JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 30, NO. 2, 349–359
https://doi.org/10.1080/10618600.2020.1814789

Taylor & Francis
Taylor & Francis Group

Improving Bayesian Local Spatial Models in Large Datasets

Amanda Lenzi[a], Stefano Castruccio[b], Håvard Rue[a], and Marc G. Genton[a]

[a]Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; [b]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN

ABSTRACT

Environmental processes resolved at a sufficiently small scale in space and time inevitably display nonstationary behavior. Such processes are both challenging to model and computationally expensive when the data size is large. Instead of modeling the global non-stationarity explicitly, local models can be applied to disjoint regions of the domain. The choice of the size of these regions is dictated by a bias-variance trade-off; large regions will have smaller variance and larger bias, whereas small regions will have higher variance and smaller bias. From both the modeling and computational point of view, small regions are preferable to better accommodate the non-stationarity. However, in practice, large regions are necessary to control the variance. We propose a novel Bayesian three-step approach that allows for smaller regions without compromising the increase of the variance that would follow. We are able to propagate the uncertainty from one step to the next without issues caused by reusing the data. The improvement in inference also results in improved prediction, as our simulated example shows. We illustrate this new approach on a dataset of simulated high-resolution wind speed data over Saudi Arabia. Supplemental files for this article are available online.

ARTICLE HISTORY
Received July 2019
Revised April 2020

KEYWORDS
Integrated nested Laplace approximation; Latent processes; Local models; Spatial models; Wind speed

1. Introduction

The rising popularity of statistical methods for environmental data calls for the development of new methods that are able to capture the underlying varying dependencies and to provide computationally efficient inference for the ever increasing amount of data. Traditional geostatistical approaches are not only computationally intensive but are also based on stationarity assumptions, which is convenient but too restrictive and rarely realistic. For instance, wind at sufficiently small temporal resolution (e.g., hourly or sub-hourly) tends to be more variable over complex terrain than over flat surfaces due to geographical features creating eddies. Additionally, failing to account for how physical processes such as weather patterns vary over time or space can lead to an unrealistic assessment of the dependence, and hence suboptimal inference and prediction.

Traditionally, methods have focused on characterizing the spatial and spatiotemporal non-stationarity explicitly via the covariance function. The deformation method in Sampson and Guttorp (1992) constructs a nonstationary covariance structure from a stationary structure by rescaling the spatial distance, which was subsequently extended to the Bayesian context in Damian, Sampson, and Guttorp (2001) and Schmidt and O'Hagan (2003). Another class of nonstationary methods is built on the process convolution or kernel smoothing method, introduced by Higdon (1998), which uses a spatially varying kernel and a white noise process to create the covariance structure. Other well-known approaches to model non-stationarity include representing the covariance function as a linear combination of basis functions and modeling the covariance matrix of the random coefficients (Nychka, Wikle, and Royle 2002), and to account for the effect of covariate information directly in the covariance function (Schmidt et al. 2011; Neto, Schmidt, and Guttorp 2014). For a review on the existing literature on nonstationary methods, see Risser (2016).

Although all of the above methods produce valid models, their computational burden for inference and prediction can be unfeasible for large datasets. Indeed, for evaluating a Gaussian likelihood in a dataset of size n, $O(n^2)$ entries need to be stored and $O(n^3)$ flops need to be computed for the log-determinant and matrix factorization. This task is feasible in modern computers only when n is at most a few tens of thousands of points. Additionally, evaluating a nonstationary model implies inference on a larger parameter space, which requires an exponentially increasing number of likelihood evaluations for frequentist inference or posterior sampling (Edwards, Castruccio, and Hammerling 2020). To address the difficulties in computation for large datasets, Nychka et al. (2018) used a multiresolution representation of Gaussian processes to represent non-stationarity based on windowed estimates of the covariance function under the assumption of local stationarity, and successfully used this idea to emulate fields from climate models. Kuusela and Stein (2018) proposed modeling Argo profiling float data using locally stationary Gaussian process regression, where parameter estimation and prediction were carried out in a moving window. Other works related to moving window methods have been developed and applied in Hammerling, Michalak, and Kawa (2012) and Tadić et al. (2015) to model remote sensing data.

CONTACT Amanda Lenzi *lenzi.amanda88@gmail.com* Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

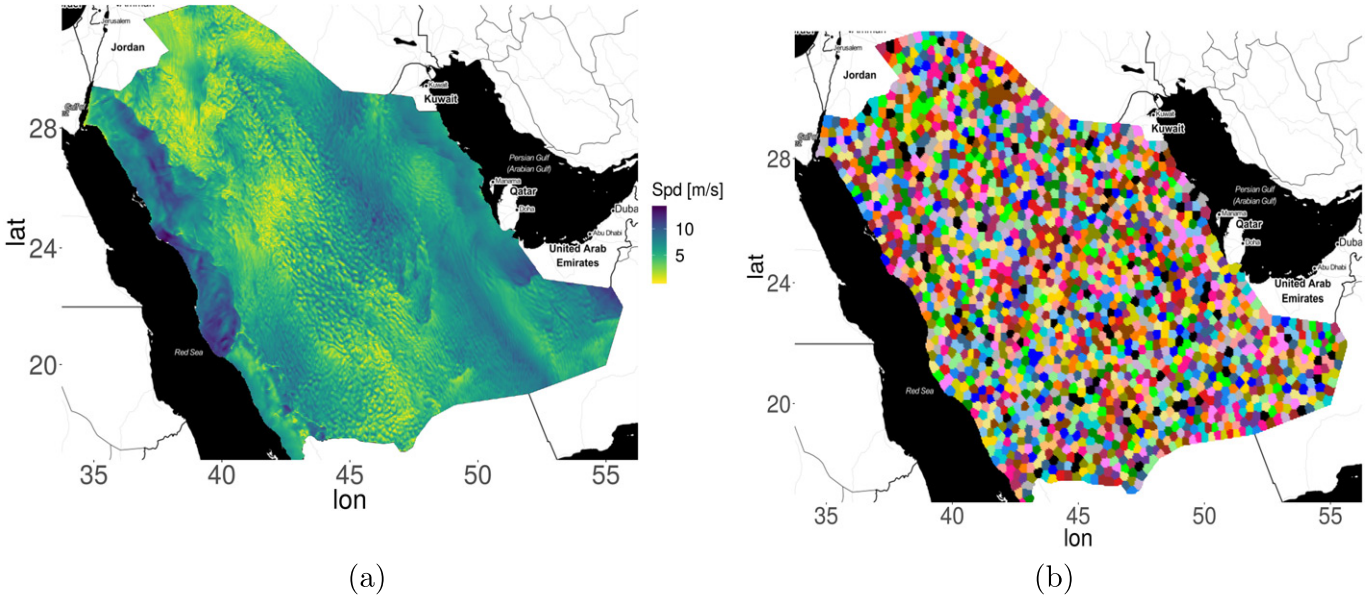


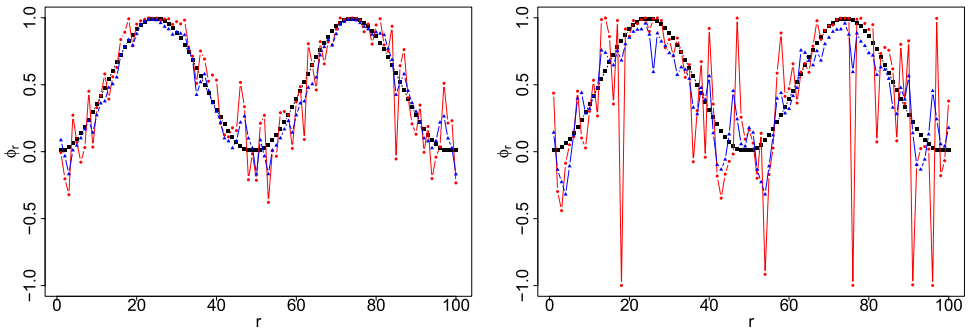
Figure 1. (a) Snapshot of 2 meters wind speed simulations at 84,494 locations over Saudi Arabia on 03/06/2010 averaged between 14:00 and 15:00 local time. The minimum wind speed is around 0.3 m/s and the maximum is 14 m/s. (b) Location of the $R = 2000$ clusters.

The seminal work of Lindgren, Rue, and Lindström (2011) predicated avoiding modeling the covariance function altogether and modeled the data via a stochastic partial differential equation (SPDE) instead. By considering a spatial field as a solution of an SPDE, and describing the covariance function only implicitly, inference is of the order $O(n^{3/2})$ (Rue et al. 2017), thus allowing inference on considerably larger datasets than covariance-based methods. The computational benefits arise from the precision matrix (inverse covariance matrix) resulting from the approximate stochastic weak solutions of the SPDE, which has a Markovian structure where only close neighbors are nonzero (Rue and Held 2005). By spatially varying the coefficients in the SPDEs, it is also possible to construct a variety of nonstationary models. Bolin and Lindgren (2011) developed such a method for global ozone mapping, whereas Bakka et al. (2019) defined a continuous solution to an SPDE with spatially varying coefficients for solving problems that involve a physical barrier to spatial correlation. By combining the SPDE representation of a stationary Matérn field with the deformation method, Hildeman, Bolin, and Rychlik (2019) modeled non-stationarity in significant wave heights. Locally nonstationary fields were considered in Fuglstad, Lindgren, et al. (2015) by letting the coefficients in the SPDE vary with position, and further discussed and generalized for spatially varying marginal standard deviations and correlation structure in Fuglstad, Simpson, et al. (2015). More recently, Fuglstad and Castruccio (2020) formulated a global SPDE model with locally varying coefficients with a change of structure across land and ocean. Another application of the SPDE approach to model non-stationarity is to include covariates directly into the model parameters; see Ingebrigtsen, Lindgren, and Steinsland (2014) for an application to annual precipitation in Norway.

