)+t} = (x_{T(k-1)+t,1}, x_{T(k-1)+t,2}, x_{T(k-1)+t,3})'$ are explanatory variables, where $x_{T(k-1)+t,1} = 1$ for all $t \in \mathcal{S}$ and $(x_{T(k-1)+t,2}, x_{T(k-1)+t,3})' = (\sin(t_j\pi/s_j), \cos(t_j\pi/s_j))'$ for $t \in \mathcal{S}_j$ to model the seasonal cyclicities and $\mathcal{S}_j$, $j = 1, \ldots, J$, are J seasons in $\mathcal{S}$ with $\mathcal{S} = \cup\mathcal{S}_j$, and s_j is the number of days in the jth season for $j = 1, \ldots, J$, and t_j is the number of days of t in $\mathcal{S}_j$ if t falls into the jth season; $\boldsymbol{\beta}_{T(k-1)+t} = (\beta_{0,k,j}, \beta_{1,k,j}, \beta_2)'$ are regression coefficients when t falls into the jth season; $\tilde{\boldsymbol{y}}_{i,T(k-1)+t} = (y_{i,T(k-1)+t-1}, y_{i,T(k-1)+t-2}, \ldots, y_{i,T(k-1)+t-\iota})'$; and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_\iota)'$. An autoregression term is included in the model to take into account the possible autocorrelation among observations at each location. Here ι denotes the number of autoregression terms in the model which is predetermined but may be chosen by an order selection. The parameter set to be estimated in model (1) is $\mathcal{H} = \{\beta_{0,k,j}, \beta_{1,k,j}, j = 1, \ldots, J, k = 1, 2, \ldots, K; \beta_2, \boldsymbol{\gamma}, \tau_1, \ldots, \tau_\kappa, \rho, \sigma^2\}$.

2.2 The Estimation

M-estimation is a maximum likelihood-type estimation (see [2]). In the development of robust methods, M-estimation can play an important and complementary role. The well-known dispersion function for the M-estimation is Huber's function defined as the following:

$$H(x) = \begin{cases} x^2, & if \ |x| \leq k, \\ 2k|x| - k^2, & if \ |x| > k, \end{cases}$$

where k is a tuning constant and usually chosen as 1.345. The EM-type algorithm given in [8] used the least squares technique. The performance of the LS estimation relies heavily on the normality assumption on the errors. Because of the complexity of spatial-temporal data, the normality assumption is easily violated in the presence of undetectable outliers and/or heavy-tailed observations. We propose to modify it by replacing the LS technique used in the algorithm by M-estimation for estimating the GSTAR model parameters, which is more stable regardless if there are outliers and/or heavy-tailed observations in the dataset.

2.2.1 Initial Values

First, we give the initial values to $\{\beta_{0,k,j},\ \beta_{1,k,j},\ j = 1,\ldots,J,\ k = 1,\ 2,\ \ldots,\ K;\ \beta_2,\ \boldsymbol{\gamma},\ \tau_1,\ \ldots,\ \tau_\kappa,\ \rho\}$. We then carry out the following:

1. We calculate the mean of the available observations for each type of stations and denote them by $a_1, a_2, \ldots, a_\kappa$. We then calculate the overall mean of the available observations and denote it by a. The initial estimates of τ_q's are thus put as $\tau_q^{(0)} = a_q - a, q = 1, \ldots, \kappa$. Let $\bar{c} = \sum_{i=1}^{L} c_i/L$. The initial estimate of $\bar{c}$ can be obtained by $\bar{c}^{(0)} = \sum_{i=1}^{L} c_i^{(0)}/L$, where $c_i^{(0)}$ takes values in $\{\tau_1^{(0)}, \ldots, \tau_\kappa^{(0)}\}$ according to different kinds of surrounding areas around the location.
2. By averaging all equations in (1), we obtain that

$$\begin{aligned}\bar{y}_{T(k-1)+t} &= \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} + \bar{\tilde{\boldsymbol{y}}}'_{T(k-1)+t}\boldsymbol{\gamma} + \bar{c} + \epsilon_{T(k-1)+t},\\ &= \beta_{0,k,j} + \beta_{1,k,j}x_{T(k-1)+t,2} + \beta_2 x_{T(k-1)+t,3} + \gamma_1\bar{y}_{T(k-1)+t-1}\\ &\quad + \ldots + \gamma_o\bar{y}_{T(k-1)+t-o} + \bar{c} + \epsilon_{T(k-1)+t},\end{aligned} \tag{2}$$

   where $\bar{y}_{T(k-1)+t}$ is the average of the observations on the $(T(k-1)+t)$th day of all spatial locations after removing all missing observations, $\bar{\tilde{\boldsymbol{y}}}_{T(k-1)+t} = (\bar{y}_{T(k-1)+t-1}, \ldots,\ \bar{y}_{T(k-1)+t-o})'$ and $\epsilon_{T(k-1)+t} = \frac{1}{L}\boldsymbol{\ell}'_L(\mathcal{I}_L - \rho W)^{-1}\varepsilon_{T(k-1)+t}$, in which $\boldsymbol{\ell}_L = (1, 1, \ldots, 1)'_{L\times 1}$.
3. Since $\sin(\pi-\theta) = \sin(\theta)$, $\sin(\pi+\theta) = \sin(2\pi-\theta)$, $\cos(\pi-\theta) = -\cos(\theta)$, and $\cos(\pi+\theta) = -\cos(\theta)$, we can remove both the constant term and the term related to $\beta_{1,k,j}$ by the difference between two properly chosen pair of the equations given in (2). By doing so, we obtain

$$y^{(1)}_{T_1(k-1)+t} = \beta_2 y^{(2)}_{T_1(k-1)+t} + \boldsymbol{\gamma}\,\tilde{\boldsymbol{y}}^{(3)}_{T_1(k-1)+t} + \tilde{\epsilon}_{T_1(k-1)+t}, \quad t \in \mathcal{S}^{(1)}. \tag{3}$$

   (A specific example of how to calculate $y^{(1)}_{T_1(k-1)+t}$, $y^{(2)}_{T_1(k-1)+t}$, $\tilde{\boldsymbol{y}}^{(3)}_{T_1(k-1)+t}$, $\tilde{\epsilon}_{T_1(k-1)+t}$, and $\mathcal{S}^{(1)}$ are given in Appendix.)

Denote $\boldsymbol{y}^{(1)} = (y_1^{(1)}, y_2^{(1)}, \ldots, y_{T_1K}^{(1)})'$, $\boldsymbol{y}^{(2)} = (y_1^{(2)}, y_2^{(2)}, \ldots, y_{T_1K}^{(2)})'$, and $\tilde{\boldsymbol{y}}^{(3)} = (\tilde{\boldsymbol{y}}_1^{(3)}, \tilde{\boldsymbol{y}}_2^{(3)}, \ldots, \tilde{\boldsymbol{y}}_{T_1K}^{(3)})'$. The M-estimates of β_2 and $\boldsymbol{\gamma}$ are given by

$$\arg\min_{\beta_2,\boldsymbol{\gamma}} H(\boldsymbol{y}^{(1)} - \beta_2 \boldsymbol{y}^{(2)} - \boldsymbol{\gamma}\tilde{\boldsymbol{y}}^{(3)}),$$

which are used as the initial estimate $\beta_2^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ of β_2 and $\boldsymbol{\gamma}$, respectively.

4. We substitute β_2 and $\boldsymbol{\gamma}$ by $\beta_2^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ in model (2). For each year k and season j, we denote $\boldsymbol{y}_j^{(1)} = (\bar{y}_{T(k-1)+t} - \beta_2^{(0)} x_{T(k-1)+t,3} - \boldsymbol{\gamma}^{(0)}\bar{\tilde{\boldsymbol{y}}}_{T(k-1)+t} - \bar{c}^{(0)},\ t \in \mathcal{S}_j)'$, and $\boldsymbol{y}_j^{(2)} = (x_{T(k-1)+t,2},\ t \in \mathcal{S}_j)'$. We derive the M-estimates of $\beta_{0,k,j}, \beta_{1,k,j}$ for season j of the kth year by

$$\arg\min_{\beta_{0,k,j},\beta_{1,k,j}} H(\boldsymbol{y}_j^{(1)} - \beta_{0,k}^j \boldsymbol{\ell}_{s_j} - \beta_{1,k}^j \boldsymbol{y}_j^{(2)})$$

for $j = 1, \ldots, J$, respectively, where $\boldsymbol{\ell}_{s_j} = (1, 1, \ldots, 1)'_{s_j \times 1}$. Therefore, we use these least square estimates of $\beta_{0,k,j}$ and $\beta_{1,k,j}$ as the initial estimates $\beta_{0,k,j}^{(0)}$ and $\beta_{1,k,j}^{(0)}$.

5. Set the initial value of $\rho^{(0)}$ as 0.5.

2.2.2 The $_M$EM-Type Algorithm

Let $\mathcal{H}^{(m-1)} = \{\beta_{0,k,j}^{(m-1)}, \beta_{1,k,j}^{(m-1)}, j = 1, \ldots, J, k = 1, 2, \ldots, K, \beta_2^{(m-1)}, \boldsymbol{\gamma}^{(m-1)}, \tau_1^{(m-1)}, \ldots, \tau_\kappa^{(m-1)}, \rho^{(m-1)}, \sigma^{2(m-1)}\}$ be the set of estimates we obtained after the $(m-1)$th iteration. The $_M$EM-type algorithm has the following three steps:

1. E-step: Estimate the observation $y_{i,T(k-1)+t}$ at the mth iteration by the following conditional expectation:

$$\begin{aligned} & y_{i,T(k-1)+t}^{(m)} \\ &= E\left(y_{i,T(k-1)+t} | y_{l,T(k-1)+t}^{(m-1)}, l = 1, 2, \ldots, L, \mathcal{H}^{(m-1)}\right) \\ &= \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t}^{(m-1)} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{(m-1)} + c_i^{(m-1)} + \rho^{(m-1)} \times \\ & \sum_{l:w_{il}\neq 0} \left(y_{l,T(k-1)+t}^{(m-1)} - \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t}^{(m-1)} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{(m-1)} - c_l^{(m-1)}\right), \end{aligned}$$

if it is missing.

2. M-step: Obtain the estimates $\boldsymbol{c}^{(m)}$, $\sigma^{2(m)}, \rho^{(m)}$, $\beta_2^{(m)}$, $\boldsymbol{\gamma}^{(m)}$, $\beta_{0,k,j}^{(m)}$, $\beta_{1,k,j}^{(m)}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ at the mth iteration sequentially as follows:
   (a) First derive the estimates $\left\{\tau_1^{(m)}, \ldots, \tau_\kappa^{(m)}\right\}$ in the same way as we obtained the estimates $\left\{\tau_1^{(0)}, \ldots, \tau_\kappa^{(0)}\right\}$. Then $\boldsymbol{c}^{(m)} = (c_1^{(m)}, c_2^{(m)}, \ldots, c_L^{(m)})$, where $c_i^{(m)}$'s take values from $\{\tau_1^{(m)}, \ldots, \tau_\kappa^{(m)}\}$ based on the types of the stations.
   (b) Similarly, we can remove both the constant term and the term related to $\beta_{1,k,j}$ by the difference between one properly chosen pair of the equations given in (1). Then we estimate σ^2 as $\sigma^{2(m)}$ by sample variances.
   (c) Find the M-estimates of ρ, β_2, and $\boldsymbol{\gamma}$ after substituting σ^2 by $\sigma^{2(m)}$ to get $\rho^{(m)}$, $\beta_2^{(m)}$, and $\boldsymbol{\gamma}^{(m)}$, respectively.
   (d) Substitute the estimates $\left\{\boldsymbol{c}^{(m)}, \rho^{(m)}, \beta_2^{(m)}, \boldsymbol{\gamma}^{(m)}\right\}$ into model (1) to obtain the M-estimates of $\beta_{0,k,j}$, $\beta_{1,k,j}$ as $\beta_{0,k,j}^{(m)}$, $\beta_{1,k,j}^{(m)}$.

3. Keep repeating the steps 1–2 until $|\boldsymbol{\gamma}^{(m)} - \boldsymbol{\gamma}^{(m-1)}| < v$, $|\beta_2^{(m)} - \beta_2^{(m-1)}| < v$, $|\beta_{0,k,j}^{(m)} - \beta_{0,k,j}^{(m-1)}| < v$ and $|\beta_{1,k,j}^{(m)} - \beta_{1,k,j}^{(m-1)}| < v$ for all k and j, where v is a predetermined small value. Then we denote $\hat{\beta}_{0,k,j} = \beta_{0,k,j}^{(m)}$, $\hat{\beta}_{1,k,j} = \beta_{1,k,j}^{(m)}$, for $j = 1, \ldots, J$, $k = 1, 2, \ldots, K$; $\hat{\beta}_2 = \beta_2^{(m)}$, $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{(m)}$; $\hat{\tau}_i = \tau_i^{(m)}$, for $i = 1, \ldots, \kappa$; $\hat{\rho} = \hat{\rho}^{(m)}$, and $\hat{\sigma}^2 = \hat{\sigma}^{2(m)}$.

The set of estimates we obtained is $\hat{\mathcal{H}} = \{\hat{\beta}_{0,k,j}, \hat{\beta}_{1,k,j}, j = 1, \ldots, J, k = 1, 2, \ldots, K, \hat{\beta}_2, \hat{\boldsymbol{\gamma}}, \hat{\tau}_1, \ldots, \hat{\tau}_\kappa, \hat{\rho}, \hat{\sigma}^2\}$.

2.3 The Change-Point Detection Procedure

We now introduce the procedure for detecting new influences that affected the measurements in the treatment area substantially by comparing with that in the control area, which is similar to the one given in [8]. We model the data collected, respectively, from the treatment and control areas of the region by two different GSTAR models using the algorithm proposed in the previous section. The main idea is that if new influences in the treatment area are not negligible, there should be detectable changes in the time-dependent regression coefficients in the GSTAR model for that area compared to those in the GSTAR model for the control area. A change-point detection method can be applied to the differences of regression

coefficient estimates from these two areas. The M-estimation-based change-point detection procedure is described below:

1. We group the stations in the treatment area of the region into group 1 and model the spatiotemporal data collected at these stations by

$$\begin{aligned} y_{i,T(k-1)+t} &= \beta^{\mathrm{I}}_{0,k,j} + \beta^{\mathrm{I}}_{1,k,j} x_{T(k-1)+t,2} + \beta^{\mathrm{I}}_{2} x_{T(k-1)+t,3} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{I}} \\ &\quad + c_i + \rho^{\mathrm{I}} \sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} - \beta^{\mathrm{I}}_{0,k,j} - \beta^{\mathrm{I}}_{1,k,j} x_{T(k-1)+t,2} \\ &\quad - \beta^{\mathrm{I}}_{2} x_{T(k-1)+t,3} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{I}} - c_l) + \varepsilon_{i,T(k-1)+t}. \end{aligned} \tag{4}$$

   Then we group the stations in the control area into group 2 and model the data from these stations by

$$\begin{aligned} y_{i,T(k-1)+t} &= \beta^{\mathrm{II}}_{0,k,j} + \beta^{\mathrm{II}}_{1,k,j} x_{T(k-1)+t,2} + \beta^{\mathrm{II}}_{2} x_{T(k-1)+t,3} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{II}} \\ &\quad + c_i + \rho^{\mathrm{II}} \sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} - \beta^{\mathrm{II}}_{0,k,j} - \beta^{\mathrm{II}}_{1,k,j} x_{T(k-1)+t,2} \\ &\quad - \beta^{\mathrm{II}}_{2} x_{T(k-1)+t,3} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{II}} - c_l) + \varepsilon_{i,T(k-1)+t}. \end{aligned} \tag{5}$$

   Note that these two models have different parameters except the effect of the station locations, c_i's.
2. First, we estimate the parameters as their initial values. Following the steps presented in Sect. 2.2.1, we derive the station type effect $\left\{\tau_1^{(0)}, \ldots, \tau_\kappa^{(0)}\right\}$ using observations collected on stations from both groups so that the same type of stations in different groups has the same station type effect. Then, we obtain $\{\beta^{\mathrm{I}(0)}_{0,k,j}, \beta^{\mathrm{I}(0)}_{1,k,j}, j = 1, 2, 3, 4, k = 1, 2, \ldots, K, \beta^{\mathrm{I}(0)}_{2}, \boldsymbol{\gamma}^{\mathrm{I}(0)}\}$ and $\{\beta^{\mathrm{II}(0)}_{0,k,j}, \beta^{\mathrm{II}(0)}_{1,k,j}, j = 1, 2, 3, 4, k = 1, 2, \ldots, K, \beta^{\mathrm{II}(0)}_{2}, \boldsymbol{\gamma}^{\mathrm{II}(0)}\}$ for two groups of stations separately. We also set the initial values of ρ^{I} and ρ^{II}as $\rho^{\mathrm{I}(0)} = \rho^{\mathrm{II}(0)} = 0.5$.
3. We apply the $_M$EM-type algorithm proposed in Sect. 2.1. In the E-step, the missing observations are filled up. In the M-step, we estimate the station type effects using data from all the stations and then estimate the other parameters sequentially for two groups of stations separately. These two steps are repeated until convergence. We obtain the estimates $\hat{\beta}^{\mathrm{I}}_{0,k,j}, \hat{\beta}^{\mathrm{I}}_{1,k,j}$ for model (4) and $\hat{\beta}^{\mathrm{II}}_{0,k,j}, \hat{\beta}^{\mathrm{II}}_{1,k,j}$ for model (5).
4. We take the difference between these two sets of parameter estimates to obtain two sets of estimates $\{d_{0,k,j} = \hat{\beta}^{\mathrm{I}}_{0,k,j} - \hat{\beta}^{\mathrm{II}}_{0,k,j}, j = 1, 2, 3, 4, k = 1, 2, \ldots, K\}$ as the difference in the intercepts of two models and $\{d_{1,k,j} = \hat{\beta}^{\mathrm{I}}_{1,k,j} - \hat{\beta}^{\mathrm{II}}_{1,k,j}, j = 1, 2, 3, 4, k = 1, 2, \ldots, K\}$ as the difference in the

slopes of two models. Then we apply the R package *change point* [3] to detect the possible mean shifts in $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$.

For convenience, we name the change-point detection procedure given in [8] as the LS-based change-point detection procedure.

It is worth mentioning that in the above procedure $d_{0,k,j}$ and $d_{1,k,j}$ describe the effect after eliminating the effects of station types, the temporal correlation, the spatial correlation, and the randomness. Therefore, after applying the proposed procedure, the estimates $\{\hat{\beta}^{\mathrm{I}}_{0,k,j},\ \hat{\beta}^{\mathrm{I}}_{1,k,j}\}$ and $\{\hat{\beta}^{\mathrm{II}}_{0,k,j},\ \hat{\beta}^{\mathrm{II}}_{1,k,j}\}$ derived, respectively, from two groups of data should behave similarly if there are no new influences in the treatment area. Then there are no changes in the means of both $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$.

3 Application

In this section, we, respectively, compare the ${}_M$EM-type algorithm with the EM-type algorithm in [8] and the M-estimation-based change-point detection procedure with the LS-based change-point detection procedure through a real data application and simulations.

3.1 A Real Data Example

The data of [8] includes measurements of the ground-level ozone concentration readings measured in parts per billion (ppb) from 36 monitoring stations in a region with longitude from $-80°$ to $-78.5°$ and latitude from $43°$ to $45°$ in Southern Ontario over the period from 1988 to 2010. Locations of the stations are shown in Fig. 1. Following [6], the data used in the examples is the log of the daily maximum 8-h moving averages of ozone concentration. There are 36 stations. Among these 36 stations, we choose 27 stations which have been monitored for more than 5 years. On average, each station has 39.4% data missing. We let $\iota = 1$ by the pre-analysis of the data. The total number of the parameters is 194. First, we obtain the estimates of the parameters in the GSTAR model using the EM-type algorithm in [8]. We name these estimates $\hat{\mathcal{H}}_{LS}$. Then the proposed ${}_M$EM-type algorithm is used to obtain the parameters in GSTAR model on the same dataset. We name these estimates $\hat{\mathcal{H}}_M$. We use the Euclidean distance to measure the differences as the following:

$$\left\|\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M\right\| = \sqrt{(\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M)'(\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M)}.$$

The distance is 0.1015, which is small enough to show that these two methods produce almost the same parameter estimates on the same dataset.

Fig. 1 The locations of 27 stations which have data for more than 5 years are shown in circle. Data source: Regional Aquatics Monitoring Program http://www.ramp-alberta.org

3.2 A Simulated Example

3.2.1 Data with Outliers

We now show that the $_M$EM-type algorithm works well in the presence of outliers. To make the outliers reasonable, we first choose an area whose latitude is less than 43.55°N. There are eight stations within this area. Then we randomly pick up a day, and for a period of 9 days after this day, we expanded the log-transformed ozone concentrations by 1.6 times. In real life, this could happen for the reasons including the machine broken, unexpected activities in this area, etc. The experiment is repeated for 500 times, we recorded the Euclidean distance for both algorithms, and in Table 1, the mean and the standard deviation (sd) of the Euclidean distance are reported.

Table 1 Mean and standard deviation of the Euclidean distances

$_M$EM-type algorithm		EM-type algorithm	
Mean	sd	Mean	sd
0.2704	0.1325	0.4392	0.0541

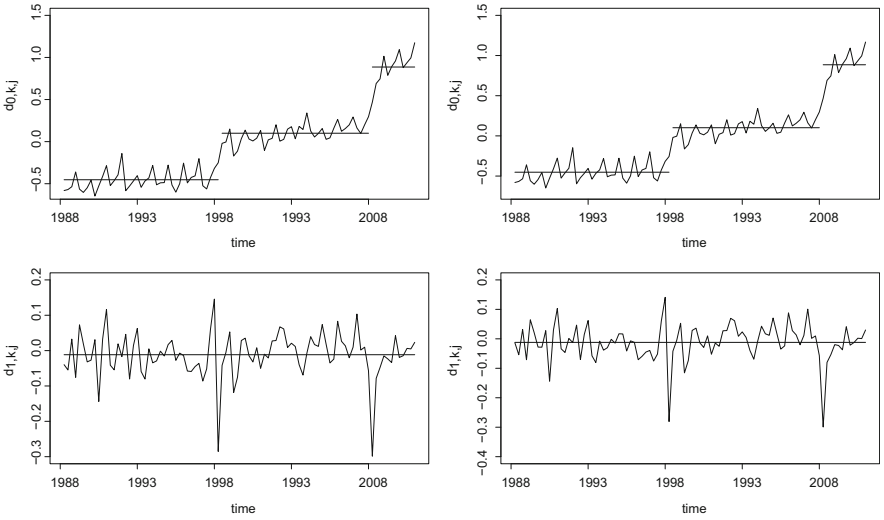

Fig. 2 Change-points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *change-point*. The left and right panels, respectively, display the results by using both LS-based and M-estimation-based change-point detection procedures

The simulation shows that when there are outliers, they have less impact on the performance of parameter estimation if the proposed $_M$EM-type algorithm is used.

3.2.2 The Change-Point Detection

In [8], they simulated the change-points under their scenario 1 in the following reasonable way. First they separated the stations into two groups by the latitude 43.65°. Then, for each station in group 1, they added a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to each observation collected from 1998 to 2010 to create the first change-point at 1998. They also added a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to the previously modified observations from 2008 to 2010 to create the second change-point at 2008. The results of detecting the change-points by using the LS-based change-point detection procedure are shown in Fig. 2. The right panel displays the results by using the M-estimation-based change-point detection procedure. Two sets of estimates, $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$, are obtained. The plot displays the change-points in $\{d_{0,k,j}\}$ (upper panel) and $\{d_{1,k,j}\}$ (lower

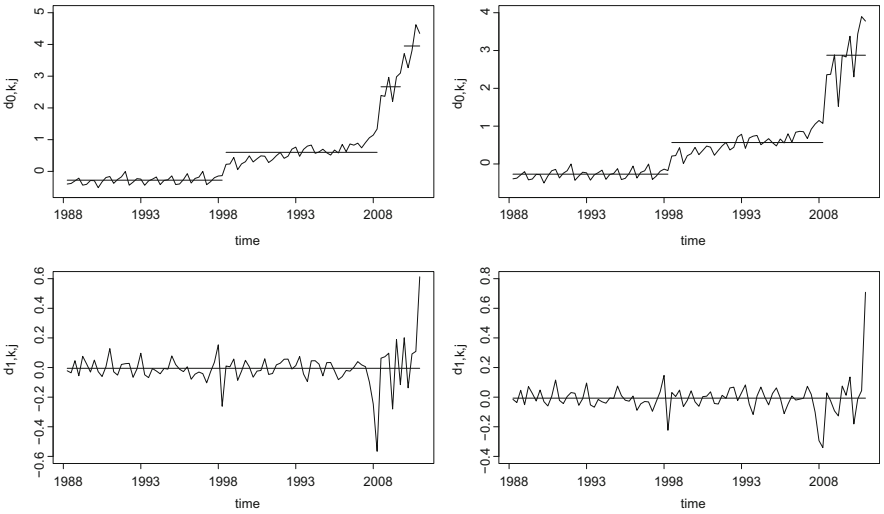

Fig. 3 Change points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *change-point* for heavy-tailed observations. The left and right panels, respectively, display the results by using both LS-based and M-estimation-based change-point detection procedures

panel) using both procedures. Figure 2 shows that both procedures capture the change-points equally well.

We now modify the random number generation by changing the variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to $\sigma^2 = 1.6\tilde{\sigma}$. This modification produces large variation in the observations after the change-points. This is a reasonable scenario because if there are some activities happening in a region, the observations would be more fluctuated than other times due to these activities. The M-estimation-based change-point detection procedure detects the change-points at 1998 and 2008 successfully using the R package *change point*; however, the LS-based method produces false change-points. The results are shown in Fig. 3, which demonstrate that the M-estimation-based change-point detection procedure is more stable than the LS-based change-point detection procedure in change-point detection in the presence of outliers and/or heavy-tailed observations.

4 Conclusions

In this paper, we improve the EM-type algorithm for the parameter estimation of the GSTAR model by replacing the least squares technique in the algorithm by M-estimation such that the modified algorithm is more stable in estimating parameters and more accurate in detecting change-points when the dataset contains outliers and/or has heavy-tailed observations. In the real data example, it is shown that

$_M$EM-type algorithm produces almost the same parameter estimates for the GSTAR model as the EM-type algorithm. In the simulation, we test the robustness of our methods in two ways. In the first case, we add some random outliers to the real data, and the parameter estimates from our method are more stable than the LS method. In the second case, we test the accuracy of detecting the change-points of our method when the observations under the same scenario as in [8]. Both methods detect the change-points equally well. Then we test the performance of our method in the case when the observations are heavy-tail distributed. We increase the variance of the observations after the change-points by 1.6 times, the result shows that the method in [8] produces false change-points, but our method still successfully detects the change-points with no false ones. Thus we conclude that our method is more robust and stable for modeling spatiotemporal data where the inconsistently behaved observations are more likely to appear as the size of data is growing rapidly.

Appendix

A specific example is given to show how to calculate $y^{(1)}_{T_1(k-1)+t}$, $y^{(2)}_{T_1(k-1)+t}$, $\tilde{y}^{(3)}_{T_1(k-1)+t}$ and $\tilde{\epsilon}_{T_1(k-1)+t}$ in Sect. 2.1.

