ORIGINAL ARTICLE

High Resolution Global Precipitation Downscaling with Latent Gaussian Models and Nonstationary SPDE Structure

Jiachen Zhang[1] | **Matthew Bonas**[1] | **Diogo Bolster**[2] | **Geir-Arne Fuglstad**[3] | **Stefano Castruccio**[1]

[1]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, USA.

[2]Department of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, USA.

[3]Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Correspondence
Stefano Castruccio, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA
Email: scastruc@nd.edu

Funding information
This research is supported by grant NSF DMS 2014166.

Obtaining high-resolution maps of precipitation data can provide key insights to stakeholders to assess a sustainable access to water resources at urban scale. Mapping a non-stationary, sparse process such as precipitation at very high spatial resolution requires the interpolation of global datasets at the location where ground stations are available with statistical models able to capture complex non-Gaussian global space-time dependence structures. In this work, we propose a new approach based on capturing the spatially varying anisotropy of a latent Gaussian process via a locally deformed Stochastic Partial Differential Equation (SPDE) with a buffer allowing for a different spatial structure across land and sea. The finite volume approximation of the SPDE, coupled with Integrated Nested Laplace Approximation ensures feasible Bayesian inference for tens of millions of observations. The simulation studies showcase the improved predictability of the proposed approach against stationary and no-buffer alternatives. The proposed approach is then used to yield high resolution simulations of daily precipitation across the United States.

KEYWORDS
Latent Gaussian Model; Precipitation; Stochastic Partial Differential Equation; Integrated Nested Laplace Approximation

1 | INTRODUCTION

Accurate high-resolution information of precipitation data is essential to effective prediction and management of water resources (Clark et al., 2015). Dramatic improvements in modeling physical processes driving precipitation have resulted in more realistic simulations from global climate models and hence more reliable predictions. The high complexity of modern climate models, however, implies a computational and storage cost which limit the spatial resolution at which global climate simulations can be performed. As such, there are significant uncertainties and mismatches with observations, due to precipitation patterns that coarse resolutions do not sufficiently represent as they cannot capture the scale of the physical processes of interest (Wood et al., 2021). The consequences can be over- or under-attribution of a particular location or incorrect timing of events, that can for example be the difference between a local flooding or not (Sapountzis et al., 2021). It is therefore of high scientific interest to refine global predictions and produce maps of both probability of rain occurrence and precipitation intensity at a high spatial scale, in order to inform impact assessment models for flood resilience and agricultural models for drought predictions.

It is in principle possible to produce high resolution precipitation using a coarse global dataset as boundary condition for a regional weather model such as the Weather and Research Forecasting (WRF, Skamarock et al. (2019)). This *dynamical downscaling* approach (Sain et al., 2011) has the appealing advantage of producing physically consistent spatial fields at high resolution, but comes with a substantial associated cost in terms of computational and storage resources, as well as expertise for model setup that only few research centers, universities or businesses could afford. A more affordable solution lies in the formulation of an empirical relationship between global data and ground observations to be fit at locations where ground data are available. Under the assumption that this relationship is at least approximately valid at unobserved locations, high resolution maps can be produced by correcting the global dataset. This *statistical downscaling* approach (Berrocal et al., 2010) is fast, computationally affordable, and has a long established track record of success in the geoscience literature. In order to work, such approach requires that the global and the ground data are co-located, which is not a priori the case since global data are defined as averages over large areas. It becomes therefore necessary to use spatial statistical models to interpolate the global simulation values at the same locations of the ground observations, and to have an assessment of the uncertainty around these estimates.

Global spatial data require the formulation of specialized models whose theoretical properties are substantially different from spatial processes on Euclidean spaces. In fact, Gneiting (2013) highlighted how a valid process on the sphere with great circle distance could be achieved only with severe restrictions on the parameter space of the most widespread covariance model, the Matérn function. In the past two decades, new modeling approaches tailored for global data have emerged. Among them, Jun and Stein (2007, 2008) proposed to embed the sphere in a three dimensional space, consider a Matérn model and apply partial derivatives to achieve more flexibility. The proposed class of models was able to capture not just an isotropic behavior, but also *axial symmetry*, i.e., a nonstationary behavior across latitude (Jones, 1963). Jun (2011) generalized this approach to multivariate global processes. A fast and flexible spectral class of axially symmetric models was proposed in the case of gridded data by Castruccio and Stein (2013). The approach was then generalized to non-parametric spectral estimation (Castruccio and Genton, 2014), three-dimensional variables (Castruccio and Genton, 2016), different land/ocean behavior (Castruccio and Guinness, 2017) and also multivariate processes (Edwards et al., 2019). On the more theoretical side, substantial progress has been made in the determination of properties of high dimensional spheres for isotropic processes via basis decomposition see, e.g., Arafat et al. (2020); Porcu et al. (2020). We refer to Jeong et al. (2017); Porcu et al. (2018) for two recent reviews on the topic.

A novel, different perspective was raised in the seminal work of Lindgren et al. (2011), where a subclass of Matérn models was associated with the solution of a diffusion-reaction Stochastic Partial Differential Equation (SPDE) with the

Markov property and inference was performed with finite elements. The key insight of this approach, as far as global models are concerned, is that the original SPDE on the plane can be just adapted to the sphere, with the additional benefit of not requiring boundary conditions. The original ideas for non-stationarity in Lindgren et al. (2011) have been explored in several directions, from nested SPDE (Bolin and Lindgren, 2011) to models with physical barriers (Bakka et al., 2019). Recently, Fuglstad et al. (2015); Fuglstad and Castruccio (2020) extended this approach by parametrizing spatially varying anisotropy on the sphere through a spatially varying scalar and vector field, which resulted in a local deformation of the SPDE. The proposed approach showed promising results, but has been so far limited to the Gaussian case and generalization to non-Gaussian data is by no means straightforward, given the challenges in modeling non-Gaussian data and the computational overhead implied by these models.

In this work, we propose a non-Gaussian, non-stationary SPDE-based global spatio-temporal model with spatially varying anisotropy and a buffer between land and sea to account for abrupt changes in spatial dependence. Non-Gaussianity is modeled via a latent Gaussian model, i.e., by assuming that the non-Gaussian marginal behavior is conditionally independent across locations, and then the spatial dependence is captured via a latent process with a Gaussian structure. Inference is still achievable for very large datasets by means of 1) a sparse precision matrix of the latent Gaussian model emerging from the finite volume solution of the SPDE and 2) a fast approximation of the high-dimensional integrals required for posterior computation via Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)). The model is ideally suited to highly non-Gaussian data such as daily global precipitation, and it is then used to 1) fit global reanalysis data, 2) provide interpolated data at the same location as the ground observations, 3) downscale precipitation using both ground and interpolated data, so that 4) high resolution maps of precipitation are provided.

The work proceeds as follows. Section 2 introduces the data which will be used in this work. Section 3 details the methodology for the latent Gaussian model, specifically the temporal and the spatial component. Section 4 shows how inference is performed and how sparsity and numerical approximations alleviate the computational burden. Section 5 assesses numerically the posterior consistency, as well as the improved predictability of the proposed model against simpler alternatives. Section 6 applies the proposed model to the precipitation data and shows it can provide high resolution maps of daily precipitation across the continental United States. Section 7 concludes with a discussion. For reproducibility, at the end of this work we provide information about the repository where the code and data are available.

2 | DATA DESCRIPTION

We focus on daily global precipitation data from the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2, Gelaro et al. (2017)) produced by the NASA Global Modeling and Assimilation Office (GMAO). MERRA-2 is a reanalysis data product that incorporates observations from satellite instruments and is considered one of the best representations of the state of the Earth's system. The data is available on a regular grid with a resolution of $0.625^\circ \times 0.5^\circ$ in longitude and latitude, respectively, for a total of $n = 207,936$ locations. We focus on the year 2021, the latest year with a continuous record available, and we use the daily Maximum Rainfall Rate (MRR, in $\text{kg/m}^2 \cdot \text{s}$). To convert the MRR into precipitation, we divided it by the water density, 1,000 (kg/m^3), and convert the unit to millimeter by multiplying by 1,000, as well as multiply by $86{,}400s$ to obtain the daily precipitation. We assume that for each location, the MRR lasts for the whole day, which leads to some overestimation, as it can be clearly seen from the two different legend scales in Figure 1. The downscaling approach in Section 6 will be able to account for this by performing a linear transformation between (interpolated) MERRA-2 and USCRN.