The aim of this article is to develop a new method for modeling large datasets with spatial dependence that not only improves local models in terms of inference and prediction, but is also computationally affordable. As a motivating example, we use the high-resolution simulated wind data from a computer model displayed in Figure 1(a). We partition this data into several small disjoint subsets of the data, which we call "regions," as shown in Figure 1(b). Modeling and predicting such variable over a large region present several challenges. First, the data structure at this high resolution is very complex, with details and features that are difficult to capture with a single model. As a consequence, the assumption of stationarity for the entire region is inappropriate. Second, because of the large number of locations, we need a method that is computationally efficient. We show that our method is able to address not only the modeling challenges arising from the inherent non-stationarity of hourly wind, but also the computational issues that are implied by the large data size.

When choosing the size of these regions, we face the conflicting issue of bias-variance trade-off in parameter estimation. Ideally, one wants to choose regions that accurately capture the features in the data (low variance), but also have high predictive out-of-sample skills (low bias). Indeed, small regions reduce the model bias and allow fast computations, at the expense of low accuracy (high variance) in the parameter estimation. Large regions instead allow a control of the variance but also imply a sub-optimal characterization of the dependence structure, hence a bias.

We propose a novel three-step approach, which simultaneously allows for small regions and low variance. The key is to allow small regions to model the local dependence, and correct the estimated parameter distribution with a smoothing step that borrows strength from neighboring regions. The smoothing step is performed so that it accounts for the uncertainty of the parameter estimates from the first step. The resulting smoothed distribution represents the adjusted uncertainty of the local parameters, which is then used for refitting the models. Allowing this adjusted uncertainty to be used as a new prior would imply the incorrect premise of the model being influenced by the data twice, hence our approach restricts the



(a) (b)

Figure 2. Values of ϕ_r used to simulate $R = 100$ time series from (1) of length 50 (black squares), estimated values of ϕ_r from fitting the AR(1) model to the simulated data (red circles), and estimates after fitting a smoothing spline (blue triangles). The left plot corresponds to simulations with fixed $\tau = 2$ and the right plot corresponds to $\tau = 1$.

information propagation by including it as the new posterior estimates instead. Crucially, the approach we propose is computationally fast and scalable to massive spatial datasets, as it can be fully parallelized across regions. We start with a simple example where the new posterior is the mode of the distribution from the smoothing step. Then, using the wind data in Figure 1(a), we show that it is possible improve the predictive performances by also allowing the uncertainty to propagate from one step to the next.

Our three-step approach is best exemplified by considering a toy dataset, where each region consists of an autoregressive process of order one, AR(1). We simulate R time series from this model, where each time series contains T observations, $\mathbf{y}_r = \{y_r(1), \ldots, y_r(T)\}^\top$. For each r, the observations $\mathbf{y}_r$ are assumed to be conditionally independent, given the latent Gaussian random field $\mathbf{x}_r = \{x_r(1), \ldots, x_r(T)\}^\top$ and the hyperparameter ϕ_r:

$$\begin{aligned} y_r(t) &= x_r(t) + \epsilon_r(t), \quad \epsilon_r(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/\tau), \\ x_r(t) &= \phi_r x_r(t-1) + \omega_r(t), \quad \omega_r(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \end{aligned} \tag{1}$$

where $t = 2, \ldots, T$ is an index for time, $|\phi_r| < 1$ and τ is the fixed precision (known and the same for all time series). Figure 2 shows the different values of ϕ_r used to simulate $R = 100$ time series from (1), where ϕ_r changes according to a series of sine squared (black squares in Figure 2). For each time series, we set $T = 50$ and two different values for the precision: $\tau = 2$ and $\tau = 1$ in Figures 2(a) and (b), respectively. In the first step, we estimate local models for each time series (red circles in Figure 2). In the second step, we apply a correction on the parameters' estimates from the first step, based on information from neighboring regions (blue triangles in Figure 2). The third step consists of refitting the model in (1) to each time series, propagating the information from the adjusted posterior estimates from the second step back into the analysis. Figure 2 shows that our correction improves the parameter estimates substantially not only for the more extreme case where $\tau = 1$ in panel (b), but also when $\tau = 2$ in panel (a). More details on this example will be provided in Section 3.

The remainder of this article is organized as follows. In Section 2, we provide an overview of the proposed methodology. Further details of our approach using the AR(1) example are given in Section 3. The application to the wind speed data in Figure 1 is presented in Section 4. A comprehensive discussion and conclusions are provided in Section 5.

2. Overview of the Proposed Methodology

2.1. Background

We consider a nonstationary and possibly very large dataset, and a partition of the domain into regions where the assumption of stationarity is plausible, defined as $\Omega_r, r = 1, \ldots, R$, where each observation is associated with exactly one Ω_r. Each region contains N_r observations, $\mathbf{y}_r = \{y_r(1), \ldots, y_r(N_r)\}^\top$. For each Ω_r, consider the following hierarchical structure:

$$\begin{aligned} \mathbf{y}_r \mid \mathbf{x}_r, \boldsymbol{\theta}_r &\sim \prod_{i=1}^{N_r} \pi\{y_r(i) \mid x_r(i), \boldsymbol{\theta}_r\}, \\ \mathbf{x}_r \mid \boldsymbol{\theta}_r &\sim \pi(\mathbf{x}_r \mid \boldsymbol{\theta}_r), \\ \boldsymbol{\theta}_r &\sim \pi(\boldsymbol{\theta}_r), \end{aligned} \tag{2}$$

where $\mathbf{x}_r = \{x_r(1), \ldots, x_r(N_r)\}^\top$ is the vector of the latent field that describes the underlying spatial dependence structure, $\boldsymbol{\theta}_r$ is the m-dimensional vector of hyperparameters and π is a generic distribution. The observations $\mathbf{y}_r$ are assumed to be conditionally independent, given $\mathbf{x}_r$ and $\boldsymbol{\theta}_r$. The resulting joint posterior distribution of $\mathbf{x}_r$ and $\boldsymbol{\theta}_r$ is given by

$$\pi(\mathbf{x}_r, \boldsymbol{\theta}_r \mid \mathbf{y}_r) \propto \pi(\boldsymbol{\theta}_r)\pi(\mathbf{x}_r \mid \boldsymbol{\theta}_r) \prod_{i=1}^{N_r} \pi\{y_r(i) \mid x_r(i), \boldsymbol{\theta}_r\}.$$

Our main goal is to extract the posterior marginal distributions for the elements of the latent field, $\pi\{x_r(i) \mid \mathbf{y}_r\}$ and hyperparameters, $\pi\{\theta_r(j) \mid \mathbf{y}_r\}$, and use them to obtain predictive distributions at unsampled locations. Calculation of these univariate posterior distributions requires integrating with respect to $\mathbf{x}_r$ and $\boldsymbol{\theta}_r$:

$$\begin{aligned} \pi\{x_r(i) \mid \mathbf{y}_r\} &= \int \pi(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)\pi(\boldsymbol{\theta}_r \mid \mathbf{y}_r)\mathrm{d}\boldsymbol{\theta}_r, \\ & \qquad i = 1, \ldots, N_r, \\ \pi\{\theta_r(j) \mid \mathbf{y}_r\} &= \int \pi(\boldsymbol{\theta}_r \mid \mathbf{y}_r)\mathrm{d}\boldsymbol{\theta}_r(-j), \quad j = 1, \ldots, m, \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_r(-j)$ is the vector of all but the jth hyperparameter component omitted. When the integrals in (3) cannot be

found analytically, approximations are typically obtained via simulation-based methods such as MCMC. Alternatively, Rue, Martino, and Chopin (2009) proposed an approximate Bayesian inference approach that has become increasingly popular in the last decade. Approximations for $\pi(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)$ and $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}_r)$ are obtained via a Laplace approximation (see Rue et al. 2017 for a comprehensive review on this approximation). The posterior $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}_r)$ is computed as

$$\pi(\boldsymbol{\theta}_r \mid \mathbf{y}_r) \approx \left.\frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})}\right|_{x=x^*(\boldsymbol{\theta})} = \tilde{\pi}(\boldsymbol{\theta}_r \mid \mathbf{y}_r),$$

where $\tilde{\pi}(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ is a Laplace approximation, and $x^*(\boldsymbol{\theta})$ is the mode of $\mathbf{x}$ for a specific value of $\boldsymbol{\theta}$. Similarly we obtain $\tilde{\pi}(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)$, the approximation of $\pi(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)$. These are then used to construct the following nested approximations

$$\begin{aligned}\tilde{\pi}\{x_r(i) \mid \mathbf{y}_r\} &= \int \tilde{\pi}(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)\tilde{\pi}(\boldsymbol{\theta}_r \mid \mathbf{y}_r)\mathrm{d}\boldsymbol{\theta}_r, \\ &\quad i = 1, \ldots, N_r, \\ \tilde{\pi}\{\theta_r(j) \mid \mathbf{y}_r\} &= \int \tilde{\pi}(\boldsymbol{\theta}_r \mid \mathbf{y}_r)\mathrm{d}\boldsymbol{\theta}_r(-j), \quad j = 1, \ldots, m.\end{aligned} \tag{4}$$

2.2. Improving the Local Estimates

We propose a new method for improving the estimation of $\tilde{\pi}\{\boldsymbol{\theta}_r \mid \mathbf{y}_r\}$ in (4) and hence also improving the estimated $\tilde{\pi}\{x_r(i) \mid \mathbf{y}_r\}$, for $i = 1, \ldots, N_r$. Since each region is selected to be small enough to approximate the local nonstationarity well, the resulting parameters' estimates are likely to have a large variance, and smoothing across the regions is used to reduce it.