Consider $t_1 = 1$, $\mathcal{S}_1 = \{1, 2, \ldots, 90, 307, 308, \ldots, 366\}$ for Winter, $\mathcal{S}_2 = \{91, 92, \ldots, 152\}$ for Spring, $\mathcal{S}_3 = \{153, \ldots, 244\}$ for Summer and $\mathcal{S}_4 = \{245, \ldots, 306\}$ for Fall. In this case, $T = 366$ and $T_1 = 181$. For winter,

$$
\begin{aligned}
y^{(1)}_{181(k-1)+t-1} &\equiv \bar{y}_{366(k-1)+t} - \bar{y}_{366(k-1)+75-t} \\
y^{(2)}_{181(k-1)+t-1} &\equiv 2\cos(2t\pi/s_1), \\
y^{(3)}_{181(k-1)+t-1} &\equiv \bar{y}_{366(k-1)+t-1} - \bar{y}_{366(k-1)+75-t-1} \\
\tilde{\epsilon}_{181(k-1)+t-1} &\equiv \epsilon_{366(k-1)+t} - \epsilon_{366(k-1)+75-t}, \quad t = 2, \ldots, 37. \\
y^{(1)}_{181(k-1)+37+t} &\equiv \bar{y}_{366(k-1)+t+75} - \bar{y}_{366(k-1)+366-t} \\
y^{(2)}_{181(k-1)+37+t} &\equiv -2\cos(2t\pi/s_1), \\
y^{(3)}_{181(k-1)+37+t} &\equiv \bar{y}_{366(k-1)+t+75-1} - \bar{y}_{366(k-1)+366-t-1} \\
\tilde{\epsilon}_{181(k-1)+37+t} &\equiv \epsilon_{366(k-1)+t+75} - \epsilon_{366(k-1)+366-t}, \quad t = 0, 1, 2, \ldots, 15. \\
y^{(1)}_{181(k-1)+37+t} &\equiv \bar{y}_{366(k-1)+t+291} - \bar{y}_{366(k-1)+366-t} \\
y^{(2)}_{181(k-1)+37+t} &\equiv -2\cos(2t\pi/s_1), \\
y^{(3)}_{181(k-1)+37+t} &\equiv \bar{y}_{366(k-1)+t+291-1} - \bar{y}_{366(k-1)+366-t-1} \\
\tilde{\epsilon}_{181(k-1)+37+t} &\equiv \epsilon_{366(k-1)+t+291} - \epsilon_{366(k-1)+366-t}, \quad t = 16, 17, \ldots, 37.
\end{aligned}
$$

For Spring,

$$y^{(1)}_{181(k-1)+74+t} \equiv \bar{y}_{366(k-1)+90+t} - \bar{y}_{366(k-1)+121-t}$$
$$y^{(2)}_{181(k-1)+74+t} \equiv 2\cos(2t\pi/s_2),$$
$$y^{(3)}_{181(k-1)+74+t} \equiv \bar{y}_{366(k-1)+90+t-1} - \bar{y}_{366(k-1)+121-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+74+t} \equiv \epsilon_{366(k-1)+90+t} - \epsilon_{366(k-1)+121-t}, \quad t = 1, 2, \ldots, 15.$$
$$y^{(1)}_{181(k-1)+90+t} \equiv \bar{y}_{366(k-1)+121+t} - \bar{y}_{366(k-1)+152-t}$$
$$y^{(2)}_{181(k-1)+90+t} \equiv -2\cos(2t\pi/s_2),$$
$$y^{(3)}_{181(k-1)+90+t} \equiv \bar{y}_{366(k-1)+121+t-1} - \bar{y}_{366(k-1)+152-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+90+t} \equiv \epsilon_{366(k-1)+121+t} - \epsilon_{366(k-1)+152-t}, \quad t = 0, 1, 2, \ldots, 15.$$

For Summer,

$$y^{(1)}_{181(k-1)+105+t} \equiv \bar{y}_{366(k-1)+152+t} - \bar{y}_{366(k-1)+198-t}$$
$$y^{(2)}_{181(k-1)+105+t} \equiv 2\cos(2t\pi/s_3),$$
$$y^{(3)}_{181(k-1)+105+t} \equiv \bar{y}_{366(k-1)+152+t-1} - \bar{y}_{366(k-1)+198-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+105+t} \equiv \epsilon_{366(k-1)+152+t} - \epsilon_{366(k-1)+198-t}, \quad t = 1, 2, \ldots, 22.$$
$$y^{(1)}_{181(k-1)+128+t} \equiv \bar{y}_{366(k-1)+198+t} - \bar{y}_{366(k-1)+244-t}$$
$$y^{(2)}_{181(k-1)+128+t} \equiv -2\cos(2t\pi/s_3),$$
$$y^{(3)}_{181(k-1)+128+t} \equiv \bar{y}_{366(k-1)+198+t-1} - \bar{y}_{366(k-1)+244-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+128+t} \equiv \epsilon_{366(k-1)+198+t} - \epsilon_{366(k-1)+244-t}, \quad t = 0, 1, 2, \ldots, 22.$$

For Fall,

$$y^{(1)}_{181(k-1)+150+t} \equiv \bar{y}_{366(k-1)+244+t} - \bar{y}_{366(k-1)+275-t}$$
$$y^{(2)}_{181(k-1)+150+t} \equiv 2\cos(2t\pi/s_4),$$
$$y^{(3)}_{181(k-1)+150+t} \equiv \bar{y}_{366(k-1)+244+t-1} - \bar{y}_{366(k-1)+275-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+150+t} \equiv \epsilon_{366(k-1)+244+t} - \epsilon_{366(k-1)+275-t}, \quad t = 1, 2, \ldots, 15.$$
$$y^{(1)}_{181(k-1)+166+t} \equiv \bar{y}_{366(k-1)+275+t} - \bar{y}_{366(k-1)+306-t}$$
$$y^{(2)}_{181(k-1)+166+t} \equiv -2\cos(2t\pi/s_4),$$

$$y^{(3)}_{181(k-1)+166+t} \equiv \bar{y}_{366(k-1)+275+t-1} - \bar{y}_{366(k-1)+306-t-1}$$
$$\tilde{\epsilon}_{181(k-1)+166+t} \equiv \epsilon_{366(k-1)+275+t} - \epsilon_{366(k-1)+306-t}, \quad t = 0, 1, 2, \ldots, 15.$$

References

1. Altieri L, Cocchi D, Greco F, Ellian JB, Scott EM (2016) Bayesian P-splines and advanced computing in R for a changepoint analysis on spatio-temporal point processes. J Stat Comput Simul 86:2531–2545
2. Huber PJ (1973) Robust regression. Ann Stat 1:799–821
3. Killick R, Eckley I (2014) Changepoint: an R package for changepoint analysis. J Stat Softw 58:1–13
4. Nappi-Choulet I, Maury T-P (2009) A spatiotemporal autoregressive price index for the Paris office property market. Real Estate Econ V37:305–340
5. Otto P, Schmid W (2016) Detection of spatial change points in the mean and covariances of multivariate simultaneous autoregressive models. Biometrical J 58:1113–1137
6. Porter PS, Rao ST, Zurbenko IG, Dunker AM, Wolff GT (2001) Ozone air quality over North America: part II-an analysis of trend detection and attribution techniques. J Air Waste Manag Assoc 51:283–306
7. Wu Y, Jin B, Chan E (2015) Detection of Changes in Ground-level ozone concentrations via entropy. Entropy 17:2749–2763
8. Wu Y, Sun X, Chan E, Qin S (2017) Detecting non-negligible new influences in environmental data via a general spatio-temporal autoregressive Model. Br J Environ Clim Chang 7(4):223–235
9. Wyse J, Friel N, Rue H (2011) Approximate simulation-free Bayesian inference for multiple changepoint models with dependence within segments. Bayesian Anal 6:501–528

Modeling Spatiotemporal Mismatch for Aerosol Profiles

Ilia Negri, Alessandro Fassò, Lucia Mona, Nikolaos Papagiannopoulos, and Fabio Madonna

Keywords Data comparison · Uncertainty · CALIOP measurements · EARLINET measurements

1 Introduction

The high variability both in space and time of tropospheric aerosols is one of the main causes of the high uncertainty related to tropospheric aerosols and their interactions with clouds. Since 2006, CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization), the LIDAR onboard CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations) specifically designed for aerosol and clouds study, is providing high-resolution vertical profiles of aerosols and clouds on a global scale. How well these CALIOP measurements represent the atmospheric conditions of a surrounding area over a longer time is an important issue to be investigated; see [19] for an overview of the CALIPSO mission and CALIOP data processing algorithms. EARLINET (European Aerosol Research Lidar NETwork) is the first LIDAR network for aerosol studies on a continental scale. EARLINET is a network of different instruments and instrumental setups with a wide variety of instrumental specifics and team expertises, so that there is not a common vertical and temporal sampling overall within the network. For a complete description of EARLINET, see [16] and references therein.

I. Negri (✉) · A. Fassò
Department of Management Information and Production Engineering, University of Bergamo, Bergamo, Italy
e-mail: ilia.negri@unibg.it; alessandro.fasso@unibg.it

L. Mona · N. Papagiannopoulos · F. Madonna
Istituto di Metodologie per l'Analisi Ambientale (IMAA), Consiglio Nazionale delle Ricerche, (CNR), Tito Scalo, PZ, Italy
e-mail: lucia.mona@imaa.cnr.it; nikolaos.papagiannopoulos@imaa.cnr.it; fabio.madonna@imaa.cnr.it

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_4

The comparison of EARLINET profiles and their CALIPSO counterpart is a straightforward procedure. Since June 2006, many EARLINET stations are providing measurements in correspondence to CALIPSO overpasses within 100 km (see [15]), according to CALIPSO validation plans. An integrated study of CALIPSO and EARLINET correlative measurements opens new possibilities for spatial (both horizontal and vertical) and temporal representativeness investigation of this set of satellite measurements.

The main aim of this work is to investigate the horizontal smoothing impact on the uncertainty term between the satellite and the ground measurement of the aerosol layers. In the current study, nine different horizontal averaging schemes for the CALIPSO data are used in order to investigate the influence of horizontal smoothing of CALIPSO data when compared against the EARLINET data. In a first analysis, we minimize the RMSE (root-mean-square error) to search for the best horizontal smoothing for CALIOP considering the whole column of aerosol from the ground to the free troposphere, in five different sites. In a second step, to take into account the differences in the vertical dimension and to exploit the vertical profiling capability of both EARLINET and CALIPSO, we split the atmosphere into three zones, below 2.5 km (as representative of local aerosol conditions), between 2.5 and 5.5 km (middle troposphere with transport of aerosols), and above 5.5 km (free troposphere), and we investigate the impact of horizontal smoothing in the three vertical zones.

Our study shows that the co-location mismatch decreases as the altitude increases, and it has its minimum around 50 km, while for peculiar situations (mountain or region surrounded by mountains), this minimum is shifted around 100–150 km.

The paper is organized as follows. The next section recalls the metrology of a data comparison and associated errors. Section 3 is devoted to the presentation of the dataset used in the paper, the CALIOP/CALIPSO description is given in Sect. 3.1, and the EARLINET description is given in Sect. 3.2. Section 4 is dedicated to the presentation of the comparison setup. The horizontal smoothing procedure and its results are presented in Sect. 5, while in Sect. 6, the method is applied to the three vertical atmosphere zones. Finally in Sect. 7, some conclusions and further remarks are given.

2 Metrology of a Data Comparison and Associated Errors

Every measurement has imperfections that give rise to an error in the result. As a consequence, a measurement y is never a perfect indicator of the instantaneous state of the measured parameter μ, but

$$y = \mu + \varepsilon$$

where ε is the error usually assumed as a random variable normally distributed with mean zero and variance σ^2. Denote $y_{sat} = y_{sat}(t, s)$ the satellite measure at time t and location s and $y_{gnd} = y_{gnd}(t', s')$ the ground measure at time t' and location s'. In perfect co-location, the true value that has to be measured is $\mu = \mu(t, s)$ for $s \in D$ and $t \in T$ where D and T are the domains of the location and time observation, respectively. The ground measurement equation is $y_{sat}(t, s) = \mu(t, s) + \varepsilon_{sat}(t, s)$, and the satellite measurement equation is $y_{gnd}(t, s) = \mu(t, s) + \varepsilon_{gnd}(t, s)$. The total co-location error is $\Delta(t, s) = y_{sat}(t, s) - y_{gnd}(t, s) = \Delta_y$ that can be written as $\Delta_y = \varepsilon_{sat} - \varepsilon_{gnd}$. Assuming ε_{sat} and ε_{gnd} are independent, we have $E(\Delta_y) = 0$, $Var(\Delta_y) = \sigma^2_{sat} + \sigma^2_{gnd}$. See [17].

In case of spatiotemporal mismatch, i.e., non-perfect co-location, we can have different times of observation with $\Delta t = t' - t$ and $t = t' + \Delta t$ and different locations of observation with $\Delta s = s' - s$ and $s = s' + \Delta s$. In case of profile measures, where the layer is given at different altitudes, we can have also different heights of observation with $\Delta h = h' - h$ and $h = h' + \Delta h$. In this case the measures are $y_{gnd}(t, s, h) = \mu(t, s, h) + \varepsilon_{gnd}$ and $y_{sat}(t', s', h') = \mu(t', s', h') + \varepsilon_{sat}$ where ε_{sat} and ε_{gnd} are the random error with mean zero and variance σ^2_{sat} and σ^2_{gnd}, respectively. They can be assumed independent, but they may depend on (t, s, h).

The total co-location error with mismatch is $\Delta(t, s, h, t', s', h') = y_{sat}(t', s', h') - y_{gnd}(t, s, h) = \Delta_y$. It can be written as $\Delta_y = \delta_{env} + \Delta_\varepsilon$ where $\delta_{env}(t, s, h, t', s', h') = \mu(t', s', h') - \mu(t, s, h)$ is the environmental component and $\Delta_\varepsilon = \varepsilon_{sat} - \varepsilon_{gnd}$ is the co-location error. The environmental component can be statistically modeled such as in [3] and [10]. In case of spatiotemporal mismatch, i.e., non-perfect co-location, we can suppose that an additional error term ε_{mis} term, with variance σ^2_{mis}, is introduced. So we have $E(\Delta_y) = \delta_{env}$ and

$$E(\Delta_y^2) = \sigma^2_{sat} + \sigma^2_{gnd} + \sigma^2_{mis} \tag{1}$$

Eq. (1) is valid if we can suppose no correlation between ε_{sat}, ε_{gnd}, and ε_{mis}. If we suppose such correlation, we have

$$\begin{aligned} E(\Delta_y^2) &= E(\varepsilon_{sat} - \varepsilon_{gnd} + \varepsilon_{mis})^2 \\ &= E(\varepsilon^2_{sat}) + E(\varepsilon^2_{gnd}) + E(\varepsilon^2_{mis}) + \\ &\quad - 2E(\varepsilon_{sat}\varepsilon_{gnd}) + 2E(\varepsilon_{sat}\varepsilon_{mis}) - 2E(\varepsilon_{gnd}\varepsilon_{mis}) \end{aligned}$$

We can suppose $E(\varepsilon_{sat}\varepsilon_{gnd}) = 0$ because we make the assumption that there is no correlation between the ground error and the satellite error measurement. So we can write

$$E(\Delta_y^2) = \sigma^2_{sat} + \sigma^2_{gnd} + \sigma^2_{mis} + 2\rho_{sat,mis} - 2\rho_{gnd,mis} \tag{2}$$

where $\rho_{sat,mis} = E(\varepsilon_{sat}\varepsilon_{mis})$ and $\rho_{gnd,mis} = E(\varepsilon_{gnd}\varepsilon_{mis})$ are covariances that we cannot suppose to be zero. Indeed it seems reasonable to suppose they are negative because the more the measurement error decreases (and this happens if you increase the time and space in the measurement), the more the mismatch error increases.

This fact is observed in data. Let l be a multi-index identifying the different observations. For example, in our study, $l = (h, d, s)$ where h is the height, d the day, and s the station of any measure. For any l the observations are $y_{sat,l} - y_{gnd,l} = \Delta_l$. Related to any of these observations, we have the given measurement uncertainty $u_{sat,l}$ and $u_{gnd,l}$ that can be considered as an estimation for σ_{sat} and σ_{gnd}, respectively, at each spatiotemporal location l. At a first analysis, fix a station s. The quantity $\hat{\sigma}^2_{TOT} = \frac{1}{N-1}\sum_l (y_{sat,l} - y_{gnd,l})^2$, where N is the total number of observations, is an estimation of $E(\Delta_y^2)$. The quantities $\hat{\sigma}^2_{sat} = \frac{1}{N-1}\sum_l u^2_{sat,l}$ and $\hat{\sigma}^2_{gnd} = \frac{1}{N-1}\sum_l u^2_{gnd,l}$ are estimation of σ^2_{sat} and σ^2_{gnd}, respectively. The correlation terms $\rho_{sat,mis}$ and $\rho_{gnd,mis}$ have to be estimated. Let $\hat{\rho}_{sat,mis}$ and $\hat{\rho}_{gnd,mis}$ be such estimators. An estimation of the variance of the mismatch component can be achieved as $\hat{\sigma}^2_{mis} = \hat{\sigma}^2_{TOT} - \hat{\sigma}^2_{sat} - \hat{\sigma}^2_{gnd} + \hat{\rho}_{sat,mis} - \hat{\rho}_{gnd,mis}$. Usually the uncertainty $\hat{\sigma}^2_{sat}$ related to CALIOP measurement is very high. This gives negative values for $\hat{\sigma}^2_{mis}$ if we consider the decomposition given by (1). For this reason is it essential to include the term $\hat{\rho}_{sat,mis}$, expected as negative, in the uncertainty budget given by (2).

3 Aerosol Profiles: Comparison of CALIOP/CALIPSO and EARLINET

The high variability both in space and time of tropospheric aerosols is one of the main causes of the high uncertainty about radiative forcing related to tropospheric aerosols and their interactions with clouds (see [4]). In particular, information about the vertical layering of aerosol and aerosol vertical distribution is a crucial point for aerosol-clouds interaction study. Moreover, the lack of information about the vertical mixing can lead also to significant horizontal inhomogeneities. These are due to large vertical concentration gradients, and it is therefore a large source of variability. Typically this source is not considered in the models. Since 2006, CALIOP is providing high-resolution vertical profiles of aerosols and clouds on a global scale. However, because of the small footprint and the revisit time of 16 days, how well these CALIOP measurements represent the atmospheric conditions of a surrounding area over a longer time is an important issue to be investigated. An integrated study of CALIPSO and EARLINET correlative measurements opens new possibilities for spatial (both horizontal and vertical) and temporal representativeness investigation of this set of satellite measurements.

3.1 CALIOP/CALIPSO Description

The NASA/CNES CALIPSO mission is designed to study aerosols and clouds (see [20]). Its aim is to provide profiling information at a global scale for improving our knowledge and understanding their climatic role. The main instrument, CALIOP, is a dual wavelength (532 and 1064 nm) elastic backscatter LIDAR with the capability of polarization sensitive observations at 532 nm [18]. The CALIPSO satellite was launched into a near sun-synchronous orbit (SSO) and low Earth orbit (LEO) at a 705 km altitude. Using active remote sensing techniques, CALIOP observes aerosols during daytime and nighttime conditions and therefore provides constant observations of aerosols and clouds. In particular, CALIPSO mission offers unprecedented observations of day and night aerosol global optical properties profiles, vital for aerosol-radiation-cloud interaction studies to understand their climatic role [21]. Instrument data is transmitted from the satellite to the ground station once per day and transferred to the level 0 processing facility to packetize, time order, and archive. The instrument data is combined with ancillary datasets such as meteorological, ephemeris, and instrument status and global reference products to enhance the quality and accuracy of the data products. The LIDAR level 1 data product contains a half orbit (day or night) of calibrated and geolocated LIDAR profiles. Apart from LIDAR data, satellite position data and viewing geometry are provided in the product. There are three types of Lidar level 2 products: layer products (cloud and aerosol), profile products (backscatter and extinction), and a vertical feature mask (cloud and aerosol locations and the corresponding type). Details of the CALIOP instrument and algorithms can be found in the companion papers of the JTECH special issue (http://journals.ametsoc.org/topic/calipso). Additional details can be found in the CALIPSO algorithm theoretical basis documents (ATBDs; available online at https://www-calipso.larc.nasa.gov/resources/project_documentation.php). The aerosol-related data are generated at a uniform horizontal resolution of 5 km. Finally, the level 3 product reports monthly mean profiles of aerosol optical properties on a uniform spatial grid. All level 3 parameters are derived from the CALIPSO level 2, 5 km aerosol profile products applying some additional quality screening filters (see [21]). The retrieval of optical profiles from CALIPSO observations is highly complex, and its detailed description is out of the scope of this document. However, the whole procedure could be briefly summarized in the following manner. Aerosol extinction and backscatter coefficients are retrieved in three steps: (1) layers are searched in the LIDAR-acquired profiles, with horizontal averaging varying from 1/3 km to 80 km; (2) these layers are flagged as clouds or aerosols; and (3) the aerosol extinction and backscatter profiles are retrieved. The succession of the abovementioned steps is described in detail in a special issue of the Journal of Atmospheric and Oceanic Technology (see [19]). Note that the above retrieval from CALIOP elastic backscatter LIDAR is underdetermined, and an additional assumption is needed. In case of an elevated aerosol layer that lies in clear air, the transmittance through the layer can be estimated from the clear-air signals [22]. This offers the needed

constraint for the extinction retrieval; however the CALIOP SNR (signal-to-noise ratio) levels do not usually permit the application of the technique. Therefore, an algorithm [13] is used to estimate the extinction-to-backscatter ratio from the 532 nm depolarization and backscatter signals, which provides the abovementioned assumption [23]. The signal calibration, which precedes the above chain of aerosol retrieval, along with the correct aerosol layer detection and the aerosol layer subtyping dictate the correct retrieval, and any errors in these parameters will lead to errors in the optical properties retrieved by CALIPSO. An extended error analysis of the aerosol extinction and backscatter retrieval can be found in [24]. As only CALIPSO level 2 aerosol profiles are used, these data are described in the following subsections.

3.1.1 CALIOP Sampling

The CALIOP sampling is dictated by the laser repetition rate, the detection configuration, and the satellite-target geometry. In particular the fundamental sampling resolution of the LIDAR is 30 m vertically and 333 m horizontally. The firing rate of the laser is 20 Hz and, according to the minimal resolution of 1/3 km (average of 15 single-shot profiles), leads to a temporal sampling of 0.75 s (https://www-calipso.larc.nasa.gov/resources/pdfs/PC-SCI-201v1.0.pdf).

3.1.2 CALIOP Smoothing

The SNR level of the CALIPSO raw signals at sampling could be very low because of many factors: CALIPSO's distance from the target, the high speed at which the LIDAR sweeps across the target space, constraints placed on the pulse energy of the laser transmitter by eye-safety requirements, the relatively low firing rate of the laser (20 Hz) relative to the velocity of the satellite, and vertical and horizontal variations in the composition of the layers being measured. Appropriate procedures are used for the CALIPSO satellite-borne aerosol measurements for improving the SNR affecting the smoothing in the vertical, horizontal, and temporal dimensions.

Vertical Smoothing

There exists a multistep averaging scheme that dominates the vertical and horizontal resolution. The spatial invariant resolution shown in Table 1 is the resolution applied to raw data already in the onboard averaging scheme. An altitude-dependent averaging scheme is used by CALIPSO and provides higher resolution in the lower troposphere where the spatial variability of cloud and aerosol is larger and lower resolution above. The degree of averaging varies with the altitude, as detailed in the mentioned Table 1 (see https://www-calipso.larc.nasa.gov/resources/pdfs/PC-SCI-202.Part1_v2-Overview.pdf). This scheme is performed before the data are downlinked to ground data processing stations and can be regarded as preprocessing.

Table 1 Spatial resolution for the CALIPSO onboard averaging scheme (altitudes are with respect to mean sea level)

Altitude range (km)	Vertical resolution (m)	Horizontal resolution (km)	Profile per 5 (km)	Samples per profile
20.2 to 30.1	180	1.7	3	55
8.2 to 20.2	60	1.0	5	200
−0.5 to 8.2	30	1/3	15	290

Table 2 The horizontal averaging applied to CALIPSO data along with the corresponding temporal sampling and number of laser shots

Level 2 product post-processing		
Horizontal averaging (km)	Temporal resolution (s)	Laser shots (number)
25	3.75	75
45	6.75	135
75	11.25	225
105	15.75	315
125	18.75	375
155	23.25	465
175	26.25	525
205	30.75	615

For the current study, only level 2 aerosol profile data in the range −0.5 to 20.2 km are used. For these data the vertical resolution is made homogenous at 60 m by averaging consecutive points in the lower range, −0.5 to 8.2 km. As a result, we have 145 and 200 samples per profile in the 0.5–8.2 km and 8.2–20.2 km ranges, respectively. Within this study no further vertical averaging has been applied to the original level 2 as released by the CALIPSO team following the ATBD reported on the CALIPSO website.

Horizontal Smoothing

The CALIPSO algorithms perform a horizontal averaging to enhance the detection of aerosol layers. The averaging is performed for 1/3, 1, 5, 20, and 80 km. For the current study, CALIPSO level 2 data are used which are spatially uniform and reported in 5 km segments. Details on how this is achieved are reported at https://www-calipso.larc.nasa.gov/resources/pdfs/PC-SCI-202.Part1_v2-Overview.pdf. For the purpose of the current study, further horizontal averaging was applied on the level 2 CALIPSO data (see Sect. 5). Different horizontal averaging schemes for the CALIPSO data are used (Table 2) in order to investigate the influence of horizontal smoothing of CALIPSO data when compared against EARLINET data. CALIPSO data at different horizontal resolutions are obtained averaging the original 5 km ones, without applying any screening criteria on them.

3.2 EARLINET Description

EARLINET (European Aerosol Research Lidar NETwork) is the first LIDAR network for aerosol studies on a continental scale. EARLINET comprises of different instrumental setups, specifics, and team expertises. Building up on substantial measurement heritage, EARLINET worked on three main aspects: (1) harmonization of the QA procedures (instrumental and algorithm), (2) establishing measurement schedule, and (3) creating a centralized dataset and homogeneous data format (see [16] and references therein). The success of EARLINET, established in 2000 and currently part of ACTRIS, the European Research Infrastructure for Aerosol, Clouds, and Trace gases observations, paved the way for a further step in the global LIDAR aerosol monitoring even if starting up from heterogeneous LIDAR networks within the global aerosol lidar network (GALION) established by the WMO.

3.2.1 EARLINET Sampling

EARLINET is a network of different instruments with a wide variety of instrumental specifics, so that there is not a common vertical and temporal sampling overall within the network. On the other hand, differences in the instrument components result in different signal-to-noise ratio throughout the network; therefore different smoothing levels are needed station by station. Spatiotemporal resolution and sampling are established at station level and vary in a significant way among the network. This aspect could mean a loss of homogeneity in the considered dataset, but on the other hand, it provides the opportunity for investigating how the different setups affect the EARLINET-CALIPSO comparison. As illustrated in Table 3, EARLINET LIDAR signals considered here are acquired with vertical sampling between 3.75 and 60 m. Moreover, temporal sampling is characterized by the fact that each LIDAR signal is acquired over a temporal window between 10 and 60 s. Of course, horizontal sampling can be considered as pointwise measurements.

Table 3 Vertical and horizontal sampling and repetition rate of the EARLINET LIDAR systems considered in this study

Station	Lidar name	Vertical sampling (m)	Temporal sampling (s)	Laser repetition rate (Hz)
Évora	Paoli	30	60	20
Granada	Raymetrics D400	7.5	10	10
Leipzig	Martha	60	30	30
Napoli	–	15	60	20
Potenza	MUSA	3.75	60	20

3.2.2 EARLINET Smoothing

Vertical Smoothing

After the LIDAR signal acquisition, the signals are integrated, thus modifying both the vertical and temporal resolution. Vertical resolution is decided at station level with the goal to improve the SNR levels. It can be variable with the altitude range: typically a finer vertical resolution is set in the lowest altitude range where the aerosol load is high and a coarser one at the upper levels where the aerosol load is low for improving the SNR. The aerosol extinction retrieval is numerically more complex with respect to aerosol backscatter because it involves the derivative of the signal. This complexity results in a coarser resolution. Even if the vertical extinction profiles are provided to the original raw resolution (typically 15 m), the effective resolution is coarser: each point is provided because the ensemble of these 'not-independent' points provides a better reconstruction of the real atmospheric feature (exactly as happens for image processing). The effective resolution is evaluated adopting interferometric criteria for peak discrimination [9]. For the EARLINET stations, typically the resolution for the aerosol backscatter is 60 m, while the resolution typically ranges between 200 and 600 m for aerosol extinction, reaching values of 1.2–1.5 km at the highest altitude ranges when no aerosol layers are identified.

Temporal Smoothing

With regard to the temporal resolution, the signals are averaged for increasing the SNR in such a way to cover the widest altitude range possible. Typically signals are averaged between 30 min and 1 h in homogeneous aerosol load conditions. Laser shots vary between 18,000 and 36,000 for a laser repetition rate of 10 Hz, between 36,000 and 72,000 for a laser repetition rate of 20 Hz, and between 54,000 and 108,000 for a laser repetition rate of 30 Hz. This resolution depends also on the aerosol content: low aerosol content means low signal, and therefore a longer temporal integration time is needed for obtaining high SNR. Table 3 reports in the last column the repetition rate of the EARLINET LIDAR systems considered in this study.