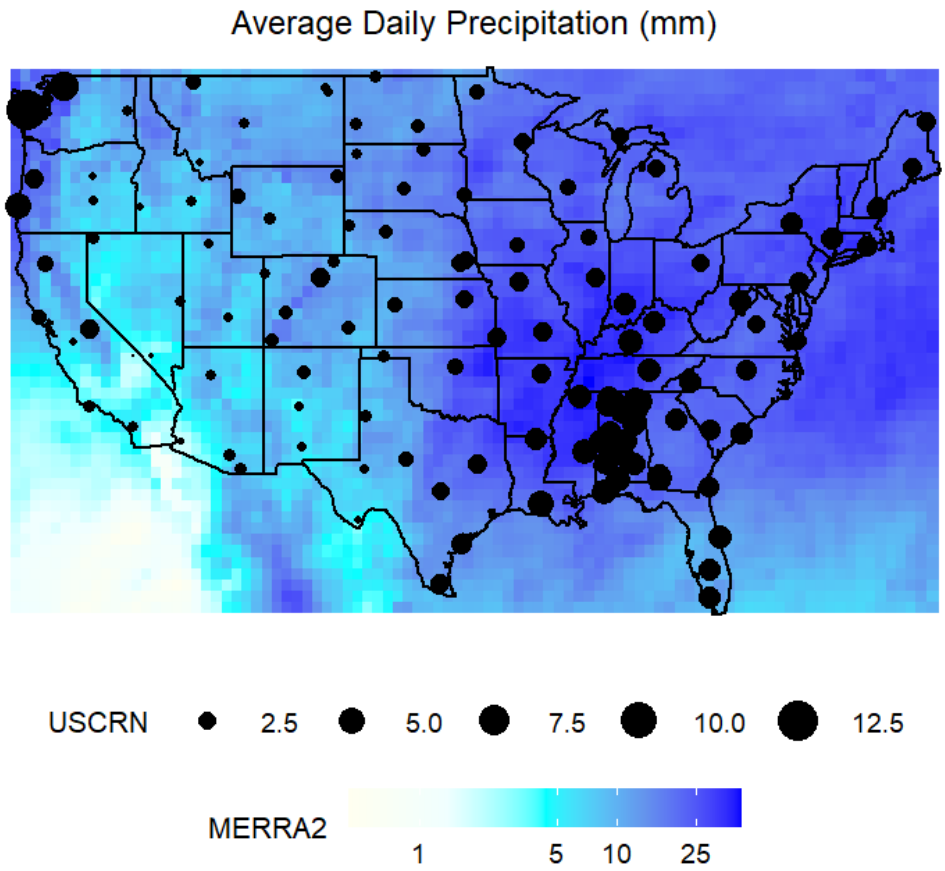

FIGURE 1 Average daily precipitation (in mm) for each USCRN site and MERRA-2 grid point from January 1st, 2021 through December 31th, 2021.

For ground observations, we consider the U.S. Surface Climate Reference Network (USCRN, NOAA (2022)), a data product containing continuous records from climate monitoring stations across the continental United States. The USCRN monitoring stations record measurements for total precipitation, measured in millimeters (mm), in real-time in 5-minute intervals. The data are collected with a Geonor T-200B precipitation gauge, whose maximum capacity is 600mm. This gauge uses a precipitation collection bucket which is surrounded by a wind/snow shield and heated in order to prevent ice buildup in cold regions. Three wires attached to this collection device vibrate with frequencies relative to the weight of the bucket, and these vibration frequencies are then converted to gauge depth (in mm). For this work, we consider data from 131 different monitoring stations post-processed to daily resolution forming a continuous record from January 1st, 2021 to December, 31th 2021. Figure 1 shows the locations of the USCRN sensors along with the average total daily precipitation throughout 2021. For comparison, the same figure also shows the average daily precipitation for the MERRA-2 grid points during the same time frame. It is readily apparent from this Figure that the regions of highest average daily precipitation are the northwest and southeast regions of the country whereas the drier region of the country spans from the eastern border of California through to the Mississippi River.

3 | METHODOLOGY

3.1 | Latent Gaussian Model

We propose a spatio-temporal latent Gaussian model (Rue et al., 2009), defined for a generic spatial point on the sphere $\mathbf{s} \in \mathbb{S}^2$ and time $t = 1, 2, \ldots$ as:

$$Y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t), \boldsymbol{\theta}_{\text{MRG}} \sim h(\mu(\mathbf{s}, t), \boldsymbol{\theta}_{\text{MRG}}), \tag{1a}$$

$$g(\mu(\mathbf{s}, t)) = \sum_{p=1}^{P} \beta_p f_p(\mathbf{s}) + f^{\text{time}}(\mathbf{s}, t) + \epsilon(\mathbf{s}), \tag{1b}$$

$$f^{\text{time}}(\mathbf{s}, t) = \sum_{k=1}^{K} \left\{ \zeta_k(\mathbf{s}) \sin\left(\frac{2\pi k t}{\delta}\right) + \zeta'_k(\mathbf{s}) \cos\left(\frac{2\pi k t}{\delta}\right) \right\}, \tag{1c}$$

where $h(\cdot)$ represents the marginal distribution of $Y(\cdot)$ conditional on the latent field and the hyperparameters, and belongs to the exponential family with some mean $\mu(\mathbf{s}, t)$, whose structure is determined by a latent Gaussian process through a link function $g(\cdot)$. The marginal parameters $\boldsymbol{\theta}_{\text{MRG}}$ characterize moments higher than the first, and could be empty. If the marginal distribution is Gaussian, we have $Y(\mathbf{s}) \sim \mathcal{N}(\mu(\mathbf{s}, t), \boldsymbol{\theta}_{\text{MRG}})$, and the link function $g(\cdot)$ is simply the identity function (Dunn and Smyth, 2018). For example, if the marginal distribution is the Bernoulli distribution instead, we have $Y(\mathbf{s}) \sim \mathcal{B}(\mu(\mathbf{s}, t))$, and the logit function can be chosen as the link function (Dunn and Smyth, 2018). We assume that the transformed mean in the latent space $g(\mu(\mathbf{s}, t))$ is modeled by a location specific time effect, $f^{\text{time}}(\mathbf{s}, t)$, $p = 1, \ldots, P$ location-specific covariates $f_p(\mathbf{s})$, and a spatial error $\epsilon(\mathbf{s})$. The time effect $f^{\text{time}}(\mathbf{s}, t)$ is described by K harmonics with parameters $\boldsymbol{\zeta}(\mathbf{s}) = (\zeta_1(\mathbf{s}), \ldots, \zeta_K(\mathbf{s}))^\top$ and $\boldsymbol{\zeta}'(\mathbf{s}) = (\zeta'_1(\mathbf{s}), \ldots, \zeta'_K(\mathbf{s}))^\top$, and the number of harmonics is chosen via a model selection metric, see the application and the supplementary material. If we assume that we have a sample observed at $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the total number of temporal parameters in equation (1c) is $\boldsymbol{\theta}_{\text{time}} = \{\boldsymbol{\theta}_{\text{time}}(\mathbf{s}_1), \ldots, \boldsymbol{\theta}_{\text{time}}(\mathbf{s}_n)\}$, where $\boldsymbol{\theta}_{\text{time}}(\mathbf{s}_i) = \{\boldsymbol{\zeta}(\mathbf{s}_i), \boldsymbol{\zeta}'(\mathbf{s}_i)\}$, for a total of $2Kn$ parameters. The period $\delta \in \{365, 366\}$ depends on the leap/no-leap year considered. We assume that the spatial random effect $\epsilon(\mathbf{s})$ is a realization from a mean-zero Gaussian random field independent in time, whose covariance function depends on some parameters $\boldsymbol{\theta}_{\text{space}}$ which will be specified in the next Section.

3.2 | Spatial Correlation Structure

The simplest models for the spatial dependence of $\epsilon(\mathbf{s})$ are stationary and isotropic, i.e., they assume that the dependence is a function of $\|\mathbf{s}_1 - \mathbf{s}_2\|$. Among them, one of the most popular choices is arguably the Matérn model, whose correlation between two locations $\mathbf{s}_1, \mathbf{s}_2$ is defined as (Stein, 1999)

$$\text{Corr}(\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2)) = C(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{\rho}\right)^{\nu} K_\nu \left(\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{\rho}\right),$$

where K_ν is the modified Bessel function of the second kind with smoothness parameter $\nu > 0$ (i.e., controlling the degree of mean squared differentiability) and range parameter $\rho > 0$. If inference is sought for a large dataset, a matrix comprising of the covariance among all locations could not be stored, and likelihood evaluation could become computationally challenging or just impossible. Instead of operating directly with the covariance matrix, a popular solution in the past decade has been to rely on the identification of a Gaussian process with Matérn covariance as the

(unique) stationary solution of the following fractional reaction diffusion SPDE (Whittle, 1954):

$$\left(\frac{1}{\rho^2} - \Delta\right)^{\nu/2+1/2} \epsilon(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \ \mathbf{s} \in \mathbb{R}^2, \tag{2}$$

where Δ is the Laplacian operator and $\mathcal{W}(\mathbf{s})$ is a spatial Gaussian white noise. By exploiting an 'explicit link' between a continuous Markov process when ν is integer in (2) and a discrete Gaussian Markov Random Field (GMRF), Lindgren et al. (2011) proved that if all locations are arranged on a 2D lattice, then the covariance structure of the GMRF could be approximated by applying the convolution of a sparse precision matrix. Moreover, any location that is not on the lattice could also be interpolated and approximated by means of a triangulation over the domain. Ultimately, this implies that the Matérn covariance can be approximated by a sparse precision matrix, and hence allow faster and feasible inference on the spatial structure of $\epsilon(\cdot)$. In this work, we rely on a similar SPDE defined on a sphere defined as

$$\left(\frac{1}{\rho^2} - \Delta_{\mathbb{S}^2}\right)^{\nu/2+1/2} \epsilon(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \ \mathbf{s} \in \mathbb{S}^2, \tag{3}$$

where $\Delta_{\mathbb{S}^2}$ is the Laplacian operator.