The method is based on two extra steps in the estimation procedure from the previous section. In Step 2, we apply a correction to the posteriors $\tilde{\pi}(\boldsymbol{\theta}_r \mid \mathbf{y}_r)$ by smoothing the mode of this distribution across r. In Section 3, we show a one-dimensional example with a smoothing spline, while in Section 4.3, we describe the two-dimensional case with a spatial model. We denote by $\tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y})$ the resulting smoothed distribution for region r in Step 2 of our approach, where $\mathbf{y}$ is the combined datasets from all regions, that is, $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_R^\top)^\top$. In Step 3, the correction from Step 2 is propagated back into the analysis as the posterior for each region:

$$\begin{aligned}\tilde{\pi}_{\text{smooth}}\{x_r(i) \mid \mathbf{y}_r\} &= \int \tilde{\pi}(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)\tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y})\mathrm{d}\boldsymbol{\theta}_r, \\ &\quad i = 1, \ldots, N_r, \\ \tilde{\pi}_{\text{smooth}}\{\theta_r(j) \mid \mathbf{y}_r\} &= \int \tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y})\mathrm{d}\boldsymbol{\theta}_r(-j), \quad j = 1, \ldots, m,\end{aligned} \tag{5}$$

where $\tilde{\pi}(x_r(i) \mid \mathbf{y}_r, \boldsymbol{\theta}_r)$ is obtained by plugging values of $\boldsymbol{\theta}_r$ from $\tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y}_r)$ obtained in Step 2. Step 3 is very computationally efficient, since the posteriors for the hyperparameters have already been estimated, and as in Step 1 the models for each region can be fully parallelized. Also, as the posterior marginals in (5) are the basis to derive the predictive distributions, the proposed correction will also have a direct impact in prediction performance.

Here, the vector $\boldsymbol{\theta}_r$ contains the hyperparameters that need to be smoothed, while the ones that do not require the smoothing are included in $\mathbf{x}_r$. In practice, it is more important to smooth hyperparameters that have a higher variability and are harder to estimate.

Our approach has a crucial difference compared to empirical Bayes methods. The key is to account for the information from the smoothing in Step 2 directly into the posterior distribution in Step 3, as opposed to introducing it through priors as in empirical Bayes methods. By doing so, we prevent the estimation in Step 3 to be influenced by the likelihood of the data that was already used in Step 1, and thus avoiding using the data twice. Moreover, our approach allows for uncertainty propagation from Step 2 to Step 3.

3. Simulation With Spatially Varying AR(1) Process

3.1. Model Description

In Section 1, we briefly introduced our method on a simulated example (see Figure 2) using the AR(1) model in (1). Here, we provide all the details about the methodology in light of the steps proposed in the previous section. For the ease of exposition, we fix the precision τ in (1), so that for each region r only the hyperparameter ϕ_r needs to be estimated. No covariates or additional random effects have been included in (1), but the steps below can be easily adapted to account for them.

The model is a special case of the hierarchical framework proposed in (2). Indeed for the first equation of the hierarchy, the likelihood of the data $\mathbf{y}_r$ given the latent field $\mathbf{x}_r$ and the hyperparameter ϕ_r is given by

$$\mathbf{y}_r \mid \mathbf{x}_r, \phi_r \sim \mathcal{N}_T(\mathbf{x}_r, \tau^{-1}\mathbf{I}_T),$$

where $\mathbf{I}_T$ is the $T \times T$ identity matrix and τ is the fixed precision, while $\mathcal{N}_T$ is a T-dimensional normal distribution. For the latent process $\mathbf{x}_r$, we assume that the marginal distribution of $x_r(1)$ is Gaussian with mean zero and variance $1/(1-\phi_r^2)$ to have a stationary process. The joint distribution can be written as

$$\pi(\mathbf{x}_r \mid \phi_r) \sim \mathcal{N}_T(\mathbf{0}, \mathbf{Q}_{x,r}^{-1}),$$

where $\mathbf{Q}_{x,r}$ is the tridiagonal precision matrix of an AR(1) process.

The three steps of our approach can be summarized as follows:

Step 1: The model fitted to each region. Fit the AR(1) model in (1) with fixed known τ to each time series, $\mathbf{y}_r$, separately. Following the notation in Section 2, we define the variance-stabilizing transformation $\boldsymbol{\theta}_r = \theta_r = \log\left(\frac{1+\phi_r}{1-\phi_r}\right)$, where θ_r has a normal prior with mean zero and precision 0.15, independent across r. We then obtain the posterior marginal distributions for the latent field and for the hyperparameter θ_r, which we denote by $\tilde{\pi}\{x_r(t) \mid \mathbf{y}_r\}$ and $\tilde{\pi}(\theta_r \mid \mathbf{y}_r)$, respectively, for $t = 1, \ldots, T$ and $r = 1, \ldots, R$. Inference is performed using the *R-INLA* package (Rue, Martino, and Chopin 2009).

Step 2: Smoothing the hyperparameter. As in Lindgren and Rue (2008), we assume a continuous spline on a discrete set of knots with a second-order random walk RW(2). We denote by $\hat{\theta}_r$ the mode for $\tilde{\pi}(\theta_r \mid \mathbf{y}_r)$ from Step 1, and we assume a normal distribution: $\hat{\theta}_r \sim \mathcal{N}(u_r, \tau_{\theta;r}^{-1})$, where $\tau_{\theta;r}$ is the precision and is such that $\log(\tau_{\theta;r}) = \log(1/\widehat{\text{SD}}_r^2)$, where $\widehat{\text{SD}}_r$ is the estimated

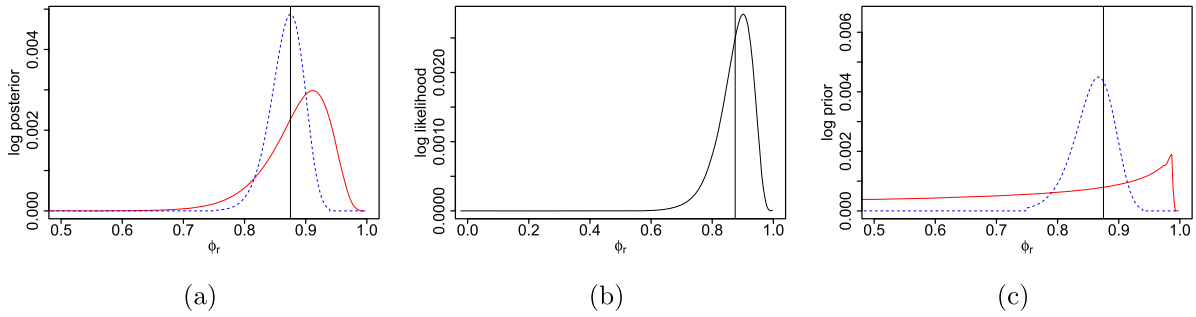


Figure 3. Comparison between nonsmoothing and smoothing changes in the posterior distribution of ϕ_r for the model (1). (a) Scaled log posterior distributions from Step 1 (solid red) and Step 2 (dashed blue). (b) Scaled log-likelihood function. (c) Scaled log prior distributions from Step 1 (solid red) and Step 2 (dashed blue). The vertical line is the true value $\phi_r = 0.88$.

standard deviation of the posterior distribution $\tilde{\pi}(\theta_r \mid \mathbf{y}_r)$. The vector $\boldsymbol{u} = (u_1, \ldots, u_R)^\top$ is assumed to have independent second-order increments:

$$\Delta^2 u_r = u_r - 2u_{r+1} + u_{r+2} \sim \mathcal{N}(0, \tau_u^{-1}), \quad r = 1, \ldots, R-2, \tag{6}$$

where τ_u is the precision parameter and can be used to control the degree of smoothing across regions. Section 3.2 discusses a method for choosing the optimal value of τ_u.

Step 3: Refit the model to each region using the estimated mode. For each region r, we assume that the posterior distribution for the hyperparameters, namely $\tilde{\pi}_{\text{smooth}}(\theta_r \mid \mathbf{y}_r)$, is a point mass concentrated at the mode of $\hat{\theta}_r$ from Step 2. Our choice was dictated by ease of exposition, and in the wind data application in Section 4.3 we will show a more general approach with integration points and weights instead of just the mode. The marginal posterior for the latent process $\mathbf{x}_r$ is then obtained from the first equation in (5). Because here there are no hyperparameters that need to be re-estimated in this example, refitting the model is equivalent to updating the posterior for $\mathbf{x}_r$ given the data under the smoothing from Step 2.

Step 3 implies a change of the original posterior in Step 1, and hence a change in the prior of the model. While retrieving the appropriate prior is not relevant for our method, it is still however possible, and in the Appendix we show the steps to do so. Figure 3 shows (a) the log posterior distributions, (b) log-likelihood function, and (c) log prior distributions from Step 1 (solid red) and Step 2 (dashed blue), for $\phi_r = 0.88$. The log prior distributions were obtained simply by subtracting the log-likelihood from the log posterior distributions, and the vertical line represents the true value. The proposed smoothing in Step 2 concentrates the posterior (and consequently the prior) considerably closer to the true value ϕ_r than a standard approach with no smoothing. Similar results can be observed for other choices of ϕ_r; the mean absolute error across r of the estimated mode posterior distributions from Steps 1 and 2 are 0.23 and 0.08, whereas the mean absolute error of the estimated mode priors for these steps are 0.61 and 0.09, respectively.

