

Spatial Statistical Downscaling for Constructing High-Resolution Nature Runs in Global Observing System Simulation Experiments

Pulong Ma

The Statistical and Applied Mathematical Sciences Institute
and Duke University

Emily L. Kang

Department of Mathematical Sciences, University of Cincinnati

Amy J. Braverman and Hai M. Nguyen

Jet Propulsion Laboratory, California Institute of Technology

Abstract

Observing system simulation experiments (OSSEs) have been widely used as a rigorous and cost-effective way to guide development of new observing systems, and to evaluate the performance of new data assimilation algorithms. Nature runs (NRs), which are outputs from deterministic models, play an essential role in building OSSE systems for global atmospheric processes because they are used both to create synthetic observations at high spatial resolution, and to represent the “true” atmosphere against which the forecasts are verified. However, most NRs are generated at resolutions coarser than actual observations from satellite instruments or predictions from data assimilation algorithms. Our goal is to develop a principled statistical downscaling framework to construct high-resolution NRs via conditional simulation from coarse-resolution numerical model output. We use nonstationary spatial covariance function models that have basis function representations to capture spatial variability. This approach not only explicitly addresses the change-of-support problem, but also allows fast computation with large volumes of numerical model output. We also propose a data-driven algorithm to select the required basis functions adaptively, in order to increase the flexibility of our nonstationary covariance function models. In this article we demonstrate these techniques by downscaling a coarse-resolution physical numerical model output at a native resolution of 1° latitude \times 1.25° longitude of global surface CO_2 concentrations to 655,362 equal-area hexagons.

Keywords: Basis functions; Change of support; Conditional simulation; Nonstationary covariance function; Observing system simulation experiment; Statistical downscaling

1 Introduction

Observing system simulation experiments (OSSEs) are widely used in atmospheric studies and climate monitoring to guide development of new observing systems including satellite missions and ground-based monitoring networks, and to evaluate performance of new data assimilation algorithms (e.g., Edwards et al., 2009; Zoogman et al., 2011; Errico et al., 2013; Atlas et al., 2015; Hoffman and Atlas, 2016); see Figure 1 for a diagram of a basic OSSE. In an OSSE, a simulated atmospheric field from a numerical model is used as the “truth” (termed a Nature Run, or NR) to produce synthetic observations by adding suitable measurement errors and other representative errors such as cloud mask (e.g., Atlas et al., 2015; Hoffman and Atlas, 2016). These synthetic observations are then fed to a data assimilation algorithm. Here, data assimilation refers to the process of fusing ground-based or airborne observations from observing systems such as satellites together with numerical model output, to infer the true state of geophysical processes; see Wikle and Berliner (2007) for a formal definition of data assimilation from a statistical perspective. The estimated true states are typically called forecasts in atmospheric sciences. Since OSSEs deal entirely with simulations, they provide a cost-effective approach to evaluating the impact of new observing systems and performance of new data assimilation algorithms, and can be used when actual observational data are not available. In particular, OSSEs can be employed to compare competing observing system designs (e.g., Atlas et al., 2015; Hoffman and Atlas, 2016). Moreover, unlike comparison against in-situ observations, the “truth” in an OSSE (that is, the NR) is known and uncontaminated, and thus can be directly used to better determine the accuracy and precision of forecasts. OSSEs are also extremely useful in understanding and quantifying capabilities of new satellite mission designs. For instance, Abida et al. (2017) use OSSEs to evaluate the potential improvement in estimating global surface carbon monoxide with the proposed Geostationary Coastal and Air Pollution Events Mission (GEO-CAPE). Liu et al. (2017) use OSSEs to investigate the potential for high spatial resolution satellite NO₂ observations to estimate surface NO₂ emissions.

The nature run (NR) is an essential component of an OSSE, since it provides the “true” state of the atmosphere, from which synthetic observations are constructed. As such, it provides a standard for evaluating the quality of forecasts from data assimilation algorithms (Figure 1). For OSSEs to be useful, it is essential that their NRs reflect characteristics of the real atmosphere.

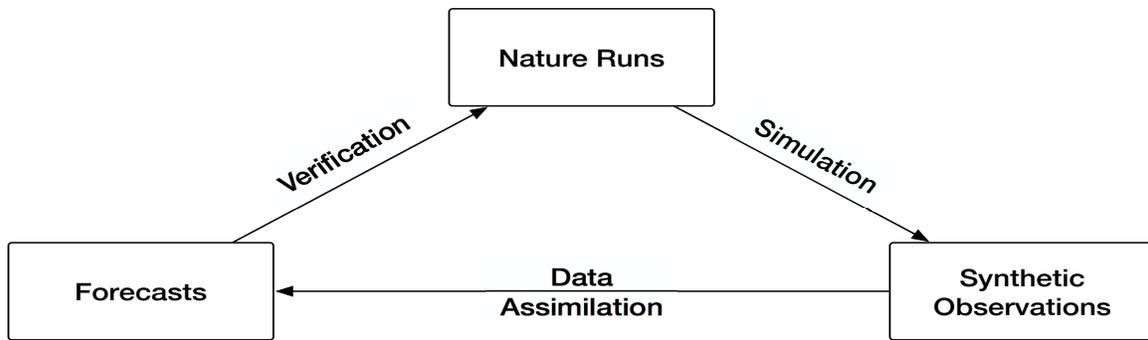


Figure 1. An OSSE system. The “nature runs” (NRs) are generated by or downscaled from numerical model outputs, and represent the assumed “true” state of the atmosphere. Synthetic observations are simulated by adding error components to NRs to mimic realistic cloud coverage and measurement errors. Forecasts from data assimilation algorithms are compared with NRs to evaluate the predictive performance of these algorithms.

Hence, NRs are often generated by state-of-the-art numerical atmospheric models driven by a set of ordinary and partial differential equations describing atmospheric chemistry and dynamics. When such a numerical model is run for the entire globe, or for a large geographical region, model complexity and computational limitations demand that many complex geophysical processes be simplified, which is referred to as parameterization (e.g., Brasseur and Jacob, 2017, pp. 342-398), and that the output be produced at low spatial and temporal resolution. On the other hand, with advancements in sensing technologies, new ground-based and space-based instruments are expected to provide observations with higher and higher spatial resolutions. Studies of local or regional air quality, or anthropogenic emissions and their impacts, require high spatial resolutions in order to be relevant for natural resource management and environmental policy decisions. As a result, NRs in OSSEs need to be generated at increasingly high spatial resolutions. The difference in resolution between numerical model output and that required to evaluate new sensor designs, or for local/regional scales analyses, motivates the need to construct finer resolution outputs from coarse-resolution numerical model outputs when generating NRs. This is called downscaling in remote sensing and atmospheric science (e.g., Gutmann et al., 2012; Atkinson, 2013; Glotter et al., 2014), and is an example of what is known as the change-of-support problem in spatial statistics (Cressie, 1993; Gotway and Young, 2002; Banerjee et al., 2014).

In the OSSE literature, heuristic methods are often used to downscale (sometimes called “preprocess”) model output to higher resolutions (e.g., Eskes et al., 2003; Errico et al., 2013; Tsai et al., 2014). These methods are computationally fast, but generally *ad hoc*. The NASA

Global Modeling and Assimilation Office (GMAO) has run a global non-hydrostatic dynamical core to perform cloud-system resolving experiments at resolutions as fine as 3.5 km within the NASA Goddard Earth Observing System global atmospheric model version 5 (GEOS-5) (Putman and Suarez, 2011). Simulations from this study provide high resolution NRs for geophysical variables, but they are computationally expensive. For example, generating two-year GEOS-5 simulations requires 61 days of computing with 7,200 cores and a total of 10.5×10^6 core hours. These requirements make it difficult for such an approach to be used widely (Webster and Duffy, 2015).

There is a vast literature on statistical downscaling and its applications in atmospheric studies. Most of this is devoted to developing methods for comparing, correcting, or calibrating numerical model output using observed data from physical experiments and monitoring stations (e.g., Lenderink et al., 2007; Kloog et al., 2011; Zhou et al., 2011; Berrocal et al., 2012; Reich et al., 2014). These methods require *both* numerical model output and adequate in-situ observations *together* in order to fit a statistical model that relates coarse-resolution model outputs to observations. Without explicitly dealing with the change-of-support problem, many methods use observations as response variables in simple linear regressions (or some variants), with the model outputs as explanatory variables. For instance, Guillas et al. (2008) develop a two-step linear regression procedure to downscale numerical model outputs, and adjust them to fit monitoring station data. There are also methods to extend these downscaling ideas to handle multiple variables in space and time (e.g., Berrocal et al., 2010). In contrast, Fuentes and Raftery (2005) address the change-of-support problem directly by expressing coarse-resolution output as the integral/average over high-resolution grid cells, and build models using observations as ground truth.

We also note that extensive work has been done on Gaussian process (GP) modeling for computer models from which output can be viewed as spatial (or spatio-temporal) data. However, the goals of computer model calibration (e.g., Sacks et al., 1989; Kennedy and O’Hagan, 2001), are different from those of spatial downscaling in this paper. In computer model calibration, a primary interest lies in estimating context-specific inputs by building a statistical model for *both* computer model outputs and physical observations *together*, where coarse-resolution model outputs are treated as low-accuracy data and a discrepancy term is included to model the gap between computer models and physical reality (e.g., Kennedy and O’Hagan, 2000, 2001; Higdon

et al., 2004). Since OSSEs are typically used to evaluate new observing systems or data assimilation algorithms when actual observational data are *not* available, our downscaling problem uses a *single* data source, that is, our method uses the numerical model output *solely, without any* additional observations. As such, we focus here on developing a principled model-based framework to construct high-resolution NRs directly from coarse-resolution numerical model output.

The problem presents the following challenges. (1) Atmospheric processes usually present nonstationary spatial structures. It is not realistic to model them with simple parametric spatial covariance functions such as the Matérn family. More flexible covariance function models are required. (2) Although the numerical model output is obtained at coarse spatial resolution, the size of the numerical model output can still be large. For our application, we construct global NRs of atmospheric CO₂ concentration at high spatial resolution using output from the PCTM/GEOS-4/CASA-GFED atmospheric model, which is coupled with biospheric, biomass burning, oceanic, and anthropogenic CO₂ flux estimates (Kawa et al., 2004, 2010). This model is referred to as PCTM hereafter. The PCTM output is generated at 1° latitude × 1.25° longitude, resulting in a dataset of size $M = 52,128$. Such a large dataset will cause computational bottlenecks for traditional spatial statistics methods due to the computational cost of the Cholesky factorization for the associated large covariance matrix, and memory limitations. This is the well-known “big n ” problem in spatial statistics (e.g., Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Nychka et al., 2015). (3) The difference between spatial resolutions of the numerical model output and the desired high-resolution NRs needs to be resolved carefully, and taken into account in the downscaling procedure. Since the numerical model output is assumed to be the “truth” at its corresponding resolution, it is important that the resulting downscaled NRs maintain certain essential properties of the coarse-resolution numerical model output. Previous work addresses some of these issues. For example, Datta et al. (2016) propose a model to handle massive spatial data, focusing on the second issue in particular. Gramacy and Apley (2015) provide a computationally efficient way to model large computer model outputs with nonstationary covariance, thus addressing both the first and second issues. To our knowledge, all three issues have not been discussed within a unified general modeling strategy in previous work.