4 Comparison Setup

Since June 2006, many EARLINET stations are providing measurements in correspondence to CALIPSO overpasses within 100 km [15], according to CALIPSO validation plans. Additionally, simultaneous measurements are planned in order to study the aerosol temporal variability or in the case of special events to study specific aerosol types and to investigate the geographical representativeness of the

observations [15]. The measurement schedule is centrally distributed among the stations, and measurements are performed under weather favorable conditions and conditioned by the station's manpower availability. Up to July 2016, the EARLINET database reports more than 9000 files related to CALIPSO overpasses (EARLINET publishing group 2014; https://data.earlinet.org). In the following, we consider only files related to overpasses within a 100 km radius from the station and excluding special events (Case B and Case C described in [15]). In particular, 143 aerosol backscatter coefficient profiles are compared against their CALIPSO counterparts. This parameter has been selected for investigating the balance difference between the two observations because it was demonstrated that, among the CALIPSO optical properties, the aerosol backscatter is less affected by the inversion assumptions [14]. Further only nighttime measurements are considered because of the larger calibration uncertainty for daytime CALIPSO measurements.

The comparison of EARLINET profiles and their CALIPSO counterpart is a straightforward procedure. Both EARLINET and CALIPSO make use of active remote sensing instruments, yet, the nature and the needs of the satellite mission require special care for any validation study. EARLINET is performing correlative measurements since CALIPSO started its life cycle (April 2006), based on a schedule established before the satellite mission. The strategy followed by the member stations is as follows: the observations occur during the satellite overflight within 100 km distance of the satellite ground track from the station and are performed for at least 60 min.

Only CALIPSO measurements synchronous to the EARLINET measurements are used here. The CALIPSO data are searched for the closest in distance point. This point corresponds to a 5 km CALIPSO profile. In this study more 5 km CALIPSO profiles are also averaged in order to assess the spatiotemporal satellite's performance. Apart from the original 5 km profile, further eight horizontal resolutions are used: 25, 45, 75, 105, 125, 155, 175, and 205 km.

To investigate dependence on the specific site, only the EARLINET stations with a large enough number of co-located observations are considered and analyzed. Namely, the stations are Évora, Granada, Napoli, Potenza, and Leipzig giving 19, 21, 40, 37, and 26 co-located observations, respectively. Table 4 reports the localization parameters of each station.

Prior to the aforementioned analysis, the following identified cirrus cases have been screened out from EARLINET data (CALIPSO aerosol data are already screened in this sense). The cloud screening of EARLINET data is not an automatic procedure for the current version of the database. This was done for this work

Table 4 Latitude, longitude, and altitude (m.o.s) of the location of EARLINET stations considered in this study

Station	Latitude	Longitude	Altitude (m.o.s)
Évora	38.568 N	7.912 W	293
Granada	37.164 N	3.605 W	680
Leipzig	51.353 N	12.435 E	90
Napoli	40.838 N	14.183 E	293
Potenza	40.601 N	15.724 E	760

Table 5 Cirrus clouds detected and discarded in EARLINET stations involved in this study

Station	Date	Altitude range (km)
Naples	19 Sep 2008	Above 6.5
	17 May 2009	Above 6
	24 Nov 2009	Above 8
	18 Apr 2010	Above 5.4
Potenza	03 Apr 2007	At 8–9.5
	21 Apr 2008	Above 7.5
	14 Nov 2008	Above 7
	27 Apr 2010	Above 8

taking advantage of the available labeling of the data. Cirrus clouds are high clouds and are predominantly of ice. These clouds attenuate the laser pulse, and their pronounced structures can be easily discriminated. The cirrus category and the reported information about the cirrus cloud altitude range found in the comment field of the data file were used. However, in some cases, manual inspection of the data was needed for identifying the cirrus cloud in the profiles, since the inclusion of this information within the EARLINET file is not mandatory at this stage. As a result the data reported in Table 5 have been removed from the analysis. In addition, the 27 data point in profiles where the uncertainties above 6923 m resulted to be greater than $1\,\mathrm{m}^{-1}\mathrm{sr}^{-1}$ at Granada have been also removed from subsequent uncertainty analyses.

In order to homogenize the EARLINET aerosol profiles in terms of altitude levels (i.e., the same altitude points for all the profiles provided by the same stations), an interpolation on two points has been applied to Évora profiles. In fact this station changed instrument configuration during the considered period so that an adaptation is needed to homogenize the data from the two periods. Finally, aerosol backscatter data from the station of Naples have been reconstructed to match the altitude points of the corresponding extinction profiles in order to allow in a second step to investigate the differences of the LIDAR ratio obtained from EARLINET and CALIPSO (i.e., EARLINET LIDAR ratio). The Naples aerosol backscatter profiles vertical resolution has been modified for the backscatter to fit the coarser extinction resolution. At each extinction altitude, it has been associated to the value of backscatter which is the closest in altitude (the difference is always within the vertical effective resolution).

5 Horizontal Smoothing

As reported above, the performed analysis is done on the backscatter variable, because it is the CALIPSO product less affected by retrieval assumptions. From Eq. (1) in Sect. 2, the uncertainty term σ^2_{mis} depends on the horizontal and temporal mismatch error between y_{sat} and y_{gnd}, the satellite and ground backscatter, respectively. In order to investigate how the horizontal smoothing impacts on the term σ^2_{mis}

of the uncertainty budget, we consider the root-mean-square error (RMSE), which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k}(y_{sat,l} - y_{gnd,l})^2}$$

where $l = (h, d, s)$ is a multi-index defined in Sect. 2 and N is the total number of observations.

In this case study, we have 200 altitude levels, corresponding to CALIOP observations. The vertical range is 97.67–12,013 m, with a step of 60 m. The counter $s = 1, \ldots, 5$ identifies the five EARLINET stations (Évora, Granada, Leipzig, Napoli, and Potenza), $d = 1, \ldots, g_s$ identifies the day's profile, and g_s gives the number of profiles available for station s. In total, the analysis considers 143 EARLINET profiles. To understand how the co-location error depends on the horizontal smoothing, we have computed the RMSE of the eight different horizontal averaging schemes for CALIOP described in Table 1, which includes also the original CALIPSO 5 km data.

In Fig. 1 and Table 6, the co-location uncertainty averaged by station and CALIOP horizontal smoothing is presented in order to understand how smoothing

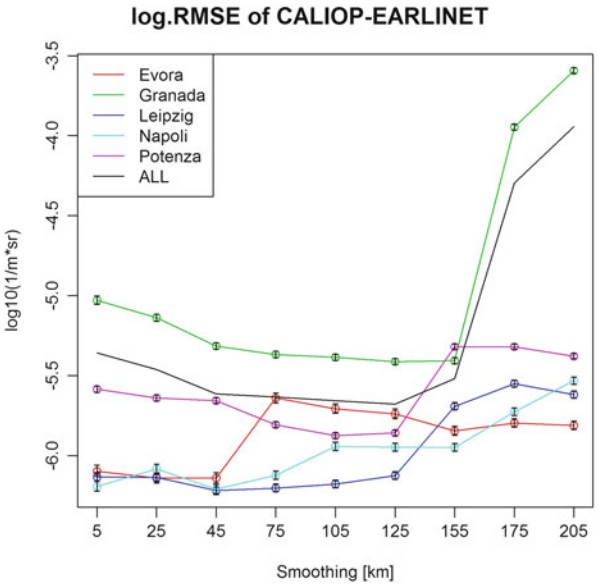

Fig. 1 Co-location uncertainty (log.RMSE) for CALIOP and EARLINET backscatter mismatch by smoothing parameter for different stations. Vertical bars represent the 95% error intervals

Table 6 Co-location uncertainty (RMSE) by smoothing and station [1/mr]

Station	Horizontal smoothing (km)								
	5	25	45	75	105	125	155	175	205
Évora	8.012E−07	7.253E−07	**7.255E−07**	2.292E−06	1.966E−06	1.825E−06	1.429E−06	1.600E−06	1.547E−06
Gran.	9.357E−06	7.278E−06	4.842E−06	4.292E−06	4.119E−06	**3.876E−06**	3.920E−06	1.131E−04	2.546E−04
Leipzig	7.352E−07	7.308E−07	**6.053E−07**	6.276E−07	6.655E−07	7.513E−07	2.034E−06	2.818E−06	2.405E−06
Napoli	6.420E−07	8.262E−07	**6.170E−07**	7.534E−07	1.141E−06	1.130E−06	1.126E−06	1.882E−06	2.941E−06
Potenza	2.602E−06	2.290E−06	2.205E−06	1.561E−06	**1.335E−06**	1.387E−06	4.794E−06	4.792E−06	4.190E−06
All	4.380E−06	3.463E−06	2.432E−06	2.327E−06	2.207E−06	**2.103E−06**	3.026E−06	5.064E−05	1.139E−04

Bold value highlights the station minima

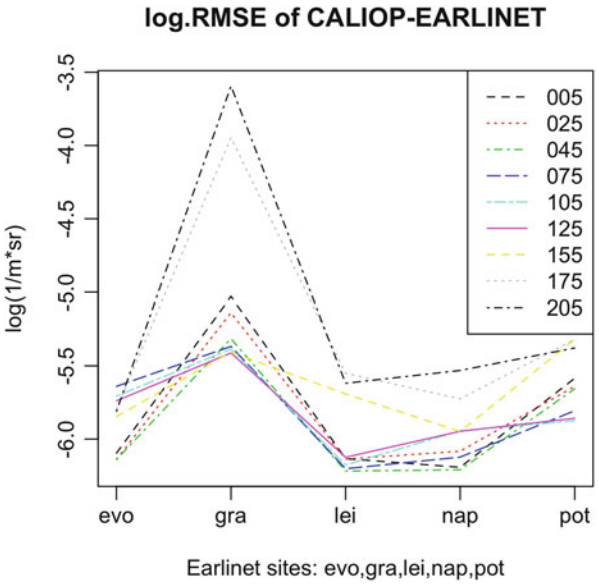

Fig. 2 Co-location uncertainty (log.RMSE) for CALIOP and EARLINET backscatter mismatch by station for different smoothing parameters

affects the comparison. The bold values correspond to station minima. The minimal co-location uncertainty is obtained at 45 km for Évora, Leipzig, and Napoli, which have small uncertainties already at 5 km. This suggests that using 45 km as standard horizontal averaging is advisable for the comparison of pointwise ground-based measurements with the CALIOP level 2 product. At Potenza and Granada instead, the minimal co-location uncertainty is obtained for 105 and 125 km, respectively. This different behavior of Granada and Potenza can be ascribed to the variant orography that affects the atmosphere sampled by the satellite LIDAR: compared to the other sites in fact, Granada and Potenza are the unique ones located at upper altitudes and surrounded by different areas (see [1] and [12]).

From Fig. 2, we see also that for all the stations, the co-location uncertainty tends to increase in the tails of smoothing range that is 175–205 km and 5–25 km. Figs. 1 and 2 indicate that the uncertainty in Granada is larger than in all the other stations, with the exception of Potenza at 105 km smoothing, where however the uncertainties are comparable. This underlines the peculiarity of Granada comparison: in this case the co-location uncertainty is higher than for all the other stations, independently from the horizontal smoothing. Granada station is located in a natural basin surrounded by mountains with the highest mountain range located to the southeast, with altitudes above 3000 m [5]. The presence of these mountains can act as a boundary for both local and free troposphere aerosol layers depending on the specific aerosol source, so that if the CALIOP track location is not favorable, very large differences are expected when compared to the ground-based measurements.

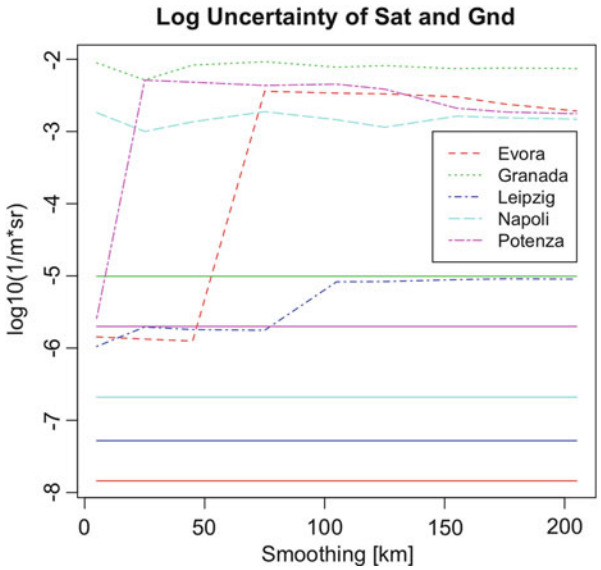

Fig. 3 Measurement uncertainty by station. Dashed lines: CALIOP uncertainty averaged by smoothing parameter. Solid lines: EARLINET uncertainties

Finally both Figs. 1 and 2 report the mean co-location uncertainty, which has the same behavior of Granada values, because it is driven by these very large values and therefore cannot be regarded as representative of the ensemble of the considered locations.

In order to better understand these results on co-location uncertainty, the measurement uncertainty has been investigated. In Fig. 3, the solid horizontal lines represent the averaged measurement uncertainty of EARLINET, while the dashed lines represent the uncertainties of CALIOP smoothed backscatter. As expected, the uncertainty of ground measurements is smaller than the satellite one for all stations. Moreover, in Table 7, the mean of the measurement uncertainties by station and smoothing are reported, while in Table 8, the relative measurement uncertainty (coefficient of variation) for the backscatter values is reported.

Generally speaking, the satellite uncertainty is larger or equal to ground uncertainty. Granada has the largest measurement uncertainty, for both EARLINET and CALIOP. This appears also from Table 8 where Granada has the largest values of relative uncertainty both for EARLINET and CALIOP when averaged over smoothing. Comparing Figs. 1 and 2, we can conclude that horizontal smoothing affects in a different way measurement uncertainty and co-location uncertainty.

Table 7 Average measurement uncertainties for ground and smoothed satellite backscatter [1/msr]

	Horizontal smoothing (km)									
Station	Earlinet	5	25	45	75	105	125	155	175	205
E	1.45E−08	1.43E−06	1.33E−06	1.25E−06	3.60E−03	3.40E−03	3.32E−03	3.02E−03	2.40E−03	1.92E−03
G	9.88E−06	8.90E−03	5.18E−03	8.30E−03	9.24E−03	7.78E−03	8.16E−03	7.41E−03	7.57E−03	7.44E−03
L	5.23E−08	1.05E−06	1.96E−06	1.81E−06	1.77E−06	8.29E−06	8.34E−06	8.88E−06	9.12E−06	9.02E−06
N	2.09E−07	1.82E−03	1.00E−03	1.36E−03	1.88E−03	1.45E−03	1.15E−03	1.63E−03	1.55E−03	1.48E−03
P	1.99E−06	2.58E−06	5.14E−03	4.82E−03	4.34E−03	4.52E−03	3.86E−03	2.10E−03	1.86E−03	1.76E−03

Table 8 Relative measurement uncertainty of ground and smoothed satellite backscatter

	Horizontal smoothing (km)									
Station	Earlinet	5	25	45	75	105	125	155	175	205
E	0.1	1.1	1.2	1.3	2106.6	2199.6	2247.2	2578.7	2273.2	2010.4
G	8.1	1422.1	1270.2	3325.5	4591.2	4897.2	4654.4	4381.4	1174.7	592.6
L	0.2	1.2	2.3	2.3	2.2	5.9	5.8	4.8	4.4	4.5
N	0.6	1794.8	1031.6	1331.7	1674.2	1410.4	1278.1	1863.5	1741.8	1479.4
P	5.1	1.9	2547.4	2499.9	2196.3	2237.3	2117.8	1468.9	1335.4	1327.9

6 Vertical Splitting

All values reported in the analysis in Sect. 5 are related to the whole column: uncertainties related to altitude regions with high aerosol content are mixed with altitude regions where the aerosol content is very low and uncertainties are expected to be high because of a low SNR. To take into account the differences in the vertical dimensions and to exploit the vertical profiling capability of both EARLINET and CALIPSO, we split the atmosphere into three zones: below 2.5 km (as representative of local aerosol conditions), between 2.5 and 5.5 km (middle troposphere with transport of aerosols), and above 5.5 km (free troposphere).

Figure 4 (left panel) corresponds to the lowest part of the troposphere, that is, the planetary boundary layer, where, normally, lower agreement is expected between the ground station and the satellite swath. All the values are grouped around $-6.0\log_{10}(\mathrm{m}^{-1}\mathrm{sr}^{-1})$, with the exception of Granada probably because of the already mentioned variant topography surrounding this EARLINET station (see [1] and [14]). Regarding Potenza station, there is a decrease in the RMSE for smoothing up to 100 km, while it is approximately constant and comparable with Évora, Leipzig, and Napoli stations for larger smoothing. This effect could be attributed on one hand to the mountainous area where Potenza is situated and, on the other hand, to the variety of surfaces that the satellite encounters (e.g., land, sea). In particular, Potenza is on a mountain close to the sea but also to big cities, so the difference with smaller horizontal smoothing can be larger (e.g., CALIPSO ground track lies over the sea) than at higher smoothing where smoothing procedure merges different conditions (comparing EARLINET mountain sampling versus an average of sea, cities, and mountain). This finding is in agreement with the smaller discrepancies observed in the PBL (planetary boundary layer, the lowest part of the atmosphere in contact with the Earth's surface) at Potenza between EARLINET and corresponding CALIPSO observations for the overpasses at about 80 km distance with respect to the closer overpasses at 40 km distance because of topographic and local effects (see [12]). The opposite behavior is found for Leipzig because of the more homogeneous topography. For the remaining stations, the RMSE is lower at original horizontal smoothing and gradually increases with increasing smoothing. Generally, the increasing smoothing tends to increase the RMSE as the satellite contains a vast geographic area, and therefore the aerosol fields can be dramatically

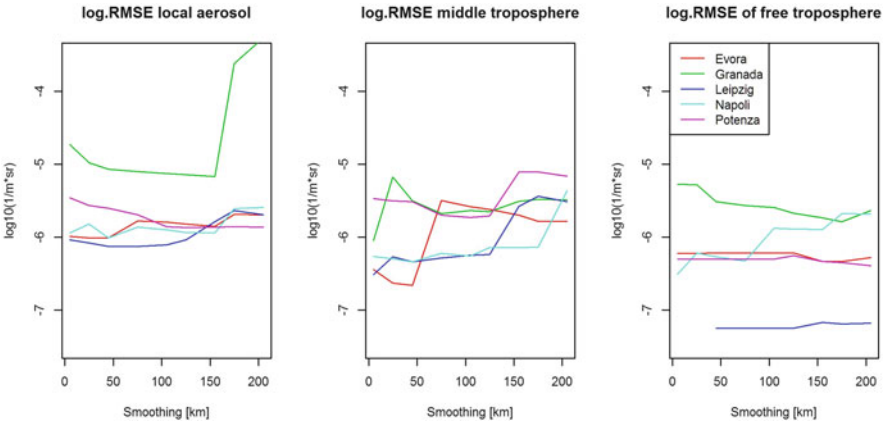

Fig. 4 RMSE for CALIOP and EARLINET measure of aerosol backscatter at different smoothing parameters for different stations and for different zones of atmosphere

different. This effect has been documented and reported in several studies (see, e.g., [2, 15]).

The mid-troposphere plot (Fig. 4; middle panel) corresponds to the height range typically free from local sources and indicates the transboundary motion of aerosol, and therefore this plot is the most relevant to assess the CALIPSO representativeness. At first sight the lines are contained in the range between -6.5 and $-5.0\log_{10}(\mathrm{m}^{-1}\mathrm{sr}^{-1})$ which shows a better agreement with respect to the previous figure implying that the effect of local sources is greatly reduced, especially at Granada. For the first 50 km and for the stations of Napoli, Leipzig, and Évora, the RMSE is the lowest and then either gradually or steeply increases for increasing values of the smoothing parameter. On the other hand, for the station of Potenza, the situation is reversed in the range 0–50 km, and behind that, it shows a behavior which is similar to the other locations. For Granada, the situation is more complex as the RMSE peaks at 25 km and then decreases until 75 km, to follow the behavior of the other stations for the next smoothing ranges.

The free troposphere behavior of Fig. 4 (right panel) corresponds to predominantly aerosol-free area without large variations of RMSE w.r.t. the smoothing parameter. As a case in point, Évora, Potenza, and Leipzig produce a constant value for the whole smoothing range, showing that in free troposphere, the smoothing parameter loses its importance in explaining the co-location error. We observe a finer structure at Granada and Napoli that can be attributed to aerosol structures not observed by either instruments or in case of CALIPSO cloud misclassification (e.g., sub-visual, thin cirrus clouds). These opaque clouds in the higher altitude levels frequently penetrate the CALIOP aerosol retrievals and alter the CALIPSO-provided atmospheric description (see [6–8, 11]).

7 Conclusions

A first effort has been undertaken toward identifying the main contributions of co-location mismatch uncertainties in comparisons of CALIOP and EARLINET aerosol backscatter profiles. The comparison is not trivial because of two main reasons: (1) the small footprint of CALIPSO measurements comparing to the distance from EARLINET sites and (2) the high uncertainty of CALIPSO products (see [24]). The comparison is then even more complex because of the fine vertical structure of the aerosol field and its variability. Co-location mismatch has been investigated as a function of the observational site and of the horizontal smoothing of the CALIOP data. Furthermore, the investigated altitude range [90 m–12 km asl] has been split in three regions corresponding to the PBL, the mid-troposphere, and the free troposphere range. The co-location mismatch decreases as the altitude increases: $10^{-6} - 5 \times 10^{-5}$ m^{-1} sr^{-1} in the PBL, $5 \times 10^{-7} - 5 \times 10^{-6}$ $m^{-1} sr^{-1}$ in the middle troposphere, and typically lower than $10^{-6} m^{-1}$ sr^{-1} in the free troposphere. An influence of the smoothing on co-location mismatch is found for the two lowest atmospheric ranges, while in the free troposphere, its influence can be disregarded. This shows that above 5.5 km, LIDAR pointwise measurements can be typically considered representative even over horizontal scale as large as 200 km, because of the low variability of the aerosol field at these altitudes. In the middle troposphere, the LIDAR data are representative for distances up to 100 km in agreement with [15]. Finally in the lowest troposphere where the orography and local source play a relevant role, the representativeness strongly depends on the site characteristics. Typically the co-location mismatch has its minimum around 50 km, while for peculiar situations (mountain or region surrounded by mountains), this minimum is shifted around 100–150 km. As a further remark, we have to say that we were not able to estimate the covariance component due to the mismatch of the observation with the data in this form. This is only a first attempt to understand how the co-location error is affected by the horizontal smoothing.

Acknowledgements This research is partially funded by GAIA-CLIM, the project funded from the European Union's Horizon 2020 research and innovation program under grant agreement No 640276.

References

1. Alados-Arboledas L, Muller D, Guerrero-Rascado JL, Navas-Guzman F, Perez-Ramirez D, Olmo FJ (2011) Optical and microphysical properties of fresh biomass burning aerosol retrieved by Raman LIDAR, and star-and sun-photometry. Geophys Res Lett 38. https://doi.org/10.1029/2010GL045999
2. Anderson T, Charlson R, Winker D, Ogren J, Holmén K (2003) Mesoscale variations of tropospheric. Aerosols J Atmos Sci 60:119–136. https://doi.org/10.1175/1520-0469(2003)060<0119:MVOTA>2.0.CO;2

3. Fassò A, Ignaccolo R, Madonna F, Demoz B, Franco-Villoria M (2014). Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles. Atmos Meas Tech 7:1803–1816
4. Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, Fahey DW, Haywood J, Lean J, Lowe DC, Myhre G, Nganga J, Prinn R, Raga G, Schulz M, Van Dorland R (2007) Changes in atmospheric constituents and in radiative forcing. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. In: Solomon S, Qin D, Manning MR, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Cambridge University Press, Cambridge, New York, pp 131–217
5. Guerrero-Rascado JL, Ruiz B, Alados Arboledas L (2008). Multi-spectral Lidar characterization of the vertical structure of Saharan dust aerosol over southern Spain. Atmos Environ 42:2668–2681. https://doi.org/10.1016/j.atmosenv.2007.12.062
6. Huang J, Hsu NC, Tsay SC, Liu Z, Jeong MJ, Hansell RA, Lee J (2011) Evaluations of cirrus contamination and screening in ground aerosol observations using collocated lidar systems. J Geophys Res 117:D15204. https://doi.org/10.1029/2012JD017757
7. Huang J et al (2012) Susceptibility of aerosol optical thickness retrievals to thin cirrus contamination during the BASE-ASIA campaign. J Geophys Res 116:D08214
8. Huang J et al (2013) Use of spaceborne lidar for the evaluation of thin cirrus contamination and screening in the Aqua MODIS collection 5 aerosol products. J Geophys Res Atmos 118:6444–6453. https://doi.org/10.1002/jgrd.50504
9. Iarlori M, Madonna F, Rizi V, Trickl T, Amodeo A (2015) Effective resolution concepts for lidar observations. Atmos Meas Tech 8:5157–5176. https://doi.org/10.5194/amt-8-5157-2015
10. Ignaccolo R, Franco-Villoria M, Fassò A (2015) Modelling collocation uncertainty of 3D atmospheric profiles. Stoch Env Res Risk Assess 29(2):417–429
11. Kittaka C, Winker DM, Vaughan MA, Omar A, Remer LA (2011) Intercomparison of column aerosol optical depths from CALIPSO and MODIS-Aqua. Atmos Meas Tech 4:131–141. https://doi.org/10.5194/amt-4-131-2011
12. Mona L, Pappalardo G, Amodeo A, D'Amico G, Madonna F, Boselli A, Giunta A, Russo F, Cuomo V (2009) One year of CNR-IMAA multi-wavelength Raman lidar measurements in coincidence with CALIPSO overpasses: level 1 products com-parison. Atmos Chem Phys 9:7213–7228. https://doi.org/10.5194/acp-9-7213-2009
13. Omar A, Winker D, Kittaka C, Vaughan M, Liu Z, Hu YX, Trepte C, Rogers R, Ferrare R, Lee K, Kuehn R, Hostetler, C (2009) The CALIPSO automated aerosol classification and lidar ratio selection algorithm. J Atmos Ocean Tech 26:1994–2014. https://doi.org/10.1175/2009jtecha1231.1
14. Papagiannopoulos N, Mona L, Alados-Arboledas L, Amiridis V, Baars H, Binietoglou I, Bortoli D, D'Amico G, Giunta A, Guerrero-Rascado JL, Schwarz A, Pereira S, Spinelli N, Wandinger U, Wang X, Pappalardo G (2016) CALIPSO climatological products: evaluation and suggestions from EARLINET. Atmos Chem Phys 16:2341–2357. https://doi.org/10.5194/acp-16-2341-2016
15. Pappalardo G et al (2010) EARLINET correlative measurements for CALIPSO: first intercomparison results. J Geophys Res 115:D00H19. https://doi.org/10.1029/2009JD012147
16. Pappalardo G, Amodeo A, Apituley A, Comeron A, Freuden-thaler V, Linné H, Ansmann A, Bösenberg J, D'Amico G, Mattis I, Mona L, Wandinger, U, Amiridis, V, Alados-Arboledas L, Nicolae D, Wiegner M (2014) EARLINET: to-wards an advanced sustainable European aerosol lidar network. Atmos Meas Tech 7:2389–2409. https://doi.org/10.5194/amt-7-2389-2014
17. Verhoelst T, Granville J, Hendrick F, Köhler U, Lerot C, Pommereau J-P, Redondas A, Van Roozendael M, Lambert J-C (2015) Metrology of ground-based satellite validation: co-location mismatch and smoothing issues of total ozone comparisons. Atmos Meas Tech 8:5039–5062. https://doi.org/10.5194/amt-8-5039-2015
18. Winker DM, Hunt WH, McGill MJ (2007) Initial performance assessment of CALIOP. Geophys Res Lett 34:L19803. https://doi.org/10.1029/2007GL030135

19. Winker DM, Vaughan MA, Omar AH, Hu Y, Powell KA, Liu Z, Hunt WH, Young SA (2009) Overview of the CALIPSO mission and CALIOP data processing algorithms. J Atmos Oceanic Technol 26:2310–2323
20. Winker DM, Pelon J, Coakley Jr, JA, Ackerman SA, Charlson RJ, Colarco PR, Flamant P, Fu Q, Hoff R, Kittaka C, Kubar TL, LeTreut H, McCormick MP, Megie G, Poole L, Powell K, Trepte C, Vaughan MA, Wielicki BA (2010) The CALIPSO mission: a global 3D view of aerosols and clouds. Bull Am Meteor Soc 91:1211–1229. https://doi.org/10.1175/2010BAMS3009.1,2010
21. Winker DM, Tackett JL, Getzewich BJ, Liu Z, Vaughan MA, Rogers RR (2013) The global 3-D distribution of tropospheric aerosols as characterized by CALIOP. Atmos Chem Phys 13:3345–3361. https://doi.org/10.5194/acp-13-3345-2013
22. Young SA (1995) Analysis of lidar backscatter profiles in optically thin clouds Appl Optics 34:7019–7031. https://doi.org/10.1364/AO.34.007019
23. Young SA, Vaughan MA (2009) The retrieval of profiles of particulate extinction from Cloud Aerosol Lidar Infrared Pathfinder Satellite Observations (CALIPSO) data: algorithm description. J Atmos Ocean Tech 26:1105–1119. https://doi.org/10.1175/2008JTECHA1221.1
24. Young SA, Vaughan MA, Kuehn RE, Winker DM (2013) The retrieval of profiles of particulate extinction from CloudAerosol Lidar and infrared pathfinder observations (CALIPSO) data: uncertainty and error sensitivity analyses. J Atmos Ocean Tech 30:395–428

A Spatiotemporal Approach for Predicting Wind Speed Along the Coast of Valparaiso, Chile

Orietta Nicolis, Mailiu Díaz, and Omar Cuevas

Keywords Spatio-temporal model · Wind speed · WRF output · Cross-validation

1 Introduction

Despite recent improvements in weather predictions due to the use of sophisticated and complex numerical models, many problems still remain unresolved. The main drawback is that the prediction from numerical models is often affected by model errors (systematic and stochastic) and that it is not possible to have a measure of the uncertainty associated to the prediction. Despite that different methods have been proposed in the literature to understand and reduce biases, the correction of the weather prediction still remains a challenge, probably due to the large variety of possible error sources (uncertainties in initial condition, parameterizations, model errors, etc.).