3.2. Sensitivity of Prediction to Smoothing

There are different approaches to control the degree of smoothness in Step 2. This can be, for instance, dictated by the case study and prior knowledge. Here, we present one possible method, which is based on two metrics: the first focuses on the departure of the estimated posterior against the exact simulated distribution, and the second is based on cross-validation.

To assess the improved accuracy in capturing the true distribution of the latent process $\mathbf{x}_r$, for each value of ϕ_r, we calculate the Kullback–Leibler (KL) divergence, a widely used metric for comparing two probability distributions. The departure from the true posterior $\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)$ is defined as

$$\text{KL}_r = \int \tilde{\pi}(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r) \log \left\{ \frac{\tilde{\pi}(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)}{\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)} \right\} d\mathbf{x}_r, \tag{7}$$

where ϕ_r in $\tilde{\pi}$ is calculated using either Step 1 or Step 2 of our approach. A small KL_r indicates a small departure from the target posterior, and a zero KL_r indicates that the two distributions are the same.

The data are simulated from a known model, and the posterior distribution of the latent process $\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)$ can be easily obtained from the joint distribution $\pi(\mathbf{x}_r, \phi_r \mid \mathbf{y}_r)$:

$$\begin{aligned}
\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r) &\propto \pi(\mathbf{x}_r, \phi_r \mid \mathbf{y}_r) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{x}_r^\top \mathbf{Q}_{x,r}\mathbf{x}_r \right) \\
&\quad \times \exp\left\{ -\frac{1}{2}\tau(\mathbf{x}_r^\top \mathbf{x}_r - 2\mathbf{y}_r^\top \mathbf{x}_r) \right\} \\
&= \exp\left\{ -\frac{1}{2}\mathbf{x}_r^\top (\mathbf{Q}_{x,r} + \tau \mathbf{I})\mathbf{x}_r + \tau \mathbf{y}_r^\top \mathbf{x}_r \right\} \\
&= \exp\left\{ -\frac{1}{2}\mathbf{x}^\top \mathbf{P}_r \mathbf{x}_r + \mathbf{b}_r^\top \mathbf{x}_r \right\},
\end{aligned}$$

where $\mathbf{P}_r = \mathbf{Q}_{x,r} + \tau \mathbf{I}$ and $\mathbf{b} = \mathbf{y}_r^\top \tau$. This implies that $\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r) \sim \mathcal{N}_T(\boldsymbol{\mu}_{0,r}, \boldsymbol{\Sigma}_{0,r})$, with $\boldsymbol{\mu}_{0,r} = \mathbf{P}_r^{-1}\mathbf{b}_r$ and $\boldsymbol{\Sigma}_{0,r} = \mathbf{P}_r^{-1}$. We also assume that the approximated posterior in (7) is normal, that is, $\pi_{\text{appr}}(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r) \sim \mathcal{N}_T(\boldsymbol{\mu}_{1,r}, \boldsymbol{\Sigma}_{1,r})$, and we obtain $\boldsymbol{\mu}_{1,r}$ and $\boldsymbol{\Sigma}_{1,r}$ based on the sample mean vector and covariance matrix from 10,000 posterior samples. The KL divergence expression in (7) can be simplified in the case of two multivariate Gaussian distributions. Indeed, if the target distribution is $\mathcal{N}_T(\boldsymbol{\mu}_{0,r}, \boldsymbol{\Sigma}_{0,r})$ and the approximation is $\mathcal{N}_T(\boldsymbol{\mu}_{1,r}, \boldsymbol{\Sigma}_{1,r})$, we have

$$\begin{aligned}
\text{KL}_r = \frac{1}{2}\Bigg\{ &\log \frac{|\boldsymbol{\Sigma}_{1,r}|}{|\boldsymbol{\Sigma}_{0,r}|} - T + \text{tr}(\boldsymbol{\Sigma}_{1,r}^{-1}\boldsymbol{\Sigma}_{0,r}) \\
&+ (\boldsymbol{\mu}_{1,r} - \boldsymbol{\mu}_{0,r})^\top \boldsymbol{\Sigma}_{1,r}^{-1}(\boldsymbol{\mu}_{1,r} - \boldsymbol{\mu}_{0,r}) \Bigg\},
\end{aligned}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. Since the KL changes across different orders of magnitudes, we opted for a variance

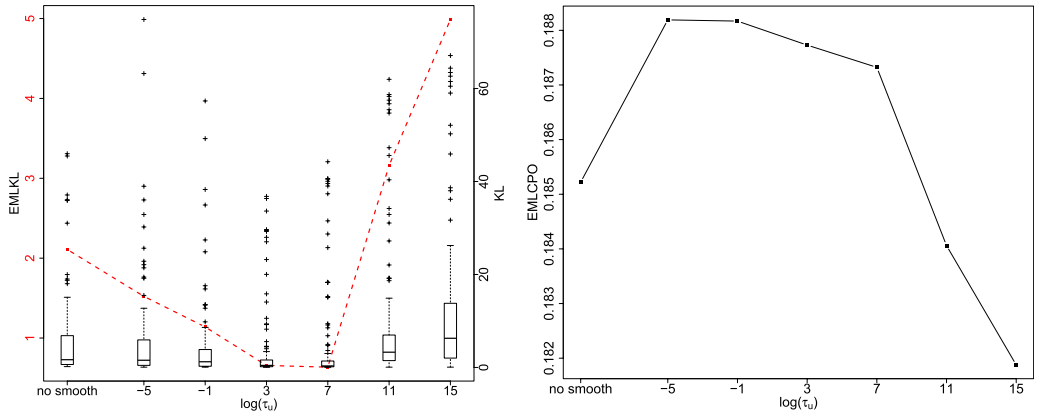


Figure 4. Comparison of inference and prediction performances between the standard non-smoothing method and the proposed smoothing approach from fitting (1). (a) EMLKL (dashed red, left y-axis) and KL_r (black, right y-axis) for no smoothing along with 6 different degrees of smoothness. Lower values of $\log(\tau_u)$ indicate less smoothing. (b) Corresponding EMLCPO for the same degrees of smoothing.

stabilizing estimator, the expected mean log KL (EMLKL) divergence, defined as $\mathrm{EMLKL} = \exp\left\{\frac{1}{R}\sum_{r=1}^{R}\log(\mathrm{KL}_r)\right\}$.

Therefore, we assess the impact of smoothing on the prediction skills of the estimated process. We use the conditional predictive ordinate (CPO) for leave-one-out cross-validation, defined as

$$\begin{aligned}\mathrm{CPO}_r(t) &= \pi\{y_r(t)|\mathbf{y}_r(-t)\}\\ &= \int \pi\{y_r(t)|\mathbf{y}_r(-t),\theta_r\}\pi\{\theta_r|\mathbf{y}_r(-t)\}\mathrm{d}\theta_r,\end{aligned}$$

where $\mathbf{y}_r(-t)$ represents the vector of observations $\mathbf{y}_r$ with the tth component omitted. In other words, $\mathrm{CPO}_r(t)$ is calculated by first obtaining the predictive distribution at t given all but the tth observation in the time series, and then evaluating it at the actual withheld value $y_r(t)$. The CPO can be interpreted as a continuous equivalent of the posterior probability that the observation is predicted from the model, so larger values are preferable. The CPO can be computed efficiently without rerunning the model $R \times T$ times (Held, Schrödle, and Rue 2010). The CPOs are then aggregated in an overall score for comparing different models by averaging across time and regions. As with the KL, we propose the expected mean log conditional predictive ordinate (EMLCPO), defined as $\mathrm{EMLCPO} = \exp\left[\frac{1}{RT}\sum_{r=1}^{R}\sum_{t=1}^{T}\log\{\mathrm{CPO}_r(t)\}\right]$, with models having relatively higher values of EMLCPO, showing a better fit.

We compare the EMLKL and the EMLCPO based on six different degrees of smoothing by changing the values of τ_u in (6): $\log(\tau_u) = \{-5,-1,3,7,11,15\}$, with lower values of $\log(\tau_u)$ indicating less smoothing. Here, $\log(\tau_u) = 15$ results in a constant value across the regions (complete smoothing), so no larger values are considered. Figure 4 shows the results based on (a) KL and on (b) CPO according to the various degrees of smoothing. The first value in the x-axis, "no smooth," corresponds to the estimates directly from Step 1 of our approach. According to the EMLKL (panel (a), left y-axis) and the EMLCPO (panel (b)), the best fit occurs when $\log(\tau_u) = 7$ and $\log(\tau_u) = -5$, respectively. For the EMLKL there is a minimal difference between the $\log(\tau_u) = 3$ and $\log(\tau_u) = 7$, and the right y-axis highlights how the first choice results in less variable KL divergences. Both scores show that there is a clear improvement against a model with no smoothing for $\log(\tau_u) = \{-5,-1,3,7\}$. After $\log(\tau_u) = 7$ the posteriors are oversmoothed and this worsens the fit compared to no smoothing (high EMLKL and low EMLCPO values). Evidence from this numerical study suggests that smoothing almost always improves the estimation of the latent process and prediction. The overall agreement between EMLKL and EMLCPO is essential, as, in a real application, the actual underlying distribution is unknown, and a cross-validation metric, such as the EMLCPO, would be used for choosing the optimal degree of smoothness.

Smoothing does not just improve the prediction and decrease the bias, but also results in less variable estimates. Figure 4(a) (right y-axis) shows the spread of KL_r for the different amounts of smoothing, displayed as a boxplot. It is readily apparent that optimal smoothing results in more stable estimates by decreasing the variance across regions.