Our approach to spatial downscaling follows the classical additive model in geostatistics, with several components that characterize different spatial variabilities. As suggested in Fuentes and Raftery (2005) and Craigmile et al. (2009), we address the change-of-support issue by building

our statistical model for the spatial process at the highest resolution of interest. To infer the process at fine resolution from coarse-resolution data, we use a spatial process model imposed with necessary assumptions to avoid an ill-posed inverse problem. Such a strategy is referred to as geostatistical regularization in geostatistics (Atkinson, 2001). To alleviate computational difficulties associated with large data volumes, we use a nonstationary covariance function model that combines a low-rank component for dimension reduction and a component with a diagonal covariance matrix and/or sparse precision matrix. We extend Fixed Rank Kriging (Cressie and Johannesson, 2008) and Fused Gaussian Process (Ma and Kang, 2018a) by making them more flexible. We capture nonstationary spatial variability with a forward stepwise algorithm for basis function selection. This includes both their locations and bandwidths of basis functions in the low-rank component.

Our basis function selection method differs from that of Tzeng and Huang (2017) in that our method is designed to learn nonstationary and localized features from the data. Tzeng and Huang (2017) uses information about data locations, but not data values, to specify basis functions. Other methods for nonstationary spatial modeling such as Katzfuss (2013) and Konomi et al. (2014) require computationally intensive reversible jump Markov chain Monte Carlo methods. In contrast, the forward stepwise algorithm we propose is simpler, more intuitive, and well-suited for parallel computing environments. The spatial downscaling procedure is also computationally efficient, and can produce not just one but many high-resolution statistical replicates from a coarse-resolution spatial field because it is based on conditional simulation.

Finally, the downscaled fields produced by our method maintain two important relationships with the coarse-resolution model output. First, the spatial dependence structure of the downscaling model is estimated, and thus inherited, from the coarse-resolution data. Second, when aggregated back to the coarse resolution, our high-resolution NRs match the coarse-resolution data exactly. Note that the numerical model output is considered as the best representation of the geophysical process of interest and is used as the “truth” at the coarse resolution. Any departure from this output solely due to the downscaling process cannot be physically justified. Therefore, we impose this aggregation requirement when generating downscaled NRs. Such a requirement has been emphasized in various environmental studies. Zhou and Michalak (2009) show that it is important to explicitly resolve the discrepancy between the coarse and fine resolutions by accounting for the relationship between the known, aggregated observations and the unknown

fine-resolution attributes. In climate science, a similar aggregation constraint (also called dynamic downscaling) is used in regional climate modeling (e.g., Wilby and Wigley, 1997), when coarse-resolution outputs from global climate models are used as boundary conditions for regional climate models. Meanwhile, from the statistical perspective, this aggregation requirement stems directly from the change-of-support property, where spatial observations at coarse resolution are defined as stochastic integrals over the fine-resolution process.

The remainder of this paper is organized as follows. In Section 2, we formulate the spatial statistical model for downscaling and inference, including parameter estimation and downscaling, via conditional simulation. The forward basis function selection algorithm is also described. In Section 3, we present simulation studies to evaluate the performance of the proposed downscaling method and basis function selection algorithm. In Section 4, the methodology is applied to surface CO₂ concentrations produced by PCTM at $M = 52,128$ grid cells to produce a high-resolution, downscaled field of $N = 655,362$ equal-area hexagons over the globe. Section 5 concludes with discussion and future work.

2 Methodology

In this section, we present our model-based spatial downscaling framework. In particular, Section 2.1 introduces the spatial statistical model and Section 2.2 presents basic derivations for parameter estimation via the EM algorithm. The downscaling procedure via conditional simulation is given in Section 2.3. Finally, Section 2.4 presents the forward basis function selection procedure.

2.1 The Spatial Statistical Model

Let $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ denote the atmospheric process of interest over a continuous spatial domain $\mathcal{D} \subset \mathbb{R}^d$ where $d \geq 1$ denotes the dimension of the spatial domain, and \mathbf{s} is a spatial location in \mathcal{D} . We consider the following additive model for the spatial process $Y(\cdot)$:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}), \mathbf{s} \in \mathcal{D}, \quad (2.1)$$

where $\mu(\cdot)$ is a trend term incorporating important covariates. The second term in (2.1), $w(\cdot)$, is assumed to be a Gaussian process with mean zero, and covariance function $C(\mathbf{s}_1, \mathbf{s}_2) \equiv$

$\text{cov}\{w(\mathbf{s}_1), w(\mathbf{s}_2)\}$ for $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$. The third term, $\epsilon(\cdot)$, is modeled as a Gaussian white-noise process in space with mean zero and variance $\sigma_\epsilon^2 > 0$, independent of $w(\cdot)$.

With a large number of observations, parameter estimation and prediction in kriging and Gaussian process regression become computationally infeasible. Many methods have been proposed to address this problem: covariance tapering (Furrer et al., 2006), composite likelihoods (Eidsvik et al., 2014), Gaussian Markov random fields (Lindgren et al., 2011) using low- and/or high-dimensional random vectors to induce covariance structures that result in low-rank covariance matrices and/or sparse precision matrices (e.g., Banerjee et al., 2008; Cressie and Johannesson, 2008; Sang and Huang, 2012; Nychka et al., 2015; Datta et al., 2016; Katzfuss, 2017), and local kriging (e.g., Hammerling et al., 2012; Gramacy and Apley, 2015; Tadić et al., 2015). Most of these methods rely on the assumption that $C(\mathbf{s}_1, \mathbf{s}_2)$ is stationary and/or has a prespecified parametric form, such as the Matérn covariance family. Cressie and Johannesson (2008) and Ma and Kang (2018a) take a different semiparametric approach. They use spatial basis functions, and allow the form of the covariance function to be flexible. We use such an approach for two reasons. First, it provides a globally valid spatial process model over the domain for *joint* inference at *all* locations. Second, it provides increased flexibility for modeling nonstationary behavior over a large spatial domain, as is typically the case for atmospheric processes.

We assume that the process $w(\cdot)$ is induced by two independent components:

$$w(\mathbf{s}) = \nu(\mathbf{s}) + \delta(\mathbf{s}), \mathbf{s} \in \mathcal{D}, \quad (2.2)$$

where the first term $\nu(\cdot)$ has a basis function representation: $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})^T \boldsymbol{\eta}$, $\mathbf{s} \in \mathcal{D}$. Here, $\mathbf{S}(\cdot) = (S_1(\cdot), \dots, S_r(\cdot))^T$ is a vector of r basis functions where r is relatively small, and so we call this component $\nu(\cdot)$ the *low-rank* component. The r -dimensional random vector $\boldsymbol{\eta}$ is assumed to follow the multivariate normal distribution with mean zero and covariance matrix \mathbf{K} . We further assume the $r \times r$ covariance matrix \mathbf{K} to be a general symmetric positive definite matrix without any pre-specified form, allowing for great flexibility in modeling spatial dependence structure. The basis functions are chosen to be compactly-supported. We describe a data-driven approach to automatically select the centers and bandwidths of these basis functions in Section 2.4. We will illustrate the procedure in Sections 3 and 4. In particular, we will show that the choice of basis functions can have a substantial impact on inference, especially predictive performance.

For the second term in (2.2), we assume that the process $\delta(\cdot)$ is induced by a high-dimensional

random vector $\boldsymbol{\xi}$: $\delta(\mathbf{s}) = \mathbf{B}(\mathbf{s})^T \boldsymbol{\xi}$, $\mathbf{s} \in \mathcal{D}$. The vector $\boldsymbol{\xi}$ is defined by a discretization of the spatial domain \mathcal{D} into a fine-resolution lattice (which can be irregular) of N grid cells $\mathcal{D} \equiv \cup\{\mathbf{s}_i \in \mathcal{A}_i : i = 1, \dots, N\}$ with $\{\mathcal{A}_i : i = 1, \dots, N\}$ called basic areal units (BAUs), as suggested in Nguyen et al. (2012). In practice, these BAUs are determined by the finest resolutions of interest that are required to construct synthetic observations and forecasts in OSSEs. Note that N can be very large, and can be much larger than the number of observed data points. We further model $\boldsymbol{\xi}$ with a Gaussian random Markov field, particularly, the spatial conditional autoregressive (CAR) model. The precision matrix of $\boldsymbol{\xi}$ is assumed to be: $\mathbf{Q} \equiv (\mathbf{I} - \gamma \mathbf{H})/\tau^2$, which is induced by the full conditional distributions $\xi_i | \{\xi_j : j \neq i\} \sim N(\gamma \sum_{j=1}^N H_{ij} \xi_j, \tau^2)$, for $i = \dots, N$. Here, γ is called the spatial dependence parameter. If $\gamma = 0$, the elements in $\boldsymbol{\xi}$ will be independent. The parameter τ^2 is called the conditional marginal variance. The matrix $\mathbf{H} \equiv (H_{ij})_{i,j=1,\dots,N}$ is an $N \times N$ proximity matrix with $H_{ii} = 0$, and $H_{ij} = 1$ if \mathcal{A}_j is a neighbor of \mathcal{A}_i and is zero otherwise, where $H_{ij} = 0$ for $i \neq j$ implies that ξ_i and ξ_j are *conditionally* independent given $\{\xi_\ell : \ell \neq i, j\}$. To specify the neighborhood structure, one can choose a threshold distance in terms of spatial adjacency; see Chapter 6 of Cressie (1993) for more details on specification of a CAR model and examples of neighborhood structures. Following Ma and Kang (2018a), we choose the basis function vector $\mathbf{B}(\cdot) \equiv (B_1(\cdot), \dots, B_N(\cdot))^T$ to be a vector of incidence functions, where $B_i(\mathbf{s}) = 1$ if the location \mathbf{s} is in \mathcal{A}_i and zero otherwise, for $i = 1, \dots, N$. For more complicated precision structures, other types of basis functions such as piecewise linear basis functions and Wendland basis functions can be used for $B_i(\cdot)$; for details see Ma and Kang (2018a).