A common practice is to calibrate the output of the numerical weather prediction models using observations collected from monitoring stations. In particular, weather station data are more accurate since, up to measurement error, they provide the actual true values. A wide literature on the spatiotemporal regression models focuses on the estimation and prediction of particular environmental variables (such as air

O. Nicolis (✉)
Faculty of Engineering, University Andres Bello, Santiago, Chile

Institute of Statistics, University of Valparaiso, Valparaiso, Chile
e-mail: orietta.nicolis@uv.cl; orietta.nicolis@unab.cl

M. Díaz
Institute of Statistics, University of Valparaiso, Valparaiso, Chile
e-mail: mailiu.diaz@postgrado.uv.cl

O. Cuevas
Institute of Physics and Astronomy, Valparaiso, Chile
e-mail: omar.cuevas@uv.cl

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_5

pollution) by using a series of exogenous variables (meteorological and land-use variables, temporal features, etc.) (see [2] and the references therein). Interesting works on this topic have been recently proposed by Craimile and Guttorp [5], Casquilho-Resende et al. [3], Fassò et al. [6], Sahu and Nicolis [13], Sahu et al. [14], Sahu and Bakar [12], Lindström et al. [10]. A short-term wind forecasting has been recently proposed by Tastu et al. [17] in order to take decisions about reliable and economic power systems.

In this work we use the spatiotemporal model proposed by Lindström et al. [10] for calibrating the wind speed forecasts coming from the WRF model. In particular, we consider the observation collected by a network of meteorological stations as response variable of the model and the WRF output as covariate. The application of the model will allow to assess the wind speed between January 1 and April 1, 2016, each 1 h at multiple sites with 1 km $\times$ 1 km spatial resolution in the coast of Valparaiso.

The paper is organized as follows. Section 2 provides a description of the data and a preliminary analysis. In Sect. 3 we shortly describe the spatiotemporal model proposed by Lindström et al. [10]. Main results on the wind predictions are given in Sect. 4. Conclusions and further developments follow in Sect. 5.

2 Preliminary Analysis

2.1 The Data Sets

We consider the hourly wind speed [m/s] data collected by 11 weather monitoring stations located along the coast of Valparaiso (see Fig. 1) and the meridional and zonal wind components (u and v, respectively) at 10 m above ground level coming from a Weather Research and Forecasting (WRF) model on small spatial resolution 1 km $\times$1 km every 1 h. Both databases are collected for the period from January 1 to April 1, 2016. The area of study is delimited by a green line in Fig. 1. Wind speed observations considered in this work can be downloaded by the web site of the National Air Quality Information System (http://sinca.mma.gob.cl/), Government of Chile. The WRF model is a computationally efficient model which offers advances in physics and numeric and provides detailed databases for land use, topography, and soil type. In this work, we used the Version 3.8.1 [15]. This model was run fully compressible, non-hydrostatic, with three domains (left panel of Fig. 2). The boundary condition was used from the Global Forecasting System (GFS) run at the National Centers for Environmental Prediction [18] that is a global operational model with 3 h of temporal forecasting resolution and 0.25° $\times$ 0.25° of horizontal resolution. Simulations were saved every 1 h from the innermost domain (d03 at 1 km horizontal resolution) centered in the coast zone between -32.5 and -33.8 S of latitude.

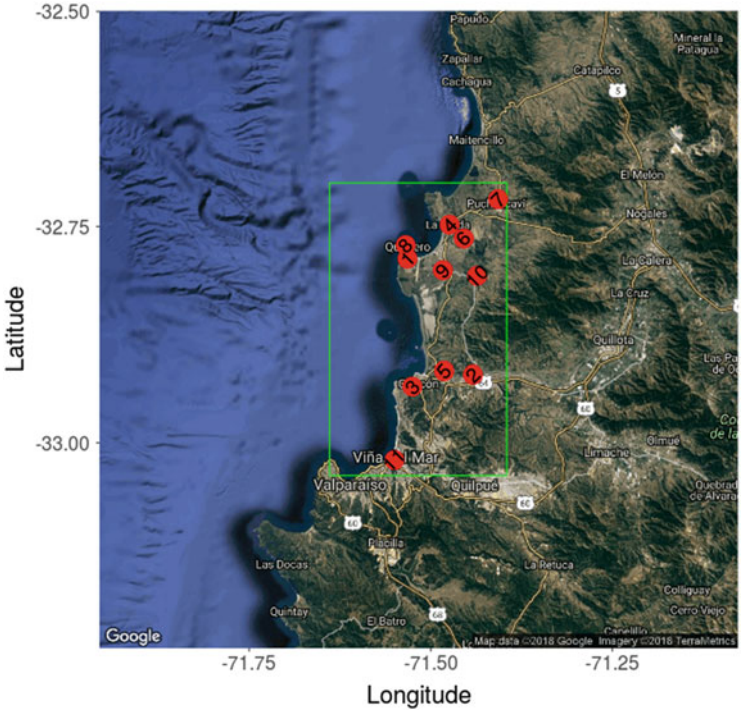

Fig. 1 Google map of the coast of Valparaiso region with the locations of the 11 weather stations (red) and the area of study (green line)

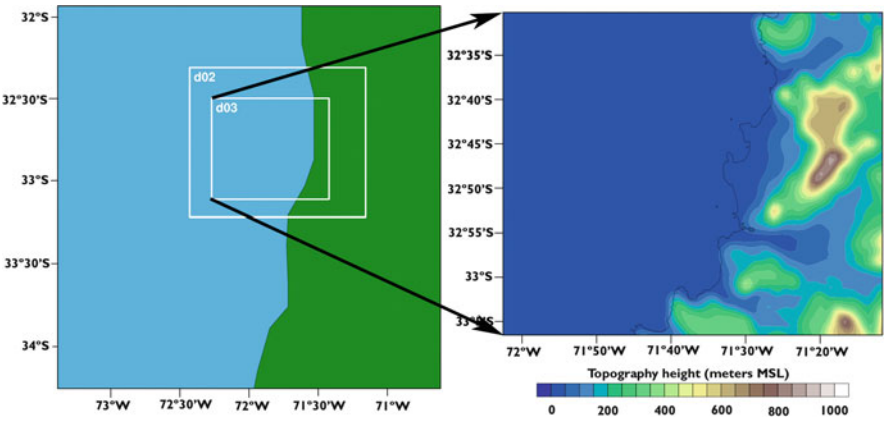

Fig. 2 Domains used in WRF model along the Valparaiso coast (left panel) and zoom of the smaller domain (right panel)

All simulations were performed using the RRTMG longwave and shortwave radiation [9], the Quasi-Normal Scale Elimination (QNSE) planetary boundary layer [16], the Noah land surface model [4], and the WRF Single-Moment 5-class scheme [8] parameterizations. Then wind speed from GFS was calculated with u and v components as $\sqrt{u^2 + v^2}$.

2.2 Preliminary Statistical Analysis

2.3 The Data Sets

Summary statistics for each weather station are described in Table 1. Figure 3a and b show the boxplots for each station and each hour, respectively. Note that station 9 (Quintero), located close to the sea, has the highest wind speed values, and the lowest values are measured by the station 10 (Valle Alegre), which is the farther station from the sea. From Fig. 3b we observe that the wind speed is higher in the afternoon.

In order to show that WRF predictions are affected by a bias and random errors, we compare the observations from the 11 monitoring stations with the output of a WRF model evaluated at the closest points to the locations of the weather stations. From Fig. 4a and b, we can see that WRF predictions are very different from the observations and the WRF model tends to overestimate the values of the wind speed. In Fig. 4b we choose the wind speed at station 6 (Los Maitenes) for the month of

Table 1 Descriptive statistics of meteorological stations

Id	Weather station	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.D.[a]	C.V.[b]
1	Centro Quintero	0.000	0.703	1.240	1.323	1.816	4.994	0.856	64.707
2	Colmo	0.343	1.119	1.766	2.300	3.370	6.922	1.446	62.860
3	Junta de Vecinos	0.195	0.406	1.044	1.398	2.243	5.406	1.128	80.636
4	La Greda	0.044	0.488	1.230	1.705	2.892	5.531	1.362	79.873
5	Las Gaviotas	0.220	1.003	1.608	1.658	2.268	5.461	0.845	50.954
6	Los Maitenes	0.015	0.496	0.951	1.176	1.811	3.759	0.804	68.383
7	Puchuncavi	0.009	0.586	1.106	1.421	2.180	6.294	1.024	72.074
8	Quintero	0.063	1.215	1.902	2.111	2.715	9.020	1.303	61.749
9	Sur	0.012	0.665	1.786	2.132	3.230	8.253	1.644	77.106
10	Valle Alegre	0.003	0.338	0.521	0.733	1.077	2.511	0.535	72.983
11	Viña del Mar	0.000	0.300	0.700	0.798	1.200	2.900	0.568	71.161

[a]Standard deviation
[b]Coefficient of variation

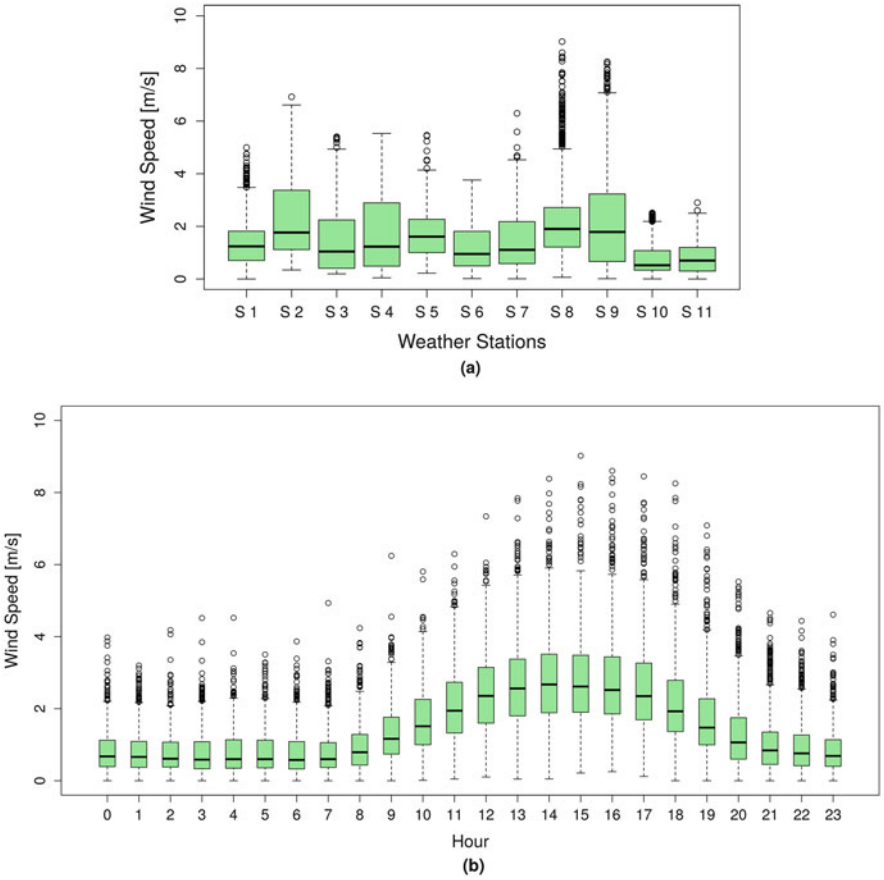

Fig. 3 Boxplots of the wind speed data for each station (**a**) and each hour (**b**)

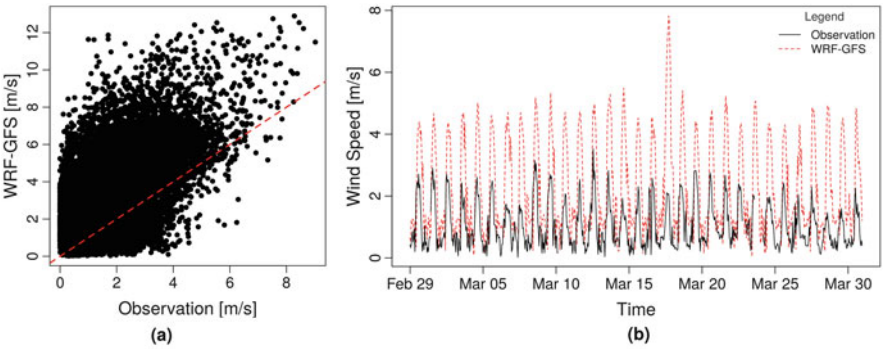

Fig. 4 Observations versus WRF outputs for all the meteorological stations (**a**) and time series of the observations (black line) with the WRF outputs (red line) for the station 6-Los Maitenes on March, 2016 (**b**)

March as example to better show the bias of the WRF predictions. Similar results have been obtained for the other stations.

3 Space-Time Regression Modeling

The spatiotemporal model for correcting the WRF predictions can be written as

$$Y(s,t) = \mu(s,t) + \nu(s,t) \tag{1}$$

where $Y(s,t)$ denotes the wind speed observed from weather stations, $s \in \mathbb{R}^2$ and $t \in \mathbb{R}$, and $\mu(s,t)$ is a structured mean field. Following the spatiotemporal approach proposed by Lindström et al. [10], the $\mu(s,t)$ component can be defined as

$$\mu(s,t) = \gamma \mathscr{M}(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t) \tag{2}$$

where the $\mathscr{M}(s,t)$ denotes the spatiotemporal output of numerical model WRF with GFS edge condition; γ_l are coefficients; $\{f_i(t)\}_{i=1}^{m}$ is a set of smooth temporal basis functions, with m as the number of temporal basis functions (including the intercept) and $f_1(t) \equiv 1$; and the $\beta_i(s)$ are spatially varying coefficients, $\beta_i(s) \sim N(X_i\alpha_i, \Sigma_{\beta_i}(\theta_i))$ for $i = 1, \ldots, m$, where X_i are design matrices (normally contain geographical covariates) and α_i are matrices of coefficients. The space-time residual field is denoted by $\nu(s,t)$, and it is distributed with the following stationary parametric spatial covariance $\nu(s,t) \sim N(0, \Sigma_\nu(\theta_\nu))$ where Σ_ν is a block matrix and the size depends on the observations at each meteorological station. The smooth temporal basis functions, $f_i(t)$, describe the temporal variability in the data. These functions can either be obtained as smoothed singular vectors as proposed by Fuentes et al. [7]. The cross-validation is then used to determine the optimal number of smooth temporal basis functions by evaluating a set of regression statistics such as the Mean Squared Errors (MSE), the coefficient of determination (R^2), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) that those describe how well the left out columns are explained by smooth temporal functions.

Then the model (1) can be written in matrix form as

$$Y = \mathscr{M}\boldsymbol{\gamma} + FB + V, \tag{3}$$

where $B \sim N(X\alpha, \Sigma_B(\theta_B))$ and $V \sim N(0, \Sigma_v(\theta_v))$ (see [10] for details). Since Eq. (3) is a linear combination of independent Gaussians and introducing the matrices $\tilde{X} = [\mathscr{M} \quad FX]$ and $\tilde{\Sigma}(\Psi) = \Sigma_v(\theta_v) + F\Sigma_B(\theta_B)F^T$, the distribution of Y can be written as $[Y|\Psi, \gamma, \alpha] \sim N\left(\tilde{X}\begin{bmatrix}\gamma \\ \alpha\end{bmatrix}, \tilde{\Sigma}(\Psi)\right)$.

The parameter estimates are obtained by maximizing the log-likelihood of

$$2l(\Psi, \alpha, \gamma \,|Y) = -N\log(2\pi) - \log\left|\widetilde{\Sigma}(\Psi)\right| - \left(Y - \tilde{X}\begin{bmatrix}\gamma \\ \alpha\end{bmatrix}\right)^{T} \widetilde{\Sigma}^{-1}(\Psi)\left(Y - \tilde{X}\begin{bmatrix}\gamma \\ \alpha\end{bmatrix}\right) \quad (4)$$

(see [10]). The spatiotemporal model of [10] can be implemented using the R packages SpatioTemporal [1].

4 Results

4.1 Spatiotemporal Estimation

In this section we estimate the spatiotemporal model proposed by Lindström et al. [10] in order to correct the WRF data using the observations coming from the 11 monitoring stations. As mentioned in the last section, we considered the WRF output as the covariate of the model and the observations as the response variable. In order to estimate the model, we considered the observations and WRF data from January 1 to March 30, leaving out the days March 31 and April 1 (equally to 48 h) for the spatiotemporal prediction. First, we selected the number of smoothed temporal bases by using the cross-validation method. Table 2 shows the outputs of cross-validated in terms of MSE, R^2, AIC, and BIC for four basis functions. As expected in any regression scenario, increasing the number of basis functions increases R^2 and decreases the MSE. Since the increment of R^2 after two basis functions was not remarkable, we decided to select two basis functions in the estimation of the model.

The estimated parameters using an exponential spatial covariance structure are shown in Table 3 (the NAs are probably due to the logarithm of standard deviations which are very close to zero).

In order to estimate the performance of the model, we implemented leave-one-out cross-validation and predicted the wind speed for the entire period (from January 1 to April 1) at each step. The results can be summarized in Table 4 where we compare the prediction errors using the WRF and the spatiotemporal model (ST). Since in all cases the prediction obtained by the spatiotemporal model is better than the WRF

Table 2 Cross-validation statistics computed for each smoothed basis function

Basis function	MSE	R^2	AIC	BIC
0	0.1879	0.0000	−3702.718	−3697.052
f_1	0.1483	0.2070	−4198.758	−4187.426
f_2	0.1451	0.2230	−4241.356	−4224.359
f_3	0.1420	0.2399	−4287.005	−4264.341
f_4	0.1416	0.2420	−4291.117	−4262.788

Table 3 Regression and log-covariance parameters

Parameters	Est.	S.D.
γ	0.0790	0.0019
α const.(intercept)	0.8923	0.1182
α (f_1 intercept)	0.1348	0.0289
α (f_2 intercept)	−0.0196	0.0253
Log range const.	−2.9991	0.9674
Log sill const.	−3.0172	0.6615
Log nugget const.	−3.0186	0.7935
Log range (f_1)	−4.9987	1.1058
Log sill (f_1)	−6.0144	1.3386
Log nugget (f_1)	−5.0390	1.1429
Log range (f_2)	−2.9964	0.8185
Log sill (f_2)	−5.9960	0.7205
Log nugget (f_2)	−7.0049	1.6869
ν Log range	−0.1906	0.0807
ν Log sill	−2.9692	0.0317
ν Log nugget (intercept)	−3.0628	0.0135

Table 4 Validation results

	MSE [a]	RMSE [b]	MAE [c]	MAPE [d]	BIAS [e]	rBIAS [f]	rMSEP [g]
Obs vs. WRF	4.3856	2.0942	1.6059	250.6468	1.3834	0.9119	0.9529
Obs vs. ST	0.7190	0.8480	0.6313	96.0073	−0.0717	−0.0473	0.4927

[a]Mean squared error
[b]Root mean squared error
[c]Mean absolute error
[d]Mean absolute percentage error
[e]Bias
[f]Relative bias
[g]Relative mean separation

forecasts, we think that the model is able to partially correct the bias and errors that affect the WRF model.

Figure 5 compares the predictions obtained by the cross-validation with the WRF predictions for each station. We can note that in all cases the spatiotemporal predictions are much better than the WRF, although for some stations (such as the numbers 2 and 9), the model slightly underestimates the true values of the wind speed. We think that this problem could be solved and the predictions could be further improved by increasing the number of weather stations for getting a more robust spatial correlation structure.

In Fig. 6 we represent the predictions for station 6-Los Maitenes for the month of March. In this case it is evident that the predictions using the spatiotemporal approach are much better than the WRF forecasts.

In Figs. 7a and 8a, we show the wind speed predictions for the day March 31 at 4:00 a.m. and 4:00 p.m., respectively. The predictions are characterized by small

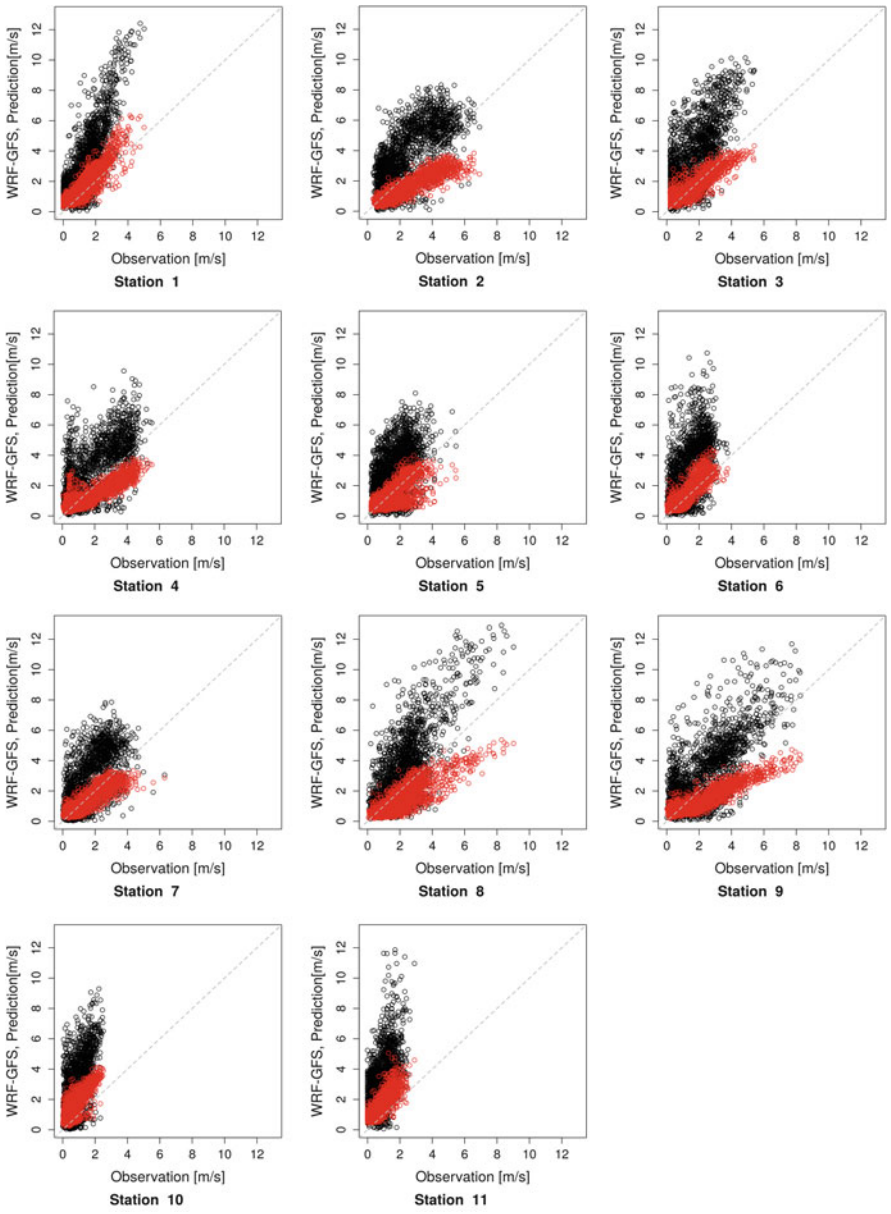

Fig. 5 Correlation plots obtained by the leave-one-out cross-validation between observations and WRF predictions (black points) and observations and ST predictions (red points) for each weather station

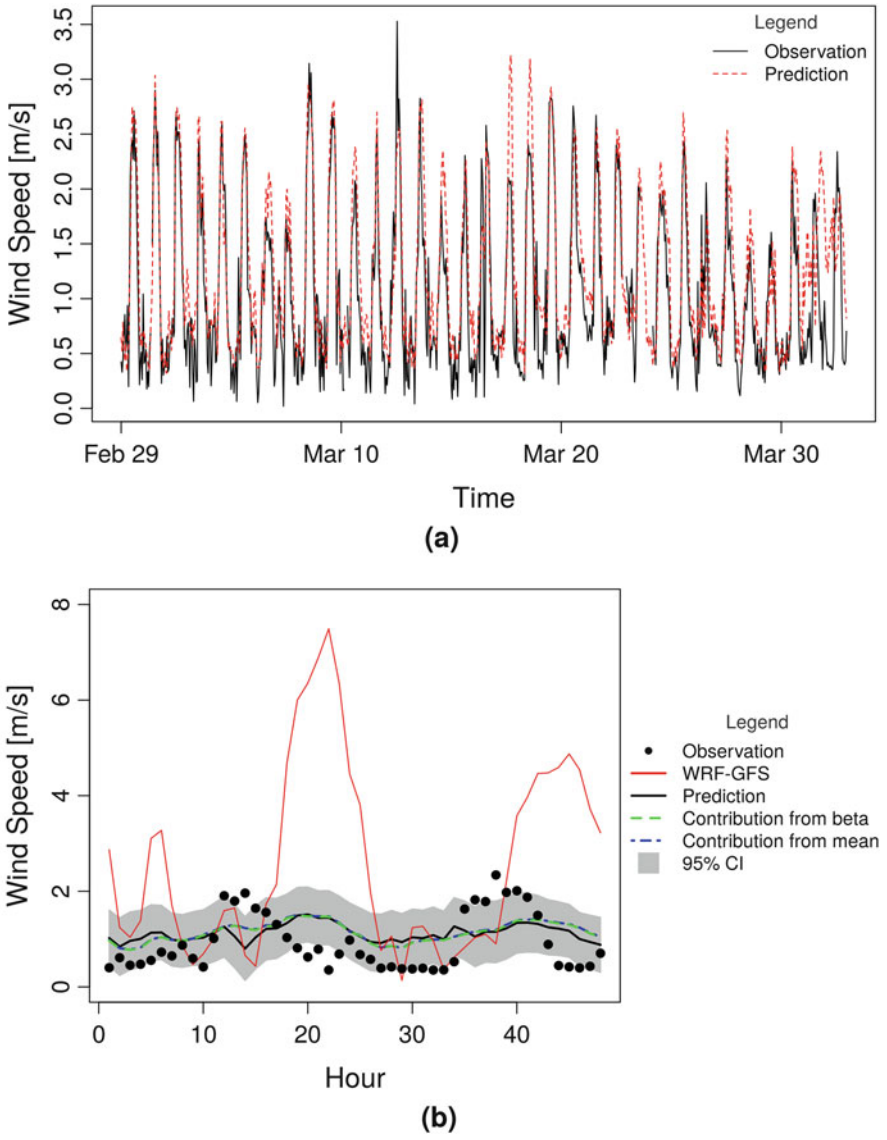

Fig. 6 Wind speed predictions for the station 6-Los Maitenes: (**a**) wind speed observations (black line) and ST predictions (red line) for the month of March, 2016; (**b**) ST predictions (black line) with the 95% interval confidence (gray area) and WRF predictions (red line) for the days March 31 and April 1. The green and blue dashed lines represent the mean (WRF data) and beta (basis function) contribution in the model