4. Application to the WRF Dataset

In this section, we apply the approaches in Section 2 to model and predict a sizeable simulated wind speed dataset in Saudi Arabia. The predictive ability at the sub-grid scale is of interest for statistical downscaling. Interpolated wind from the numerical simulation can be used as a baseline to build a mathematical relationship (e.g., pattern scaling) from in-situ ground wind data at the same location, hence allowing to generate more accurate, observation-driven wind maps

We apply our method to a spatial dataset of simulated wind speed detailed in Section 4.1. In Section 4.2, we present the local model that is fitted to each region and in Section 4.3, we explain in details each step of our approach and present the results.

4.1. The WRF Dataset

We focus on a simulation generated by Yip (2018) from the Weather Research and Forecasting (WRF) model, which is a state-of-the-art Numerical Weather Prediction model, devel-

oped at the National Center for Atmospheric Research and also recently used in wind energy assessment in Tagle et al. (2020). Mesoscale numerical models such as WRF rely on large-scale atmospheric phenomena or meteorological reanalysis to provide boundary conditions and solve physical equations driving the real processes on a fine scale. The boundary conditions used to simulate the WRF data are obtained from the Modern-Era Retrospective analysis for Research and Applications (MERRA, Rienecker et al. 2011), a reanalysis product developed at NASA's Global Modeling and Assimilation Office, using the Goddard Earth Observing System Version 5 general circulation model, together with satellite and surface observations through a data assimilation system.

Each simulation corresponds to hourly data of the zonal and meridional (U and V) wind components on a regular grid of 769 × 659 points in space (5-km resolution) bounded by 5–35°N and 30–65°E during the 2009–2014 period, at 2 meters above ground level. The full dataset comprises of 506,771 spatial locations. We select data that fall inside Saudi Arabia from 03/06/2010 between 14:00 and 15:00 local time, when wind speeds tend to peak, resulting in 84,494 points in space. The U and V components are converted into wind speed: $\sqrt{U^2+V^2}$. Figure 1(a) shows the map of the wind field.

We first partition the domain into R regions small enough so that the assumption of stationarity is plausible. The disjoint subsets are obtained using the k-means clustering method, which minimizes the sum of squares from points to the assigned region centers (Hartigan and Wong 1979). It is, in principle, possible to provide a more formal assessment of stationarity and use it as a metric for clustering, for example, by fitting directional variograms to each region. However, because these estimates are corrected with a smoothing step, the clustering method is less critical. Our partition results into $R = 2000$ regions (see Figure 1(b)), with the smallest region containing 26 locations (≈28 × 28 km) and the largest 62 (≈48 × 48 km).

4.2. The Spatial Model

The distribution of wind speed is bounded below by zero and is significantly right-skewed. Therefore, wind speed cannot be directly modeled with the Gaussian distribution. Common transformations for normalizing wind speed data include logarithmic transformation and square-root transformation (Taylor, McSharry, and Buizza 2009). Haslett and Raftery (1989) showed that square-root transformation is well suited for wind data, as the resulting transformed wind speed resembles the Gaussian distribution. Hence, for each region r we model the square-root transformed wind speed y_r at sampling locations $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{N_r})$ with a latent Gaussian model, a special case of the hierarchical framework proposed in (2). For each region r, we assume

$$y_r(\mathbf{s}_i) = \mathbf{z}_r(\mathbf{s}_i)^\top \boldsymbol{\beta}_r + u_r(\mathbf{s}_i) + \epsilon_r(\mathbf{s}_i), \quad i = 1, \ldots, N_r,$$

where $\mathbf{z}_r$ is a p-dimensional vector of covariates, and $\boldsymbol{\beta}_r$ is the vector of the linear coefficients. Here, $\{\epsilon_r(\mathbf{s}_1), \ldots, \epsilon_r(\mathbf{s}_{N_r})\} \sim \mathcal{N}_{N_r}(\mathbf{0}, \tau_{\epsilon,r}^{-1}\mathbf{I}_{N_r})$ is the iid random noise that accounts for the model uncertainty. The aforementioned model can be written in the vector form

$$\mathbf{y}_r|\boldsymbol{\beta}_r, \mathbf{u}_r, \tau_{\epsilon,r} \sim \mathcal{N}_{N_r}(\mathbf{Z}_r\boldsymbol{\beta}_r + \mathbf{u}_r, \tau_{\epsilon,r}^{-1}\mathbf{I}_{N_r}), \tag{8}$$

where $\mathbf{y}_r = \{y(\mathbf{s}_1), \ldots, y(\mathbf{s}_{N_r})\}^\top$ is the observation vector and the $N_r \times P$ design matrix is $\mathbf{Z}_r = \{\mathbf{z}_r(\mathbf{s}_1), \ldots, \mathbf{z}_r(\mathbf{s}_{N_r})\}^\top$. We consider $p = 2$, thus two covariates: elevation and distance to the coast. In terms of the hierarchical framework in Section 2.1, (8) is the first equation, that is, the data level, in (2).

The spatial field $u_r(\mathbf{s}_i)$ is assumed to be Gaussian and isotropic, with a covariance described by the Matérn function, a widely popular choice in spatial statistics. For two locations $\mathbf{s}_1$ and $\mathbf{s}_2$ at distance $h = \|\mathbf{s}_1 - \mathbf{s}_2\|$, the Matérn covariance is defined as (Stein 1999)

$$\text{cov}\{u_r(\mathbf{s}_1), u_r(\mathbf{s}_2)\} = C_r(h) = \sigma_{u,r}^2 \frac{1}{\Gamma(\nu_r)2^{\nu_r-1}}(\kappa_r h)^{\nu_r}\mathcal{K}_{\nu_r}(\kappa_r h), \tag{9}$$

where $\sigma_{u,r}^2 = 1/\tau_{u,r}$ is the marginal variance and $\mathcal{K}_{\nu_r}$ is the modified Bessel function of the second kind of order $\nu_r > 0$. The popularity of the Matérn is mainly attributable to the control of the number of mean square derivatives of the underlying process through the parameter ν_r. The range is controlled by $\kappa_r > 0$ and $\rho_r = \sqrt{8\nu_r}/\kappa_r$ represents the distance at which the spatial correlation is approximately 0.13, and we set $\nu_r = 1$.

The vector of hyperparameters to be estimated is given by the precision of the data, the precision of the latent process, and its range, so that

$$\boldsymbol{\theta}_r = (\theta_{1,r}, \theta_{2,r}, \theta_{3,r})^\top = \{\log(\tau_{\epsilon,r}), \log(\tau_{u,r}), \log(\rho_r)\}^\top.$$

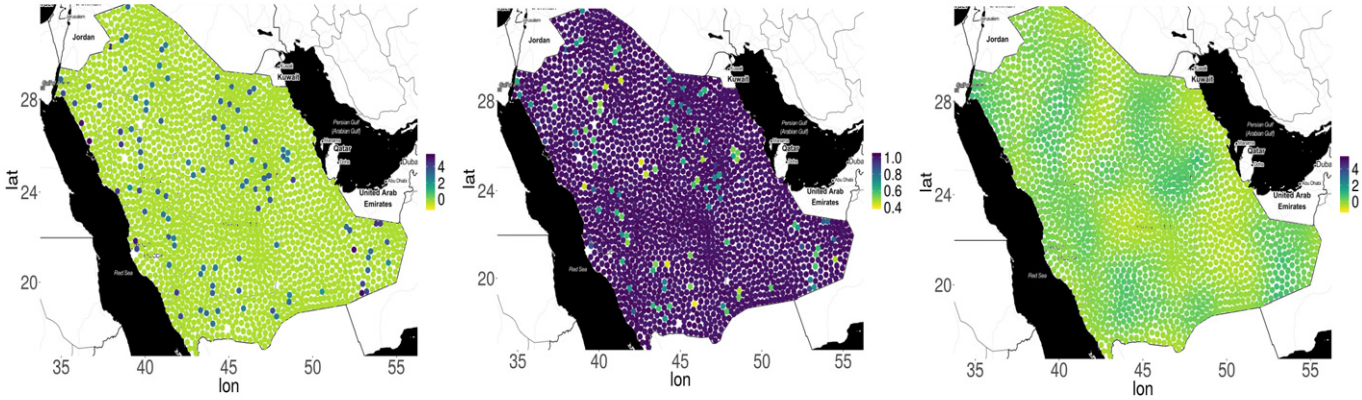
The linear coefficients $\boldsymbol{\beta}_r$ in (8) are less variable, so they are not included in the vector of hyperparameters to be smoothed.

We provide a joint distribution for the range ρ_r and the variance $\sigma_{u,r}^2$ using the concept of the penalized complexity (PC) prior that was recently introduced by Simpson et al. (2017). PC develops priors that allow shrinkage toward a base model, which is assumed to be the reference. The prior is then built by allowing a control of the KL divergence from the base to the actual model. Following Fuglstad et al. (2019), we assume a base model with infinite range and precision, that is, a constant, and we assign PC priors to ρ_r and $\tau_{u,r}$ that are able to control the tail probabilities: $\text{P}(\sigma_{u,r}^2 > \sigma_{0,r}^2) = \alpha_1$ and $\text{P}(\rho_r < \rho_{0,r}) = \alpha_2$. We choose $\alpha_1 = \alpha_2 = 0.01$, $\rho_{0,r}$ to be the 20% of the range of the observations and $\sigma_{0,r}^2$ the variance estimated from the data at region r. In other words, we assume a prior that bounds the variance to be larger than that estimated from the data with a 1% chance, and the range to be below 20% of the range of the observations with a 1% chance. For $r = 1, \ldots, R$, we assume a vague Gamma prior with parameters 1 and 0.00005 for $\tau_{\epsilon,r}$ and a vague Gaussian prior $\mathcal{N}(0, 1000)$ for $\boldsymbol{\beta}_r$. The priors are also assumed to be independent across components. The *R-INLA* package is used for model fitting and predictions (Rue, Martino, and Chopin 2009).