The model for the process $Y(\cdot)$ is thus,

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \mathbf{S}(\mathbf{s})^T \boldsymbol{\eta} + \mathbf{B}(\mathbf{s})^T \boldsymbol{\xi} + \epsilon(\mathbf{s}), \quad (2.3)$$

with covariance function

$$C_{FGP}(\mathbf{s}_1, \mathbf{s}_2) \equiv \text{cov}\{Y(\mathbf{s}_1), Y(\mathbf{s}_2)\} = \mathbf{S}(\mathbf{s}_1)^T \mathbf{K} \mathbf{S}(\mathbf{s}_2) + \mathbf{B}(\mathbf{s}_1)^T \mathbf{Q}^{-1} \mathbf{B}(\mathbf{s}_2) + \sigma_\epsilon^2 I(\mathbf{s}_1 = \mathbf{s}_2). \quad (2.4)$$

We call this model the Fused Gaussian Process model, referred to simply as FGP in the discussion that follows. Previous work in Cressie and Kang (2010) and Nguyen et al. (2012) assume that the process $\delta(\cdot)$ is a spatial white noise process with mean zero and variance σ_ξ^2 . This model is called the spatial random effects model, and the resulting method is referred to as Fixed Rank Kriging

(FRK) hereafter. The corresponding covariance function is given by,

$$C_{FRK}(\mathbf{s}_1, \mathbf{s}_2) \equiv \text{cov}\{Y(\mathbf{s}_1), Y(\mathbf{s}_2)\} = \mathbf{S}(\mathbf{s}_1)^T \mathbf{K} \mathbf{S}(\mathbf{s}_2) + \sigma_\xi^2 I(\mathbf{s}_1 = \mathbf{s}_2) + \sigma_\epsilon^2 I(\mathbf{s}_1 = \mathbf{s}_2). \quad (2.5)$$

In the case studied here, rather than observing the process $Y(\cdot)$ at BAU-level, we only have the aggregated values of this process at coarse spatial resolution, i.e., the coarse-resolution numerical output. The difference between the spatial resolution of this numerical model output and the NRs we need is a type of change-of-support problem (e.g., Cressie, 1993; Gotway and Young, 2002; Wakefield and Lyons, 2017). Suppose that the numerical model output is obtained over a total of M coarse grid cells, $\{\Delta_i \subset \mathcal{D} : i = 1, \dots, M\}$ in the spatial domain. We call the region Δ the *support* of $Y(\Delta)$ and define $Y(\Delta)$ as the average of $Y(\cdot)$ over its support:

$$Y(\Delta) := \frac{1}{|\Delta|} \int_{\mathbf{s} \in \Delta} Y(\mathbf{s}) d\mathbf{s}, \quad (2.6)$$

where $|\Delta| > 0$ is the volume of Δ . This stochastic integral is defined as a mean-square limit, which can be approximated by an appropriately weighted sum (see Cressie, 1993, Section 5.2). An alternative and more flexible definition is $Y(\Delta) \equiv \int_{\mathbf{s} \in \Delta} Y(\mathbf{s}) h(\mathbf{s}) d\mathbf{s}$, where $h(\mathbf{s})$ is called the impulse response or the point spread function in remote sensing science. Eq. (2.6) assumes additionally that $h(\mathbf{s}) = \frac{1}{|\Delta|}$, if $\mathbf{s} \in \Delta$, and 0, otherwise. It is also possible to use a non-constant impulse response (e.g., Cracknell, 1998). Although we focus on the constant case (2.6) in this work, it is straightforward to apply our method and algorithms with a more general impulse response.

To relate Δ to fine-resolution BAUs, the integral (2.6) is approximated by

$$Y(\Delta) \approx \frac{1}{\sum_{\mathbf{s} \in \mathcal{D}} I(\mathbf{s} \in \Delta)} \sum_{\mathbf{s} \in \mathcal{D}} I(\mathbf{s} \in \Delta) \cdot Y(\mathbf{s}), \quad (2.7)$$

where the summation is taken over the discretized domain \mathcal{D} with N BAUs, and $I(\mathbf{s} \in \Delta)$ is an indicator function that is equal to one if the centroid \mathbf{s} of \mathcal{A} lies in the region Δ , and is equal to zero otherwise.

Let $\tilde{\mathbf{Y}} \equiv (Y(\Delta_1), \dots, Y(\Delta_M))^T$ be a vector of the numerical model output for M coarse-resolution grid cells. We are interested in recovering the process $Y(\cdot)$ at BAU-level, $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_N))^T$, from $\tilde{\mathbf{Y}}$. We define the so-called $M \times N$ aggregation matrix \mathbf{A} whose

(i, j) -th entry a_{ij} is given by,

$$a_{ij} \equiv \frac{I(\mathbf{s}_j \in \Delta_i)}{\sum_{\mathbf{s} \in \mathcal{D}} I(\mathbf{s} \in \Delta_i)}, \quad i = 1, \dots, M; j = 1, \dots, N. \quad (2.8)$$

Recall that the FGP and FRK models are defined at the BAU resolution. Fortunately, it is straightforward to obtain the *marginal* distributions of \mathbf{Y} and $\tilde{\mathbf{Y}}$: $\mathbf{Y} \sim \mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\tilde{\mathbf{Y}} \sim \mathcal{N}_M(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$, where $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_N))^T$ is a vector of trends terms defined at BAU-level. The covariance matrix of \mathbf{Y} can be easily derived from the covariance functions given in Eq. (2.4) and Eq. (2.5):

$$\boldsymbol{\Sigma} \equiv \text{cov}(\mathbf{Y}) = \mathbf{S}\mathbf{K}\mathbf{S}^T + \boldsymbol{\Sigma}_\delta + \mathbf{V}, \quad (2.9)$$

where \mathbf{S} is the $N \times r$ matrix with its i -th row defined as the transpose of $\mathbf{S}(\mathbf{s}_i)$ for $i = 1, \dots, N$. The $N \times N$ matrix $\boldsymbol{\Sigma}_\delta$ is obtained from the process $\delta(\cdot)$, and takes the form $\boldsymbol{\Sigma}_\delta = \mathbf{Q}^{-1}$ in FGP and $\boldsymbol{\Sigma}_\delta = \sigma_\xi^2 \mathbf{I}_N$ in FRK. The last term $\mathbf{V} \equiv \sigma_\epsilon^2 \mathbf{I}_N$ is an $N \times N$ matrix resulting from the process $\epsilon(\cdot)$ in Eq. (2.1). When the spatial dependence parameter $\gamma = 0$, the covariance matrix from FGP is reduced to that of FRK. Under both models, the number of basis functions, r , is assumed to be much smaller than the number of data points, M , which provides substantial dimension reduction.

Note that both FRK and FGP inherit an additive structure widely used in modeling spatial data. The modified predictive process (Finley et al., 2009), the full scale approximation (Sang and Huang, 2012), and the multi-resolution approximation (Katzfuss, 2017), all construct additive models based the assumption of a particular parametric covariance function such as the Matérn covariance function. Ba and Joseph (2012) use a combination of two spatial covariance structures together, but this requires empirical constraints on parameters, and is not designed to handle large datasets. Comparing FRK and FGP, the latter introduces spatial dependence for the term ξ so that the resulting model can give better predictive performance than typical low-rank models including FRK. Numerical examples to demonstrate the robust predictive performance of FGP for different covariance functions can be found in Ma and Kang (2018a). They also discuss other assumptions on ξ , besides the CAR model for FGP.

2.2 Parameter Estimation

The parameters of the models proposed in Section 2.1 are estimated using likelihood-based approaches. We follow the approach of Cressie and Kang (2010) and Nguyen et al. (2012) and assume that the variance parameter of $\epsilon(\cdot)$, σ_ϵ^2 , is known from independent validation data or estimated separately by examining the empirical variograms (Kang et al., 2010). The trend term is assumed to be $\mu(\cdot) = \mathbf{X}(\cdot)^T \boldsymbol{\beta}$ with a p -dimensional vector of known covariates $\mathbf{X}(\cdot) = (X_1(\cdot), \dots, X_p(\cdot))^T$ and corresponding unknown coefficients $\boldsymbol{\beta}$. Let $\boldsymbol{\theta}$ denote the set of parameters to be estimated. For FGP, $\boldsymbol{\theta}$ consists of $\{\boldsymbol{\beta}, \mathbf{K}, \tau^2, \gamma\}$, and for FRK, $\boldsymbol{\theta} \equiv \{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\}$.

Recall that the ‘‘observed’’ data come from the numerical model output, $\tilde{\mathbf{Y}}$, which is assumed to follow the multivariate normal distribution with mean $E(\tilde{\mathbf{Y}}) = \mathbf{A}\boldsymbol{\mu}$, and covariance matrix $\text{cov}(\tilde{\mathbf{Y}}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$. Up to an additive constant, the corresponding twice-negative-marginal-log-likelihood function is,

$$-2 \ln L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) = \ln |\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T| + (\tilde{\mathbf{Y}} - \mathbf{A}\mathbf{X}\boldsymbol{\beta})^T (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1} (\tilde{\mathbf{Y}} - \mathbf{A}\mathbf{X}\boldsymbol{\beta}), \quad (2.10)$$

where \mathbf{X} is the $N \times p$ design matrix associated with the covariates, and the covariance matrix $\boldsymbol{\Sigma}$ is given in Eq. (2.9). Note that evaluating Eq. (2.10) requires inverting and calculating the log-determinant of the $M \times M$ matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$. Specifically, we have

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T = (\mathbf{A}\mathbf{S})\mathbf{K}(\mathbf{A}\mathbf{S})^T + \mathbf{D}^{-1}, \text{ where } \begin{cases} \mathbf{D} = [\mathbf{A}(\sigma_\xi^2\mathbf{I})\mathbf{A}^T + \sigma_\epsilon^2\mathbf{A}\mathbf{A}^T]^{-1} \text{ for FRK;} \\ \mathbf{D} = (\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1} \text{ for FGP.} \end{cases} \quad (2.11)$$

Although the numerical model output, $\tilde{\mathbf{Y}}$, is defined at coarse spatial resolution, the resulting number of grid cells, M , can still be large. For example, in the application of downscaling atmospheric CO₂ concentrations in Section 4, $M = 52,128$. Recall that in the definition of \mathbf{A} in (2.8), the (i, j) -th element in the product of \mathbf{A} and \mathbf{A}^T is given by

$$(\mathbf{A}\mathbf{A}^T)_{ij} = \sum_{k=1}^N a_{ik}a_{jk}; \quad i, j = 1, \dots, M.$$

Since any BAU \mathcal{A} is assumed to be uniquely associated with a single coarse-resolution grid cell, at least one of a_{ik} and a_{jk} is zero whenever $i \neq j$, and the $M \times M$ matrix $\mathbf{A}\mathbf{A}^T$ is diagonal. For FRK, the matrix $\mathbf{D} = [\mathbf{A}(\sigma_\xi^2\mathbf{I})\mathbf{A}^T + \mathbf{A}(\sigma_\epsilon^2\mathbf{I})\mathbf{A}^T]^{-1}$ is diagonal as well. Thus, the matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ can be inverted by applying the Sherman-Woodbury-Morrison formula (e.g., Henderson and Searle,

1981) as follows:

$$(\mathbf{A}\Sigma\mathbf{A}^T)^{-1} = \mathbf{D} - \mathbf{D}(\mathbf{A}\mathbf{S})[\mathbf{K}^{-1} + (\mathbf{A}\mathbf{S})^T\mathbf{D}(\mathbf{A}\mathbf{S})]^{-1}(\mathbf{A}\mathbf{S})^T\mathbf{D}. \quad (2.12)$$