The aforementioned SPDE approach has clear computational advantages and can be generalized to allow for nonstationary constructs, while still yielding sparse precision matrices (Lindgren et al., 2011). In this work we rely on a spatially varying SPDE originally formulated in Fuglstad et al. (2019) for spatially varying anisotropy, but other approaches for spatially varying parameters (Lindgren et al., 2011) or nested SPDE (Bolin and Lindgren, 2011) have been proposed. We assume a location on the sphere has polar coordinates $\mathbf{s} = (L, l)$, where L is the latitude and l is the longitude. We introduce two terms: a vector field $\mathbf{v}(\cdot) = (v_1(\cdot), v_2(\cdot))^\top$ and a positive-valued scalar field $\rho(\cdot)$. We then define the inverse deformation tensor as:

$$\mathbf{G}(\mathbf{s})^{-1} = \rho(\mathbf{s})^2 \frac{\mathbf{I}_2 + \mathbf{v}(\mathbf{s})\mathbf{v}(\mathbf{s})^\top}{\sqrt{1 + \|\mathbf{v}(\mathbf{s})\|^2}},$$

where $\mathbf{I}_2$ is a 2×2 identity matrix. One can show that with the spatially varying metric tensor defined above, the distance along the direction $\mathbf{v}(\mathbf{s})$ is scaled by $1/(\rho(\mathbf{s})(1 + \|\mathbf{v}(\mathbf{s})\|^2)^{\frac{1}{4}}$. In the orthogonal direction of $\mathbf{v}(\mathbf{s})$, the distance is scaled by $(1 + \|\mathbf{v}(\mathbf{s})\|^2)^{\frac{1}{4}}/\rho(\mathbf{s})$. Therefore, the vector field $\mathbf{v}(\cdot)$ specifies the direction of the local anisotropic effect at each location, while $\rho(\cdot)$ represents its strength. After specifying the metric tensor $\mathbf{G}(\mathbf{s})$, it case be shown that an appropriate change of variable in the SPDE (3) yields (Fuglstad and Castruccio, 2020):

$$[|\mathbf{G}(\mathbf{s})|^{\frac{1}{2}} - \nabla \cdot |\mathbf{G}(\mathbf{s})|^{\frac{1}{2}} \mathbf{G}(\mathbf{s})^{-1} \nabla] \epsilon(\mathbf{s}) = |\mathbf{G}(\mathbf{s})|^{\frac{1}{4}} \mathcal{W}(\mathbf{s}), \ \mathbf{s} \in \mathbb{S}^2. \tag{4}$$

3.3 | Spherical Harmonics

Both the vector field $\mathbf{v}(\cdot)$ and the scalar field $\rho(\cdot)$ can be specified through basis decomposition such as spherical vector harmonics and spherical harmonics, respectively. However, a more flexible approach is necessary for global models, which must account not just for slowly changing nonstationarity, but also for abrupt changes dictated by large geographical descriptors such as land and ocean (Castruccio and Guinness, 2017). In order to formulate a valid

model via SPDE while still accounting for abrupt changes, we consider the buffering approach proposed by Bakka et al. (2019). More specifically, we use a buffer area along coastlines with a separate parameter that describes the multiplicative drop $d \in [0,1]$ in the strength of dependence in the buffer area for all triangles at the boundary $T_{\mathcal{B}}$, so that for each of the land/ocean domain we propose a separate spherical harmonics decomposition:

$$\log\{\rho^j(\mathbf{s})\} = \sum_{l=0}^{\mathcal{L}} \sum_{m=-l}^{l} \alpha_{ml}^j Y_l^m(\mathbf{s}) + d \times I(\mathbf{s} \in T_{\mathcal{B}}),$$

where α_{ml}^j are real-valued coefficients and $Y_l^m(\mathbf{s})$ are Laplace's spherical harmonic of degree l and order m and $j = \{\text{land, ocean}\}$ specifies the geographical descriptor where $\mathbf{s}$ is located. Similarly, the vector field $\mathbf{v}(\cdot)$ can be described as:

$$\mathbf{v}^j(\mathbf{s}) = \sum_{l=1}^{\mathcal{L}} \sum_{m=-l}^{l} \{E_{lm}^{(1,j)} \nabla Y_m^l(\mathbf{s}) + E_{lm}^{(2,j)} \hat{\mathbf{r}}(\mathbf{s}) \nabla \times Y_m^l(\mathbf{s})\},$$

where $\hat{\mathbf{r}}$ is the unit vector in the positive radial direction, $E_{lm}^{(1,j)}$ and $E_{lm}^{(2,j)}$ are real coefficients, $\mathcal{L}$ is the highest order in the bases. Additionally, in order to account for micro-scale variability, we assume that the process for both land and sea also has a nugget τ_j^2. In summary, the spatial parameters of the model are $\boldsymbol{\theta}_{\text{space}} = \left\{d, \left\{\tau_j^2, j \in \{\text{land, sea}\}\right\}, \left\{\alpha_{ml}^j, E_{lm}^{(1,j)}, E_{lm}^{(2,j)}, m = -l, \ldots, l; l = 1, \ldots, \mathcal{L}, j \in \{\text{land, sea}\}\right\}\right\}$, for a total of $6(\mathcal{L}^2 + 2\mathcal{L}) + 3$ parameters.

We use a priori independent standard normal distributions as priors for all parameters, with log transformation if they are constrained to be positive. The same setting is applied to the parameters used in simulation study and application. Given the overall large amount of data, the posterior results are not expected to substantially deviate for (reasonable) changes in the prior. Nevertheless, one could in principle use other more sophisticated choices such as penalized complexity priors (Simpson et al., 2017), even though the implementation with a user-defined model such as ours is not straightforward. We have added this remark in the prior discussion.

4 | INFERENCE

We propose a stepwise inference approach to reduce the overall dimension of the parameter space in each step. We first estimate $\boldsymbol{\theta}_{\text{time}}$ at each location independently, then $\boldsymbol{\theta}_{\text{space}}$ conditionally on the temporal parameters. In Edwards et al. (2020) it was shown that the stepwise approach results in an asymptotically consistent inference, and Castruccio and Guinness (2017) showed that uncertainty and bias propagation have small impact for large yet finite datasets such as the one we work with here.

4.1 | Step 1: Temporal Structure

In the first step, the inference is performed at each location independently. We redefine equation (1) as the following:

$$\begin{aligned} Y(\mathbf{s},t) &\sim h(\mu(\mathbf{s},t), \boldsymbol{\theta}_{\text{MRG}}), \\ g(\mu(\mathbf{s},t)) &= \textstyle\sum_{p=1}^{P} \beta_p f_p(\mathbf{s}) + \sum_{k=1}^{K} \left\{\zeta_k(\mathbf{s}) \sin\left(\frac{2\pi k t}{\delta}\right) + \zeta_k'(\mathbf{s}) \cos\left(\frac{2\pi k t}{\delta}\right)\right\}. \end{aligned} \tag{5}$$

The vector of temporal parameters $\boldsymbol{\theta}_{\text{time}}$ and the linear parameters $\beta_1, \ldots, \beta_p$ are estimated using least-squares and the parameters are considered fixed in the following inference steps. Once $\hat{\boldsymbol{\theta}}_{\text{time}}, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are obtained, conditional on them the spatial parameters $\boldsymbol{\theta}_{\text{space}}$ of the spatial process $\epsilon(\mathbf{s})$ can be estimated.