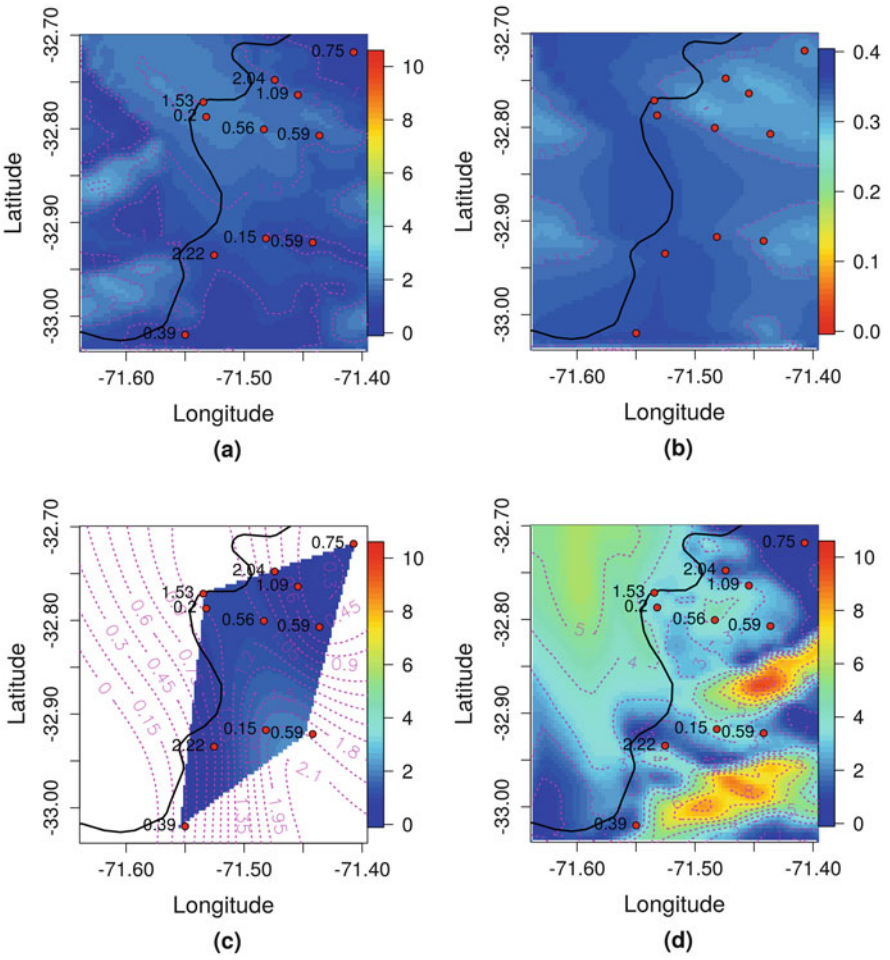

Fig. 7 Wind speed predictions with their contour lines for the day March 31, 4:00 a.m.: (**a**) ST prediction map and (**b**) its standard deviation. (**c**) Interpolated observed wind speed data using the thin spline regression and (**d**) WRF predictions. The red points indicate the monitoring stations with their observed values

standard predictions (Figs. 7b and 8b), and the range of the wind speed values are very close to the observations. In order to compare the predicted values with the observations, we use a thin plate spline (TPS) regression (see [19]) for interpolating the data where there are no monitoring stations (Figs. 7c and 8c). For implementing the TSP, we use the package `fields` [11] of the R software. If we compare the ST predictions with the WRF, we can note a significative difference. However, although the model seems very good for correcting the WRF predictions in the locations close to the weather stations, we cannot evaluate its performance in the sea where there are no weather stations and the meteorological conditions could be different.

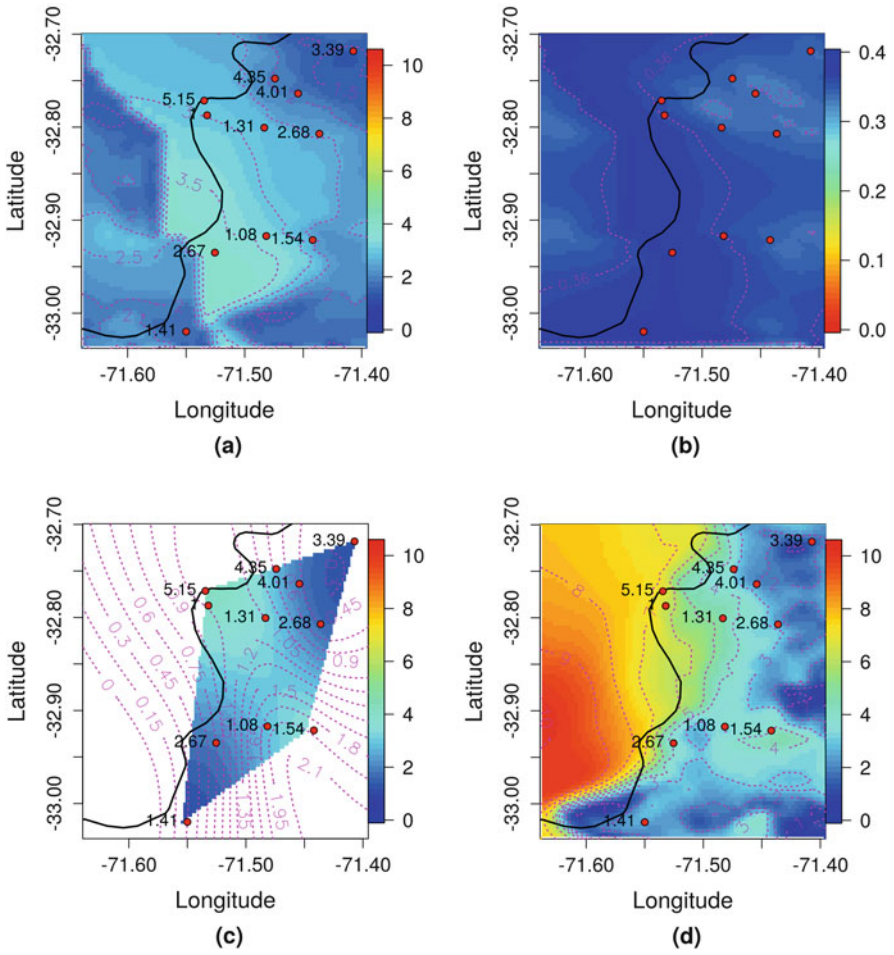

Fig. 8 Wind speed predictions with their contour lines for the day March 31, 4:00 a.m.: (**a**) ST prediction map and (**b**) its standard deviation. (**c**) Interpolated observed wind speed data using the thin spline regression and (**d**) WRF predictions. The red points indicate the monitoring stations with their observed values

5 Conclusions and Further Developments

In this work we use the spatiotemporal model proposed by Lindström et al. [10] for correcting the WRF predictions. The estimation of the model and the cross-validation study show that the proposed methodology is able to improve the predictions provided by the WRF model by reducing the bias and the random errors. However, it is very difficult to assess the performance of the method where there are not monitoring stations such as in the sea. We think that better results could be achieved if more weather stations were available on the entire area of study. Also, we

are going to improve the spatiotemporal model by considering different covariance structures and by including local information (such as land use or others).

Acknowledgements This work is partially supported by the Interdisciplinary Center for Atmospheric and Astro Statistical Studies.

References

1. Bergen S, Lindström J (2013) Comprehensive tutorial for the spatio-temporal R-package. Retrieved from https://mran.microsoft.com/web/packages/SpatioTemporal/vignettes/ST_tutorial.pdf
2. Carlin BP, Clark JS, Gelfand AE (2006) Hierarchical modelling for the environmental sciences: statistical methods and applications. Oxford University Press, Inc., New York
3. Casquilho-Resende CM, Le ND, Zidek JV (2016) Spatio-temporal modelling of temperature fields in the pacific northwest. ArXiv e-prints
4. Chen F, Dudhia J (2001) Coupling an advanced land surface–hydrology model with the penn state–ncar mm5 modeling system. Part I: model implementation and sensitivity. Mon Weather Rev 129(4):569–585
5. Craimile PF, Guttorp P (2011) Space-time modelling of trends in temperature series. J Time Ser Anal 32:378–395
6. Fassò A, Cameletti M, Nicolis O (2007) Air quality monitoring using heterogeneous networks. Environmetrics 18(3):245–264
7. Fuentes M, Guttorp P, Sampson PD (2006) Using transforms to analyze space-time processes. In: Finkenstadt B, Held L, Isham V (eds) Statistical methods for spatio-temporal systems. CRC Chapman and Hall, Boca Raton
8. Hong S-Y, Dudhia J, Chen S-H (2004) A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. Mon Weather Rev 132(1):103–120
9. Iacono MJ, Delamere JS, Mlawer EJ, Shephard MW, Clough SA, Collins WD (2008) Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models. J Geophys Res 113:D13103
10. Lindström J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson TV, Sheppard L (2014) A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ Ecol Stat 21(3):411–433
11. Nychka D, Furrer R, Paige J, Sain S (2017) fields: Tools for spatial data. R package version 9.6. http://doi.org/10.5065/D6W957CT
12. Sahu SK, Bakar K (2012) A comparison of Bayesian models for daily ozone concentration levels. Stat Methodol 9:144–157
13. Sahu SK, Nicolis O (2009) An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network. Environmetrics 20(8):943–961
14. Sahu S, Gelfand A, Holland D (2007) High-resolution space-time ozone modeling for assessing trends. J Am Stat Assoc 102(480):1221–1234
15. Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, Powers JG (2008) A description of the advanced research WRF version 3. NCAR Technical note -475+STR
16. Sukoriansky S, Galperin B, Perov V (2005) Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. Bound-Layer Meteorol 117(2):231–257
17. Tastu J, Pinson P, Trombe PJ, Madsen H (2014) Probabilistic forecasts of wind power generation accounting for geographically dispersed information. IEEE Trans Smart Grid 5(1):480–489

18. The GFS Atmospheric Model (2004) Office note. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service National Centers for Environmental Prediction
19. Wood SN (2003) Thin plate regression splines. J R Stat Soc Ser B Stat Methodol 65(1):95–114

Spatiotemporal Precipitation Variability Modeling in the Blue Nile Basin: 1998–2016

Yasmine M. Abdelfattah, Abdel H. El-Shaarawi, and Hala Abou-Ali

Keywords Blue Nile basin · Precipitation · El Niño-Southern Oscillation · Empirical orthogonal function · Dynamic harmonic regression · Anthropogenic signals

1 Introduction

The Nile basin is an ecological system under severe tension. Any future variations in the river flow will add to an already existing pressure on basin inhabitants and will elevate water stress. Precipitation over the Ethiopian highlands at the Greater Horn of Africa in East Africa is feeding the headwaters of the Blue Nile basin (BNB), which supplies approximately 60% of the main flow of the mighty Nile [6, 44]. The Grand Ethiopian Renaissance Dam (GERD), currently under construction on the Ethiopian-Sudanese border, is agitating the water conflict because of its negative influence on downstream countries' water share especially Egypt. Consequently, it is essential to review the effect of climate change and climate variability on BNB precipitation, as Nile water resources management and planning are necessary to its riparian countries [15, 23]. This paper, therefore, assesses how much of the seen decline in BNB precipitation is due to natural climate variability or to human-induced effects. Identification of the factors controlling precipitation trends is vital

Y. M. Abdelfattah (✉)
The British University in Egypt, Cairo, Egypt
e-mail: yasmine.mohamed@bue.edu.eg

A. H. El-Shaarawi
Canadian National Water Research Institute, Burlington, ON, Canada

Cairo University, Giza, Egypt
e-mail: elshaarawi@feps.edu.eg

H. Abou-Ali
Cairo University and Economic Research Forum (ERF), Giza, Egypt
e-mail: habouali@feps.edu.eg

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_6

to the understanding of contemporary precipitation variability and to the prediction of future precipitation changes in BNB.

The East African monsoon is a key atmospheric phenomenon leading the precipitation regime in the Greater Horn of Africa. There are considerable indications proposing that the monsoon of East Africa is strongly connected to the spatiotemporal changes in global sea surface temperatures (SSTs). The SST exhibits intense variabilities due to the heterogeneity of oceanic hydrographic features affecting it. An example of SSTs are El *Niño*-Southern Oscillation (ENSO) and the Indian summer monsoon, which are significant influencers of large-scale modes of natural climate variability. There is tremendous year-to-year differences in timing and magnitude of the Ethiopian precipitation. The seasonal north-south intertropical convergence zone (ITCZ) movement is the most known driver of underlying physical mechanisms of precipitation in the headwaters of the BN [6, 25]. ITCZ is the region where the exchange winds from the two hemispheres converge. Berhane et al. [5] advocate that in the summer months [June–September (JJAS)], the ITCZ brings lots of humid air from the Indian Ocean in the south, the Gulf of Guinea and the equatorial Atlantic Ocean across the Congo, and the Sahel in the west as well as possibly from the Mediterranean Sea in the north and the Red and Arabian Sea in the east. Then, ITCZ moves southward and dry conditions start to take place from October through May. Accordingly, there are two rainy seasons: the pre-monsoon season called the Belg in Ethiopia (short rainy season from March to May) which paves the way to the Kirmet (rainy season from June to September) followed by Bega (dry season from October to February). ITCZ fuels Kirmet season by precipitation events. Uncertainty surrounding climate change projections is causing the prediction of precipitation variability in the BNB to be rather challenging especially for water resources management and planning for society and for regional economies.

There is an ongoing literature studying the teleconnections between large-scale ocean-atmosphere interactions and precipitation in the River Nile basin as a whole and BNB precipitation in particular with the intention of predicting Nile flow. For instance, Siam and Eltahir [45] assessed the flow and rainfall patterns of the Upper BNB, Sobat and Atbara, and projected increases in the inter-annual variability of the Nile River flow as a consequence of climate change. Jury [22] examined the determinants of southeast Ethiopian seasonal rainfall in the September–November season from year 1980 to 2010 using satellite observations and elevation models. Siam and Eltahir [44] estimated that Pacific and Indian oceans SSTs indices can jointly describe around 84% of the Nile flow inter-annual variation. Berhane et al. [5] found links between Upper BNB boreal summer precipitations and large-scale atmospheric and global SST field, a strong relationship between the Indian monsoon and the eastern stimuli-ENSO in September. Elsanabary and Gan [14] adopted wavelet principal component analysis to predict rainfall at the Upper BNB and identified a correlation between the first wavelet principal component of the June to September seasonal rainfall and selected sectors of the Atlantic, Indian, and Pacific Oceans SSTs. Block and Rajagopalan [6] proposed an ensemble forecast framework for analyzing Kirmet precipitation for the Upper BNB. It is constructed using a

nonparametric approach grounded on local polynomial regression. They concluded that ENSO phenomenon is the foremost determinant of the inter-annual variability of seasonal precipitation in the Upper BNB. Remarkably, almost all precipitation studies on the Blue Nile have concentrated on Kirmet precipitation as the most relevant response variable. Given the large number of large-scale drivers involved in describing and predicting precipitation in the BNB, an extra thorough temporal modeling is needed. Therefore, this paper studies teleconnections and potential drivers of precipitation variability at different temporal scales (inter-annual, intra-annual, inter-seasonal, and intra-seasonal) over the whole of Ethiopia, not only Upper BNB as the rest of the other studies.

In recent years, investigating the effect of climate variability on hydrometeorology involved matrix methods for statistical analysis of structures in large datasets (as considered in this paper). The methods ranged from simple correlation analysis and multiple linear regression [13, 16, 18] to linear multivariate methods such as empirical orthogonal function (EOF) analysis (see, e.g., [2, 4, 12, 14, 25, 37, 40, 52, 57]). For example, Zeleke and Damtie [57] studied Upper BNB rainfall seasonal variability and annual cycle from year 1979 to 2014 using rotated EOFs and wavelet analysis via station and satellite data. Similarly, linear multivariate methods such as EOF analysis are now frequently used by regional attribution studies to study the impact of natural climate variability on precipitation in other regions such as Australia [59], the United States [50], Hawaii [17], Alaska, [27], India [31], Peru [41], Oman [48], China [51, 54, 55], Zimbabwe [30], and West Africa [1, 47]. As for human-induced climate change, also known as the anthropogenic effect, its impact on global precipitation has been identified [35, 58]. In climate science, including linear trend term in linear regression model is most commonly used to quantify the change in a climate variable (e.g., temperature and precipitation) over time [3]. These climate variables exhibit linear trends in response to climate change. Regional attribution studies have not perceived an anthropogenic sign in precipitation trends beyond natural forcing [42]. Alternatively, Frazier et al. [17] detected a significant linear trend term in the multiple linear regression models only in the dry season in Hawaii. Therefore, estimating the effect of anthropogenic signal on precipitation trends is problematic, as a lot of studies try to eliminate the effect of natural forcing and test if the trend is significant [19].

From the studies above, we find a research gap regarding BNB precipitation patterns and its spatiotemporal variability. In sum, the region exhibits high spatial precipitation variability and intensified extreme climatic conditions giving importance in applying spatiotemporal techniques to BNB precipitation magnitude. This paper aims at measuring the total influence of natural climate changeability to precipitation variations over the time span of the study and evaluates whether an anthropogenic effect can be distinguished from natural forcing. To reach this end, the paper uses an empirical orthogonal function (EOF) analysis to define the leading precipitation structures and their amplitude variability modes in the BNB. Then, the relative impact of natural forcing and human-induced effect on the leading precipitation amplitudes are estimated by dynamic harmonic regression (DHR). This is the first precipitation study in BNB to use such type of regression which

takes into account BNB precipitation periodic modes. By using these techniques, the current paper provides a better understanding of the driving forces behind the existing precipitation trends in BNB. This information can aid Nile countries' decision-makers in mitigating possible catastrophes. The paper embarks on datasets description and potential predictors of precipitation. Then, the adopted method of EOFs and DHR models are presented, followed by results of the study. The final section extends the study conclusions.

2 Data

2.1 Tropical Precipitation Measuring Mission (TRMM)

The precipitation data employed in this study are obtained from the Tropical Precipitation Measuring Mission (TRMM) Multisatellite Precipitation Analysis 3B42 V7 [21]. TRMM 3B42 V7 is a gauge-calibrated satellite product which provides estimates of precipitation magnitude [7], and it is validated using in situ data in West Africa. No bias is revealed with the root mean square error (RMSE) in the order of 0.7 and 0.9 mm/day for the seasonal and August precipitation, respectively [34]. The coverage of the TRMM 3B42 V7 is global (i.e., 50°S and 50°N), and the estimates of monthly precipitation with a spatial resolution of $0.25° \times 0.25°$ are provided [32]. Owing to its high spatial resolution, TRMM 3B43 is utilized to estimate seasonal and monthly precipitation in the BNB. The study covers monthly means over the period 1998–2016. TRMM 3B42 V7 dataset is accessible at the National Aerospace and Space Administration (NASA) Goddard Space Flight Center (GSFC) website.[1] The lack of in situ measurements restricts the robustness and large-scale monitoring possibility of major hydrological variables in the BNB. Routine readings of these hydrological quantities are unavailable because of the limited number of gauge stations and deteriorating condition of available facilities. Most of the date is unobtainable to the public or to the relevant research institutions because of government policies. Incomplete data records affect the proper assessment of the region's hydrological conditions. Nevertheless, the existing satellite climate data allows for the examination of hydro-climatic conditions.

2.2 Large-Scale Atmospheric and Climate Indices

Because of the importance of BNB precipitation for the region, the mechanism behind the modes of large-scale climate variability and its teleconnection with BNB precipitation has been studied by a number of researchers. Sir Gilbert Walker, in

[1] https://disc.gsfc.nasa.gov/datasets/TRMM_3B42_V7/summary.

1910, was the first to find a positive correlation between the summer Indian monsoon rainfall and the Nile floods [9]. Several studies have shown that ENSO and its related indices are the most connected large-scale climate variability with inter-annual and seasonal BNB precipitation patterns (e.g., [5, 39, 44–46]). El *Niño*-Southern Oscillation (ENSO) comprises of El *Niño*, which is a periodic fluctuation in sea surface temperature (SST) (e.g., every 2–7 years), and Southern Oscillation which is air pressure of the overlying atmosphere across the equatorial Pacific Ocean. El *Niño* is interconnected with dry waves, while it's opposite event, La *Niña*, is associated with high precipitation. Following the general approach, precipitation is modeled as a function of climate indices and atmospheric fields as predictors. Previous studies have guided our indices selection. The indices are retrieved from the National Oceanic and Atmospheric Administration/Climate Prediction Center (NOAA/CPC).[2] A range of indices considered to measure SSTs anomalies are as follows:

1. East Central Tropical Pacific SST "*Niño* 3.4" (5N-5S, 170W-120W).
2. Southern Oscillation Index "SOI" is calculated as a standardized monthly-mean sea level pressure alteration among Darwin and Tahiti.
3. Atlantic Meridional Mode "AMM" (21S-32N, 74W-15E) described as the tropical Atlantic basin leading maximum covariance analysis mode.

3 Methods

Precipitation and large-scale ocean-atmosphere interactions are characterized by complexity, and their spatiotemporal interactions cannot be disregarded. In this context, empirical orthogonal functions (EOFs) are implemented to understand the precipitation dynamical/physical behavior (in terms of space-time covariance/correlation structure). This will allow us to obtain a considerably smaller number of renowned modes of variability. Then, dynamic harmonic regression (DHR) is applied to estimate the partial influence of selected global SSTs as well as linear time trend on the BNB precipitation time series amplitude.

3.1 Space-Time Empirical Orthogonal Function Analysis

Empirical orthogonal functions (EOFs) are extensively used and important multivariate statistical analysis in climate and atmospheric sciences including meteorology, climatology, and oceanography. Lorenz [29] was the first study to introduce EOFs into meteorological literature. The early review of EOFs mathematical

[2] https://www.esrl.noaa.gov/psd/data/climateindices/list.

derivation using eigenvector representation is by Kutzbach [26]. EOFs is a geophysicist's term for the eigenvectors in the classical eigenvalue decomposition analysis of covariance/correlation matrix. It can also be obtained by the singular value decomposition algorithm. EOF analysis is basically principal component (PC) analysis and Karhunen-Loeve (KL) in its spatially continuous representation (e.g., [28]). Thus, EOF/PC analysis is a dimension reduction (spatially and/or temporally) technique which is an important part of spatiotemporal modeling in large spatiotemporal datasets like high-resolution satellite datasets such as TRMM dataset. EOFs/PCs is an exploratory (i.e., non-model orientated) method. We focus on the discrete setting, where the dimensionality of spatiotemporal dataset is reduced by obtaining the foremost dominant modes of variability which are the EOFs. These modes explain most of the observed variance from a spatiotemporal precipitation field via a linear combination of the original variables. The first few EOFs describe most of the original dependencies between variables while at the same time reducing noise. Therefore, EOFs have multiple uses: dimension reduction and patterns extraction [4, 11, 52].

BNB precipitation data have three dimensions—one dimension in time and two dimensions in space—it is a random field F. The latter is a function of time t, longitude ϕ, and latitude θ. The latitudes θ_j are the horizontal coordinates, $j = 1, \ldots, p_1$, and the longitudes ϕ_k are the vertical coordinates, $k = 1, \ldots, p_2$. Thus, the total number of grid points is $p = p_1 p_2$ and the random field is read as follows:

$$F_{ijk} = F(t_i, \theta_j, \phi_k) \tag{1}$$

with $1 \leq i \leq n$, $1 \leq j \leq p_1$, and $1 \leq k \leq p_2$. The random field F is transformed into a data matrix Z where longitude and latitude are combined together to represent a space-time field Z(t, s). Thus, let $Z_t = \left[\{(Z(s_1; t), \ldots, Z(s_p; t))'\}; (s, t) \in D \times T\right]$ denotes BNB monthly precipitation observation at a discrete time t_i $(i = 1, \ldots, n)$ and grid point s $(s = 1, \ldots, p)$ where $D \subseteq \Re^2$and $T \subseteq \Re_+$. It is depicted by the data matrix:

$$Z = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1p} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & z_{np} \end{bmatrix} \tag{2}$$

In order to calculate EOFs, the empirical lag τ spatial covariance matrix has to be obtained first, and it is represented by

$$\hat{C}_Z^{\tau} = \frac{1}{n - \tau} \sum_{t=\tau+1}^{n} (Z_t - \hat{\mu}_Z)(Z_{t-\tau} - \hat{\mu}_Z)' \tag{3}$$

where lag $\tau = (0, 1, \ldots, n-1)$ and the time average $\hat{\mu}_Z$ is equal to

$$\hat{\mu}_Z = \frac{1}{n} \sum_{t=1}^{n} Z_t \tag{4}$$

EOFs are obtained by using spectral decomposition of the empirical lag zero covariance matrix as follows:

$$\hat{C}_Z^{\tau} = \Psi \Lambda \Psi^{'} \tag{5}$$

where Ψ is eigenvectors matrix and Λ is the diagonal matrix of eigenvalues, specifically, $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$ where $r \leq min(n, p)$ which is the rank of Z. The kth eigenvectors ψ_k is the kth column of Ψ which is the kth empirical orthogonal function (EOF), is analogous to the PC loadings. Thus, the PCs are also called EOFs expansion coefficients, EOFs amplitudes, PCs time series, and PCs scores. Here, EOFs and PCs terminologies stand for the spatial and temporal patterns, respectively. Each EOF represents a spatial map. The time-varying amplitude function which is PC time series is calculated by the projection method using the formula:

$$a_t(k) = \Psi_k^{'} Z_t, \qquad k = 1, \ldots, p. \tag{6}$$

The EOFs eigenvectors are orthogonal, $\Psi^{'}\Psi = 1$. This orthogonality constraint enables us to maximize $var(a_t(1))$ then $var(a_t(2))$ to be maximized, etc. as in the classical PCA where $var(a_t(k)) = \lambda_k,\ k = 1, \ldots, p$. The eigenvalues are normally written in descending order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. Usually, the variance accounted for in percentage is written as follows:

$$\frac{100\lambda_k}{\sum_{k=1}^{p} \lambda_k}\% \tag{7}$$

Finally, in order to test the significance of the EOFs modes, North et al. [36] rule of thumb is used to calculate the uncertainty of a given eigenvalue as:

$$\Delta\lambda_k \approx \lambda_k \sqrt{\frac{2}{n^*}} \tag{8}$$

$$\Delta\psi_k \approx \frac{\Delta\lambda_k}{\lambda_j - \lambda_k} \psi_j \tag{9}$$

where λ_j is the nearest eigenvalue to λ_k and n^* is the independent sampling grid points. EOFs have a shortcoming of not being able to explain physical patterns, and they are used to obtain simple structures. This is due to the fact that the physical processes are known to be non-orthogonal in nature. Rotated EOFs (REOFs) solve

this problem of physical interpretation by further rotating the leading EOFs. In this study, Varimax criterion is used—the most well-known rotation algorithm. Varimax rotation secures physical variability patterns as well as the orthogonality of EOFs [39, 40]. The normalized Varimax criterion was developed by Kaiser [24] as follows

$$max \left\{ \frac{1}{n} \sum_{j=1}^{m} \left\{ \sum_{i=1}^{n} (\psi_{ij})^4 - \frac{1}{n} \left(\sum_{i=1}^{n} (\psi_{ij})^2 \right)^2 \right\} \right\} \tag{10}$$

where m is the number of chosen rotated EOFs. The Varimax rotation leads to maximizing the leading EOFs, so they are close to one, and minimizing the rest of EOFs, so they are close to zero. This process yields to an extremely localized pattern. Therefore, the different REOFs modes are spatially orthogonal, and the corresponding RPCs (Rotated Principal Components) are temporally uncorrelated.