This only requires inverting diagonal and low-rank ($r \times r$) matrices. To calculate the determinant $|\mathbf{A}\Sigma\mathbf{A}^T|$ for FRK, we use Sylvester's determinant identity (see Akritas et al., 1996):

$$|\mathbf{A}\Sigma\mathbf{A}^T| = |(\mathbf{A}\mathbf{S})\mathbf{K}(\mathbf{A}\mathbf{S})^T + \mathbf{D}^{-1}| = |\mathbf{K}^{-1} + (\mathbf{A}\mathbf{S})^T\mathbf{D}(\mathbf{A}\mathbf{S})||\mathbf{K}||\mathbf{D}^{-1}|, \quad (2.13)$$

which involves determinants of diagonal and $r \times r$ matrices. For FGP, $\mathbf{D} = (\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1}$, where \mathbf{Q} is a sparse matrix. The Sherman-Morrison-Woodbury formula can be used to calculate \mathbf{D} in (2.12) as well,

$$\mathbf{D} = (\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1} - (\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}[\mathbf{Q} + \mathbf{A}^T(\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}]^{-1}\mathbf{A}^T(\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1}. \quad (2.14)$$

This only requires solving a sparse linear system. To calculate $|\mathbf{D}^{-1}|$ in (2.13) for FGP, Sylvester's determinant identity can be used again:

$$|\mathbf{D}^{-1}| = |\mathbf{Q} + \mathbf{A}^T(\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}||\mathbf{Q}^{-1}||\sigma_\epsilon^2\mathbf{A}\mathbf{A}^T|. \quad (2.15)$$

So, in both FRK and FGP, the twice-negative-marginal-log-likelihood function in (2.10) can be computed efficiently.

Here, we adapt the EM algorithms used in Katzfuss and Cressie (2011) and Ma and Kang (2018a) to obtain maximum-likelihood estimates of the parameters $\boldsymbol{\theta}$ using data $\tilde{\mathbf{Y}}$. Specifically, the random vector $\boldsymbol{\eta}$ is treated as ‘‘missing data’’, and the ‘‘complete data’’ likelihood $L_C(\boldsymbol{\eta}, \tilde{\mathbf{Y}})$ can be obtained. Up to an additive constant, the twice-negative-complete-data-log-likelihood function is given by

$$\begin{aligned} -2 \ln L_C(\boldsymbol{\eta}, \tilde{\mathbf{Y}}) &= \ln |\mathbf{D}^{-1}| + [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta} - \mathbf{A}\mathbf{S}\boldsymbol{\eta}]^T \mathbf{D} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta} - \mathbf{A}\mathbf{S}\boldsymbol{\eta}] \\ &\quad + \ln |\mathbf{K}| + \boldsymbol{\eta}^T \mathbf{K}^{-1} \boldsymbol{\eta}. \end{aligned} \quad (2.16)$$

In the E-step of the EM algorithm, the conditional distribution of $\boldsymbol{\eta}$ given $\tilde{\mathbf{Y}}$ under parameters $\boldsymbol{\theta}$ is multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}},\boldsymbol{\theta}} = \mathbf{K}(\mathbf{A}\mathbf{S})^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}(\tilde{\mathbf{Y}} - \mathbf{A}\mathbf{X}\boldsymbol{\beta})$ and covariance matrix $\Sigma_{\boldsymbol{\eta}|\tilde{\mathbf{Y}},\boldsymbol{\theta}} = \mathbf{K} - \mathbf{K}(\mathbf{A}\mathbf{S})^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{S})\mathbf{K}^T$. Then the conditional expectation of $\ln L_C(\boldsymbol{\eta}, \tilde{\mathbf{Y}})$ with respect to the distribution $[\boldsymbol{\eta}|\tilde{\mathbf{Y}}]$, referred to as the Q function, can be derived. In the M-step, parameters are updated by finding the maximum of this Q function with respect to $\boldsymbol{\theta}$. For FRK,

all parameters in θ have closed-form updates, while in FPG numerical optimization algorithms are used to update τ^2 and γ . Note that the results in (2.12), (2.13), (2.14), and (2.15) allow efficient computation in the execution of the EM algorithm. Details of the parameter estimation scheme via the EM algorithm are given in Appendix A.

2.3 Statistical Downscaling via Conditional Simulation

Recall that numerical model output $\tilde{\mathbf{Y}}$ is defined at coarse spatial resolution over grid cells $\{\Delta_i : i = 1, \dots, M\}$. In Section 2.1, we gave the resulting distribution of $\tilde{\mathbf{Y}}$ under FRK and FGP, respectively. Our goal is to simulate the process $Y(\cdot)$ at fine-resolution, i.e., over all BAUs $\{\mathcal{A}_i : i = 1, \dots, N\}$, given the numerical model output. To avoid introducing additional notation, we also use $\tilde{\mathbf{Y}}$ to represent a realization of this random vector: the observed numerical model output. In OSSEs, numerical models represent state-of-the-art understanding of atmospheric processes, and so their output is assumed to be the “truth” at its corresponding resolution. Therefore, when we downscale $\tilde{\mathbf{Y}}$ to construct NRs at the resolution of the BAUs, we impose the hard constraint that fine-resolution NRs should match the numerical model output exactly when the NRs are aggregated from BAU-level back up to the coarse-resolution grid cells.

The conditional distribution of \mathbf{Y} given $\tilde{\mathbf{Y}}$ is derived as follows. To ensure that the simulated \mathbf{Y} match $\tilde{\mathbf{Y}}$ after aggregation, we impose the constraint that $\mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}}$. Using the standard result for conditional distributions of multivariate normal distributions (e.g., Ravishanker and Dey, 2002, pp. 156-157), the conditional distribution of \mathbf{Y} given $\tilde{\mathbf{Y}}$ is:

$$\mathbf{Y} \mid \mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}} \sim \mathcal{N}_N(\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}(\tilde{\mathbf{Y}} - \mathbf{A}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}). \quad (2.17)$$

To efficiently compute the conditional mean vector for FRK and FGP, we use the results in Eq. (2.12) and Eq. (2.14) to evaluate $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}$. The conditional mean vector, $\boldsymbol{\mu}_{\mathbf{Y}|\tilde{\mathbf{Y}}} \equiv \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}(\tilde{\mathbf{Y}} - \mathbf{A}\boldsymbol{\mu})$, gives the optimal spatial predictions of $Y(\cdot)$ at BAU-level, given data $\tilde{\mathbf{Y}}$, under squared-error loss. The associated prediction uncertainties, i.e., the prediction variances, are the corresponding *diagonal* elements in the conditional covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\tilde{\mathbf{Y}}} \equiv \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}$, and can also be calculated efficiently using Eq. (2.12) and Eq. (2.14).

To construct an ensemble of NRs we simply draw samples from the conditional distribution of \mathbf{Y} given $\tilde{\mathbf{Y}}$. Directly sampling from this conditional distribution in (2.17) requires storing the $N \times N$ covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\tilde{\mathbf{Y}}}$ and performing a Cholesky decomposition on it, which results in

$O(N^2)$ memory cost and $O(N^3)$ flops. Note that with BAUs defined at fine-resolution, N is very large. For example, in the application of downscaling surface CO_2 concentrations in Section 4, $N = 655,362$. Therefore, directly sampling from the distribution in (2.17) is prohibitive. To circumvent this problem, we devised a step-by-step procedure to generate a sample, denoted by \mathbf{Y}_{CS} , from the conditional distribution. The procedure is given by Algorithm 1. The resulting random vector, \mathbf{Y}_{CS} , has some desirable properties, given in Proposition 1. See Appendix B for the proof.

Proposition 1 *The conditional sample \mathbf{Y}_{CS} generated via Algorithm 1 has the following properties:*

- (1) *The sample \mathbf{Y}_{CS} satisfies the hard constraint: $\mathbf{A}\mathbf{Y}_{\text{CS}} = \tilde{\mathbf{Y}}$.*
- (2) *The conditional distribution of \mathbf{Y}_{CS} given $\mathbf{A}\mathbf{Y}$ is multivariate normal with mean $E(\mathbf{Y}_{\text{CS}} | \mathbf{A}\mathbf{Y}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\boldsymbol{\mu})$ and covariance matrix $\text{cov}(\mathbf{Y}_{\text{CS}} | \mathbf{A}\mathbf{Y}) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}$. Thus, given $\mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}}$, the random vector \mathbf{Y}_{CS} follows the same distribution as \mathbf{Y} given in Eq. (2.17).*
- (3) *The marginal distribution of \mathbf{Y}_{CS} is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.*
- (4) *With the parameters $\boldsymbol{\theta}$ known, if we define the mean-squared-error of \mathbf{Y}_{CS} to be $E[\mathbf{Y}_{\text{CS}} - \mathbf{Y}]^2$, then $E[\mathbf{Y}_{\text{CS}} - \mathbf{Y}]^2 = 2[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}]$.*

High-resolution NRs can be constructed efficiently by drawing samples from the conditional distribution of \mathbf{Y} given $\tilde{\mathbf{Y}}$ using Algorithm 1. As stated in Proposition 1, the constraint $\mathbf{A}\mathbf{Y}_{\text{CS}} = \tilde{\mathbf{Y}}$ is satisfied, implying that when the high-resolution NRs are aggregated back to coarse resolution, exactly match the numerical model output, $\tilde{\mathbf{Y}}$. In addition, the high-resolution NR, \mathbf{Y}_{CS} , has a marginal distribution based on the models defined for the process $Y(\cdot)$ at the finest scale as discussed in Section 2.1. Parameters are assumed known in Algorithm 1. In practice they are estimated from coarse-resolution output, $\tilde{\mathbf{Y}}$, using the EM algorithm as described in Section 2.2.

We conclude this section with remarks about computational complexity related to the downscaling procedures for the two models FRK and FGP in Section 2.1. As shown in Cressie and Johannesson (2008) and Ma and Kang (2018a), both FRK and FGP have desirable change-of-support properties, since the integration in Eq. (2.6) and summation in Eq. (2.7) for basis functions can be done offline. With sparse matrices \mathbf{A} and \mathbf{S} , the product of \mathbf{A} and \mathbf{S} can be com-

Algorithm 1 Generate a sample, \mathbf{Y}_{CS} , from the conditional distribution of \mathbf{Y} given the $\mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}}$ in (2.17). Here, the subscript ‘‘CS’’ stands for conditional simulation.