4.2 | Step 2: Spatial Covariance Structure

We define a collection of triangles $T_1, \ldots, T_{n_T}$ on the sphere, and use a finite volume method to discretize the SPDE in (4). We redefine the inverse matrix tensor as $\mathbf{G}(\mathbf{s})^{-1} = \rho(\mathbf{s})^2\mathbf{H}(\mathbf{s})$, where $|\mathbf{H}(\mathbf{s})| = 1$, and we integrate it over triangles T_i generated on a global mesh and seek for a piece-wise constant solution to the SPDE. For all triangles T_i, we have the following equality in distribution:

$$\left[\int_{T_i} \frac{1}{\rho(\mathbf{s})^2} - \nabla \cdot \mathbf{H}(\mathbf{s})\nabla\right] \epsilon(\mathbf{s}) \mathrm{d}V \overset{d}{=} \int_{T_i} \frac{1}{\rho(\mathbf{s})} \mathcal{W}(\mathbf{s}) \mathrm{d}V. \tag{6}$$

Here $\nabla\cdot$ is the divergence operator, ∇ is the gradient operator, and $\mathbf{H}(\cdot)$ is a 2×2 piecewise continuously differentiable diffusion tensor and $\mathrm{d}V$ is the surface measure on the triangles. This allows to translate the SPDE into a set of linear equations for a Gaussian vector that is assumed to be constant across each triangle.

Similarly to Bertolazzi and Manzini (2007); Fuglstad and Castruccio (2020), let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ be the vector of values at triangle center, then the following $n \times n$ matrix $\mathbf{A_H}$ could be calculated to describe a discrete approximation:

$$\left(\sum_{j=1}^{3} \int_{\sigma_{i,j}} (\mathbf{H}(\mathbf{s})\nabla\epsilon(\mathbf{s}))^\top \mathbf{n}_{i,j} \mathrm{d}\mathbf{s}\right)_{i=1}^{n} \approx \mathbf{A_H}\boldsymbol{\epsilon}.$$

Here, $\sigma_{i,j}$ represents the three faces of the triangle T_i and $\mathbf{n}_{i,j}$ is its outward-facing vector. Then, we combine this with a $n \times n$ diagonal matrix $\mathbf{D}$, in which $d_{ii} = |T_i|/\rho(x_i)^2$, so that we have:

$$\left(\int_{T_i} \frac{\epsilon(\mathbf{s})}{\rho(\mathbf{s})^2} \mathrm{d}\mathbf{s} - \sum_{j=1}^{3} \int_{\sigma_{i,j}} (\mathbf{H}(\mathbf{s})\nabla\epsilon(\mathbf{s}))^T \mathbf{n}_{i,j} \mathrm{d}\mathbf{s}\right)_{i=1}^{n} \approx (\mathbf{D} - \mathbf{A_H})\boldsymbol{\epsilon}.$$

With this approximation, the equality in distribution expressed in equation (6) can now be expressed as:

$$(\mathbf{D} - \mathbf{A_H})\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{L}),$$

where $\mathbf{L}$ is a $n \times n$ diagonal matrix with elements $l_{ii} = |T_i|/\rho(x_i)^2$. This implies that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{Q}^{-1})$, and $\mathbf{Q}$ is a sparse precision matrix defined as:

$$\mathbf{Q} = (\mathbf{D} - \mathbf{A_H})^\top \mathbf{L}^{-1} (\mathbf{D} - \mathbf{A_H}).$$

Therefore, the finite volume method ensures a sparse precision matrix, which mitigates the computational burden for large global data and boosts the computing speed of the nonstationary model during inference.

4.3 | Inference for Latent Gaussian model

In order to perform inference on the latent Gaussian Model, in this work we make use of the Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)) a method for Bayesian inference alternative to traditional Markov Chain Monte Carlo (MCMC), which could further ease the computational burden. INLA is a deterministic method for fast approximation of high dimensional integrals which takes advantage of computational properties of models that can be expressed as a latent GMRF. Thus, the INLA approach is used for performing the inference in this study. Under the proposed latent Gaussian Model structure, we have the observed data vector denoted here as $\boldsymbol{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^\top$ at locations $\mathbf{s}_i$ that can be described by hyperparameter vector $\boldsymbol{\theta}_{\text{space}}$. For simplicity, throughout this section, we will use $\boldsymbol{\theta}$ to represent hyperparameter vector $\boldsymbol{\theta}_{\text{space}}$. If conditioned on latent spatial field $\mathbf{X}$, the observations are marginally independent with likelihood:

$$\pi(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(Y(\mathbf{s}_i)|X(\mathbf{s}_i), \boldsymbol{\theta}),$$

where $\boldsymbol{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))^\top$ is a Gaussian field with mean zero and modeled by a SPDE approach with precision matrix $\mathbf{Q}(\boldsymbol{\theta})$. Therefore, the joint distribution of latent effect and hyperparameters can be written as:

$$\begin{aligned} \pi(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}) &\propto \pi(\boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta}) \textstyle\prod_{i=1}^{n} \pi(Y(\mathbf{s}_i)|X(\mathbf{s}_i), \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|Q(\boldsymbol{\theta})|^{1/2}\exp\{-\tfrac{1}{2}\mathbf{X}^\top Q(\boldsymbol{\theta})\mathbf{X}\} \textstyle\prod_{i=1}^{n} \pi(Y(\mathbf{s}_i)|X(\mathbf{s}_i), \boldsymbol{\theta}), \end{aligned}$$

where $|\boldsymbol{Q}(\boldsymbol{\theta})|$ is the determinant of the precision matrix. The main goal is to approximate the posterior marginals $\pi(X(\mathbf{s}_i)|\mathbf{Y})$, $\pi(\boldsymbol{\theta}|\mathbf{Y})$ and $\pi(\theta_j|\mathbf{Y})$. The marginal posterior distributions of interest can be written as:

$$\begin{aligned} \pi(X(\mathbf{s}_i)|\boldsymbol{Y}) &= \int \pi(X(\mathbf{s}_i)|\boldsymbol{\theta}, \boldsymbol{Y})\pi(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta} \\ \pi(\theta_j|\boldsymbol{Y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta}_{-j}. \end{aligned}$$

The key idea of INLA approach is to use the form above to construct nested approximations. The approximations of the marginals for the latent field $\pi(X(\mathbf{s}_i)|\mathbf{Y})$ are computed by approximating $\pi(\boldsymbol{\theta}|\mathbf{Y})$ and $\pi(X(\mathbf{s}_i)|\boldsymbol{\theta}, \mathbf{Y})$, and using numerical integration to integrate out $\boldsymbol{\theta}$. In other words, the posterior marginals of the latent parameter would be obtained by:

$$\tilde{\pi}(X(\mathbf{s}_i)|\boldsymbol{Y}) = \sum_k \tilde{\pi}(X(\mathbf{s}_i)|\boldsymbol{\theta}_k, \boldsymbol{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{Y}) \times \Delta_k,$$

where Δ_k are the weights associated with a vector $\boldsymbol{\theta}_k$ of hyperparameters in a grid.

5 | SIMULATION STUDIES

Throughout this section, we denote with NS-LS the proposed nonstationary latent Gaussian model (4) with land/sea effect with NS the nonstationary model with no land/sea effect. We further consider the stationary SPDE model (3), and denote with S-LS the model with land/sea effect and with S without it. In Section 5.1, we perform simulations from the Gaussian marginal distribution for NS-LS to numerically assess posterior consistency for both the hyperpa-

rameters and the resulting covariance matrix. In Section 5.2 and Section 5.3, we perform simulations from Gaussian and Bernoulli marginal distributions with identity and logit link, respectively, to assess the interpolation (kriging) performance of the NS-LS against NS, S-LS and S.

Since the key contribution of this work lies in the spatial component of the model, throughout this section we will assume a purely spatial process with no covariates. In other words, model (1) simplifies to

$$Y(\mathbf{s}) \sim h(\mu(\mathbf{s}), \boldsymbol{\theta}_{\text{MRG}}), \tag{7a}$$

$$g(\mu(\mathbf{s})) = \epsilon(\mathbf{s}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_{\text{space}})). \tag{7b}$$

In the Gaussian case we also have $\boldsymbol{\theta}_{\text{MRG}} = \sigma^2 = 0.05$, while in the Bernoulli case no marginal parameters are defined, so that $\boldsymbol{\theta}_{\text{MRG}} = \varnothing$.

For each simulation, we sample $n = 2,000$ data points on the unit sphere, and then draw the parameters of $\boldsymbol{\theta}_{\text{space}}$ from a Normal distribution with mean 1 and standard deviation 0.5, assume them fixed. Each simulation comprises of $n_r = 100$ replicates from the resulting covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{\text{space}})$, and could be intuitively interpreted as the number of independent replicates in time. We simulate data from a NS-LS model with $\mathcal{L} = 1$, so that there is a total of $6(\mathcal{L}^2 + \mathcal{L}) + 3 = 21$ hyperparameters. In other terms, we perform n_s independent simulations, each one comprising n_r replicates to aid the identifiability of the parameters.