4.3. Results

We now detail our approach with the data and the model described in the previous sections. The three steps are described as follows:

Step 1: The model fitted to each region. We fit the model outlined in Section 4.2 to each of the $R = 2000$ regions in Figure 1(b) separately, and obtain estimates of the posterior distribution for the kth element of $\boldsymbol{\theta}_r$ for $k = 1, 2, 3$, which



(a) $\hat{\theta}_{3,r}$ (b) $\widehat{\mathrm{SD}}_{3,r}$ (c)

Figure 5. Map of the estimated (a) mode and (b) standard deviation of $\tilde{\pi}(\theta_{3,r} \mid \mathbf{y}_r)$ for the $R = 2000$ regions shown in Figure 1(b). (c) posterior mode of $\tilde{\pi}_{\text{smooth}}(\theta_{3,r} \mid \mathbf{y})$.

we denote by $\tilde{\pi}(\theta_{k,r} \mid \mathbf{y}_r)$. We denote as $\hat{\theta}_{k,r}$ the mode of $\tilde{\pi}(\theta_{k,r} \mid \mathbf{y}_r)$, while the posterior standard deviation is denoted as $\widehat{\mathrm{SD}}_{k,r}$. We show the results for $\theta_{3,r} = \log(\rho_r)$, since the range is the hardest parameter to identify, and hence the most variable across regions. Figure 5(a) shows the maps of $\hat{\theta}_{3,r}$. Many regions have a considerably higher estimated posterior mode than the neighboring regions, hence smoothing is necessary. Figure 5(b) shows the map of the posterior standard deviation $\widehat{\mathrm{SD}}_{3,r}$, and it is apparent how the locations with large range values correspond to the ones with low posterior variance. The high variance in the estimates of Figure 5(a) is a consequence of the small region size needed to accommodate the nonstationarity. The region size is another tuning parameter of the method, and cross-validation could have been used to choose the optimal region size.

Step 2: Smoothing the hyperparameters. The modes $\hat{\theta}_{k,r}$ from Step 1 are smoothed here independently across k for simplicity and are normalized by subtracting the mean and dividing by its standard deviation computed across $r = 1, \ldots, R$, for each k. For a general Matérn in region r with known smoothness ν_r, under infill asymptotics, only $\sigma_r^2 \kappa_r^{2\nu_r}$ can be consistently estimated (Zhang 2004). As the likelihood and the posteriors can be concentrated around nonlinear manifolds of the parameter space, modeling the hyperparameters in the log-scale alleviates the problem of smoothing them separately. With an abuse of notation, we now refer to $\boldsymbol{\theta}_r$ and their components as their normalized version. We assume an additive model for smoothing: $\hat{\theta}_{k,r}(\mathbf{s}_c) = u_k(\mathbf{s}_c) + \varepsilon_{k,r}(\mathbf{s}_c)$, where the locations $\mathbf{s}_c$ are the centroids of each region r. The process $u_k(\mathbf{s}_c)$ is assumed to be Gaussian and modeled with the Matérn covariance in (9), with marginal variance $\tilde{\sigma}^2_{u,k}$, range $\tilde{\rho}_k$, and the iid noise is $\varepsilon_{k,r} \sim \mathcal{N}(0, \tilde{\tau}^{-1}_{\epsilon;k,r})$, for $k = 1, 2$ and 3.

We assume $\tilde{\tau}_{\epsilon;k,r}$ to be fixed at the value of $1/\widehat{\mathrm{SD}}^2_{k,r}$, $r = 1, \ldots, R$, from Step 1. This ensures that the same degree of smoothness is applied to all three additive models, that is, the hyperparameters with a larger standard deviation will be smoothed more than the ones with a smaller standard deviation. Here, $\tilde{\rho}_k$ is fixed to half of the domain of the study region. A choice of considerably different values, such as the size of the domain, would result in oversmoothing. The choice of $\tilde{\tau}_u = 1/\tilde{\sigma}^2_{u,k}$ is performed via cross-validation and will be discussed later. Because $\hat{\theta}_{k,r}, k = 1, 2, 3$, are at the same scale after normalization, we can use the same smoothness and therefore $\tilde{\tau}_u$ will not be strongly dependent on k. We use six equally spaced values for $\log(\tilde{\tau}_u)$, varying from -7.5 to 5. The fitted values from the smoothing are then transformed back from the normalized to the original scale. Figure 5(c) is an example of the estimated posterior mode of $\tilde{\pi}_{\text{smooth}}(\theta_{3,r} \mid \mathbf{y}_r)$ with $\log(\tilde{\tau}_u) = -5$.

Step 3: Refit the model to each region using integration points. In the AR(1) simulated example in Section 3, the smoothed hyperparameter posterior was assumed to be a point mass concentrated at the smoothed posterior mode from Step 2, so that calculation of $\tilde{\pi}_{\text{smooth}}\{x_r(i) \mid \mathbf{y}_r\}$ in (5) was trivial. In this application, we propose a more articulated method which numerically approximates the integral in the first equation of (5).

We use the Gauss–Hermite quadrature, a numerical scheme to approximate integrals of the form $\int e^{-\xi^2} f(\xi) \mathrm{d}\xi \approx \sum_{l=1}^{L} f\{\xi^{(l)}\}\Delta^{(l)}$ for a fixed L. The abscissas for the quadrature of order L, which are given by the roots of the Hermite polynomials $\xi^{(l)}$, and the weights $\Delta^{(l)}$, both have a closed form expression (Abramowitz and Stegun 1964).

We operate under the assumption that $\tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y})$ can be well approximated by a product of marginal normal distributions $\tilde{\pi}_{\text{smooth}}(\boldsymbol{\theta}_r \mid \mathbf{y}) \approx \prod_{k=1}^{3} \mathcal{N}(\mu_{k,r}, \sigma^2_{k,r})$, where $\mu_{k,r}$ and $\sigma^2_{k,r}$ are the posterior mean and variance of $\tilde{\pi}_{\text{smooth}}(\theta_{k,r} \mid \mathbf{y})$, respectively. The independence implied by the product is made for convenience, although empirically, we found a relatively low correlation between the components of $\boldsymbol{\theta}_r$. Also, because of the log-scale, the posteriors can be well approximated by a Gaussian distribution, and the first expression in (5) becomes

$$\tilde{\pi}_{\text{smooth}}\{x_r(i) \mid \mathbf{y}_r\} = \int \tilde{\pi}_{\text{smooth}}(\mathbf{x}_r \mid \mathbf{y}_r, \boldsymbol{\theta}_r) \prod_{k=1}^{3} \frac{1}{\sqrt{2\pi\sigma^2_{k,r}}}$$

$$\times \exp\left\{-\frac{(\theta_{k,r}-\mu_{k,r})^2}{2\sigma_{k,r}^2}\right\} \mathrm{d}\boldsymbol{\theta}_r$$
$$= \frac{1}{\sqrt{\pi}} \int \tilde{\pi}_{\text{smooth}}\left(\mathbf{x}_r \mid \mathbf{y}_r, \sum_{k=1}^{3}\mu_{k,r} + \sqrt{2}\xi_{k,r}\sigma_{k,r}\right)$$
$$\times \exp\left(-\sum_{k=1}^{3}\xi_{k,r}^2\right) \mathrm{d}\xi_{1,r}\mathrm{d}\xi_{2,r}\mathrm{d}\xi_{3,r}$$
$$\approx \frac{1}{\sqrt{\pi}} \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\sum_{l_3=1}^{L} \tilde{\pi}_{\text{smooth}}\left(\mathbf{x}_r \mid \mathbf{y}_r, \sum_{k=1}^{3}\mu_{k,r} + \sqrt{2}\xi_r^{(l_k)}\sigma_{k,r}\right)$$
$$\times \Delta^{(l_1)}\Delta^{(l_2)}\Delta^{(l_3)},$$

where the latent field $\mathbf{x}_r = (\mathbf{u}_r^\top, \boldsymbol{\beta}_r^\top)^\top$ contains the linear coefficients and the spatial process in (8). Using a change of variables, we obtain $\xi_{k,r} = \frac{\theta_{k,r}-\mu_{k,r}}{\sqrt{2}\sigma_{k,r}} \Leftrightarrow \theta_{k,r} = \mu_{k,r} + \sqrt{2}\xi_{k,r}\sigma_{k,r}$. For this case study, $L = 5$ integration points in each of the three dimensions provide an approximation that is sufficiently accurate. Thus, the required number of configurations to evaluate the integral is $L^3 = 5^3 = 125$. Since each configuration can be evaluated independently, the computations can be easily parallelized.

4.4. Choice of the Smoothing Parameter

There is no true underlying model here, so the EMLKL in Section 3.2 is not applicable and we only focus on the cross-validation score EMLCPO. We compare the leave-one-out predictive performance using the different degrees of smoothing, as explained previously in Section 3. Figure 6 shows this comparison: lower values of $\log(\tilde{\tau}_u)$ indicate more smoothing than higher values. The highest value corresponds to the results obtained directly from Step 1. The EMLCPO value attains its maximum at $\log(\tilde{\tau}_u) = -5$, and any of the smoothing levels improves the original estimates from Step 1. Differently from the AR(1) case in Section 3, where at some point the smoothing becomes excessive and the scores progressively deteriorate, here the performance is significantly improved even for a large smoothing. We also compare the predictive performances of the integration method against the approach using only the mode as in the AR(1) case. The Gauss–Hermite integration shows marginal improvement, especially for low degrees of smoothing. For higher degrees of smoothing, the estimated posterior distribution is more narrow, and the effect of the integration is less apparent.