- 1: **Input:** The numerical model output $\tilde{\mathbf{Y}}$, the $M \times N$ aggregation matrix \mathbf{A} , σ_ϵ^2 , and $\boldsymbol{\theta}$. For FGP, $\boldsymbol{\theta}$ consists of $\{\boldsymbol{\beta}, \mathbf{K}, \tau^2, \gamma\}$; for FRK, $\boldsymbol{\theta} \equiv \{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\}$.
// Generate \mathbf{Y}_{NS} , a sample from the marginal distribution of \mathbf{Y} . Here, the subscript ‘‘NS’’ stands for marginal or unconditional simulation.
 - 2: Generate a sample $\boldsymbol{\eta}_{\text{NS}}$ from $\mathcal{N}_r(\mathbf{0}, \mathbf{K})$, requiring Cholesky decomposition of the $r \times r$ matrix \mathbf{K} .
 - 3: Generate a sample $\boldsymbol{\epsilon}_{\text{NS}}$ from $\mathcal{N}_N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$, i.i.d. normal random variables with mean zero and variance σ_ϵ^2 .
 - 4: Generate a sample $\boldsymbol{\delta}_{\text{NS}}$ from $\mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$:
 - For FRK, $\boldsymbol{\Sigma}_\delta = \sigma_\xi^2 \mathbf{I}$, thus sampling $\boldsymbol{\delta}_{\text{NS}}$ as i.i.d. random variables with mean zero and variance σ_ξ^2 .
 - For FGP, $\boldsymbol{\Sigma}_\delta = \mathbf{Q}^{-1}$; this sampling step requires Cholesky decomposition of the sparse matrix $\mathbf{Q} \equiv (\mathbf{I} - \gamma \mathbf{H})/\tau^2$.
 - 5: **Return** $\mathbf{Y}_{\text{NS}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta}_{\text{NS}} + \boldsymbol{\delta}_{\text{NS}} + \boldsymbol{\epsilon}_{\text{NS}}$, a sample from the marginal distribution.
// Adjust \mathbf{Y}_{NS} to obtain \mathbf{Y}_{CS} , a sample from the distribution in (2.17) conditional on $\mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}}$.
 - 6: Calculate $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\mathbf{Y}_{\text{NS}})$ with $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ given in Eq. (2.11):
 - For FRK, use the Sherman-Woodbury-Morrison formula as in Eq. (2.12).
 - For FGP, use both Eq. (2.12) and Eq. (2.14).
 - 7: **Return** $\mathbf{Y}_{\text{CS}} = \mathbf{Y}_{\text{NS}} + \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\mathbf{Y}_{\text{NS}})$.
// If more than one sample is needed, repeat Step 2 through Step 7, noting that calculating Cholesky decompositions of matrices only needs to be done once.
-

puted efficiently with, at most, $O(Mrb_0)$ flops for both FRK and FGP, where b_0 is the maximum number of BAUs falling into a single coarse-resolution grid cell. To generate a unconditional sample \mathbf{Y}_{NS} from its marginal distribution, FRK requires $O(Nr + r^3)$ flops, while FGP requires $O(Nr + N^{1.5})$ flops. To adjust \mathbf{Y}_{NS} to obtain \mathbf{Y}_{CS} , evaluation of $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{b}$ for a vector \mathbf{b} of length M is needed. Solving $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{b}$ requires $O(Mr^2)$ flops for FRK, since it only needs to calculate inversions of $r \times r$ matrices and $n \times n$ diagonal matrices, as well as multiplication of $M \times r$ and $r \times r$ matrices. Therefore, the overall computational cost is $O(Mrb_0 + Mr^2 + Nr)$ for spatial downscaling based on FRK. For FGP, solving $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{b}$ requires Cholesky decomposition of sparse matrix \mathbf{Q} and $\mathbf{Q} + \mathbf{A}^T(\sigma_\epsilon^2 \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$. As discussed in Rue and Held (2005), the Cholesky decomposition of \mathbf{Q} has $O(N^{1.5})$ computational cost for a two-dimensional domain. Notice that the matrix $\mathbf{A}^T(\sigma_\epsilon^2 \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$ is a block diagonal matrix with at most b_0^2 nonzero elements for each of $\lfloor N/b_0 \rfloor$ block matrices. Hence, the sparse matrix $\mathbf{Q} + \mathbf{A}^T(\sigma_\epsilon^2 \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$

has $2Nb_0$ nonzero elements at most given the fact that the number of nonzero elements in \mathbf{Q} is smaller than Nb_0 . The Cholesky decomposition of $\mathbf{Q} + \mathbf{A}^T(\sigma_c^2\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$ has computational cost $O(N(p_0^2 + 3p_0))$ after appropriate reordering to obtain a band matrix with its bandwidth $p_0 \ll N$, though solving the sparse linear system associated with matrix $\mathbf{Q} + \mathbf{A}^T(\sigma_c^2\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$ can cost more computationally than solving that associated with \mathbf{Q} . Therefore, the overall computational cost is $O(Mrb_0 + 2N(r^2 + p_0^2) + N^{1.5}r)$ at most for spatial downscaling based on FGP. Note that in practice, the fine-resolution grid and corresponding BAUs are available based on the scientific goals of the observing systems or data assimilation algorithms. Because its computational complexity is related to N , it is neither practically necessary nor computationally economical to define the spatial process at a finer spatial resolution than what is needed in OSSEs. For the memory cost, both FRK and FGP require storage of sparse matrices \mathbf{A} , \mathbf{S} , $\mathbf{A}\mathbf{S}$, and the diagonal matrix $\mathbf{A}\mathbf{A}^T$, which have $O(Nr)$ memory cost at most. FGP also requires storage of the Cholesky factor of an $N \times N$ sparse matrix and $N \times r$ sparse matrix, which has $O(N \log N + Nr)$ memory cost at most. As $\log N \ll r$, $O(N \log N)$ is upper bounded by $O(Nr)$. Therefore, the overall memory cost is $O(Nr)$ in both FRK and FGP.

2.4 The Forward Basis Function Selection Algorithm

Basis function selection has been investigated for low-rank methods used in analyzing large spatial data sets. Many methods (e.g., Banerjee et al., 2008; Sang and Huang, 2012; Katzfuss, 2017) require pre-specification of a parametric covariance function that, in practice, is usually chosen to be stationary. For these methods, basis functions are determined by the locations of knots, and a pre-specified parametric covariance function. Finley et al. (2009) propose an algorithm to sequentially find such knot locations. For semiparametric methods including FRK and FGP, while they avoid the assumption of a specific parametric covariance function, basis functions do need to be specified. Cressie and Johannesson (2008) recommend compactly-supported multiresolution functions, where the centers and bandwidths of basis functions for each resolution need to be specified. Specifically, a fixed number of resolutions (typically two or three) is chosen first. Then, for the i -th resolution, users specify the total number of basis functions, say, r_i . The centers of these r_i basis functions are then regularly placed over the spatial domain, and all use the same bandwidth. We call such basis functions *equally-spaced* basis functions hereafter. Readers

are referred to Cressie and Johannesson (2008) and Nguyen et al. (2012, 2014) for examples of these equally-spaced basis functions. Zhu et al. (2015) and Zammit-Mangion and Cressie (2017) discuss the use of these equally-spaced basis functions, and suggest removing those where data are rare in order to achieve stable estimation. To the best of our knowledge, no method has been suggested to select both the centers and bandwidths of basis functions in a data-driven way.

Our method sequentially adds new basis function centers and specifies their bandwidths based on their potential ability to improve spatial prediction. This protocol directly addresses one of the primary purposes of spatial data analysis, i.e., spatial prediction, and tends to perform very well in capturing nonstationary spatial variability. See our simulation study in Section 3.2. For the numerical examples in this paper, we focus on one particular form of Wendland basis functions (Wendland, 1995): $S(\mathbf{u}) = (1 - \|\mathbf{u} - \mathbf{c}\|/r)^4 I(\|\mathbf{u} - \mathbf{c}\| \leq r)$, $\mathbf{u} \in \mathcal{D}$, where \mathbf{c} is the center and r is the bandwidth. This form of the Wendland basis function belongs to the family of compactly-supported basis functions. Compactly-supported basis functions have been widely adopted in previous studies (e.g., Cressie and Johannesson, 2006, 2008; Nguyen et al., 2012, 2014; Nychka et al., 2015; Shi and Kang, 2017; Ma and Kang, 2018a). For example, Cressie and Johannesson (2008) use the bisquare basis functions, and Nychka et al. (2015) use the Wendland basis functions that are different from ours. Zammit-Mangion and Cressie (2017) discuss other types, such as Gaussian basis functions, that also require specification of center and bandwidth. The method we propose can be used for other types of basis functions as well.

The basic idea is as follows. First, we define a finite set of r_* locations spread out across the entire domain \mathcal{D} of interest. This set is referred to as a set of *candidate centers*, denoted by $\mathcal{S}^* \equiv \{\mathbf{s}_i \in \mathcal{D} : i = 1, \dots, r_*\}$. Suppose that we pre-specify a set of $r^{(1)}$ basis functions with centers $\mathcal{C}^{(1)} \equiv \{\mathbf{c}_{1,1}, \dots, \mathbf{c}_{1,r^{(1)}}\}$ and bandwidths $\mathcal{B}^{(1)} \equiv \{b_{1,1}, \dots, b_{1,r^{(1)}}\}$ as the initial sets of centers and corresponding bandwidths, where the superscript stands for the iteration of the algorithm. A possible choice is a small set of equally-spaced basis functions over the domain. The centers and bandwidths of new basis functions are added automatically at each iteration of the forward algorithm. At the beginning of the i -th iteration ($i \geq 1$), the current set of basis functions is used to fit data with the FRK model, which gives the estimated trend $\hat{\mu}(\cdot)$ and the estimated random effects, $\hat{\boldsymbol{\eta}}$. We use the pseudo-residuals defined as $D^{(i)}(\mathbf{s}) \equiv Y(\mathbf{s}) - \hat{\mu}(\mathbf{s}) - \mathbf{S}^{(i)}(\mathbf{s})\hat{\boldsymbol{\eta}}$, where $Y(\mathbf{s})$ is the observation at location \mathbf{s} and $\mathbf{S}^{(i)}$ denotes the matrix resulting from the current set of basis functions, to assess where basis functions should be added to improve the fit. As in

classical geostatistics (Cressie, 1993), these pseudo-residuals can be used to carry out the local semivariogram analysis for each observation location. The empirical-local-mean-squared error (ELMSE) is defined for each point in the candidate centers \mathcal{S}^* : $L^{(i)}(\mathbf{s}) \equiv \text{var}\{D^{(i)}(\mathbf{u}) : \mathbf{u} \in \mathcal{N}(\mathbf{s})\} + (\text{ave}\{D^{(i)}(\mathbf{u}) : \mathbf{u} \in \mathcal{N}(\mathbf{s})\})^2$, where $\mathcal{N}(\mathbf{s})$ denotes a local neighborhood surrounding the location \mathbf{s} and is chosen based on the effective range obtained from the semivariogram analysis. The effective range is defined as the distance at which the semivariogram value achieves 95% of the sill. New basis functions are placed where the ELMSE is large, and the bandwidths of these basis functions are chosen corresponding to the effective range. We also impose a separation criterion to avoid substantial overlap among the supports of the newly-added basis functions at each iteration. We repeat these steps until the upper bound of the number of basis functions, r_{\max} , is reached, or the ELMSE does not change substantially. In practice, we recommend using both together as a stopping criterion, and r_{\max} can be chosen as large as computational constraints allow. When computational limits are less constrained, users can set r_{\max} larger, or even let the stopping criterion depend solely on the change in ELMSE.