5.1 | Posterior consistency in the Gaussian case

In order to numerically assess posterior consistency, for each simulation we consider an increasing number of replicates $n_r = 10, \ldots, 100$. Inference is performed assuming the same model (7) and with a mesh of $n_T = 2,000$ triangles. The choice of the number/size of triangles is dictated mostly by computational constraints. While it would be desirable to have a mesh as fine as possible, this would require in larger matrices and hence more challenging likelihood evaluation. On the other hand, a mesh too coarse would loose some fundamental structure in the spatial field, so there is a tradeoff. Our choice allowed for challenging yet not impossible inference. For varying levels of n_r, the hyperparameters' posterior distributions is retrieved and is compared with the true value. Posterior consistency can be empirically verified in the extent to which the hyperparameters' posterior distributions converges to the true parameters $\boldsymbol{\theta}_{\text{space}}$ as n_r increases.

TABLE 1 Median MSE (IQR) between the true hyperparameter and the posterior distribution across all simulations n_s for Gaussian case.

n_r	20	40	60	80	100
Median MSE (IQR)	0.32 (0.13)	0.25 (0.07)	0.14 (0.05)	0.05 (0.05)	0.01 (0.007)

Figure 2 shows the functional boxplot (Sun and Genton, 2011) for all n_s of the posterior distributions, for two hyperparameters for increasing values of realizations n_r. It is readily apparent how the posterior mean aligns to the true parameter value and the posterior standard deviations decreases as the replicates increase. While results are shown for NS-LS, similar patterns have been observed across all other models (NS, S-LS and S). Table 1 shows the median MSE and InterQuartile Range (IQR) of the hyperparameters posterior means estimated from the NS-LS model and the true values across all hyperparameters and across all $n_s = 100$ simulations. The median MSE decreases as the

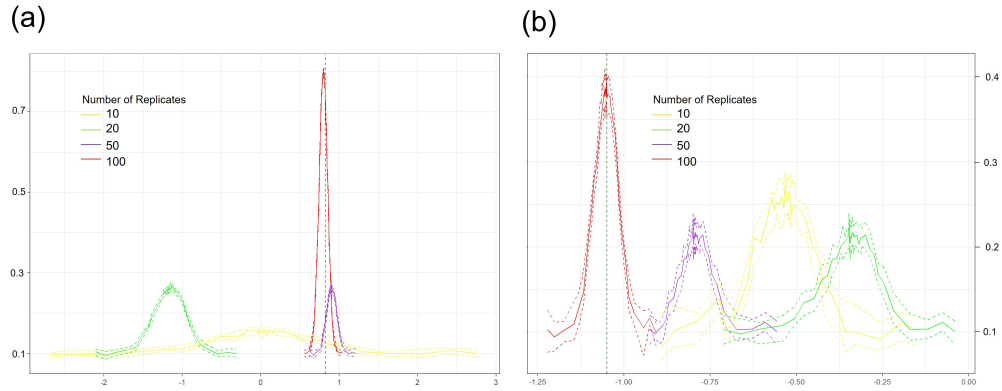

FIGURE 2 Functional boxplots (Sun and Genton, 2011) across n_s simulations of the posterior distribution of two hyperparameters (a) α_{11}^2 and (b) $E_{10}^{(2,2)}$ for different number of replicates n_r. The vertical dashed lines represent the true hyperparameter values.

replicates increases.

In order to perform a uniform comparison across all hyperparameters, whose number quickly becomes overbearing (e.g., with $\mathcal{L} = 4$ we would have $6(4^2 + 4) + 3 = 123$ hyperparameters), we also compare the covariance matrix implied by the hyperparameters with the true one. We assess the discrepancy in the covariances via the Kullback-Leibler Divergence (KLD), which in the case of an n-dimensional Gaussian distributions with mean $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ simplifies to:

$$\frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - n + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + \ln\left(\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_0}\right)\right).$$

In our case $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 = 0$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_{\text{space}})$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{\text{space}})$, so that the KLD measures the distance between the true and estimated covariance. The results as shown in Figure 3 for NS-LS (panel (a)) and S (panel (b)), where the functional boxplot (Sun and Genton, 2011) of KLD across all $n_s = 100$ simulations for an increasing number of realizations n_r is shown. The functional boxplot is used to report the envelope of the 50% central region (pink area), the median curve (black line) and the maximum non-outlying envelope (outer blue line). As in the case of the estimated parameters, we observe how even with a relatively small number of replicates in the training set, the estimated covariance is converging to the true one. In particular, after 40 replicates the estimated covariance is practically indistinguishable from the true one.

5.2 | Interpolation performance in the Gaussian case

In order to assess the interpolation performance, we perform inference on the hyperparameters for all four models and use them to interpolate at specified locations. We consider two cases (1) all n data points are used in the training set and interpolation is performed at the same sites (2) 92% of the n locations are considered in the training set, and the others 8% are withheld for crossvalidation. The test locations are located in within three selected areas indicated in Figure S1. Interpolation performance is measured with the MSE.

Results for both cases are reported in Table 2, and it is readily apparent how the MSE of NS-LS model is the small-

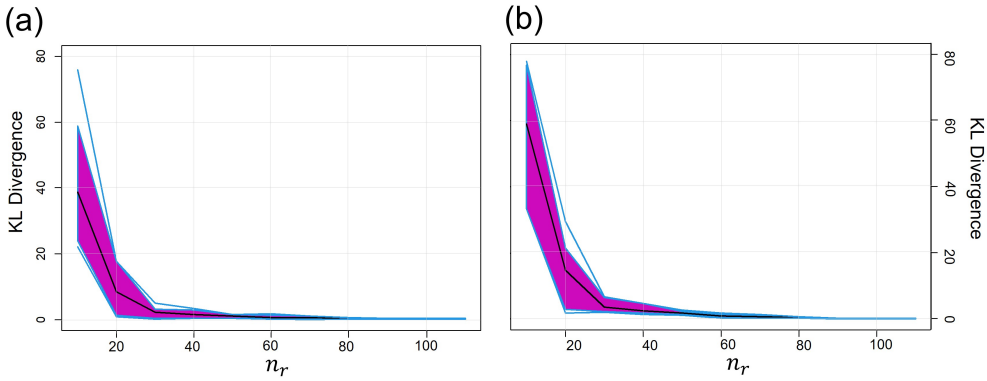

FIGURE 3 Functional boxplot across $n_s = 100$ simulations of the KLD between the true covariance matrix and the estimated one according to (a) NS-LS and (b) S-LS.

est among all four models for the both the all location case (1) and the cross-validation setting (2). More specifically, compared to the S-LS model, the NS-LS model shows an improvement of the median MSE across all locations by 14.6%. The NS-LS model also shows an appreciable improvement in MSE by 10.4% and 16.7%, compared with the NS and S models respectively. From these results it is clear how the land/sea effect and buffer area construction yield significant improvement when used in conjunction with the NS model.

TABLE 2 Comparison of interpolation performance across models. The first two rows show the median MSE (IQR) across all $n_s = 100$ simulations in the Gaussian case for both (1) all locations and (2) for crossvalidation. The last two rows show the median AUC (IQR) for the Bernoulli case across the same two cases.

Model	locations	NS-LS	S-LS	NS	S
Gaussian	All locations	90.12 (8.17)	105.47 (9.94)	100.55 (11.25)	108.25 (11.23)
Gaussian	Crossvalidation	9.11 (1.04)	21.88 (1.18)	19.35 (1.62)	21.39 (1.59)
Bernoulli	All locations	0.824 (0.048)	0.769 (0.074)	0.782 (0.051)	0.753 (0.050)
Bernoulli	Crossvalidation	0.707 (0.072)	0.676 (0.081)	0.672 (0.079)	0.641 (0.079)

5.3 | Interpolation performance in the Bernoulli case

We now assess predictability in the case of a Bernoulli distribution with logit link, and as in Section 5.2 we assess both the case where all locations are used as training set, as well as cross-validation with the same testing locations as before. Figure 4 shows the average differences across all $n_s = 100$ simulations between receiver operating characteristic curve (ROC) for NS-LS and S-LS, using S as reference for all locations and validation locations. The ROC for NS are visually indistinguishable to that of the S-LS model, so the results associated to that model are not show. The ROC difference in both cases show how the NS-LS model is uniformly better than the stationary S model (as the ROC difference is always positive), and also uniformly better than the S-LS model, especially in the middle of the curve. As expected, the extent of improvement of NS-LS is larger in the case of cross-validation (panel (b)), where the added

value of the model at unobserved locations is more apparent.

In order to have a comprehensive assessment across all possible choice of thresholds, we consider the area under the curve (AUC) of the ROC for all models and we report it in Table 2. In the best case of a perfect prediction, i.e., 100% true positive rate uniformly across the choice the threshold the AUC should equal 1, and in the worst case of a random guess it should be 0.5. The extent to which the AUC is close to 1 is a measure of predictive performance in this case. As it is shown in Table 2, the NS-LS outperforms every other model in both cases. More specifically, across all locations, the NS-LS yields an improvement by 7.2%, 5.3% and 9.4% for the S-LS, NS and S models respectively. These results agree with those presented in Section 5.2, for the use of the land/sea effect and buffer area construction definitively yields improved performance when included in the NS model.