3.2 Dynamic Harmonic Regression (DHR)

In the 1980s, dynamic harmonic regression (DHR) model was first introduced by Young and fellow workers [56] to deal with time series data that has periodic or quasi-periodic behavior. DHR decomposes the BNB precipitation amplitudes obtained in Sect. 3.1 and express it mathematically as an algebraic sum of harmonic components. These components are represented as sine and cosine waves of precipitation curve. The two waves are representing a single cosine wave. Each harmonic component has its wave size (amplitude) and the wave offset (wave phase angle) [8, 20]. The first harmonic component represents a curve with frequency one (i.e., one maximum and one minimum). This accounts for annual variation of the precipitation curve. The difference between the annual precipitation maximum and minimum is represented by its amplitude. The second harmonic represents a curve with frequency two (i.e., two maxima and minima). This harmonic describes any semiannual variation in precipitation curve. Likewise, the third harmonic represents 4-month variation and so on [43]. Thus, the purpose of using DHR in our study is to detect hidden periodicities in the BNB precipitation amplitudes which is well known for exhibiting strongly periodic behavior. Harmonic components represent the process of BNB precipitation seasonal variations throughout the year.

DHR is basically a multivariate time series technique. As, DHR is a regression model with Fourier components and a time series residuals. This is due to the fact that estimating regression with time series variables such as precipitation and large-scale atmospheric and climate indices will impose the model's residuals to follow a time series process too. DHR may be a better choice than multiple linear regression (MLR) in this case. As regression with time series residuals follow an autoregressive moving average (ARMA) process which are more likely to violate the ordinary least squares (OLS) independent errors assumption if MLR is estimated instead. This leads to incorrect standard errors, confidence interval, and tests as well. The

residuals of DHR are modeled as an ARMA process which is estimated by Box and Jenkins methodology. This kind of models accounts for the serial correlation of the residuals. The appropriate ARMA process relies on the autocorrelation function (ACF) and partial autocorrelation function (PACF). DHR is defined as follows:

$$y_t = \beta_0 + \sum_{k=1}^{K} [\alpha_k \sin(2\pi kt/m) + \gamma_k \cos(2\pi kt/m)] + \beta_1 x_{t,1} + \ldots + \beta_s x_{t,s} + \eta_t \tag{11}$$

$$\varphi_p(B)\eta_t = \theta_q(B)\varepsilon_t \tag{12}$$

$$AR(p) : \varphi_p(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p \tag{13}$$

$$MA(q) : \theta_q(B) = 1 + \theta_1 B + \ldots + \theta_q B^q \tag{14}$$

where a response variable $y_t = y_1, \ldots, y_n$ which is EOF precipitation amplitudes and several predictors (x_s), which are global climate indices and trend, $\beta_0, \beta_1, \ldots, \beta_s$ are $(s + 1)$ parameters, α_k and γ_k are the amplitudes of the sine and cosine coefficients of the kth harmonic Fourier mode, k is the total number of harmonics and is chosen based on the lowest Akaike information criterion (AIC) value, m is the seasonal period, B stands for the backward shift operator, $B^a y_t = y_{t-a}$, η_t are residuals following ARMA process, ε_t are white noise errors, and n is the number of observations. The stationarity of time series is one of the time series models crucial assumption. The latter entails that the series exhibits constant mean, variance, and autocorrelation function over the time period. Augmented Dickey-Fuller (ADF) test is used to test stationarity of the resulting residuals. Finally, the best DHR model is selected based on some criteria which are the largest value for the likelihoods and the lowest AIC value.

4 Results

In order to depict the whole picture of BNB precipitation, first the spatial distribution at different temporal scales (inter-annual, intra-annual, inter-seasonal, intra-seasonal) of TRMM precipitation dataset are inspected all over Ethiopia. Then, EOF/PC modes and DHR are estimated. This will give us a chance to have an in-depth knowledge of contemporary precipitation modes of variability together with the projected impact of climate change on BNB precipitation.

4.1 Spatial and Temporal Variation of BNB Precipitation Patterns

The mapping of inter-annual precipitation variation is among the most essential tools in climate research. Thereby, yearly climatology maps of precipitation over the study years 1998–2016 is used to quantify year-to-year changes in the yearly climatological mean precipitation (Fig. 1). Ethiopia has an intense tendency toward exhibiting drought and flooding episodes that dominate the inter-annual variation. Figure 1 shows wave of drought in the following years: 1999, 2000, 2002, as well as the period from 2008 to 2011. These drought waves are the same ones which Viste et al. [49] have found in their study. In their paper, they investigated drought all over Ethiopia from year 1971 to 2011, and they concluded that year 1984 was the driest year followed by year 2009, which is very obvious in Fig. 1. In 2016, deadly flooding in Ethiopia made headlines in news media [33], and this is clearly shown as well in the mapped climatological mean BNB precipitation.

BNB precipitation has three dominant intra-annual variations which constitute the three seasons (Kirmet, Belg, and Bega) as discussed in Sect. 1. Thereby, monthly TRMM precipitation magnitude distribution supports the seasonality of BNB precipitation and the existence of three seasons. Moreover, TRMM precipitation in the BNB has the same monthly spatial distribution in the Ethiopian precipitation literature [10]. In March, precipitation starts in the southwest of Ethiopia with an average precipitation of 45 mm. Precipitation moves toward western region in

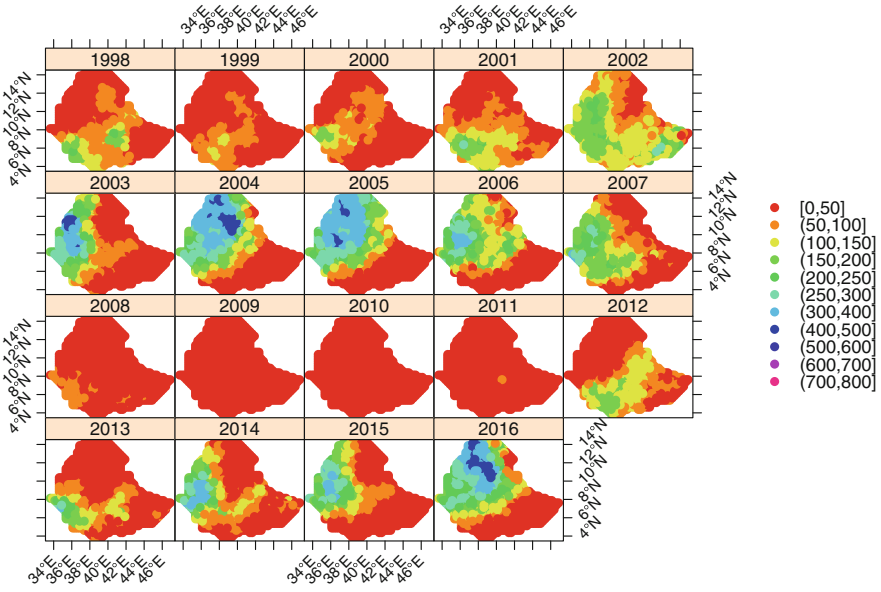

Fig. 1 Blue Nile Basin mean annual precipitation (mm) from year 1998 to 2016

May, which is the end of Belg known as the short rainy season with approximately 113 mm. In June, the Kirmet rainy season accelerates precipitation magnitude in the western region. The highest average precipitation magnitude in BNB is 155 mm in August, where precipitation covers the whole of the basin except for the southern part of Ethiopia. Kirmet ends in September where precipitation declines and reaches the same average as in the end of the Belg. The Bega dry season starts in October, with low precipitation declining and moving toward the southern region of Ethiopia. Thereafter, it starts the same cycle again from the beginning. Finally, there is barely any precipitation over all of Ethiopia from November to February. Accordingly, maximum precipitation magnitude occurs between June and September, while the lowest magnitude occurs between November and February in the BNB. Seasonal-to-inter-annual precipitation maps in Figs. 2, 3, and 4 display the three different seasons, Bega, Belg, and Kiremt, respectively, for each year of the study. These maps give initial insights about the inter-annual and intra-annual BNB precipitation

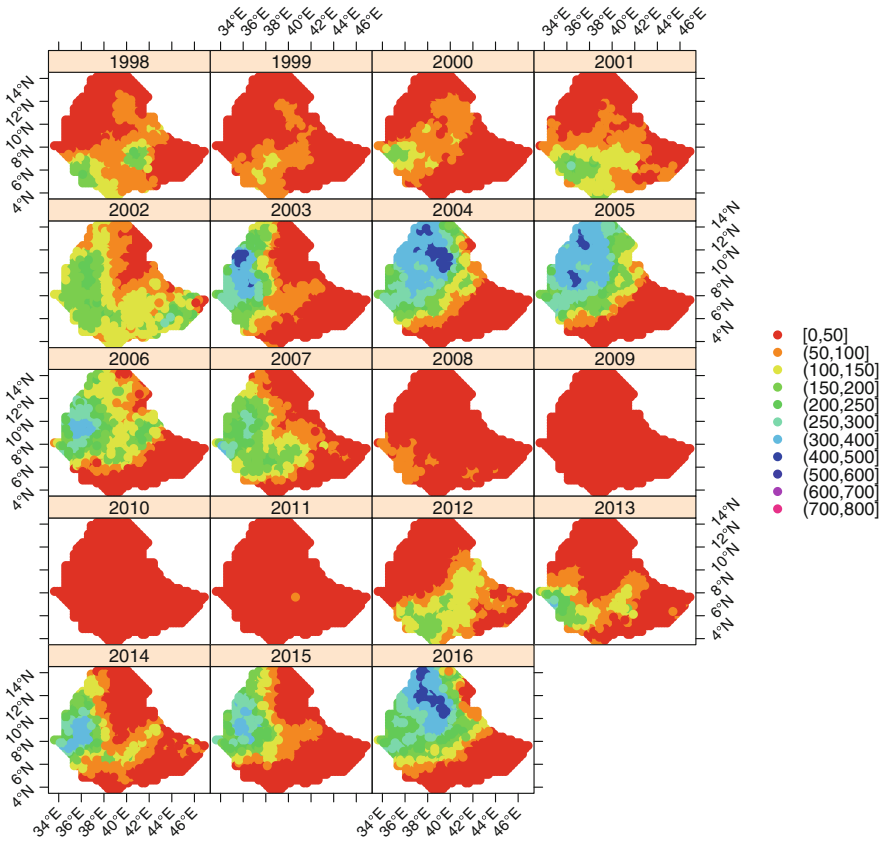

Fig. 2 Seasonal-to-inter-annual BNB precipitation (mm) from 1998 to 2016: Bega dry season (October–February)

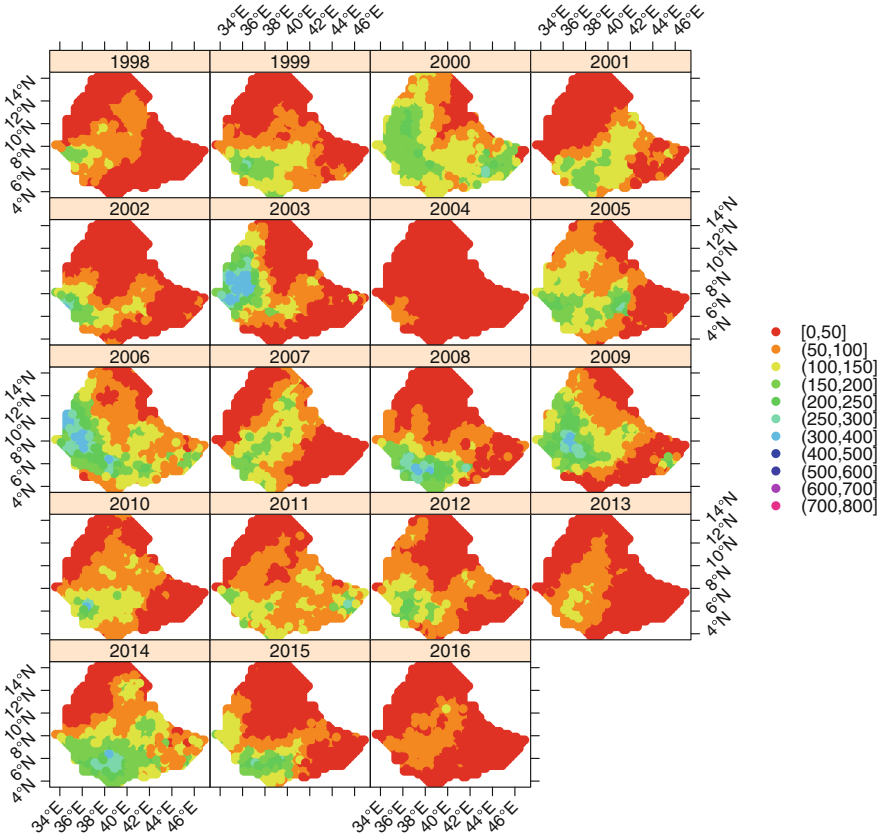

Fig. 3 Seasonal-to-inter-annual BNB precipitation (mm) from 1998 to 2016: Belg short rainy season (March–May)

variations to help detecting if there is seasonal variations over BNB. A shift in precipitation distribution over the three different seasons has been observed over the years. The drought wave during the period 2008–2011 seems to be due to extreme dryness of both Belg and Bega seasons. Viste et al. [49] and Williams et al. [53] studied the reasons that contributed to the drought in Ethiopia during this period. Both studies agreed that the drought was due to repetition of a dry Belg season. Also, Viste et al. [49] added that the year-to-year variation of the Belg precipitation is greater than the Kirmet ones and this affects resource-poor farms in Ethiopia. All of the studies on Ethiopia's and BNB precipitation focus only on Kirmet season, and only few studies considered the impact of Belg season as well, but none have included the Bega. Here, we study the effect of the three seasons simultaneously to discover this seasonal imbalance. Another example of this observed climate change effect is in year 2004 and 2016, where extreme dry Belg and wet Bega seasons are detected. This repeated phenomenon in year 2004 and 2016 contrasts the normal

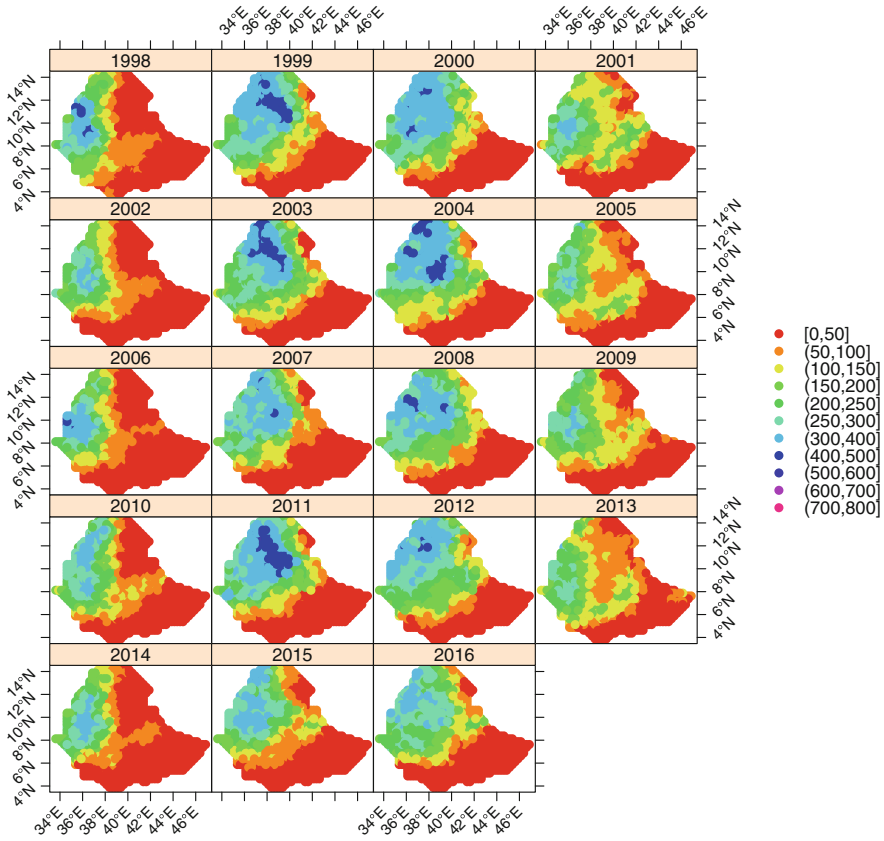

Fig. 4 Seasonal-to-inter-annual BNB precipitation (mm) from 1998 to 2016: Kirmet rainy season (June–September)

seasonal precipitation climatology maps of Ethiopia. This shift in BNB precipitation would lead to 50% increase in the twenty-first-century Nile flow as compared to the twentieth century as concluded by Siam and Eltahir [45]. This is a result of Nile-ENSO teleconnection and the frequent occurrence of El *Niño* and La *Niña* phenomena.

4.2 Spatiotemporal Patterns of Precipitation by EOFs

In this paper, we have used the EOF/PC analysis. First, the BNB precipitation anomaly is calculated according to Onyutha [38], to deseasonalize the monthly

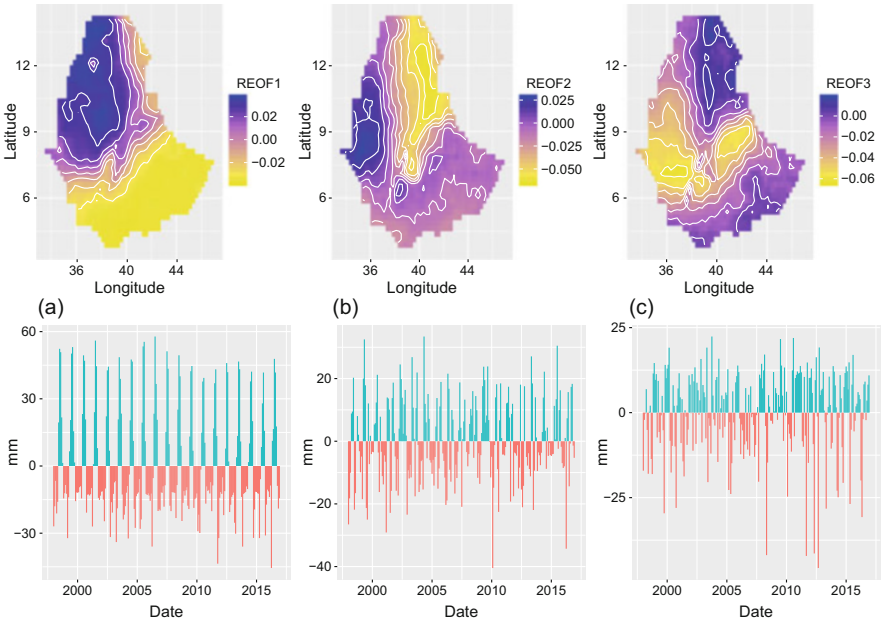

Fig. 5 REOFs and the corresponding RPCs time-varying amplitude functions of monthly precipitation anomalies. (**a**) Ethiopian northern region's precipitation pattern, (**b**) Ethiopian western region's precipitation pattern and, (**c**) Ethiopian eastern and southern region's precipitation pattern

series.[3] Then, the leading BNB precipitation structures and their amplitude mode of variability are extracted from the decomposition of the correlation matrix $\hat{C}_Z^{\tau}$ depicted in Eq. (5). North et al. [36] rule of thumb imposed a restriction on retaining only the first three EOFs modes. The first three EOFs/PCs are extracted and rotated using Varimax criterion. The first 100 EOFs modes account for approximately 99% of the BNB precipitation anomaly variation, while the first three EOFs account for over 74% of the BNB precipitation anomaly variation. The BNB precipitation spatial pattern for the three rotated EOFs was plotted together with corresponding rotated PC time amplitudes in Fig. 5. The figure defines spatially distinct regions. The spatial structure is simplified by rotation through separating regions with similar temporal variation. The REOF1 precipitation mode has the highest portion of variance; it accounts for 47% of the total spatial precipitation variance. The precipitation anomalies in the REOF1 mode reflect higher loading

[3]The steps to remove the seasonality from monthly precipitation time series are as follows: compute the mean for each month, and then repeat the monthly mean values for every year of the dataset, followed by deseasonalize the data by subtracting the monthly mean, the seasonal component of each month, from the original series. This renders precipitation residual which can be utilized in the analysis.

over the Ethiopian northern region. Positive time coefficients indicate that the precipitation has increased in the northern area, while negative coefficients indicate that the precipitation has decreased. Positive values are observed in the middle of each year during Kirmet season. The REOF1 reflects the existence of intra-seasonal variation as the different months of Kirmet season have different positive values. The REOF1 also reflects clear inter-annual variation, and this is also clear when we compare the time coefficients of the first REOF1 across the years of study. The largest positive value is in the year 2006 associated with the flooding as mentioned before. Negative loadings corresponds to the Belg and Bega seasons and the dry waves in 2011 and 2016. RPC1 also shows that there are years where there are waves of extremes (drought and flooding) in the same year such as in year 2006 and year 2016. These results were consistent with that shown in Sect. 4.1. The REOF2 mode depicts the precipitation mode of variability in the Ethiopian western region, which is represented by REOFs and the corresponding RPCs time-varying amplitude functions in Fig. 5b. It explains 13% of the total variance where precipitation starts to cover during the Belg season. The RPC2 time amplitudes take smaller values than the RPC1 ones. The worst dry waves were in 2011 as well which has the largest negative value of RPC2. The precipitation variations in the REOF3 mode account for 12% of the total variance covering the eastern and southern Ethiopia where there are few or barely any precipitation in the southern region all year, which is represented by REOFs and the corresponding RPCs time-varying amplitude functions in Fig. 5c. The RPC3 scores show greater inter-annual variation than they did for inter-seasonal ones, a trend of negative values. Still, year 2011 was the worst dry wave, covering Ethiopia.

4.3 Dynamic Harmonic Regression with Global Atmospheric and Climate Indices

The dynamic harmonic regression is used to model the three RPC time amplitudes over as a function of atmospheric, oceanic, and anthropologic predictors using standard Box and Jenkins methodology. The dynamic behavior of the stochastic trend and seasonal subcomponents of RPC time amplitudes are needed to have a complete model. Therefore, DHR models are estimated to find hidden sinusoids of precipitation. Cosines and sines found at the harmonic Fourier frequencies generate an orthogonal set of regressors. Another addition to the estimated models is including a dummy variable presenting the occurrence of significant effect of El *Niño* and La *Niña* events.

The DHR modeling scheme in this paper starts with estimating mean model for each of the RPC time amplitude. Then we estimated the model using independent variables without El *Niño* and La *Niña* events dummies and Fourier components. Subsequently, the events dummies are added to the model, followed by the estimation of the full model with gradual inclusion of Fourier components. The

determination of the number of Fourier components to be included depends on two criteria: the shape of RPC time amplitude curve and the contributed variance of each component. Likelihood-ratio test is adopted to compare between these nested models for each dependent variable separately. Finally, the residuals of each model were checked using diagnostic tests related to fitting and model assumptions. In addition to that, models are characterized by having Gaussian white noise residual distribution and minimum AIC value.

The estimated parameters of the DHR for the three RPC amplitudes models are shown in Table 1. The first RPC amplitude represents the precipitation mode of variability in northern Ethiopia. It is regressed as a function of six predictors ($Niño3.4$, linear trend, Kirmet and Belg season dummies, October 2011 dummy and May 2016 dummy). Two Fourier components are used for this first dependent variable as the precipitation curve of this region exhibit two maxima and minima. RPC1 model is represented by ARIMA (2, 0, 3) process. The western Ethiopia full model (model 11) is displaying statistical significant relationship with all the incorporated covariates. In addition to that, model (11) is the best performing model for western Ethiopia with the smallest log likelihood and AIC equal to 817.9 and 1659.8, respectively. This entail several important results. First, linear trend term has a significant negative relationship with northern precipitation. Meaning that northern precipitation has a downward trend, which is due to human made effect other than natural forcing. Second, El $Niño$ event in October 2011 had a significant negative impact, while La $Niña$ event in May 2016 had a significant positive impact. Third, the results shows a negative relationship between precipitation over Northern Ethiopia and $Niño3.4$. Finally, strong positive relationship between precipitation mode of variability in this region and both of Belg and Kirmet season is depicted.

The second RPC amplitude acts for the precipitation mode of variability in western Ethiopia which is regressed as a function of three predictors (AMM, SOI, and July 2015 dummy). Three Fourier components are used for this second dependent variable as the precipitation curve of this region has more than two maxima and minima. RPC2 model is represented by ARIMA (1,0,0) process. The western Ethiopia full model (model 11) is displaying statistical significant relationship with all the incorporated covariates. In addition to that, model (11) is the best performing model for western Ethiopia with the smallest log likelihood and AIC equal to 817.9 and 1659.8, respectively. Furthermore, important results are also reached. They show a negative relationship between precipitation over Western Ethiopia and AMM. Besides a positive relationship between precipitation mode of variability in this region and SOI is observed. Moreover, La $Niña$ event in July 2015 had a significant positive impact.

The third RPC amplitude symbolizes the precipitation mode of variability in eastern and southern Ethiopia. It is regressed as a function of four predictors ($Niño3.4$, May 2003 dummy, September 2011 dummy, and September 2012 dummy). Three Fourier components are added to the model explaining this dependent variable since the precipitation curve of this region has more than two maxima and minima. RPC3 model is represented by a white noise process. The eastern and southern Ethiopia full model (model 17) is showing statistical significant relationship with all the

Table 1 Dynamic harmonic regression (DHR) results

	RPC1					RPC2						RPC3					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
Constant	−0.016	−14.722***	−14.294***	−4.979**	−4.994***	−0.048	0.516	0.318	−0.088	0.250	0.243	0.000	−0.068	0.242	0.227	0.208	0.146
	(0.544)	(0.955)	(0.788)	(2.261)	(1.749)	(1.142)	(1.225)	(1.213)	(1.100)	(0.852)	(0.728)	(0.827)	(0.824)	(0.777)	(0.732)	(0.688)	(0.647)
Sin $\left(2\pi\frac{t}{12}\right)$				4.418***	−21.942***				−3.228**	−3.528***	−3.612***				−0.263	−0.278	−0.216
				(1.127)	(1.399)				(1.409)	(1.110)	(0.945)				(1.037)	(0.975)	(0.918)
Cos $\left(2\pi\frac{t}{12}\right)$				−9.813***	−12.569***				−5.554***	−5.005***	−4.856***				5.581***	5.617***	5.651***
				(2.413)	(1.708)				(1.496)	(1.180)	(1.004)				(1.030)	(0.969)	(0.911)
Sin $\left(2\pi\frac{2t}{12}\right)$					22.569***					−8.532***	−8.589***					5.291***	5.325***
					(1.070)					(1.021)	(0.857)					(0.969)	(0.911)
Cos $\left(2\pi\frac{2t}{12}\right)$					4.801***					1.570	1.550*					0.237	0.299
					(0.783)					(1.017)	(0.853)					(0.975)	(0.918)
Sin $(2\pi\frac{3t}{12})$											5.913***						−4.492***
											(0.821)						(0.921)
Cos $(2\pi\frac{3t}{12})$											−5.976***						2.188**
											(0.817)						(0.908)
AR(1)	1.270***	−0.377	−0.434	0.005	0.032**	0.257***	0.243***	0.243***	0.171***	0.053	0.084						
	(0.080)	(0.363)	(0.340)	(0.004)	(0.016)	(0.064)	(0.065)	(0.065)	(0.066)	(0.067)	(0.067)						
AR(2)	−0.632***	0.180	0.144	−0.999***	−0.978***												
	(0.074)	(0.210)	(0.198)	(0.003)	(0.013)												
MA(1)	−0.423***	0.065	0.048	−0.516***	0.031												
	(0.096)	(0.362)	(0.338)	(0.065)	(0.079)												
MA(2)	−0.165**	−0.566***	−0.632***	0.990***	0.957***												
	(0.073)	(0.104)	(0.078)	(0.024)	(0.042)												
MA(3)	−0.206***	0.041	0.033	−0.534***	0.156*												
	(0.078)	(0.172)	(0.193)	(0.065)	(0.083)												
Linear trend		−0.017***	−0.018***	−0.017***	−0.017**												
		(0.005)	(0.004)	(0.005)	(0.007)												
Niño 3.4		−1.366***	−1.740***	−1.666***	−1.643***								−1.353	−1.337	−1.230	−1.171	−1.200*
		(0.404)	(0.319)	(0.342)	(0.477)								(0.867)	(0.815)	(0.767)	(0.721)	(0.679)

(continued)

Table 1 (continued)

	RPC1					RPC2						RPC3					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
AMM							−0.700*	−0.661	−0.294	−0.587*	−0.676**						
							(0.419)	(0.414)	(0.443)	(0.351)	(0.298)						
SOI							0.702	0.779	0.808	0.793*	1.011***						
							(0.599)	(0.593)	(0.549)	(0.452)	(0.383)						
Belg dummy		2.157	1.413	−19.966***	18.345***												
		(1.495)	(1.347)	(3.664)	(3.300)												
Kirmet dummy		48.151***	47.926***	36.086***	7.113**												
		(1.297)	(1.147)	(4.147)	(2.800)												
Mar. 1999 dummy			−10.284	−15.784*	−22.831***												
			(8.961)	(8.242)	(6.368)												
May 2003 dummy														18.240	23.271**	27.821***	32.357***
														(11.677)	(11.039)	(10.426)	(9.853)
Sept. 2011dummy														−43.379***	−43.549***	−43.263***	−38.608***
														(11.686)	(11.049)	(10.436)	(9.865)
Oct. 2011 dummy			−40.324***	−34.977***	−21.226***												
			(8.585)	(8.155)	(7.160)												
Sept. 2012 dummy														−45.257***	−45.560***	−45.348***	−40.656***
														(11.681)	(11.042)	(10.431)	(9.860)
Jul. 2015 dummy								28.092**	24.913**	30.924***	36.980***						
								(12.287)	(11.958)	(10.708)	(8.882)						
May 2016 dummy			29.918***	19.322**	22.272***												
			(9.040)	(8.120)	(6.480)												
Log likelihood	−926.543	−863.151	−847.742	−829.030	−736.016	−905.297	−903.512	−900.925	−890.840	−861.052	−817.901	−899.116	−897.903	−883.253	−869.429	−855.378	−841.460
Akaike inf. crit.	1867.086	1748.302	1723.484	1690.060	1508.033	1816.594	1817.024	1813.849	1797.679	1742.105	1659.802	1802.231	1801.807	1778.506	1754.858	1730.756	1706.919

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

applied covariates with the smallest log likelihood and AIC equal to 841.5 and 1706.9, respectively. East Central Tropical Pacific SST *Niño*3.4 remains a dominant global climatic field. A statistically significant negative relationship exists between eastern and southern Ethiopia precipitation mode of variability and *Niño*3.4. Second, El *Niño* event in October 2011 also hit this region as well as northern Ethiopia, followed by another El *Niño* event in 2012 at the same time of the year. La *Niña* event in May 2003 had a significant positive impact.