The step-by-step procedure is described in Algorithm 2. In local variogram analysis, a parametric variogram function, such as the exponential function, is fitted using data in a small neighborhood as in Hammerling et al. (2012) via weighted least squares or maximum likelihood estimation (Cressie, 1985). Tadić et al. (2015) give a less user-specified way to define the neighborhood, but their method is computationally more complicated due to sampling based on pairwise distances. We nominally set the separation distance to be two-thirds of the effective range, such that the shortest distance between centers of two basis functions added within the same iteration will be no less than the 1.5 times either of their bandwidths. This is motivated by the suggestion in Cressie and Johannesson (2008) regarding overlap between supports of basis functions. To specify the candidate centers \mathcal{S}^* , we choose a set of grid points that covers the spatial domain, or simply set them to be the observation locations, \mathcal{S}_O . In our simulation study presented in Section 3.2, we set $\mathcal{S}^* = \mathcal{S}_O$, and the empirical results show that this choice is robust against large gaps in observations, and gives improved predictive performance. In practice, to maintain computational efficiency, only a small number, typically no more than a few hundred, of basis functions are used (Cressie and Johannesson, 2008). The ξ term in the CAR component of FGP is designed to capture the remaining variation. If the set of basis functions were able to capture the spatial variation completely, the CAR model in ξ would reduce to the special case with the spa-

Algorithm 2 Forward basis function selection.

- 1: **Input:** Observed data $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{S}_O\}$ with \mathcal{S}_O denoting the set of observation locations, and candidate centers $\mathcal{S}^* \equiv \{\mathbf{s}_i \in \mathcal{D} : i = 1, \dots, r_*\}$ with fixed and finite r_* .
// Notice that the location \mathbf{s} is a generic notation to denote a location where a data value is obtained. The data are not necessarily defined at BAU levels. When data are at coarse spatial resolution, the residuals and related calculations are obtained at the same resolution correspondingly.
 - 2: **Initialization:** $i \leftarrow 1$; a starting set of $r^{(i)}$ basis functions with centers $\mathcal{C}^{(i)} \equiv \{\mathbf{c}_{1,1}, \dots, \mathbf{c}_{1,r^{(i)}}\}$ and corresponding bandwidths $\mathcal{B}^{(i)} \equiv \{b_{1,1}, \dots, b_{1,r^{(i)}}\}$.
 - 3: **while** the stopping criterion is not satisfied **do**
 - 4: Fit the FRK model with the current basis functions.
 - 5: Calculate the pseudo-residuals: $D^{(i)}(\mathbf{s}) = Y(\mathbf{s}) - \hat{\mu}(\mathbf{s}) - \mathbf{S}^{(i)}(\mathbf{s})\hat{\boldsymbol{\eta}}$, for all $\mathbf{s} \in \mathcal{S}_O$.
 - 6: **for all** $\mathbf{s} \in \mathcal{S}^*$ **do**
 - 7: Perform a local semivariogram analysis to obtain the effective range, $d(\mathbf{s})$.
 - 8: Define the neighborhood $\mathcal{N}(\mathbf{s}) \equiv \{\mathbf{u} : \mathbf{u} \in \mathcal{S}_O, \text{ and } \|\mathbf{s} - \mathbf{u}\| \leq d(\mathbf{s})\}$ and calculate the empirical local mean squared error (ELMSE): $L^{(i)}(\mathbf{s}) \equiv \text{var}\{D^{(i)}(\mathbf{u}) : \mathbf{u} \in \mathcal{N}(\mathbf{s})\} + (\text{ave}\{D^{(i)}(\mathbf{u}) : \mathbf{u} \in \mathcal{N}(\mathbf{s})\})^2$.
 - 9: **end for**
 - 10: Calculate $L_0^{(i)}$, a cutoff based on $\{L^{(i)}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}^*\}$. For example, the 90th percentile of $\{L^{(i)}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}^*\}$.
 - 11: Define the set of *potential* centers in the i -th iteration: $\mathcal{PC}^{(i)} = \{\mathbf{s} : L^{(i)}(\mathbf{s}) \geq L_0^{(i)}, \mathbf{s} \in \mathcal{S}^*\}$, and define the set of *new* centers $\mathcal{NC}^{(i)} \leftarrow \emptyset$, the empty set. Correspondingly, the set of *new* bandwidths $\mathcal{NB}^{(i)} \leftarrow \emptyset$.
 - 12: **for all** $\mathbf{s} \in \mathcal{PC}^{(i)}$ **do**
 - 13: **if** $\|\mathbf{s} - \mathbf{u}\| \geq \gamma(\mathbf{u})$ for *all* $\mathbf{u} \in \mathcal{NC}^{(i)}$, where $\gamma(\mathbf{u})$ denotes the corresponding separation distance, here chosen to be two-thirds of the effective range, $\gamma(\mathbf{u}) \equiv 2d(\mathbf{u})/3$. **then**
 - 14: Add \mathbf{s} into $\mathcal{NC}^{(i)}$, with its corresponding bandwidth, the effective range $d(\mathbf{s})$, added into $\mathcal{NB}^{(i)}$.
 - 15: **end if**
 - 16: **end for**
 - 17: Update: $\mathcal{C}^{(i+1)} \leftarrow \mathcal{C}^{(i)} \cup \mathcal{NC}^{(i)}$, and $\mathcal{B}^{(i+1)} \leftarrow \mathcal{B}^{(i)} \cup \mathcal{NB}^{(i)}$.
 - 18: $i \leftarrow i + 1$
 - 19: **end while**
-

tial dependence parameter $\gamma = 0$, or approximately zero. By placing basis functions at locations where the current model fit is poor, and choosing bandwidths adaptively based on information in their local semivariograms, we expect to iteratively adapt to nonstationary spatial variability. Compared to methods in Katzfuss (2013) and Konomi et al. (2014) that require computationally intensive procedures including reversible jump Markov Chain Monte Carlo, our method is simple, intuitive and well-suited to parallel computing environments, since local variogram fitting

can be done in parallel.

In summary, the following inputs are required to implement Algorithm 2: the initial set of basis functions, the set of candidate centers \mathcal{S}^* , the cutoff value for the ELMSEs, and the stopping criterion. The initial set of basis functions can be chosen to be a small number of equally-spaced basis functions whose centers are from a regular coarse grid over the spatial domain. A default choice of \mathcal{S}^* is the set of observation locations. In our numerical studies, we choose the 90th percentiles of the ELMSEs as the cutoff value. This cutoff will affect how many basis functions will be added at each iteration of the forward basis function selection algorithm, but the predictive accuracy is not sensitive to the choice of this cutoff. For the stopping criterion, we recommend stopping the algorithm either when the upper bound on the number of basis functions, r_{\max} , is reached or when the ELMSEs do not change significantly. In the numerical examples, we set the threshold to be 0.01. Moreover, Algorithm 2 requires the user to choose the type of basis function. In all numerical studies, we use one particular form of the Wendland basis function, but others such as bisquare can also be used. The impact of using different types of basis functions can be assessed via model selection criteria including BIC or cross validation. However, a thorough empirical and theoretical comparison is beyond the scope of this work.

3 Simulation Studies

We present two simulation studies in this section. Section 3.1 presents an illustration of the downscaling framework, and the importance of handling the change-of-support problem. In Section 3.2, we demonstrate how our algorithm for forward basis function selection works, and its superior performance compared to widely equally-spaced basis functions.

3.1 A Synthetic Example for Statistical Downscaling

In this section, we use synthetic data to illustrate our the statistical downscaling method. First, we simulate a fine-resolution dataset, and aggregate it to form a coarse-resolution dataset. Then, we downscale this coarse-resolution dataset to obtain a fine-resolution field, which we compare to the originally simulated fine-resolution data. Specifically, we simulate from a Gaussian process with mean $\mu(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta}$ and exponential covariance function $c(h) = \sigma^2 \exp(-h/\rho) + \sigma_\epsilon^2 I(h = 0)$

at $N = 100 \times 100$ regular BAUs in a $[0, 100] \times [0, 100]$ domain with $\mathbf{s} \equiv (x, y)^T$. Here, $\mathbf{X}(\mathbf{s}) = (1, x, y)^T$ with $\boldsymbol{\beta} = (2, 0.5, 0.2)^T$, $\sigma^2 = 2$, $\rho = 5$, and $\sigma_\epsilon^2 = 0.2$. The simulated dataset is denoted by $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_N))^T$ at the N BAUs. Then, the fine-resolution dataset \mathbf{Y} is upscaled to the coarse-resolution dataset at $M = 50 \times 50$ regular grid cells in the same domain, using the change-of-support property in Eq. (2.7). The resulting coarse-resolution data are referred to as $\tilde{\mathbf{Y}}$. To assess the quality of the downscaling approach, 10% of coarse-resolution data are randomly held out. These are shown in white regions in Figure 2. The remaining 90% of coarse-resolution data are treated as “synthetic observations”.

Based on synthetic observations, we implement four methods:

- (1) kriging using the true parameters in an exponential covariance function model that accounts for change of support;
- (2) kriging based on estimated parameters in the exponential covariance function model *without* handling change-of-support problem. That is, the coarse-resolution data are treated as point-level ones at the centers of coarse-resolution grid cells;
- (3) kriging based on FRK presented in Section 2.1;
- (4) kriging based on FGP presented in Section 2.1.

These methods are referred to as EK, PK, FRK, FGP, respectively. Note that the kriging predictor is indeed the conditional mean of the distribution of the true field given the data. The parameters in FRK and FGP are estimated using maximum likelihood methods given in Section 2.2. The basis functions are chosen at three different resolutions with $152 = 4^2 + 6^2 + 10^2$ equally-spaced centers. The proximity matrix in FGP is chosen based on first-order neighborhood structure.

Spatial predictions are made at all fine-resolution BAUs, and we compare the predictions from the four methods with the “truth”, \mathbf{Y} , by calculating the mean-squared-prediction errors (MSPEs). We see from Table 1 that EK gives the best results as expected, since it uses the true parameters and covariance function model, and handles change-of-support problem. PK does not handle change-of-support problem, and treats coarse-resolution data incorrectly as being at fine resolution. The predictions from PK are much worse than those from FRK and FGP which do account for change of support. Between FRK and FGP, the latter gives better predictive performance, as expected, since its model is more flexible. Figure 2 visualizes the fine-resolution true fields and synthetic observations at coarse-resolution, together with the downscaled fields

from FRK and FGP, over the entire region. Figure 3 shows the results zoomed-in on a $[0, 20] \times [0, 30]$ region. The downscaled fields from both FRK and FGP mimic the pattern presented in the true fine-resolution field, but the FGP field is closer to \mathbf{Y} .