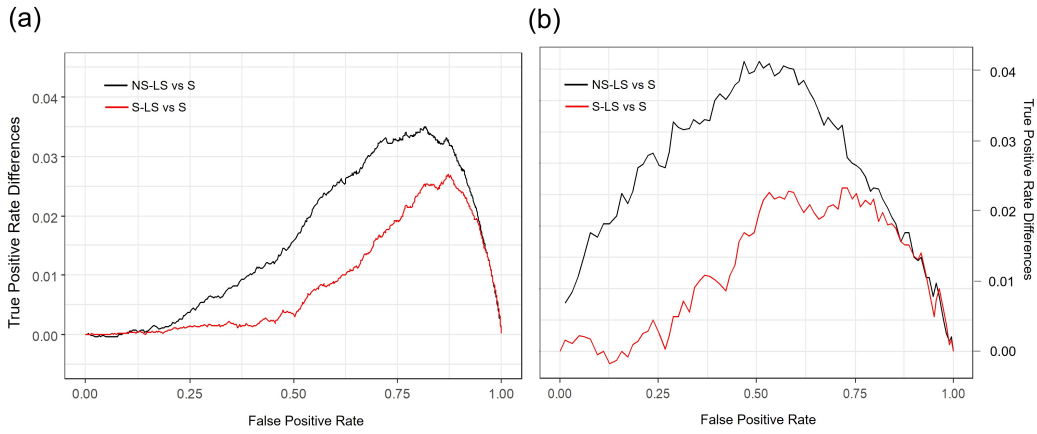

FIGURE 4 Average differences across all $n_s = 100$ simulations between ROC curves of NS-LS and S (black line), and S-LS and S (red line) for (a) all locations and (b) cross-validation. The ROC for NS are visually indistinguishable to that of the S-LS model, so the results associated to that model are not show.

6 | APPLICATION

In this section, we use the data detailed in Section 2 and the proposed latent Gaussian model with nonstationary SPDE introduced in Section 3 to estimate the global probability of a rain event and the precipitation intensity. In section 6.1, we discuss both the fit of the global MERRA-2 dataset and the downscaling approach to adjust interpolated MERRA-2 data with ground USCRN precipitation measurement. In section 6.2, we provide evaluation metrics to assess the model performance.

6.1 | Modeling global precipitation and downscaling

We initially focus on the MERRA-2 data and consider two global data sets 1) a binary rain occurrence event and 2) in case of rain, the actual rain intensity. We then fit the latent Gaussian model (1) with nonstationary SPDE (4) with $\mathcal{L} = 1$, using a Bernoulli marginal distribution with a logit link function $g(\cdot)$ for rain occurrence and a Gamma distribution with negative inverse link function for rain intensity. Validation for the choice of the marginal distribution can be found in

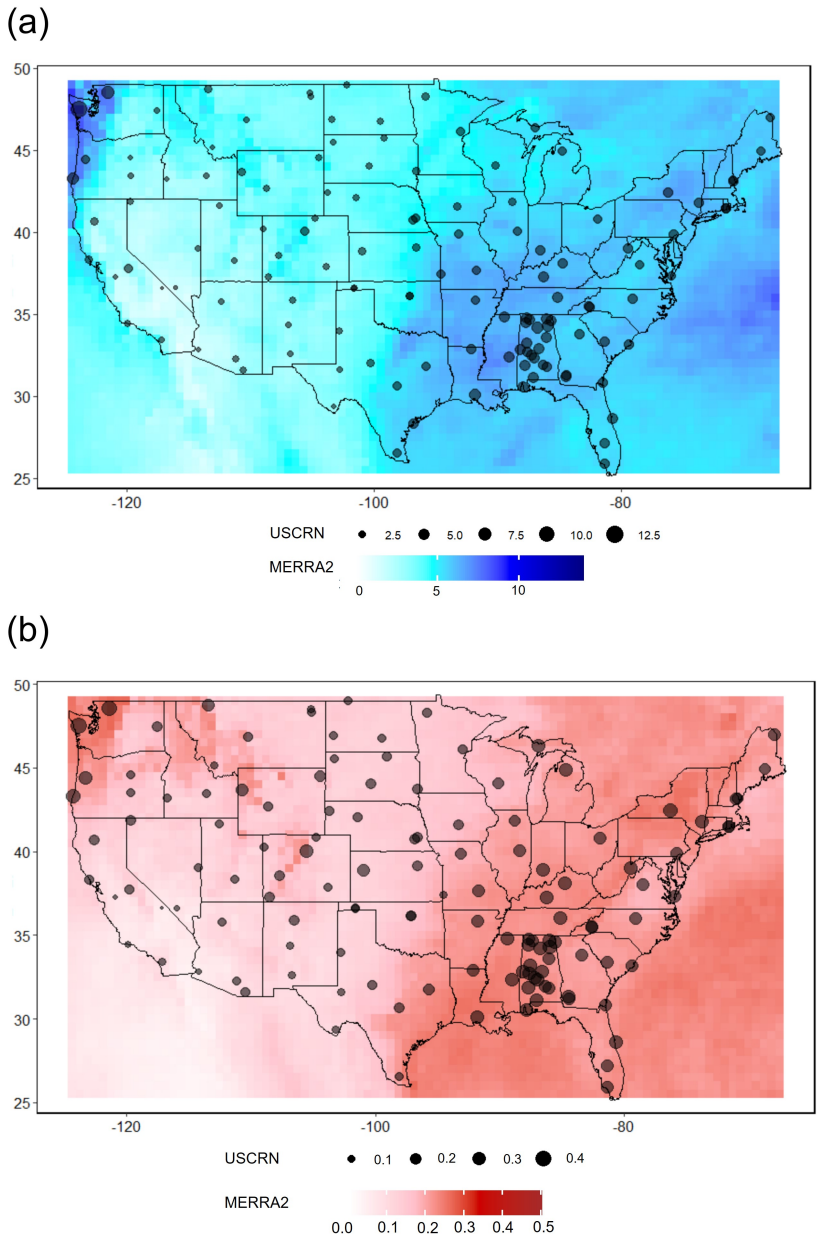

FIGURE 5 Average (a) daily precipitation and (b) precipitation probability. The global dataset is interpolated at the same sites as the ground observations according to the nonstationary global SPDE model (4), the linear model (9) is fit, and the resulting relationship is used to produce the downscaled maps.

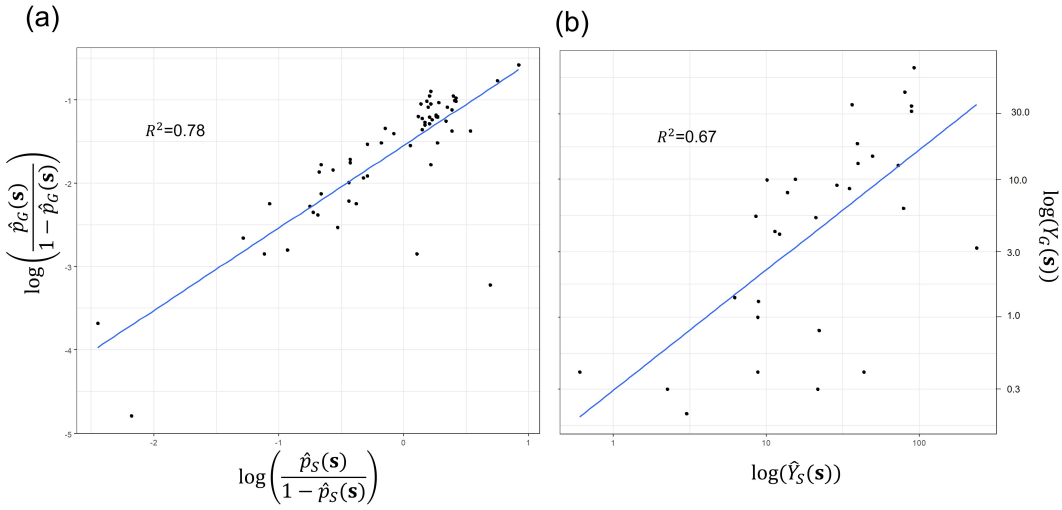

FIGURE 6 The fitted lines using downscaling models described in (a) equation (9a) and (b) equation (9b) on February 1st, 2021.

the supplementary along with Figure S3 showing the histogram of precipitation at 456 sample locations (resolution of $18.75° \times 15°$ in longitude and latitude) with estimated Gamma density. The sample locations are sparse in space to mitigate any spatial influences.