5. Discussion

In this work, we developed a new three-step approach for analyzing large datasets with spatial dependence that improves local models in terms of inference and prediction. The method is scalable to extremely large spatial datasets and can properly propagate the uncertainty across steps in a Bayesian framework. In Step 1, the domain is partitioned into regions, and local models are fit to each region. The size of these regions is a bias-variance trade-off; larger regions will have a smaller variance and more substantial bias, whereas smaller regions will have higher variance and lower bias. We choose to use smaller regions, thus allowing the capture of local non-stationarities, followed by a correction for the high variance, based on borrowing information from neighboring regions in Step 2, while accounting for the uncertainty of the parameter estimates from Step 1. Finally, in Step 3, the model is refitted to each region, propagating the uncertainty from the smoothing back into the analysis as the new posterior, thus avoiding problems of using the data twice. The approach allows flexible modeling of complex dependence structures, but is at the same time computationally affordable, as the proposed adjustment is amenable to full parallelization across regions.

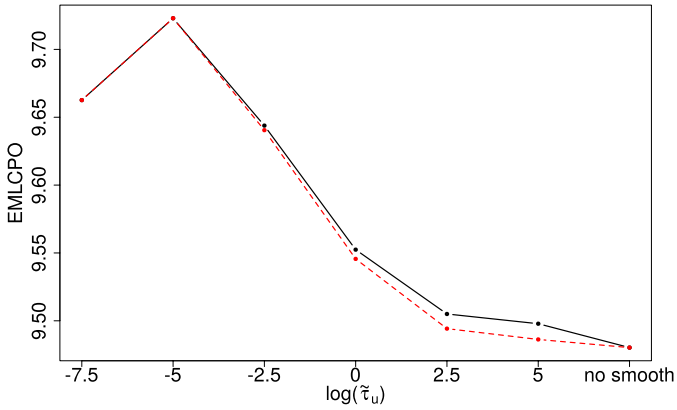


Figure 6. EMLCPO values for 6 different degrees of smoothness as well as no smoothing, the last being the results directly from Step 1. The dashed red indicates marginal posteriors computed with a point mass and the black solid with the Gauss–Hermite quadrature. From left to right: very smooth to no smoothing.

In both the AR(1) simulated data and the application, the improvement from our method compared to fitting local models to each region is apparent. Indeed, the smoothing adjustment allows us to better recover the actual posterior distribution in the simulation study, and most importantly, it enables a superior predictive skill. The smoothing can be chosen to achieve the best possible advantage over the uncorrected model. Ad-hoc sensitivity analysis shows that our method is robust concerning the smoothing technique, with improved results for a wide range of smoothings. The existing methods for the nonstationary case involve highly complex model fitting strategies, and only a few software is available (see, e.g., Gramacy 2007; Risser and Calder 2017). However, a detailed comparison with other methods was not feasible due to the large size of our simulated wind dataset, which could not be presently handled by either of these packages.

Our method is general and can be applied to many settings: space, time, space/time, and different domains, as long as a partition is provided. It relies on local models defined through a hierarchical latent process framework, a class large enough to allow a wide range of applications. If better local models are provided, our method can still be used to correct the variance of the estimated parameters.

A limitation of this approach lies in the assumption of a domain partition. For some applications such as wind, the regions imply a discontinuity at the border, and hence prediction at unsampled locations at the border might be suboptimal. Partition-based approaches that do not imply independence across blocks of the partition with a globally valid model are

available. Castruccio and Stein (2013) and Castruccio and Guinness (2017) proposed an evolutionary spectrum model to capture temperature across latitude and day and night regimes, respectively. If the nature of the problem suggests a change in spatial dependence dictated by some geographical features such as mountain range as proposed in Jeong et al. (2018), then this strategy could be naturally employed with appropriate likelihood approximation. For our domain and wind, however, the partition must be provided by a clustering scheme such as the k-means.

An application of our model to spatiotemporal data is possible. Still, it would likely require additional approximations and a careful choice of the regions as the data size and the hyperparameter space will be considerably larger.

Appendix: Retrieving the Priors

The refitting procedure in Step 3 of our approach uses the information from Step 2 as the new posterior distribution. We show how to retrieve the prior distribution that corresponds to the posterior for the toy example in Section 3.

For each ϕ_r and corresponding data $\mathbf{y}_r$, with $r = 1, \ldots, R$, let $\pi(\mathbf{y}_r \mid \phi_r)$ be the likelihood of observing data $\mathbf{y}_r$ given the hyperparameter ϕ_r. We denote by $\tilde{\pi}(\phi_r \mid \mathbf{y}_r)$ and $\tilde{\pi}_{\text{smooth}}(\phi_r \mid \mathbf{y}_r)$ the posterior distributions from Steps 1 and 3, respectively, and $\tilde{\pi}(\phi_r)$ and $\tilde{\pi}_{\text{smooth}}(\phi_r)$ are the corresponding priors.

From Bayes' theorem, the prior distributions are given by

$$\begin{aligned} \log\{\tilde{\pi}(\phi_r)\} &= A + \log\{\tilde{\pi}(\phi_r \mid \mathbf{y}_r)\} - \log\{\pi(\mathbf{y}_r \mid \phi_r),\} \\ \log\{\tilde{\pi}_{\text{smooth}}(\phi_r)\} &= A + \log\{\tilde{\pi}_{\text{smooth}}(\phi_r \mid \mathbf{y}_r)\} - \log\{\pi(\mathbf{y}_r \mid \phi_r)\}, \end{aligned} \tag{A.1}$$

where A is the normalizing constant.

Recall that the posteriors $\tilde{\pi}(\phi_r \mid \mathbf{y}_r)$ and $\tilde{\pi}_{\text{smooth}}(\phi_r \mid \mathbf{y}_r)$ in the right-hand sides of (A.1), are readily available from Steps 1 and 2, respectively. Therefore, to evaluate $\tilde{\pi}(\phi_r)$ and $\tilde{\pi}_{\text{smooth}}(\phi_r)$, what remains to be computed is the likelihood term $\pi(\mathbf{y}_r \mid \phi_r)$, which is the same in both equations given in (A.1). To compute this term, we start by writing:

$$\pi(\mathbf{y}_r \mid \phi_r) = \frac{\pi(\mathbf{y}_r, \mathbf{x}_r \mid \phi_r)}{\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)}, \tag{A.2}$$

and then compute (A.2) in two steps:

1. The joint distribution $\pi(\mathbf{y}_r, \mathbf{x}_r \mid \phi_r)$:

   We assume that the marginal distribution of $x_r(1)$ is Gaussian with mean zero and variance $1/(1-\phi_r^2)$. Then, we can express the joint distribution of $\mathbf{x}_r$, $\pi(\mathbf{x}_r \mid \phi_r) = \pi\{x_r(1)\}\pi\{x_r(2) \mid x_r(1)\}, \ldots, \pi\{x_r(T) \mid x_r(T-1)\}$, as

   $$\pi(\mathbf{x}_r \mid \phi_r) \sim \mathcal{N}_T(\mathbf{0}, \mathbf{Q}_{x,r}), \tag{A.3}$$

   where $\mathbf{Q}_{x,r}$ is the tridiagonal precision matrix of an AR(1) process

   $$\mathbf{Q}_{x,r} = \begin{pmatrix} 1 & -\phi_r & & & & \\ -\phi_r & 1+\phi_r^2 & -\phi_r & & & \\ & \cdots & \cdots & \cdots & & \\ & & & -\phi_r & 1+\phi_r^2 & -\phi_r \\ & & & & -\phi_r & 1 \end{pmatrix}.$$

   It follows that the joint posterior distribution is

   $$\begin{aligned} \pi(\mathbf{x}_r, \phi_r \mid \mathbf{y}_r) &\propto \pi(\phi_r)\pi(\mathbf{x}_r \mid \phi_r) \prod_{t=1}^{T} \pi\{y_r(t) \mid x_r(t), \phi_r\} \\ &\propto \pi(\phi_r)|\mathbf{Q}_{x,r}|^{1/2}\tau^{1/2} \\ &\quad \times \exp\left[-\frac{1}{2}\left\{ \mathbf{x}_r^\top \mathbf{Q}_{x,r}\mathbf{x}_r + \tau(\mathbf{y}_r - \mathbf{x}_r)^\top(\mathbf{y}_r - \mathbf{x}_r) \right\} \right]. \end{aligned} \tag{A.4}$$

2. The conditional distribution $\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r)$:

   We use the fact that the conditional distribution of $\mathbf{x}_r$ is just the joint distribution between $\mathbf{x}_r$ and $\mathbf{y}_r$, without the terms that do not depend on $\mathbf{x}_r$ since $\mathbf{y}_r$ and ϕ_r are fixed:

   $$\begin{aligned} \pi(\mathbf{x}_r \mid \mathbf{y}_r, \theta_r) &\propto \pi(\mathbf{y}_r, \mathbf{x} \mid \phi_r) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{x}_r^\top \mathbf{Q}_{x,r}\mathbf{x}_r \right) \\ &\quad \times \exp\left\{ -\frac{1}{2}\tau(\mathbf{x}_r^\top \mathbf{x}_r - 2\mathbf{y}_r^\top \mathbf{x}_r) \right\} \\ &= \exp\left\{ -\frac{1}{2}\mathbf{x}_r^\top(\mathbf{Q}_{x,r} + \tau\mathbf{I})\mathbf{x}_r + \tau\mathbf{y}_r^\top \mathbf{x}_r \right\}. \end{aligned} \tag{A.5}$$