Table 1. Summary of results for spatial predictions based on coarse-resolution data using four methods: EK, PK, FRK, and FGP, respectively. “COS” stands for change of support, which indicates whether the method deals with change of support or not.

Method	EK	PK	FRK	FGP
COS	Yes	No	Yes	Yes
MSPE	0.457	0.623	0.570	0.477

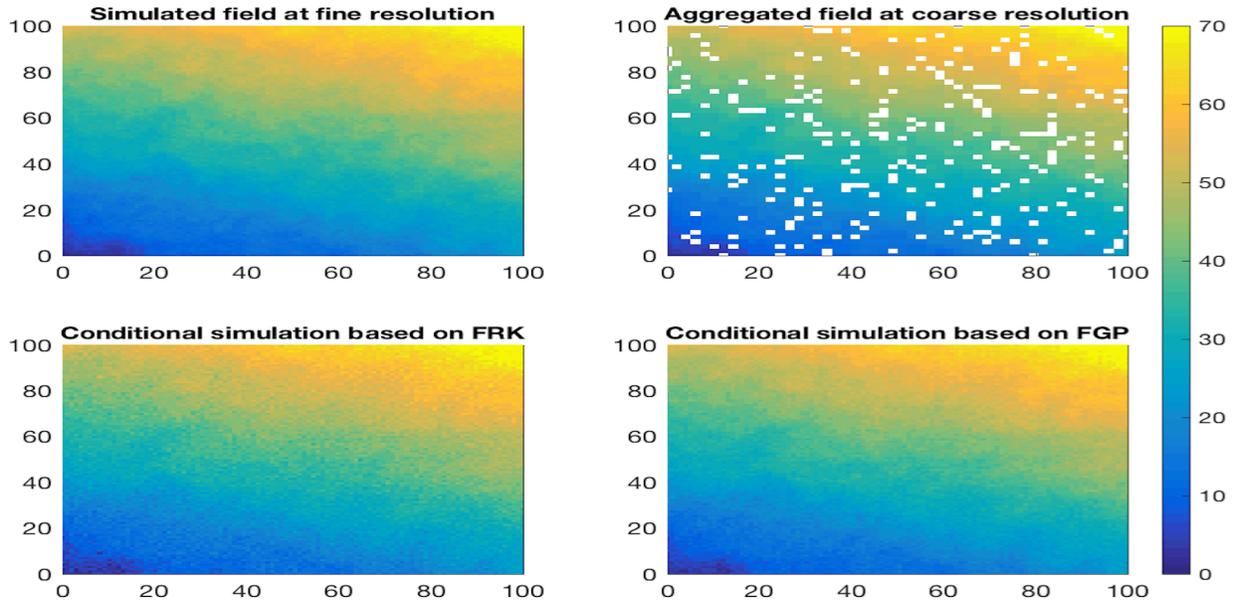


Figure 2. Simulated data and downscaling results from FRK and FGP over the entire domain $[0, 100] \times [0, 100]$. The top-left panel shows \mathbf{Y} , the simulated data at fine-resolution. The top-right panel plots the “synthetic observations” at coarse resolution, after randomly taking out 10% of \mathbf{Y} in the white regions. Bottom panels show the downscaled fields from FRK (bottom-left) and FGP (bottom-right).

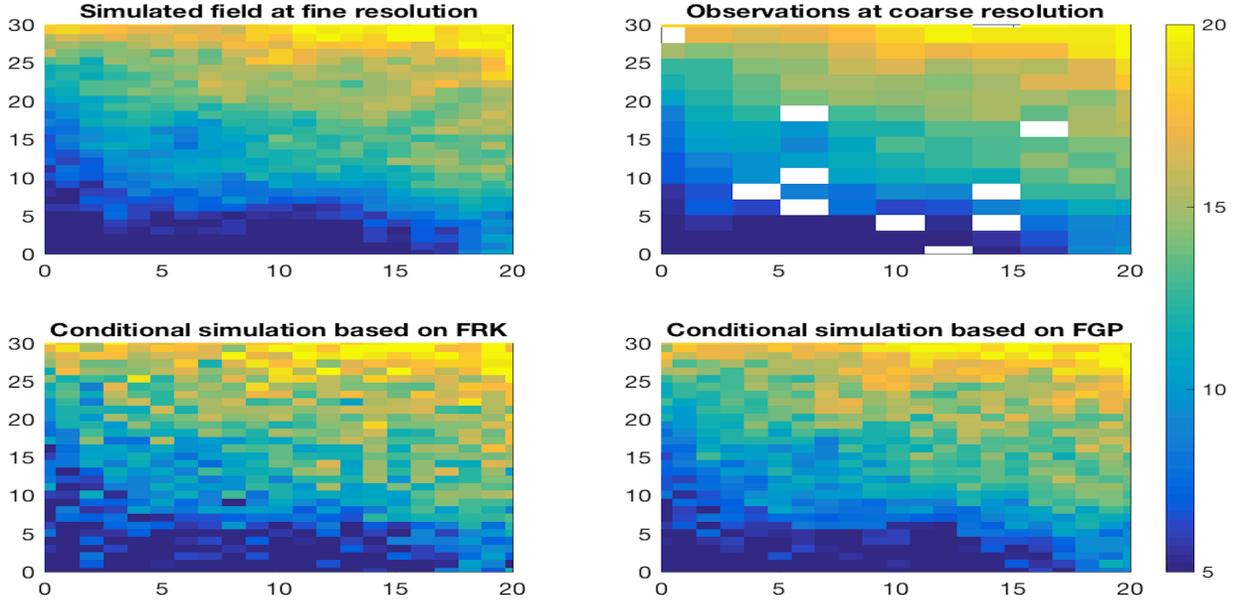


Figure 3. Simulated data and downscaling results from FRK and FGP, with a zoomed-in view over the subregion $[0, 20] \times [0, 30]$. The top-left panel shows \mathbf{Y} , the simulated data at fine-resolution. The top-right panel shows the “synthetic observations” at coarse resolution. Bottom panels show the downscaled fields from FRK (bottom-left) and FGP (bottom-right).

3.2 A Toy Example for the Forward Basis Function Selection Algorithm

Here we give an example to illustrate the method and performance of our algorithm for forward basis function selection. Consider the deterministic function $f(\mathbf{s}) = 50x \exp(-x^2 - y^2)$ for $\mathbf{s} \equiv (x, y)^T$ in the domain $\mathcal{D} \equiv [-2, 6] \times [-2, 6]$. This function has two localized features inside the subregion $[-2, 2] \times [-2, 2]$, and is almost zero everywhere else; see the top-left panel in Figure 4. To create the synthetic true field, we generate function values of $f(\cdot)$ on the 100×100 regular grid covering the domain. We add a Gaussian white noise term with mean zero and variance $\sigma_\epsilon^2 = 0.01\hat{\sigma}_f^2$ at each grid cell, where $\hat{\sigma}_f^2$ is the empirical variance of the function $f(\cdot)$ evaluated at these 100×100 regular grid cells. We hold out data on a small block region $\mathcal{S}_1 \equiv [-0.5, 0.5] \times [-1.5, 1.5]$, referred to as “missing-by-design” locations. We also hold out data at randomly selected 10% of the remaining grid cells \mathcal{S}_2 , referred to as “missing-at-random” locations. The remaining 90% of data are treated as “observations” for which their locations are denoted by $\mathcal{D}_o \equiv \mathcal{D} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$; see the top-right panel in Figure 4.

The initial set of basis functions is chosen to be a set of 25 Wendland basis functions with centers equally spaced over the domain, as shown in the top-left panel of Figure 5. The corre-

sponding bandwidths are 1.5 times the shortest distance among these 25 centers, as suggested in Cressie and Johannesson (2008). In the forward selection algorithm, we set the maximum number of basis functions to be $r_{\max} = 200$, and set the stopping criterion to be the time that the absolute difference between the cutoff values $L_0^{(i)}$ at two consecutive iterations is at or below 0.01. In our example, the forward selection algorithm stops after 12 iterations, resulting in a total of 191 basis functions (including the original 25). The middle-left and bottom-left panels of Figure 5 plot the centers of basis functions added at the eighth and 12th iterations, respectively. The corresponding pseudo-residuals for these two iterations are shown in the middle-right and bottom-right panels, respectively. We see that basis functions are placed where the variabilities of pseudo-residuals are large. After adding the new basis functions, the model fits the data more closely. Figure 4 shows spatial predictions and associated standard errors after the eighth and 12th iterations of the procedure. It is obvious that the predicted field is closer to the true field using 191 basis functions (i.e., after all 12 iterations), compared to those from only the eighth iteration. This demonstrates that our algorithm adds basis functions in a way that improves predictive performance. Figure 6 shows how the associated ELMSEs $\{L^{(i)}(\mathbf{s})\}$ and the cutoffs, $L_0^{(i)}$, change with iteration for $i = 1, \dots, 12$. Observe that both ELMSE and $L_0^{(i)}$ decrease as basis functions are added.

To further demonstrate the advantage of our adaptive basis function approach, we compare our results with the more commonly used, equally-spaced basis function approach. Cressie and Johannesson (2008), Nguyen et al. (2012), and Nguyen et al. (2014) suggest using compactly-supported basis functions at several resolutions, but choose basis functions at the same resolution to be equally-spaced. When we fit a model with equally-spaced basis functions, we call the method “simple”, and we call the method “adaptive” if we instead use the adaptive basis functions from our forward algorithm. Here, we use equally-spaced basis functions from three resolutions, giving a total of $151 (= 5^2 + 7^2 + 9^2 - 4)$ basis functions, with four basis functions removed where data are sparse (Zhu et al., 2015; Zammit-Mangion and Cressie, 2017). We also add additional equally-spaced basis functions at the next-finer resolution, resulting in a total of $264 (= 5^2 + 7^2 + 9^2 + 11^2 - 12)$ basis functions.