In both cases, no additional covariates are assumed, and we assume $K = 2$ harmonics for the temporal component, as it was shown to be the optimal choice according to the model selection in Figure S2. Formally, model (1) now specializes in the following two models:

$$\log\left(\frac{\mu(\mathbf{s},t)}{1-\mu(\mathbf{s},t)}\right) = f^{\text{time}}(\mathbf{s},t) + \epsilon_{\text{pr}}(\mathbf{s}), \quad \text{precipitation probability,} \tag{8a}$$

$$-\mu(\mathbf{s},t)^{-1} = f^{\text{time}}(\mathbf{s},t) + \epsilon_{\text{in}}(\mathbf{s}), \quad \text{precipitation intensity,} \tag{8b}$$

where $\epsilon_{\text{pr}}(\mathbf{s})$ and $\epsilon_{\text{in}}(\mathbf{s})$ are independent processes between them and in time. The histogram shows that precipitation intensity follows a Gamma distribution with shape parameter 0.826 and scale parameter 0.184. Inference is performed with a global triangulation of $n_T = 2,340$ triangles, of which $1,134$ are within the area of interest (contiguous United States), while the remaining $1,206$ cover the rest of the world.

The hyperparameters' posterior distributions is obtained and used to predict both the precipitation probability and intensity at the locations where the 131 USCRN ground observations locations are located, see Figure 1. These predictions are then adjusted (downscaled) to point resolution via linear regression. Since we perform downscaling independently for every time point, for simplicity we now drop the time dependence, and we denote as $Y_G(\mathbf{s})$ and $Y_S(\mathbf{s})$ the precipitation intensity for USCRN and MERRA2, respectively (G=ground, S=simulation), and with $p_G(\mathbf{s})$ and $p_S(\mathbf{s})$ the probability of precipitation occurrence. We further denote as $\hat{Y}_S(\mathbf{s})$ and $\hat{p}_S(\mathbf{s})$ the estimated intensity and probability of occurrence, respectively, according to the proposed SPDE model. Finally, we estimate the probability of precipitation occurrence for the USCRN data by fitting the latent Gaussian model (1) for each location independently as a time series model, i.e., assuming no spatial dependence and denote the estimate as $\hat{p}_G(\mathbf{s})$. We further assume a

linear relationship between USCRN and MERRA2 precipitation occurrence probability and intensity:

$$\log\left(\frac{\hat{p}_G(\mathbf{s})}{1-\hat{p}_G(\mathbf{s})}\right) = \beta_0^{(O)} + \beta_1^{(O)} \log\left(\frac{\hat{p}_S(\mathbf{s})}{1-\hat{p}_S(\mathbf{s})}\right) + \xi_O(\mathbf{s}), \quad \text{precipitation probability} \tag{9a}$$

$$\log\left(Y_G(\mathbf{s})\right) = \beta_0^{(I)} + \beta_1^{(I)} \log\left(\hat{Y}_S(\mathbf{s})\right) + \xi_I(\mathbf{s}), \quad \text{precipitation intensity} \tag{9b}$$

where $\xi_j(\mathbf{s}) \sim \mathcal{N}(0, \sigma_j^2), j \in \{O, I\}$ independent and identically distributed in space. A functional boxplot of the variogram of the residuals in Figure S4 (with each curve representing a different time point) lends support to the assumption of spatial independence of the error. The downscaling parameters $\beta_0^{(I)}$ and $\beta_1^{(I)}$ for precipitation intensity are then estimated using the ordinary least squares.

6.2 | Results and Evaluation

Downscaled probabilities of precipitation occurrence and precipitation intensity according to the aforementioned model are displayed in Figure 5(a) and (b), respectively, with the dark bubbles representing average values from the USCRN data. The prediction maps of the United States show high daily precipitation and high precipitation intensity around Seattle, while the lowest values can be found near Las Vegas, and overall the model prediction resembles the ground observation values across the United States. To evaluate the model performance, we calculate the root mean squared error (RMSE) for both probability of precipitation occurrence and precipitation intensity. The RMSE for intensity and probability of precipitation occurrence is 2.01 mm and 0.14 mm, respectively. In order to assess the value added by the smoothing of our SPDE model, we also perform downscaling with the linear models in (9), but assuming that no spatial model is fit, i.e., that the MERRA-2 data are not interpolated at the locations of the USCRN sites. Instead, we consider MERRA-2 data at their original resolution, and attribute to each USCRN site the value in the same cell. In other words, we consider as covariates $p_S(\mathbf{s}, t)$ and $Y_S(\mathbf{s}, t)$. The resulting RMSE for this model in the case of precipitation intensity and probability of precipitation occurrence is 82.74 mm and 0.28 mm, respectively. Therefore, the proposed SPDE approach has narrowed the discrepancy between MERRA-2 and USCRN significantly, as it has reduced the RMSE for precipitation intensity and probability of precipitation occurrence by 97.6% and 50%, respectively. Figure 6 shows the fitted lines using downscaling model in (9a) and (9b) on February 1$^{\text{st}}$, 2021. The R^2 for the two linear models are 0.78 and 0.67 for precipitation probability and intensity, respectively.

We also evaluate the model uncertainty by crossvalidation. First, we remove the data from one ground observation location and fit the model using the remaining observations. Next, we construct the 95% credibility interval for the posterior mean of the probability of precipitation occurrence or precipitation intensity at the removed location with the estimated posterior distributions of the hyperparameters of the model. Then, we repeat the same procedure for all the 131 locations in USCRN. Finally, we determine how many intervals among the 131 the 95% credibility intervals cover the true value. For precipitation, 93.1% (122/131) of the 95% credibility intervals cover the true value, while for probability of raining, 91.6% (120/131) of the 95% credibility intervals cover the true value.

7 | CONCLUSION AND DISCUSSION

In this work, we have proposed a novel non-stationary spatio-temporal SPDE model able to smooth both probability of precipitation occurrence and probability intensity from a global datasets. Such interpolated dataset is then used in conjuction with ground observation to produce high resolution (downscaled) precipitation maps, which allow to

predict what would ground observations would look like in unsampled location with a higher degree of accuracy compared to the original simulated data (i.e., the global data at their native resolution). One may in principle use MERRA-2 as a boundary condition to drive regional simulations with models such as WRF to obtain precipitation maps at equally high spatial resolution, with the added benefit of being able to produce predictions compliant with physical laws. Such dynamical downscaling approach is however considerably more involved as it require substantial computational and storage resources, as well as considerable expertise to set up WRF properly. As such, our proposed *statistical downscaling* approach is considerably faster and easier to implement without specialized computational resources. The proposed method of adjustment of a simulation via ground observation can also be seen as a bias correction approach, i.e., a method to correct simulations (see, e.g., Yuan et al. (2019); Kim et al. (2015) and Ho et al. (2012); Hawkins et al. (2013) for a general review). While a large body of literature in geoscience focuses on bias correction as a means to adjust the first (Hemer et al., 2012; Chen et al., 2012) and possibly the second moment (Teutschbein and Seibert, 2012; Li et al., 2019) of the marginal distribution, such approach can be used also to adjust non-Gaussian features, similarly to other recent efforts (Piani and Haerter, 2012; Vrac and Friederichs, 2014).

The proposed statistical model is scalable to future reanalysis data products with even higher spatial resolution, owing to the finite volume approximation of the SPDE generating the spatial model. Even more realistic downscaled patters could be generated if additional physical variables such as temperature and humidity could be considered as covariates. An incorporation of covariates could be performed either as the latent Gaussian model in (1b), as suggested in this work, or as as additional input of the scalar or vector field which dictate the deformation of the SPDE model. This could be implemented assuming either a linear contribution, or a non-linear one by means of neural networks (Hu et al., 2022). In principle, multiple variables could be modeled jointly. However, this would considerably increase both the methodological challenge and the computational overhead, as fast, flexible, multivariate and non-Gaussian global models are currently an active area of investigation (Genton and Kleiber, 2015).

While the proposed approach has many advantages over the chosen alternative models, there are also limitations which are ultimately inherited by and inextricably linked with the general modeling strategy chosen. *In primis*, the use of latent Gaussian models for non-Gaussian data has a long history in statistics, allows flexible hierarchical modeling while retaining computational affordability but by its own nature does not allow explicit control over some basic statistical properties such as the moments. Secondly, the use of SPDE requires a discretization, whose resolution is limited by how many triangles can be used in the domain: more triangles result in a more accurate solution but imply larger matrices. Finally, for this particular model, the choice of basis function for the scalar and vector field may require a lot of parameters, and other choices are possible.