5 Conclusions

The main purpose of this study is to explain the precipitation spatiotemporal distributions in the Blue Nile basin and their responses to large-scale modes of climate variability and anthropologic impact. The monthly scale BNB precipitation is decomposed into spatial variation and temporal patterns using EOF/PC analysis. BNB precipitation is obtained from TRMM dataset covering Ethiopia during the period 1998–2016. The teleconnections between three global climatic events *Niño*3.4 , SOI, and AMM as well as anthropogenic effect, namely, linear trend term, harmonic Fourier frequencies, El *Niño* and La *Niña* events dummies and the RPCs time amplitudes were modeled using dynamic harmonic regression. This regression modeling approach is implemented with the intention of determining the dominant climatic driving factors while taking account of periodic seasonality of BNB precipitation. Thereby, the two methods of analysis (EOFs/PCs and DHR) compliment each other. Given the analysis results in Sect. 4, it can be concluded that most studies on precipitation over Ethiopia which discuss natural climate variability focus only on Kirmet and Belg as the dominant seasons, and Bega is regularly overlooked. Examining Bega, in this study, has given insights about the seasonal shifts that is causing BNB extreme precipitation events. The EOF analysis yields three BNB precipitation modes of variability with cumulative variance of approximately 74%. Our results confirm the strong influence of distant SST anomalies on the observed Ethiopian precipitation trend patterns via varying large-scale circulation features in various periodicities. In the dynamic harmonic regression, *Niño* 3.4 is the dominant term for RPC1 and RPC3. AMM and SOI are significant in RPC2 model. Those regression results have demonstrated that the influences of distant tropical Atlantic climatic events must be considered. The SOI has a positive relationship with RPC2. To address whether a detectable anthropogenic influence on BNB precipitation has been found, a line of evidence shows that the linear trend term in RPC1 model is significant. The RPC1 model is the only model where the trend term exhibited statistical significance. This study provides a modeling scheme which allows for seasonal periodicity, accounting for serial correlation of residuals and El *Niño* and La *Niña* events. All of the previous literature on BNB precipitation ignores all of these modeling essentials to have a well-constructed analysis that gives reliable insights about BNB precipitation-large-scale atmospheric teleconnections nexus. Furthermore, it represents the first regional attribution study that perceives

an anthropogenic sign in BNB precipitation trends beyond natural forcing using linear trend term. The latter is a popular trend analysis method in climate research. Last but not least, two policy implications can be inferred: first, there has to be global awareness on the negative impact of anthropogenic influence on precipitation and, second, the need of regional cooperation between Egypt, Sudan, and Ethiopia on several issues such as Grand Ethiopian Renaissance Dam-High Aswan Dam safeguard policy. This policy allows the uses of GERD storage to ensure that the High Aswan Dam lowest power pool elevation (147 m) is secured. There is an urgent need for regional cooperation to increase total water storage in Eastern Nile basin to accommodate future increase of extreme drought/flood events which both have disastrous impact on the economies of these three countries. A caveat, this study lacks the insights on the BNB intra-decadal precipitation variability with the associated dominant climatic factors. Moreover, a similar analysis of longer time spam of BNB precipitation is needed. Finally, future research exploiting statistical climate prediction models will also be essential to predict monthly BNB precipitation at different lead time for proactive water risk management.

Acknowledgements We would like to express our gratitude to guest editors Michela Cameletti and Francesco Finazzi and two anonymous reviewers for critically reviewing the manuscript and offering useful comments. Special thanks go to John Paul Dunne, Jen Schmidt, Dina Rabie and Rahma Ali for their remarks, support, technical support and review of former drafts.

References

1. Antoniadis A, Helbert C, Prieur C, Viry L (2012) Spatio-temporal metamodeling for West African monsoon. Environmetrics 23:24–36
2. Awange JL, Ferreira VG, Forootan E, Khandu, Andam-Akorful SA, Agutu NO, He XF (2016) Uncertainties in remotely sensed precipitation data over Africa. Int J Climatol 36:303–323
3. Barnes EA, Barnes RJ (2015) Estimating linear trends: simple linear regression versus epoch differences. J Clim 28:9969–9976
4. Beltrán F, Sansó B, Lemos RT, Mendelssohn R (2012) Joint projections of North Pacific sea surface temperature from different global climate models. Environmetrics 23:451–465
5. Berhane F, Zaitchik B, Dezfuli A (2014) Subseasonal analysis of precipitation variability in the Blue Nile River Basin. J Clim 27:325–344
6. Block P, Rajagopalan B (2007) Interannual variability and ensemble forecast of upper Blue Nile basin Kiremt season precipitation. J Hydrometeorol 8:327–343
7. Boers N, Bookhagen B, Marwan N, Kurths J (2016) Spatiotemporal characteristics and synchronization of extreme rainfall in South America with focus on the Andes Mountain range. Clim Dyn 46:601–617
8. Bowman KP, Fowler MD (2015) The diurnal cycle of precipitation in tropical cyclones. J Clim 28:5325–5334
9. Camberlin P (1997) Rainfall anomalies in the source region of the Nile and their connection with the Indian summer monsoon. J Clim 10:1380–1392
10. Conway D (2000) The climate and hydrology of the Upper Blue Nile River linked references are available on JSTOR for this article : the climate and hydrology of the Upper Blue Nile River. Geogr J 166:49–62
11. Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, New York

12. Eden JM, Widmann M, Evans GR (2014) Pacific SST influence on spring precipitation in Addis Ababa, Ethiopia. Int J Climatol 34:1223–1235
13. Eldaw AK, Salas JD, Garcia LA (2003) Long-range forecasting of the Nile River flows using climatic forcing. J Appl Meteorol 42:890–904
14. Elsanabary MH, Gan TY (2014) Wavelet analysis of seasonal rainfall variability of the upper Blue Nile basin, its teleconnection to global sea surface temperature, and its forecasting by an artificial neural network. Mon Weather Rev 142:1771–1791
15. Elsanabary MH, Gan TY (2015) Evaluation of climate anomalies impacts on the Upper Blue Nile Basin in Ethiopia using a distributed and a lumped hydrologic model. J Hydrol 530:225–240
16. Eltahir EA (1996) El niño and the natural variability in the flow of the Nile River. Water Resour Res 32:131–137
17. Frazier A, Elison Timm O, Giambelluca T, Diaz H (2017) The influence of ENSO, PDO and PNA on secular rainfall variations in Hawai'i. Clim Dyn 51(5–6):2127–2140
18. Gissila T, Black E, Grimes DI, Slingo JM (2004) Seasonal forecasting of the Ethiopian summer rains. Int J Climatol 24:1345–1358
19. Hamlington BD, Leben RR, Strassburg MW, Nerem RS, Kim KY (2013) Contribution of the Pacific Decadal Oscillation to global mean sea level trends. Geophys Res Lett 40:5171–5175
20. Hu H, Duan Y, Wang Y, Zhang X (2017) Diurnal cycle of rainfall associated with landfalling tropical cyclones in China from rain gauge observations. J Appl Meteorol Climatol 56:2595–2605
21. Huffman GJ, Bolvin DT (2015) TRMM and other data precipitation data set documentation. TRMM 3B42_3B43 documentation: 1–44
22. Jury MR (2016) Determinants of southeast Ethiopia seasonal rainfall. Dyn Atmos Oceans 76:63–71
23. Kahsay T, Kuik O, Brouwer R, Van der Zaag P (2017) The economy-wide impacts of climate change and irrigation development in the Nile Basin: a computable general equilibrium approach. Clim Change Econ 08:1750004
24. Kaiser HF (1958) The varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187–200
25. Korecha D, Barnston AG (2007) Predictability of June–September Rainfall in Ethiopia. Mon Weather Rev 135:628–650
26. Kutzbach JE (1967) Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. J Appl Meteorol 6:791–802
27. L'Heureux ML (2004) Atmospheric circulation influences on seasonal precipitation patterns in Alaska during the latter 20th century. J Geophys Res 109:1–17
28. Loeve M (1963) Probability theory, 3rd edn. Van Nostrand, New York
29. Lorenz EN (1956) Empirical orthogonal functions and statistical weather prediction
30. Mamombe V, Kim WM, Choi YS (2017) Rainfall variability over Zimbabwe and its relation to large-scale atmosphere–ocean processes. Int J Climatol 37:963–971
31. Moron V, Robertson AW, Pai D (2017) On the spatial coherence of sub-seasonal to seasonal Indian rainfall anomalies. Clim Dyn 49:3403–3423
32. Ndehedehe CE, Agutu NO, Okwuashi O, Ferreira VG (2016) Spatio-temporal variability of droughts and terrestrial water storage over Lake Chad Basin using independent component analysis. J Hydrol 540:106–128
33. News B (2016) Ethiopians die in floods and landslides after heavy rain. http://www.bbc.com/news/world-africa-36266310
34. Nicholson SE (2013) The West African Sahel: a review of recent studies on the rainfall regime and its interannual variability 2013
35. Noake K, Polson D, Hegerl G, Zhang X (2012) Changes in seasonal land precipitation during the latter twentieth-century. Geophys Res Lett 39:1–5
36. North GR, Bell TL, Cahalan RF, Moeng FJ (1982) Sampling errors in the estimation of empirical orthogonal functions. Mon Weather Rev 110:699–706
37. Onyutha C (2016) Geospatial trends and decadal anomalies in extreme rainfall over Uganda, East Africa. Adv Meteorol 2016:Article ID 6935912

38. Onyutha C (2017) Variability of rainfall and river flow in the Nile Basin. Doctoral dissertation. Retrieved from https://www.researchgate.net/profile/Charles_Onyutha/publication/320146885_Variability_of_Rainfall_and_River_Flow_in_the_Nile_Basin/links/59d0d57aaca2721f43672973/Variability-of-Rainfall-and-River-Flow-in-the-Nile-Basin.pdf
39. Onyutha C, Willems P (2017) Influence of spatial and temporal scales on statistical analyses of rainfall variability in the River Nile basin. Dyn Atmos Oceans 77:26–42
40. Sanabria J, Bourrel L, Dewitte B, Frappart F, Rau P, Solis O, Labat D (2018) Rainfall along the coast of Peru during strong El Niño events. Int J Climatol 38(4):1737–1747
41. Sanabria J, Bourrel L, Dewitte B, Frappart F, Rau P, Solis O, Labat D (2017) Rainfall along the coast of Peru during strong El Niño events. Int J Climatol
42. Sarojini BB, Stott PA, Black E (2016) Detection and attribution of human influence on regional precipitation. Nat Clim Change 6:669–675
43. Scott CM, Shulman MD (1979) An areal and temporal analysis of precipitation in the Northeastern United States. J Appl Meteorol 18:627–633
44. Siam MS, Eltahir EAB (2015) Explaining and forecasting interannual variability in the flow of the Nile River. Hydrol Earth Syst Sci 19:1181–1192
45. Siam MS, Eltahir EAB (2017) Climate change enhances interannual variability of the Nile river flow. Nat Clim Change 7:350–354
46. Siam MS, Wang G, Demory ME, Eltahir EaB (2014) Role of the Indian Ocean sea surface temperature in shaping the natural variability in the flow of Nile River. Clim Dyn 43:1011–1023
47. Tchakoutio A, Nzeukou A (2016) On the differences in the intraseasonal rainfall variability between Western and Eastern Central Africa: Case of 25–70-Day Oscillations. J Geosci Environ Protect 4:141–158
48. Varikoden H, Al-Shukaili HSA, Babu C, Samah A (2016) Rainfall over oman and its teleconnection with el niño southern oscillation. Arab J Geosci 9:520
49. Viste E, Korecha D, Sorteberg A (2013) Recent drought and precipitation tendencies in Ethiopia. Theor Appl Climatol 112:535–551
50. Wang H, Schubert S, Suarez M, Chen J, Hoerling M, Kumar A, Pegion P (2009) Attribution of the seasonality and regionality in climate trends over the United States during 1950–2000. J Clim 22:2571–2590
51. Wang X, Cui G, Wu F, Li C (2015) Analysis of temporal-spatial precipitation variations during the crop growth period in the Lancang River basin, southwestern China. Ecol Eng 76:47–56
52. Wikle CK (2015) Modern perspectives on statistics for spatio-temporal data. Wiley Interdiscip Rev Comput Stat 7:86–98
53. Williams AP, Funk C, Michaelsen J, Rauscher SA, Robertson I, Wils TH, Koprowski M, Eshetu Z, Loader NJ (2012) Recent summer precipitation trends in the greater horn of Africa and the emerging role of Indian ocean sea surface temperature. Clim Dyn 39:2307–2328
54. Xiao M, Zhang Q, Singh VP (2015) Influences of ENSO, NAO, IOD and PDO on seasonal precipitation regimes in the Yangtze River basin, China. Int J Climatol 35:3556–3567
55. Xing W, Wang B, Yim Sy (2016) Long-Lead Seasonal Prediction of China Summer Rainfall Using an EOF – PLS Regression-Based Methodology. Am Meteorol Soc 1:1783–1796
56. Young PC, Pedregal DJ, Tych W (1999) Dynamic harmonic regression. J Forecast 18:369–394
57. Zeleke T, Damtie B (2017) Temporal and spatial climate variability and trends over Abay (Blue Nile) River basin. In: Social and ecological system dynamics, Chap 6. Springer International Publishing, Berlin, pp 59–75
58. Zhang X, Zwiers FW, Hegerl GC, Lambert FH, Gillett NP, Solomon S, Stott PA, Nozawa T (2007) Detection of human influence on twentieth-century precipitation trends. Nature 448:461–465
59. Zhu Z (2018) Breakdown of the relationship between Australian summer rainfall and ENSO caused by tropical Indian Ocean SST warming. J Clim 31(6):2321–2336

A Hidden Markov Random Field with Copula-Based Emission Distributions for the Analysis of Spatial Cylindrical Data

Francesco Lagona

Keywords Cylindrical data · Copulas · Hidden Markov random fields · Marine currents

1 Introduction

Cylindrical spatial series are bivariate vectors of angles and intensities that are simultaneously observed at a number of sites in an area of interest. The name *cylindrical* is motivated by the special domain of these data, because the pair of an angle and an intensity can be described as a point on a cylinder. Cylindrical spatial series arise frequently in environmental and ecological studies. Examples include hurricane wind satellite data [25], wave directions and heights that are generated by deterministic wave models [31, 32], speeds and directions of marine currents recorded by a network of high-frequency radars [20, 28], as well as telemetry data of animal movement [8]. Further examples of cylindrical spatial series can be found in specific case studies of image analysis [15, 27].

The analysis of cylindrical spatial series is complicated by the special topology of the support on which the measurements are taken (the cylinder) and by the difficulties in modeling the cross-correlations between angular and linear measurements across space. Additional complications arise from the multimodality of the marginal distribution of the data, which are often observed under heterogeneous, space-varying conditions.

We describe a cylindrical hidden Markov random field (MRF) model that parsimoniously accounts for the specific features of cylindrical spatial series. More precisely, we approximate the data distribution with a mixture of copula-based cylindrical densities, whose parameters vary across space according to a latent Potts model. The Potts model [29] is a categorical MRF, i.e., a multinomial process

F. Lagona (✉)
University of Roma Tre, Rome, Italy
e-mail: francesco.lagona@uniroma3.it

M. Cameletti, F. Finazzi (eds.), *Quantitative Methods in Environmental and Climate Research*, https://doi.org/10.1007/978-3-030-01584-8_7

in discrete space, which fulfills a spatial Markovian property: the conditional distribution at each site given the rest of the field is independent of the field values outside a neighborhood of the site. It segments an area of interest according to an interaction parameter that captures the correlation between adjacent observations and controls the smoothness of the segmentation.

Cylindrical hidden MRFs have been already proposed in the literature [19, 28], by exploiting the Abe-Ley density [1] as emission distribution. The Abe-Ley density is a five-parameter bivariate density on the cylinder. A mixture of Abe-Ley densities therefore provides a distributional extension to allow for multimodal cylindrical data. Assuming that the mixture parameters vary according to the segmentation provided by a Potts MRF is a further extension to capture unobserved spatial heterogeneity and to allow for spatial correlation.

We extend these proposals by considering copula-based cylindrical densities. Copulas allow the marginal densities and the joint dependence structure to be modeled separately. As a result, they provide a general method for binding any pair of univariate marginal distributions together to form a bivariate distribution. This is particularly advantageous in the cylindrical setting, because a copula can be exploited to bind two marginal densities that do not necessarily have the same support. In this work, we take this approach by binding a Weibull and a circular wrapped Cauchy together to form a cylindrical density. However, this proposal can be promptly adapted with different marginal densities, if desired.

Hidden MRFs are popular models in spatial statistics, since the seminal paper by Besag [3]. They can be seen as an extension of hidden Markov models, exploited in time series analysis, to the spatial setting. Hidden Markov models have been recently proposed for the analysis of cylindrical time series [21, 22]. This paper extends this approach to the analysis of cylindrical spatial series.

Special computational issues arise in the estimation of the parameters of the proposed cylindrical hidden MRF model. When the spatial interaction parameter of the Potts model is equal to zero, the cylindrical hidden MRF reduces to a latent class model for independent cylindrical data, and a standard expectation-maximization (EM) algorithm can be exploited for likelihood maximization [17, 18, 23]. EM algorithms are based on the definition of a complete-data likelihood function and, under regularity conditions [33], provide a sequence of estimates that converges to a local maximum of the likelihood function by iteratively updating and maximizing the expected value of the complete-data log-likelihood function. When, however, the interaction parameter of the Potts model is not equal to zero, the computation of the expected complete-data log-likelihood is unfeasible, and special approximation strategies are needed. For Gaussian MRFs, Celeux et al. [5] suggest a mean-field approximation of the conditional distribution of the segmentation labels given the data that is optimal in the sense of the Kullback-Leibler divergence. By extending this method to a cylindrical setting, we propose a numerically efficient EM algorithm for estimating the parameters of the cylindrical hidden MRF.

Unfortunately, the proposed EM algorithm does not provide information on the uncertainty of the estimates. In principle, standard errors could be obtained by

numerically approximating the observed information matrix, but reliable results are obtained only when the sample size is extremely large. However, we show that simulation of the proposed cylindrical MRF is straightforward, and, as a result, we recommend parametric bootstrap as the most convenient method to obtain quantiles of the distribution of the estimates.

The rest of the paper is organized as follows. Section 2 illustrates the data that motivated this study. The copula-based cylindrical hidden MRF is presented in Sect. 3, while Sect. 4 describes the EM algorithm that we propose for maximizing the likelihood and the routine that we exploit to compute bootstrap quantiles of the estimates. Section 5 summarizes the results obtained from the model when it was used to segment a vector field of sea currents in the Adriatic sea. A list of relevant discussion points is finally included in Sect. 6.

2 Sea Currents in the Adriatic Sea

Sea current segmentation is useful in several research areas which include studies of the drift of floating objects and oil spills [10], sediment transport [11], and coastal erosion [26]. Surface current measurements are often obtained by high-frequency (HF) radars installed along the coast of the area of interest. HF radars measure surface currents by detecting the Doppler shift of an electromagnetic wave transmitted at a certain frequency. A single HF radar station determines only the radial component of the surface currents relative to that station. Therefore, two or more radar stations are needed to reconstruct the surface currents field in an area of overlapping coverage. HF radars extract the radial components of the ocean surface currents from the analysis of the Doppler spectrum of reflected signals, by combining their radial components, and produce vector maps of the currents. A vector map (or field) decomposes the currents field into the u and v components (Cartesian representation) of the sea surface at each observation point in time and space, where u corresponds to the W-E and v to the N-S current component. The data considered in this paper are based on a network of three HF radars. Two are installed on the western coast of Istria (Zub and Savudrija, Cratia), while the third station (Bibione—Punta Tagliamento, Italy) is located on the Italian coast (Fig. 1). The entire network was created in the framework of the NASCUM (North Adriatic Surface Current Mapping) project [6, 24]. We illustrate the proposed cylindrical hidden MRF model on a surface current field, observed in wintertime across a regular grid of 297 points having a horizontal resolution of about 2 km $\times$ 2 km. For each grid point, we computed the speed $\omega = \sqrt{u^2 + v^2} \in (0, +\infty)$ of the current (m/s) and its direction $x = \arctan(u/v) \in (0, 2\pi]$, hence obtaining a cylindrical spatial series with linear and circular components. The proposed model requires the definition of a neighborhood structure among the sites. The neighborhood of each observation site i was in this study defined by including all the sites j at a distance

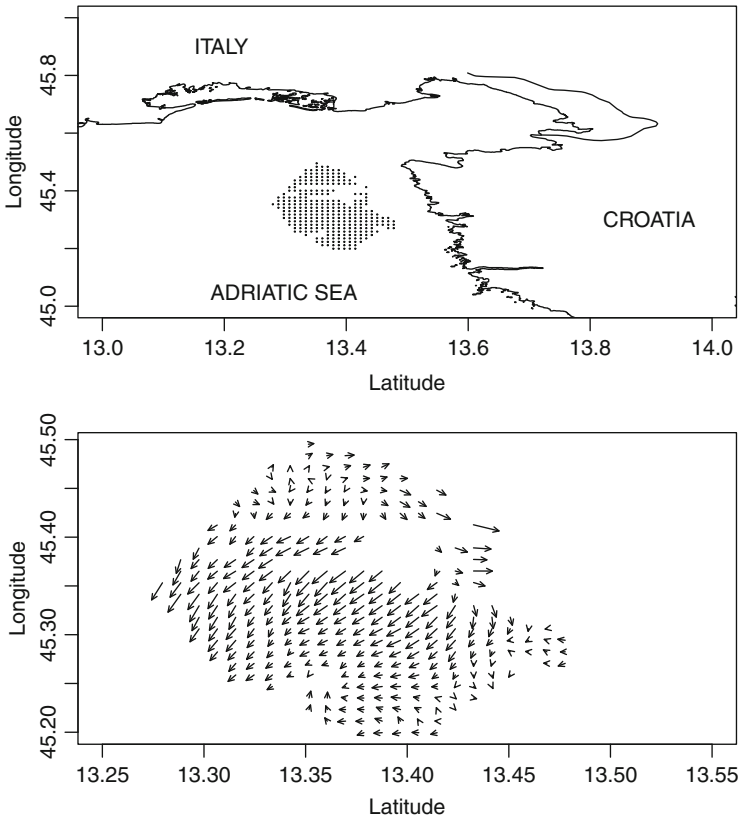

Fig. 1 Top: the spatial grid of 297 sites that partition the study area in the northern Adriatic. Bottom: vector field of the observed data; the orientation of each arrow indicates the current direction at each observation site; the arrow length is proportional to the current speed at that site

$d(i, j) < 4$ km. Figure 1 displays both the location of the study and the observed data. Figure 2 shows the marginal distribution of both the directions and the speeds. The observed speeds range between 0.009 and 0.966 m/s, while directions are distributed around a main southwestern direction. Figure 2 further includes the joint distribution of the data. This planar scatterplot should be interpreted by recalling that points are actually on a cylinder, obtained by wrapping the picture along the x-axis. It shows that the fastest currents flow along the modal southwestern direction.

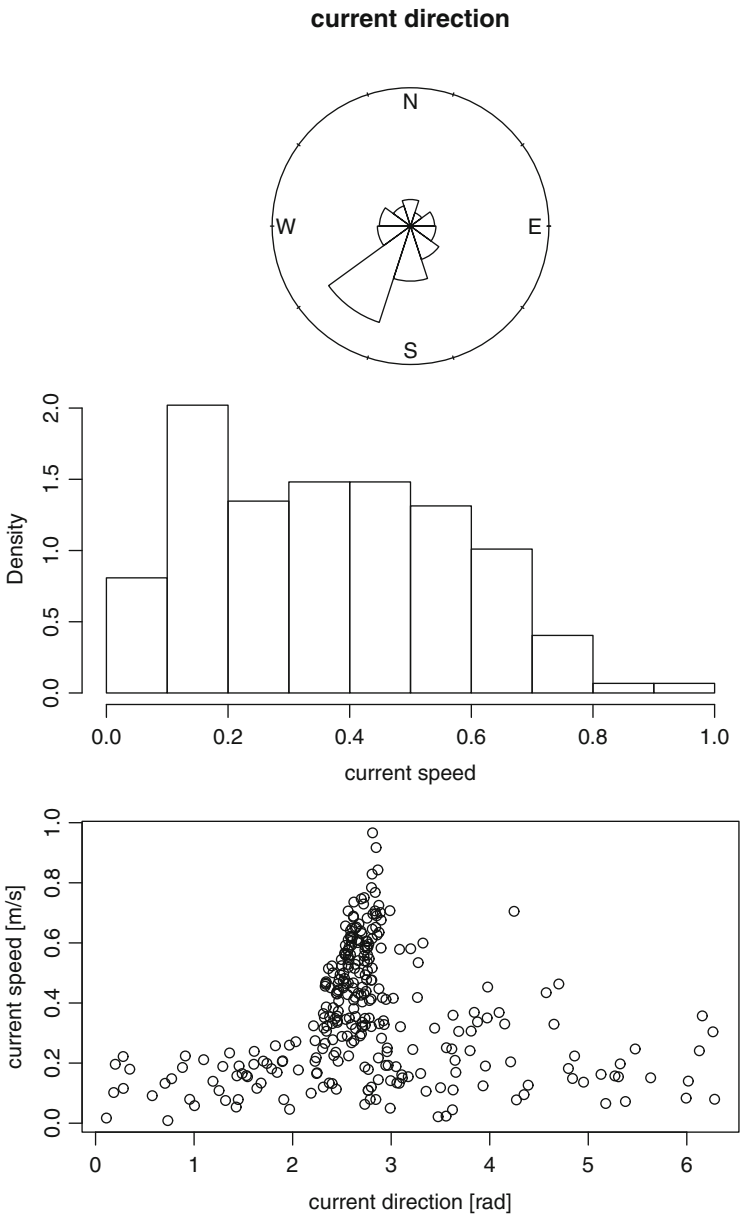

Fig. 2 Top: marginal distribution of current direction. Middle: marginal distribution of current speed. Bottom: joint distribution of current directions (0, $\pi/2$, π, and $3\pi/4$ indicate north, west, south, and east, respectively) and speeds

3 A Copula-Based Hidden Markov Field

The proposed model integrates copula-based cylindrical densities with a latent Potts model. In the following, Sect. 3.1 introduces the proposed family of cylindrical densities, while Sect. 3.2 is devoted to the Potts model. Finally, Sect. 3.3 introduces the proposed hidden MRF.