   Using the canonical form of the multivariate Gaussian distribution, we can write

   $$\pi(\mathbf{x}_r \mid \mathbf{y}_r, \phi_r) \propto \exp\left(-\frac{1}{2}\mathbf{x}_r^\top \mathbf{P}_r\mathbf{x}_r + \mathbf{b}_r^\top \mathbf{x}_r \right),$$

   where $\mathbf{P}_r = \mathbf{Q}_{x,r} + \tau\mathbf{I}$ and $\mathbf{b}_r = \mathbf{y}_r^\top \tau$. It follows that

   $$\mathbf{x}_r \mid \mathbf{y}_r, \phi_r \sim \mathcal{N}_T(\mathbf{P}_r^{-1}\mathbf{b}_r, \mathbf{P}_r).$$

Finally, from (A.4) and (A.5), we can write $\pi(\mathbf{y}_r \mid \phi_r)$ in (A.2) evaluated at $\mathbf{x}_r = 0$ as

$$\begin{aligned} \pi(\mathbf{y}_r \mid \phi_r)\Big|_{\mathbf{x}_r=0} &\propto \frac{|\mathbf{Q}_{x,r}|^{1/2}\exp\left(-\frac{1}{2}\tau\mathbf{y}_r^\top \mathbf{y}_r \right)}{|\mathbf{P}_r|^{1/2}\exp\left\{ -\frac{1}{2}(-\mathbf{P}_r^{-1}\mathbf{b}_r)^\top \mathbf{P}_r(-\mathbf{P}_r^{-1}\mathbf{b}_r) \right\}} \\ &= \frac{|\mathbf{Q}_{x,r}|^{1/2}}{|\mathbf{P}_r|^{1/2}\exp\left\{ -\frac{1}{2}(\mathbf{b}_r^\top \mathbf{P}_r^{-1}\mathbf{b}_r - \tau\mathbf{y}_r^\top \mathbf{y}_r) \right\}}. \end{aligned} \tag{A.6}$$

Next, from the posteriors $\tilde{\pi}(\phi_r \mid \mathbf{y}_r)$ and $\tilde{\pi}_{\text{smooth}}(\phi_r \mid \mathbf{y}_r)$ on the right-hand side of (A.1) that are computed in Steps 1 and 2, respectively, together with the likelihood term in (A.6), we can obtain the corresponding priors in (A.1). The right-hand side plot of Figure 3 shows these exact scaled log prior distributions.

Supplementary Materials

The supplementary material provides the R code to reproduce the example in Section 3.

Funding

This publication is based on research supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under award no. OSR-2018-CRG7-3742.

References

Abramowitz, M., and Stegun, I. A. (1964), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, New York: Dover, Equation (25.4.46). [356]

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., and Rue, H. (2019), "Non-Stationary Gaussian Models With Physical Barriers," *Spatial Statistics*, 29, 268–288. [350]

Bolin, D., and Lindgren, F. (2011), "Spatial Models Generated by Nested Stochastic Partial Differential Equations, With an Application to Global Ozone Mapping," *The Annals of Applied Statistics*, 5, 523–550. [350]

Castruccio, S., and Guinness, J. (2017), "An Evolutionary Spectrum Approach to Incorporate Large-Scale Geographical Descriptors on Global Processes," *Journal of the Royal Statistical Society*, Series C, 66, 329–344. [358]

Castruccio, S., and Stein, M. L. (2013), "Global Space–Time Models for Climate Ensembles," *The Annals of Applied Statistics*, 7, 1593–1611. [358]

Damian, D., Sampson, P. D., and Guttorp, P. (2001), "Bayesian Estimation of Semi-Parametric Non-Stationary Spatial Covariance Structures," *Environmetrics*, 12, 161–178. [349]

Edwards, M., Castruccio S., and Hammerling, D. (2020), "Marginally Parametrized Spatio-Temporal Models and Stepwise Maximum Likelihood Estimation," *Computational Statistics and Data Analysis*, 151, 107018. [349]

Fuglstad, G.-A., and Castruccio, S. (2020), "Compression of Climate Simulations With a Nonstationary Global Spatio-Temporal SPDE Model," *Annals of Applied Statistics*, 14, 542–559. [350]

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015), "Exploring a New Class of Non-Stationary Spatial Gaussian Random Fields With Varying Local Anisotropy," *Statistica Sinica*, 25, 115–133. [350]

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015), "Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields?," *Spatial Statistics*, 14, 505–531. [350]

——— (2019), "Constructing Priors That Penalize the Complexity of Gaussian Random Fields," *Journal of the American Statistical Association*, 114, 445–452. [355]

Gramacy, R. B. (2007), "tgp: An R package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models," *Journal of Statistical Software*, 19, 6. [357]

Hammerling, D. M., Michalak, A. M., and Kawa, S. R. (2012), "Mapping of CO_2 at High Spatiotemporal Resolution Using Satellite Observations: Global Distributions From OCO-2," *Journal of Geophysical Research: Atmospheres*, 117, D06306. [349]

Hartigan, J. A., and Wong, M. A. (1979), "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society*, Series C, 28, 100–108. [355]

Haslett, J., and Raftery, A. E. (1989), "Space-Time Modelling With Long-Memory Dependence: Assessing Ireland's Wind Power Resource," *Journal of the Royal Statistical Society*, Series C, 38, 1–50. [355]

Held, L., Schrödle, B., and Rue, H. (2010), "Posterior and Cross-Validatory Predictive Checks: A Comparison of MCMC and INLA," in *Statistical Modelling and Regression Structures*, eds. T. Kneib and G. Tutz, Berlin: Springer, pp. 91–110. [354]

Higdon, D. (1998), "A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190. [349]

Hildeman, A., Bolin, D., and Rychlik, I. (2019), "Spatial Modeling of Significant Wave Height Using Stochastic Partial Differential Equations," arXiv no. 1903.06296. [350]

Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014), "Spatial Models With Explanatory Variables in the Dependence Structure," *Spatial Statistics*, 8, 20–38. [350]

Jeong, J., Castruccio, S., Crippa, P., and Genton, M. G. (2018), "Reducing Storage of Global Wind Ensembles With Stochastic Generators," *Annals of Applied Statistics*, 12, 490–509. [358]

Kuusela, M., and Stein, M. L. (2018), "Locally Stationary Spatio-Temporal Interpolation of Argo Profiling Float Data," *Proceedings of the Royal Society A*, 474, 20180400. [349]

Lindgren, F., and Rue, H. (2008), "On the Second-Order Random Walk Model for Irregular Locations," *Scandinavian Journal of Statistics*, 35, 691–700. [352]

Lindgren, F., Rue, H., and Lindström, J. (2011), "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach," *Journal of the Royal Statistical Society*, Series B, 73, 423–498. [350]

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014), "Accounting for Spatially Varying Directional Effects in Spatial Covariance Structures," *Journal of the Royal Statistical Society*, Series C, 63, 103–122. [349]

Nychka, D., Hammerling, D., Krock, M., and Wiens, A. (2018), "Modeling and Emulation of Nonstationary Gaussian Fields," *Spatial Statistics*, 28, 21–38. [349]

Nychka, D., Wikle, C., and Royle, J. A. (2002), "Multiresolution Models for Nonstationary Spatial Covariance Functions," *Statistical Modelling*, 2, 315–331. [349]

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., and Bloom, S. (2011), "MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications," *Journal of Climate*, 24, 3624–3648. [355]

Risser, M. D. (2016), "Nonstationary Spatial Modeling, With Emphasis on Process Convolution and Covariate-Driven Approaches," arXiv no. 1610.02447. [349]

Risser, M. D., and Calder, C. (2017), "Local Likelihood Estimation for Covariance Functions With Spatially-Varying Parameters: The convoSPAT Package for R," *Journal of Statistical Software*, 81, 1–32. [357]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman and Hall/CRC. [350]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [352,355]

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017), "Bayesian Computing With INLA: A Review," *Annual Review of Statistics and Its Application*, 4, 395–421. [350,352]

Sampson, P. D., and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure," *Journal of the American Statistical Association*, 87, 108–119. [349]

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011), "Considering Covariates in the Covariance Structure of Spatial Processes," *Environmetrics*, 22, 487–500. [349]

Schmidt, A. M., and O'Hagan, A. (2003), "Bayesian Inference for Non-Stationary Spatial Covariance Structure via Spatial Deformations," *Journal of the Royal Statistical Society*, Series B, 65, 743–758. [349]

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017), "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors," *Statistical Science*, 32, 1–28. [355]

Stein, M. L. (1999), *Statistics for Spatial Data: Some Theory for Kriging*, New York: Springer. [355]

Tadić, J. M., Qiu, X., Yadav, V., and Michalak, A. M. (2015), "Mapping of Satellite Earth Observations Using Moving Window Block Kriging," *Geoscientific Model Development*, 8, 3311–3319. [349]

Tagle, F., Genton, M. G., Yip, A., Mostamandi, S., Stenchikov, G., and Castruccio, S. (2020), "A High-Resolution Bi-Level Skew-*t* Stochastic Generator for Assessing Saudi Arabia's Wind Energy Resources" (with discussion), *Environmetrics* (in press). [355]

Taylor, J. W., McSharry, P. E., and Buizza, R. (2009), "Wind Power Density Forecasting Using Ensemble Predictions and Time Series Models," *IEEE Transactions on Energy Conversion*, 24, 775. [355]

Yip, C. M. A. (2018), "Statistical Characteristics and Mapping of Near-Surface and Elevated Wind Resources in the Middle East," Ph.D. thesis, King Abdullah University of Science and Technology. [354]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [356]