Code and Data Availability

The code for this work is available at the GitHub repository: `https://github.com/Env-an-Stat-group/23.Zhang.public`. The MERRA-2 data are freely available from the Global Modeling and Assimilation Office (GMAO) at `https://disc.gsfc.nasa.gov/datasets/M2SDNXSLV_5.12.4/summary` and the ground observation data from the USCRN is also freely available at `https://www.ncei.noaa.gov/access/crn/qcdatasets.html`.

references

Arafat, A., Gregori, P. and Porcu, E. (2020) Schoenberg coefficients and curvature at the origin of continuous isotropic positive definite kernels on spheres. *Statistics & Probability Letters*, **156**, 108618.

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. URL: https://www.sciencedirect.com/science/article/pii/S221167531830099X.

Berrocal, V., Gelfand, A. and Holland, D. (2010) A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural Biological and Environmental Sciences*, **15**, 176–197.

Bertolazzi, E. and Manzini, G. (2007) On vertex reconstructions for cell-centered finite volume approximations of 2d anisotropic diffusion problems. *Mathematical Models and Methods in Applied Sciences*, **17**, 1–32. URL: https://doi.org/10.1142/S0218202507001814.

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. URL: https://doi.org/10.1214/10-AOAS383.

Castruccio, S. and Genton, M. G. (2014) Beyond axial symmetry: An improved class of models for global data. *Stat*, **3**, 48–55.

— (2016) Compressing an ensemble with statistical models: An algorithm for global 3d spatio-temporal temperature. *Technometrics*, **58**, 319–328.

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 (Applied Statistics)*, **66**, 329–344. URL: http://www.jstor.org/stable/44682577.

Castruccio, S. and Stein, M. L. (2013) Global space–time models for climate ensembles. *The Annals of Applied Statistics*, **7**, 1593 – 1611. URL: https://doi.org/10.1214/13-AOAS656.

Chen, L., Pryor, S. C. and Li, D. (2012) Assessing the performance of intergovernmental panel on climate change ar5 climate models in simulating and projecting wind speeds over china. *Journal of Geophysical Research: Atmospheres*, **117**, D24102. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD017533.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S. et al. (2015) Improving the representation of hydrologic processes in earth system models. *Water Resources Research*, **51**, 5929–5956.

Dunn, P. K. and Smyth, G. K. (2018) *Generalized Linear Models With Examples in R*. Springer New York, NY, 1 edn.

Edwards, M., Castruccio, S. and Hammerling, D. (2019) A multivariate global spatio-temporal stochastic generator for climate ensembles. *Journal of Agricultural, Biological and Environmental Sciences*, **24**, 464–483.

— (2020) Marginally parametrized spatio-temporal models and stepwise maximum likelihood estimation. *Computational Statistics and Data Analysis*, **151**, 107018.

Fuglstad, G.-A. and Castruccio, S. (2020) Compression of climate simulations with a nonstationary global SpatioTemporal SPDE model. *The Annals of Applied Statistics*, **14**, 542 – 559. URL: https://doi.org/10.1214/20-AOAS1340.

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. URL: http://www.jstor.org/stable/24311007.

Fuglstad, G.-A., Simpson, D., Lindgren, F. and Rue, H. (2019) Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, **114**, 445–452. URL: https://doi.org/10.1080/01621459.2017.1415907.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M. and Zhao, B. (2017) The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, **30**, 5419–5454. URL: https://doi.org/10.1175/JCLI-D-16-0758.1.

Genton, M. G. and Kleiber, W. (2015) Cross-Covariance Functions for Multivariate Geostatistics. *Statistical Science*, **30**, 147 – 163.

Gneiting, T. (2013) Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, **19**, 1327 – 1349. URL: https://doi.org/10.3150/12-BEJSP06.

Hawkins, E., Osborne, T. M., Ho, C. K. and Challinor, A. J. (2013) Calibration and bias correction of climate projections for crop modelling: An idealised case study over europe. *Agricultural and Forest Meteorology*, **170**, 19–31. Agricultural prediction using climate model ensembles.

Hemer, M., McInnes, K. and Ranasinghe, R. (2012) Climate and variability bias adjustment of climate model-derived winds for a southeast australian dynamical wave model. *Ocean Dynamics*, **62**, 87–104. URL: http://search.ebscohost.com.proxy.library.nd.edu/login.aspx?direct=true&db=aph&AN=70162084&site=ehost-live.

Ho, C. K., Stephenson, D. B., Collins, M., Ferro, C. A. T. and Brown, S. J. (2012) Calibration strategies: a source of additional uncertainty in climate change projections. *Bulletin of the American Meteorological Society*, **93**, 21–26.

Hu, W., Fuglstad, G.-A. and Castruccio, S. (2022) A stochastic locally diffusive model with neural network-based deformations for global sea surface temperature. *Stat*, **11**, e431.

Jeong, J., Jun, M. and Genton, M. G. (2017) Spherical Process Models for Global Spatial Statistics. *Statistical Science*, **32**, 501 – 513. URL: https://doi.org/10.1214/17-STS620.

Jones, R. (1963) Stochastic processes on a sphere. *Annals of Mathematical Statistics*, **34**, 213–218.

Jun, M. (2011) Non-stationary cross-covariance models for multivariate processes on a globe. *Scandinavian Journal of Statistics*, **38**, 726–747.

Jun, M. and Stein, M. L. (2007) An approach to producing space–time covariance functions on spheres. *Technometrics*, **49**, 468–479. URL: https://doi.org/10.1198/004017007000000155.

— (2008) Nonstationary covariance models for global data. *The Annals of Applied Statistics*, **2**, 1271 – 1289. URL: https://doi.org/10.1214/08-AOAS183.

Kim, K. B., Kwon, H.-H. and Han, D. (2015) Bias correction methods for regional climate model simulations considering the distributional parametric uncertainty underlying the observations. *Journal of Hydrology*, **530**, 568–579. URL: http://www.sciencedirect.com/science/article/pii/S002216941500774X.

Li, D., Feng, J., Xu, Z., Yin, B., Shi, H. and Qi, J. (2019) Statistical bias correction for simulated wind speeds over cordex-east asia. *Earth and Space Science*, **6**, 200–211. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018EA000493.

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 (Statistical Methodology)*, **73**, 423–498. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x.

NOAA (2022) U.S. climate reference network. https://www.ncei.noaa.gov/access/crn/. Last accessed 2022/10/25.

Piani, C. and Haerter, J. O. (2012) Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophysical Research Letters*, **39**, L20401.

Porcu, E., Alegria, A. and Furrer, R. (2018) Modeling temporally evolving and spatially globally dependent data. *International Statistical Review*, **86**, 344–377.

Porcu, E., Senoussi, R., Mendoza, E. and Bevilacqua, M. (2020) Reduction problems and deformation approaches to nonstationary covariance functions over spheres. *Electronic Journal of Statistics*, **14**, 890 – 916. URL: https://doi.org/10.1214/19-EJS1670.

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 (Statistical Methodology)*, **71**, 319–392. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x.

Sain, S. R., Nychka, D. and Mearns, L. (2011) Functional anova and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics*, **22**, 700–711.

Sapountzis, M., Kastridis, A., Kazamias, A., Karagiannidis, A., Nikopoulos, P. and Lagouvardos, K. (2021) Utilization and uncertainties of satellite precipitation data in flash flood hydrological analysis in ungauged watersheds. *Glob. Nest J*, **23**, 388–399.

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. URL: https://doi.org/10.1214/16-STS576.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, Powers, J. G., Duda, M. G., Barker, D. M. and X.-Y., H. (2019) A description of the advanced research wrf version 4. *Tech. rep.*, NCAR Tech. Note NCAR/TN-556+STR.

Stein, M. (1999) *Interpolation for Spatial Data: Some Theory for Kriging*. Springer, NY.

Sun, Y. and Genton, M. G. (2011) Functional boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316–334. URL: https://doi.org/10.1198/jcgs.2011.09224.

Teutschbein, C. and Seibert, J. (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, **456-457**, 12–29. URL: http://www.sciencedirect.com/science/article/pii/S0022169412004556.

Vrac, M. and Friederichs, P. (2014) Multivariate—intervariable, spatial, and temporal—bias correction. *Journal of Climate*, **28**, 218–237. URL: https://doi.org/10.1175/JCLI-D-14-00059.1.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449. URL: https://doi.org/10.1093/biomet/41.3-4.434.

Wood, R. R., Lehner, F., Pendergrass, A. G. and Schlunegger, S. (2021) Changes in precipitation variability across time scales in multiple global climate model large ensembles. *Environmental Research Letters*, **16**, 084022.

Yuan, Q., Thorarinsdottir, T. L., Beldring, S., Wong, W. K., Huang, S. and Xu, C.-Y. (2019) New approach for bias correction and stochastic downscaling of future projections for daily mean temperatures to a high-resolution grid. *Journal of Applied Meteorology and Climatology*, **58**, 2617–2632. URL: https://doi.org/10.1175/JAMC-D-19-0086.1.