3.1 Copula-Based Cylindrical Densities

A cylindrical sample is a pair $\mathbf{z} = (x, y)$, where $x \in [0, 2\pi)$ is a point in the circle and y is a point in the positive semi-line $[0, +\infty)$. Let $f(x; \boldsymbol{\alpha})$ be a density on the circle, known up to a parameter $\boldsymbol{\alpha}$, with cumulative distribution function (cdf) $F(x; \boldsymbol{\alpha})$, defined with respect to a fixed, although arbitrary, origin. Moreover, let $f(y; \boldsymbol{\beta})$ be a density on the semi-line, known up to a parameter $\boldsymbol{\beta}$, with cdf $F(y; \boldsymbol{\beta})$. Finally, let $g(u; \boldsymbol{\gamma}), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter $\boldsymbol{\gamma}$. Then,

$$f_q(\mathbf{z}; \boldsymbol{\theta}) = 2\pi g\left(2\pi\left(F(x; \boldsymbol{\alpha}) - qF(y; \boldsymbol{\beta})\right)\right) f(x; \boldsymbol{\alpha}) f(y; \boldsymbol{\beta}) \quad q = \pm 1 \tag{1}$$

is a parametric cylindrical density with support $[0, 2\pi) \times (0, +\infty)$, known up to the parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \boldsymbol{\alpha})$ and $f(y; \boldsymbol{\beta})$ [12]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by de-coupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then Eq. (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$). Additional details on copula-based methods that use a circular binding density to specify bivariate and multivariate densities can be found in [13].

3.2 The Potts Model

Often introduced as an extension of the more popular spatial autologistic model [7], the K-colors Potts model is a multinomial process in discrete space with K classes. Given a lattice that divides an area of interest according to n observation sites $i = 1, \ldots, n$, a sample that is drawn from a spatial multinomial process is a segmentation of this area, obtained by associating each site with a segmentation label $k = 1, \ldots, K$. Formally, each observation site i is associated with a

multinomial random variable $\mathbf{U}_i = (U_{i1}, \ldots, U_{iK})$ with one trial and K classes, where U_{ik} is a Bernoulli random variable that is equal to 1 if i is labelled by k and 0 otherwise. A specific segmentation of the area can be accordingly represented as a sample drawn from the multinomial process $\mathbf{U} = (\mathbf{U}_1, \ldots \mathbf{U}_n)$. The Potts model is a spatial multinomial process which accounts for a neighborhood structure $N(i), i = 1, \ldots, n$ among the observation sites, which associates each site with a set $N(i)$ of neighbors. Under the simplest one-parameter form of the Potts model, each segmentation $\mathbf{u}$ is associated with a single sufficient statistic that indicates the number of neighboring sites which share the same class $k \neq K$, say

$$n(\mathbf{u}) = \sum_{i=1}^{n} \sum_{j>i: j \in N(i)} \sum_{k=1}^{K-1} u_{ik} u_{jk}.$$

Accordingly, the probability of a specific segmentation $\mathbf{u}$ is known up to a single parameter ρ, and it is given by

$$p(\mathbf{u}; \rho) = \frac{\exp(\rho n(\mathbf{u}))}{W(\rho)}, \tag{2}$$

where $W(\rho)$ is the normalizing constant. The parameter ρ is an autocorrelation parameter: if it is positive (negative), then it penalizes segmentations with a few concordant (discordant) neighbors. In the image analysis literature, it is often referred to a regularization parameter, given that large values of ρ are associated with segmentations where areas with the same label are geometrically regular. For each site i and each label k, let

$$n_k(\mathbf{u}_{\bar{N}(i)}) = u_{ik} \sum_{j \in N(i)} u_{jk}$$

be the number of sites in the neighborhood of i that are labelled by k, where $\bar{N}(i)$ indicates the neighborhood of i, completed by i. Under model (2), the conditional distribution of each site depends only on the labels taken by the neighboring sites, namely,

$$p(u_{ik} = 1 \mid \mathbf{u}_1, \ldots \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots \mathbf{u}_n) = \frac{\exp\left(\rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\rho n_k(\mathbf{u}_{\bar{N}(i)})\right)}, \tag{3}$$

Accordingly, the Potts model is a Markov random field with respect to the chosen neighborhood structure, and the autocorrelation coefficient ρ can be viewed as an auto-regression coefficient that is associated with the spatially lagged outcome $n_k(\mathbf{u}_{\bar{N}(i)})$.

3.3 A Cylindrical Hidden Markov Random Field

The specification of the cylindrical hidden MRF is completed by assuming that the cylindrical observations at the n sites of an areal partitioning are conditionally independent, given a segmentation generated by the Potts model. Formally, a cylindrical spatial series can be represented as a bivariate vector of angles x_i and intensities y_i, observed at n observation points, say $\mathbf{z} = (\mathbf{z}_i, i = 1, \ldots, n)$, $\mathbf{z}_i = (x_i, y_i)$, $x_i \in [0, 2\pi)$, and $y_i \in [0, +\infty)$. We assume that the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} f_q(\mathbf{z}_i; \boldsymbol{\theta}_k)^{u_{ik}}, \tag{4}$$

where the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_K)$ includes K label-specific parameters and $f_q(\mathbf{z}; \boldsymbol{\theta}_k), k = 1, \ldots, K$ are K copula-based densities defined in (1) and known up to the label-specific vector of parameters $\boldsymbol{\theta}_k$. Under this setting, the segmentation labels generated by the Potts model can be interpreted as latent classes, which cluster observation sites according to label-specific cylindrical distributions.

The joint density of the observed data and the unobserved class memberships is therefore given by

$$f(\mathbf{z}, \mathbf{u}; \boldsymbol{\theta}, \rho) = f(\mathbf{z}|\mathbf{u}; \boldsymbol{\theta}) p(\mathbf{u}; \rho).$$

Integrating this distribution with respect to the segmentation $\mathbf{u}$, we obtain the marginal distribution of the observed data, known up to the parameters $(\rho, \boldsymbol{\theta})$. Under this setting, the maximum likelihood estimates, $\hat{\rho}$ and $\hat{\boldsymbol{\theta}}$, of the parameters can be in principle obtained by maximizing the likelihood function

$$L(\rho, \boldsymbol{\theta}; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \rho) f(\mathbf{z} \mid \mathbf{u}; \boldsymbol{\theta}). \tag{5}$$

These parameter estimates can be usefully exploited to infer a posterior segmentation of the study area, by computing the posterior probabilities $p(u_{ik} = 1 \mid \mathbf{z}; \hat{\rho}, \hat{\boldsymbol{\theta}})$ and exploiting a maximum a posteriori (MAP) criterion: site i is associated to class k if

$$p(u_{ik} = 1 \mid \mathbf{z}; \hat{\rho}, \hat{\boldsymbol{\theta}}) > p(u_{ih} = 1 \mid \mathbf{z}; \hat{\rho}, \hat{\boldsymbol{\theta}})$$

for each $h \neq k$. According to this rule, data are clustered according to the latent class that is conditionally expected at each location, given the observed data and the estimated parameters.

When $\rho = 0$, data are independent, and the proposed hidden MRF reduces to a latent class model that involves K cylindrical densities. In this setting, standard

EM algorithms for mixture models can be exploited to maximize the likelihood function, and maximum likelihood estimates can be exploited to compute posterior class membership probabilities. However, by assuming $\rho = 0$, we take a latent class approach to spatial segmentation, and the cylindrical observations are clustered according to similarities in the variable space, i.e., the cylinder $[0, 2\pi) \times (0, +\infty)$. More generally, by allowing $\rho \neq 0$, we account for the redundancy of the data which is due to spatial correlation. As a result, on the one side, taking a hidden MRF approach to segmentation, data clustering is not only based on similarities in the variables space but also on similarities that occur in a spatial neighborhood. On the other side, assuming spatial dependence complicates maximum likelihood estimation and requires special approximation methods.

4 Parameter Estimation

4.1 Mean-Field Approximation and EM Algorithm

Maximum likelihood estimates of the proposed cylindrical hidden MRF model can in principle be obtained by maximizing the log-likelihood function

$$l(\boldsymbol{\theta}, \rho) = \log \sum_{\mathbf{u}} f(\mathbf{z} \mid \mathbf{u}; \boldsymbol{\theta}) p(\mathbf{u}; \rho).$$

However, direct maximization of this likelihood is unfeasible due to the summation over all the possible segmentations $\mathbf{u}$. In this setting, EM algorithms offer a viable maximization strategy. By treating the segmentation labels $\mathbf{u}$ as missing values, a complete-data log-likelihood function can be defined as follows:

$$l_{\text{comp}}(\boldsymbol{\theta}, \rho) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \log f(z_i; \boldsymbol{\theta}_k) - \log W(\rho) + \rho n(\mathbf{u}). \tag{6}$$

Starting with an initial parameter estimate, say $\boldsymbol{\theta}_0$, ρ_0, the EM algorithm alternates an expectation (E) step and a maximization (M) step, generating a sequence of estimates $(\boldsymbol{\theta}_s, \rho_s)$, $s = 1, 2 \ldots$ that converges to a local maximum of the likelihood, under suitable regularity conditions [33]. During the sth E step, the expected value of the complete-data log-likelihood $l_{\text{comp}}(\boldsymbol{\theta}, \rho)$ is computed with respect to the conditional distribution $p(\mathbf{u} \mid \mathbf{z}; \boldsymbol{\theta}_{s-1}, \rho_{s-1})$ of the unobserved segmentation given the data and the estimates available from the previous step $s - 1$ of the algorithm. During the sth M step, this expected value is then maximized, and an update $(\boldsymbol{\theta}_s, \rho_s)$ of the estimates is provided. However, these steps require the computation of the normalizing constant $W(\rho)$, which is unfeasible.

Following Celeux et al. [5], a possible solution relies on a mean-field approximation of the posterior distribution $p(\mathbf{u} \mid \mathbf{z})$. This approach has been successfully

exploited in the estimation of complex Markov random field models [2, 15]. Given the estimate $(\boldsymbol{\theta}_{s-1}, \rho_{s-1})$, we first compute for each location a mean-field value

$$\hat{u}_{iq} = \frac{\sum_{k=1}^{K} u_{ik} \exp\left(\rho_{s-1} \sum_{j \in N(i)} \hat{u}_{jk,s-1} + \log f(\mathbf{z}_i \mid \boldsymbol{\theta}_{k,s-1})\right)}{\sum_{k=1}^{K} \exp\left(\rho_{s-1} \sum_{j \in N(i)} \hat{u}_{jk,s-1} + \log f(\mathbf{z}_i \mid \boldsymbol{\theta}_{k,s-1})\right)}, \tag{7}$$

where $\hat{u}_{ik,s-1}$ is the mean-field estimate at location i, available from the previous step. We then approximate the posterior segmentation distribution by

$$p(\mathbf{u} \mid \mathbf{z}; \boldsymbol{\theta}_s, \rho_s) \approx \prod_{i=1}^{n} p(\mathbf{u}_i \mid \hat{\mathbf{u}}_{N(i),s}),$$

where $\hat{\mathbf{u}}_{N(i),s} = (\hat{u}_{j,s}, j \in N(i))$ is the mean-field configuration in the neighborhood of i, and $p(\mathbf{u}_i \mid \hat{\mathbf{u}}_{N(i),s})$ is the univariate conditional distribution under the Potts model, obtained by setting the neighborhood as equal to the mean-field configuration. This is the best approximation of the posterior distribution $p(\mathbf{u} \mid \mathbf{z})$ with respect to the Kullback-Leibler divergence [5].

By taking this approximation into account, the expected value of the complete-data log-likelihood at the sth iteration reduces to the sum of the following two functions

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} p(u_{ik} = 1 \mid \hat{\mathbf{u}}_{N(i),s-1}) \log f(\mathbf{z}_i \mid \boldsymbol{\theta}_k) \tag{8}$$

$$Q(\rho) = \sum_{i=1}^{n} \sum_{k=1}^{K} p(u_{ik} = 1 \mid \hat{\mathbf{u}}_{N(i),s-1}) \log p(\mathbf{u}_i \mid \hat{\mathbf{u}}_{N(i),s-1}; \rho). \tag{9}$$

Because (8) and (9) depend on separate sets of parameters, the M step of the EM algorithm reduces to the separate maximization of the two functions. We can maximize (8) with respect to all the parameters, or, more efficiently, we can take a IFM (inference function for margins [14]) approach. Precisely, (8) can be written as the sum of three components, namely,

$$\sum_{i=1}^{n} \hat{\pi}_{ik} f(\mathbf{z}_i; \boldsymbol{\theta}_k) = \sum_{i=1}^{n} \hat{\pi}_{ik} \log g(2\pi(F(x_i; \alpha) - q F(y_i; \beta)); \boldsymbol{\gamma}) \tag{10}$$

$$+ \sum_{i=1}^{n} \hat{\pi}_{ik} f(x_i; \boldsymbol{\alpha}_k) \tag{11}$$

$$+ \sum_{i=1}^{n} \hat{\pi}_{ik} f(y_i; \boldsymbol{\beta}_k), \tag{12}$$

where

$$\hat{\pi}_{ik} = \hat{p}(u_{ik} = 1 \mid \hat{\mathbf{u}}_{N(i),s-1}).$$

Accordingly, IFM proceeds by finding the parameter values $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ that respectively maximize (11) and (12) and then maximizing function (10), evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, to obtain an estimate of γ. Maximization of $Q(\rho)$ instead reduces to a traditional one-dimensional optimization routine such as that provided by the function `optimize` of `R`.

4.2 Computational Aspects

The EM algorithm is quite sensitive to the choice of the starting values that are exploited at the initialization step. Depending on the initial conditions, the EM algorithm may converge to local maxima of the log-likelihood function or diverge to singularities at the edge of the parameter space, where the log-likelihood is unbounded [33]. The presence of multiple local and spurious maxima is well documented in the literature on mixture models and several strategies have been proposed to select a local maximizer and detect a spurious maximizer [23]. We follow the strategy developed by Bulla et al. [4]. Precisely, we pursue a short-run strategy, by running the EM algorithm from a number of random initializations and stopping the algorithm without waiting for full convergence. We have observed that convergence to spurious maxima is fast and can be detected within short EM runs by monitoring the class proportions. We selected the ten outputs of the EM short run maximizing the log-likelihood and checked for spurious solutions, where this effect did not occur. Then, these ten parameter sets were used to initialize longer runs of the EM algorithm. We stopped the optimization when the increase of two successive log-likelihoods fell below $10^{-4}\%$, as this stopping criterion produced stable parameter estimates in preliminary experiments.

In the analysis of the marine data illustrated in Sect. 2, carried out by using a i7 processor (2.50 GHz), the computational time of a single short run was rarely greater than 30 s, whereas a single long run could take up to 200 s. Therefore the computational cost of the proposed estimation strategy essentially depends on the number of short runs. Computational speed can be improved by choosing a small number of short runs, at the price of a high risk of convergence to a local maximum.

The procedure outlined above does not produce standard errors of the estimates, because approximations based on the observed information matrix often require a very large sample size [23]. By alternatively taking a parametric bootstrap approach, we re-fitted the model to $R = 200$ bootstrap samples, which were simulated from the estimated model parameters. In this case, the EM was initialized at the estimated model parameters, and only one long run was executed. As a result, for each bootstrap sample, convergence was mostly achieved within 30 s, by using a

i7 processor (2.50 GHz), and the total procedure required less than 2 h. We finally computed the 2.5% and the 97.5% quantiles of the empirical distribution of each bootstrap estimate.

Simulation of the hidden MRF proposed in this paper is straightforward. We first simulate a spatial segmentation from a Potts distribution. Several are the routines available to simulate this model. We exploit the Swendsen-Wang algorithm [30] that is available in the R package `potts`. Given a configuration of segmentation labels, a cylindrical observation at site i is obtained as follows. First, a sample η_1 is drawn from the uniform circular distribution. Then a sample η is drawn from the binding distribution g, evaluated at $\gamma = \hat{\gamma}_k$, where k is the class of the random field at site i. Finally, by setting $\eta_2 = (\eta + q\eta_1)(\text{mod}2\pi)$, the cylindrical sample at site i is obtained as $(x_i, y_i) = (F^{-1}(\eta_1/2\pi; \hat{\alpha}_k), F^{-1}(\eta_2/2\pi; \hat{\beta}_k))$, where k is the class that the field takes at site i.

5 Segmenting Sea Current Fields

We have segmented the study area by estimating a number of cylindrical hidden MRFs from the data illustrated above, by varying the number K of latent classes from 2 to 5. The BIC statistic is an excellent tool to detect the order of complex hidden Markov random fields [16]. In this application, it suggested a model with $K = 2$ components. Table 1 displays the estimates under this model, along with bootstrap percentiles.

This table provides two general pieces of evidences that support the distributional choices of this paper. First, within each state, the copula dependence parameter is significant. This supports the choice of a cylindrical density and indicates that, at least in this case study, a conditional independence assumption between univariate distributions of circular and linear variables is unrealistic. Second, the spatial dependence parameter is significant. This supports the inclusion of a spatial process to account for spatial autocorrelation and indicates that the assumption of spatial independence is unrealistic.

The rest of Table 1 should be interpreted with the help of Fig. 3. This figure displays the threefold output of the model. First, the data are clustered according to the posterior state-membership probabilities $p(u_{ik} = 1 \mid \mathbf{z}; \hat{\rho}, \hat{\boldsymbol{\theta}})$ (black indicates $p(u_{ik} = 1 \mid \mathbf{z}; \hat{\rho}, \hat{\boldsymbol{\theta}}) = 1$) under each state (Fig. 3, top). Second, the data distribution is described in terms of two conditional distributions under each state (Fig. 3, middle). Third, the cylindrical spatial series is segmented according to two spatial patterns, each associated with a specific latent state (Fig. 3, bottom). The interpretation of these results is intuitively appealing. State 1 is associated with anticyclonic circulation flows. The moderate currents that travel along the Istrian coast are clustered under this regime. State 2 is instead associated with Bora wind jets that blow northeasterly. Bora blows when a polar high-pressure area sits over the snow-covered mountains of the interior plateau behind the coastal mountain range and a calm low-pressure area lies further south over the warmer Adriatic. Under a

Table 1 Parameter estimates and bootstrap quantiles of a two-state cylindrical hidden Markov random field

	Parameter	Estimate	2.5% Quantile	97.5% Quantile
State 1				
Wrapped Cauchy	Location	2.586	0.654	4.763
	Concentration	0.399	0.182	0.718
Weibull	Shape	2.015	1.798	2.301
	Scale	0.228	0.183	0.710
Copula	Dependence	0.042	0.015	0.054
	q	-1		
State 2				
Wrapped Cauchy	Location	2.655	2.481	2.791
	Concentration	0.895	0.777	0.902
Weibull	Shape	4.075	3.371	4.345
	Scale	0.581	0.210	1.045
Copula	Dependence	0.077	0.040	0.110
	q	-1		
Markov field	Spatial dependence	0.401	0.341	0.521

Bora episode, most of the wind energy is transferred to the sea surface, and, as a result, most of the currents that flow at the fastest rates in the sample are clustered within this regime. The high value of the concentration parameter further indicates that currents are highly concentrated around one modal direction.

6 Discussion

We have illustrated a novel hidden Markov random field for the analysis of cylindrical spatial series. The model can be exploited to cluster spatially correlated cylindrical data according to a finite number of latent classes, associated with specific copula-based cylindrical densities that describe the distribution of the data under each class. The proposed approach was motivated by segmentation issues that arise in marine studies, but it can be easily adapted to a wide range of real-world data, including ecological studies of animal behavior, where direction and speed of movements are recorded across space [9], as well as environmental studies that involve spatial patterns of wind speeds and directions [25].

From a methodological viewpoint, the model offers a number of advantages. It flexibly accommodates spatial correlation, linear-circular correlation, and multimodality by means of parameters that can be easily interpreted in terms of traditional concepts such as location, shape, scale, and concentration. In the examined case study of marine data, the model offered a parsimonious description of current dynamics by capturing the plasticity of current fields in terms of intuitively appealing copula-based distributions that represent specific environmental regimes.

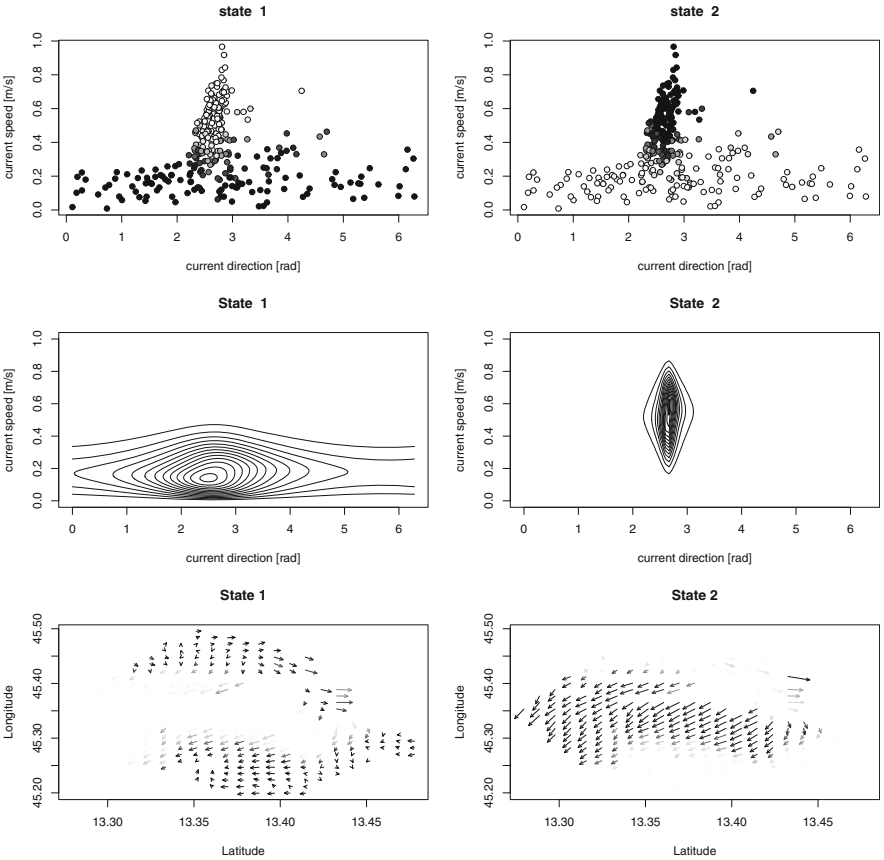

Fig. 3 Model-based clustering and state-specific data distributions according to a cylindrical hidden Markov random field model with two states. Top: clusters of the observed data in a planar plot (points are colored with gray levels according to the estimated posterior membership probabilities—black indicates a probability equal to 1). Middle: state-specific cylindrical densities. Bottom: spatial clusters of the observed data (arrows are colored with gray levels according to the estimated posterior membership probabilities)

Acknowledgements F. Lagona was supported by the 2015 PRIN-supported project "Environmental processes and human activities: capturing their interactions via statistical methods," funded by the Italian Ministry of Education, University and Scientific Research.

References

1. Abe T, Ley C (2017) A tractable, parsimonious and flexible model for cylindrical data, with applications. Econ Stat 4:91–104
2. Alfò M, Nieddu L, Vicari D (2008) A finite mixture model for image segmentation. Stat Comput 18:137–150

3. Besag J (1986) On the statistical analysis of dirty pictures. J R Stat Soc B 48:259–302
4. Bulla J, Lagona F, Maruotti A, Picone M (2012) A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. J Agric Biol Environ Stat 17:544–567
5. Celeux G, Forbes F, Peyrard N (2003) EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recogn 6:131–144
6. Cosoli S, Gacic M, Mazzoldi A (2012) Surface current variability and wind influence in the north-eastern Adriatic Sea as observed from high-frequency (HF) radar measurements. Cont Shelf Res 33:1–13
7. Guyon X (1995) Random fields on a network. Modeling, statistics, and applications. Springer, New York
8. Hanks EM, Hooten MB, Alldredge MW (2015) Continuous-time discrete-space models for animal movement. Ann Appl Stat 9:145–165
9. Holzmann H, Munk A, Suster M, Zucchini W (2006) Hidden Markov models for circular and linear-circular time series. Environ Ecol Stat 13:325–347
10. Huang G, Wing-Keung Law A, Huang Z (2011) Wave-induced drift of small floating objects in regular waves. Ocean Eng 38:712–718
11. Jin KR, Ji ZG (2004) Case study: modeling of sediment transport and wind-wave impact in lake Okeechobee. J Hydraul Eng 130:1055–1067
12. Johnson, RA, Wehrly, TE (1978) Some angular-linear distributions and related regression models. J Am Stat Assoc 73:602–606
13. Jones MC, Pewsey A, Kato S (2015) On a class of circulas: copulas for circular distributions. Ann Inst Stat Math 67:843–862
14. Kim G, Silvapulle M, Silvapulle P (2007) Comparison of semiparametric and parametric methods for estimating copulas. Comput Stat Data Anal 51:2836–2850
15. Klauenberg K, Lagona F (2007) Hidden Markov random field models for TCA image analysis. Comput Stat Data Anal 52:855–868
16. Lagona F (2002) Adjacency selection in Markov random fields for high spatial resolution hyper-spectral data. J Geogr Syst 4:53–68
17. Lagona F, Picone M (2011) A latent-class model for clustering incomplete linear and circular data in marine studies. J Data Sci 9:585–605
18. Lagona F, Picone M (2012) Model-based clustering of multivariate skew data with circular components and missing values. J Appl Stat 39:927–945
19. Lagona F, Picone M (2016) Model-based segmentation of spatial cylindrical data. J Stat Comput Simul 86:2598–2610
20. Lagona F, Picone M, Maruotti A, Cosoli S (2015) A hidden Markov approach to the analysis of space-time environmental data with linear and circular components. Stoch Environ Res Risk Assess 29:397–409
21. Lagona F, Picone M, Maruotti A (2015) A hidden Markov model for the analysis of cylindrical time series. Environmetrics 26:534–544
22. Mastrantonio G, Maruotti A, Jona-Lasinio G. (2015) Bayesian hidden Markov modelling using circular-linear general projected normal distribution. Environmetrics 26:145–158
23. McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York
24. Mihanovic H, Cosoli S, Vilibic I, Ivankovic D, Dadic V, Gacic M (2011) Surface current patterns in the northern Adriatic extracted from high frequency radar data using self organizing map analysis. J Geophys Res 116:C08033
25. Modlin D, Fuentes M, Reich B (2012) Circular conditional autoregressive modeling of vector fields. Environmetrics 23:46–53
26. Pleskachevsky A, Eppel D, Kapitza H (2009) Interaction of waves, currents and tides, and wave-energy impact on the beach area of Sylt island. Ocean Dyn 59:451–461
27. Plötz T, Fink GA (2009) Markov models for offline handwriting recognition: a survey. Int J Doc Anal Recogn 12:269–298

28. Ranalli M, Lagona F, Picone M, Zambianchi E (2018) Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. J R Stat Soc C 67:575–598
29. Strauss DJ (1977) Clustering on coloured lattices. J Appl Probab 14:135–143
30. Swendsen RH, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. Phys Rev Lett 58:86–88
31. Wang F, Gelfand AE (2014) Modeling space and space-time directional data using projected Gaussian processes. J Am Stat Assoc 109:1565–1580
32. Wang F, Gelfand A, Jona-Lasinio G (2015) Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the Adriatic sea. Stat Sin 25:25–39
33. Wu C (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103