We compare the predictive performance of the simple and adaptive methods by looking at their corresponding mean square prediction errors (MSPEs). As shown in Table 2, simple FRK with more basis functions ($r = 264$) gives better predictions at both missing-by-design and

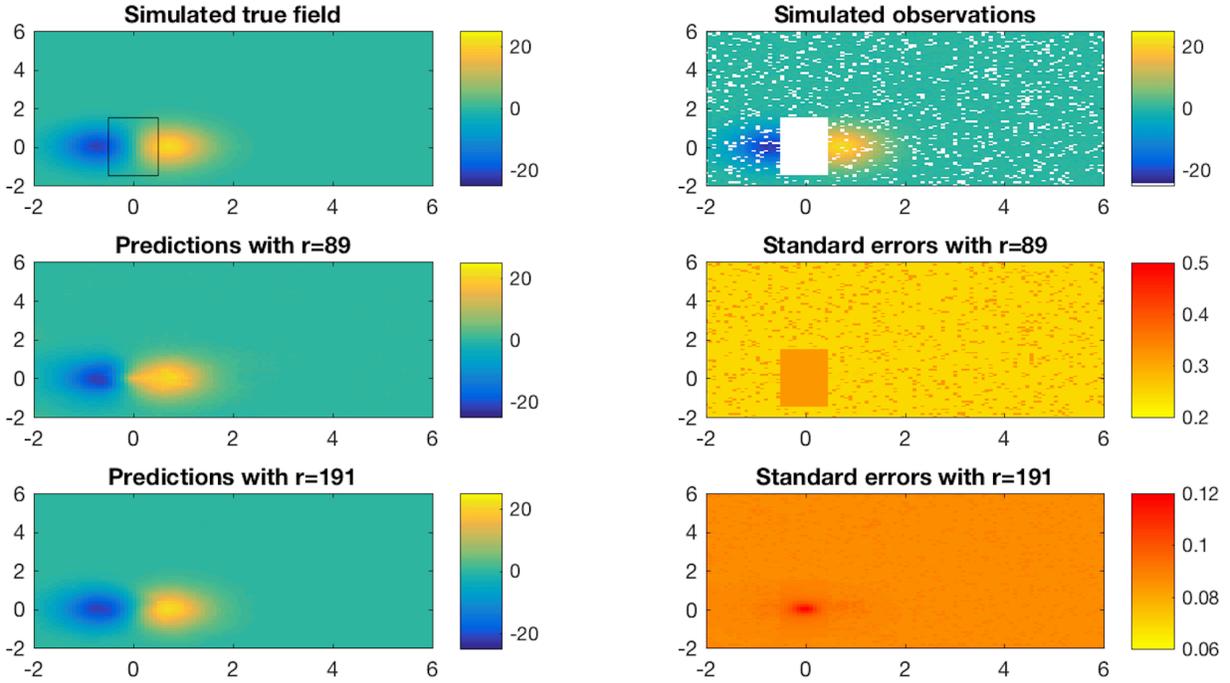


Figure 4. Predictions with adaptive basis functions from the forward selection algorithm. Top-left panel shows the true deterministic field on a 100×100 regular grid; top-right panel shows the observations after removing the prediction locations, $\mathcal{S}_1 \cup \mathcal{S}_2$. The white box in the top-right panel shows the missing-by-design locations, \mathcal{S}_1 , and other white locations show the missing-at-random locations, \mathcal{S}_2 . The middle and bottom panels are predictions for the underlying true field, and associated standard errors with $r = 89$ and 191 basis functions, respectively.

missing-at-random locations, compared to simple FRK with only $r = 151$ basis functions. However, using only $r = 191$ adaptive basis functions, adaptive FRK gives much better predictions than simple FRK with more basis functions. Specifically, the MSPE over all missing locations from adaptive FRK with 191 basis functions is only about 36% of that from simple FRK with 264 basis functions.

We then fit FGP in which a CAR model is assumed for the random vector, ξ . The proximity matrix for FGP is specified by assuming a parsimonious first-order neighborhood structure. Recall that FGP reduces to FRK when the spatial dependence structure parameter is equal to zero, and thus FGP is more flexible (Ma and Kang, 2018a). With 191 adaptive basis functions, adaptive FGP gives the best predictive performance overall. Compared to simple FRK with $r = 264$ functions, its MSPE, over the missing-by-design locations, is less than one-third that of simple FRK. We see even more improvement at the missing-at-random locations: the MSPE of adaptive FGP over these locations is 0.015, compared to 0.477 from simple FRK with 264 basis functions.

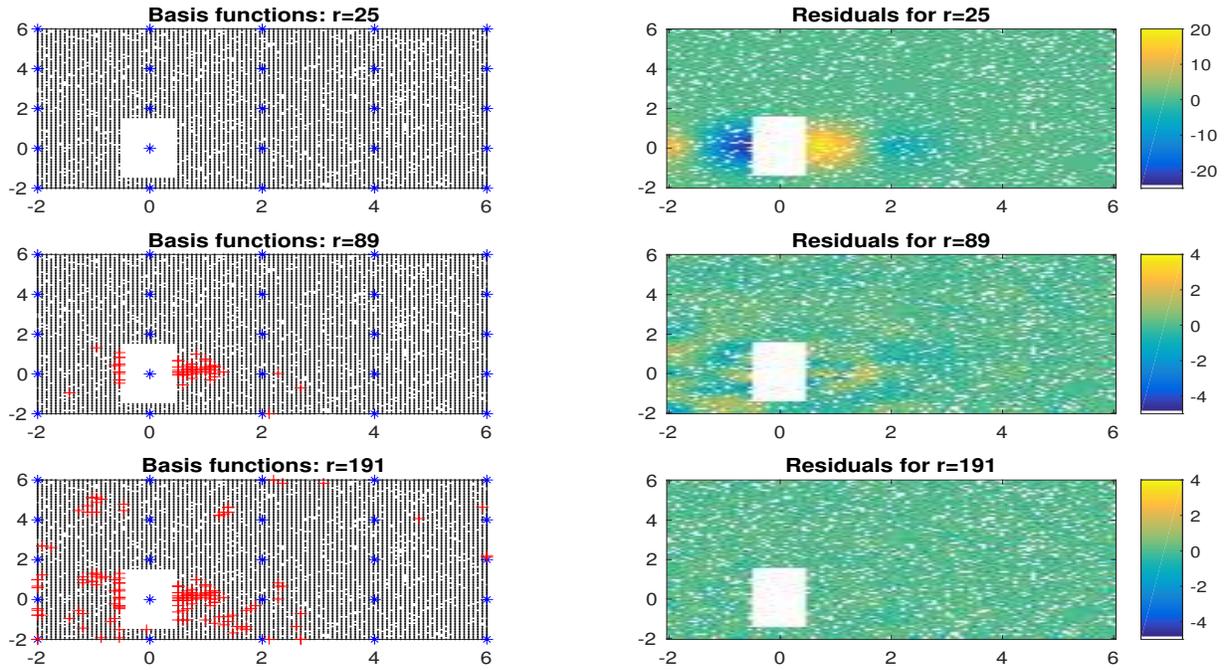


Figure 5. Adaptive basis functions and corresponding residuals for simulation example. The asterisks in three left panels represent 25 initial basis centers, and the dots show the observation locations. The plus signs represent new basis centers added using the forward selection algorithm. The three right panels show the residuals in the forward selection algorithm with different numbers of basis functions $r = 25, 89,$ and $191,$ respectively.

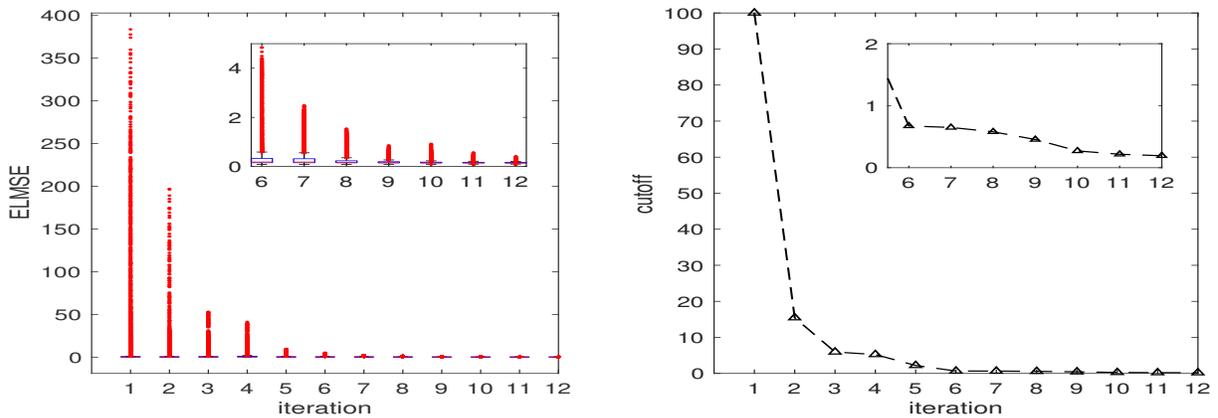


Figure 6. Diagnostics of the forward selection algorithm at each iteration for simulation example. The left panel shows the box plot of ELMSEs at each iteration of the forward selection algorithm. In the right panel, the triangles \triangle show the 90th percentile of the ELMSEs, i.e., cutoff, at each iteration of the forward selection algorithm.

For comparison, we perform local kriging (Haas, 1990; Kitanidis, 1997). We choose a neighborhood composed of 6-by-6 pixels for each location in \mathcal{S}_2 , and a neighborhood of 15-by-15 pixels for each grid cell for locations in \mathcal{S}_1 , and perform kriging with an exponential covariance function. From Table 2, we see that over \mathcal{S}_1 (i.e, data missing in a contiguous region), both adaptive FRK and adaptive FGP outperform local kriging substantially, while they give comparable results over \mathcal{S}_2 (i.e., data missing at random). Note that local kriging is based on a moving window of nearest observations. When prediction is made over a contiguous missing region, the majority of prediction locations in that region will depend mostly on the same set of nearby observations. In contrast, adaptive FRK and adaptive FGP, with appropriately chosen basis functions, are better able to capture spatial dependence structures, and give better prediction results. Our finding here is also consistent with that of Shi and Cressie (2007): globally valid, flexible models such as FRK and FGP outperform local kriging and other fast non-statistical spatial prediction methods, such as Inverse Distance Weighting (IDW) and Nearest Neighbors Smoothing (NNS) for data that have contiguous missing values. Moreover, local kriging only gives *marginal* inference at *each* location *separately*, but FRK and FGP, as global models over the entire spatial domain, are able to provide *joint* inference at *all* locations of interest. This is the key to properly quantify uncertainties with a globally valid spatial process model via conditional simulation.

We now report on computational time for methods in this simulation study. The forward basis function selection algorithm took about 70 seconds on a Macbook Pro with a 2.8-GHz Intel Core i7 processor. The parameter estimation and prediction took about 12, 137, 35 seconds for adaptive FRK, adaptive FGP, and local kriging, respectively. We can see that here FRK and local kriging are faster than FGP, but FGP gives smaller MSPE, overall.

Table 2. Numerical results for comparing local kriging, FRK, and FGP. Here, r denotes the total number of basis functions in the low-rank component in FRK and FGP. The method is called “simple” if equally-spaced basis functions are employed, and “adaptive” when basis functions are selected via Algorithm 2.

Method	Local Kriging	Simple FRK		Adaptive FRK	Adaptive FGP
		$r = 151$	$r = 264$	$r = 191$	$r = 191$
MSPE(\mathcal{S}_1)	6.252	7.451	6.753	2.431	2.042
MSPE(\mathcal{S}_2)	0.010	0.495	0.477	0.019	0.015
MSPE($\mathcal{S}_1 \cup \mathcal{S}_2$)	1.929	2.634	2.407	0.761	0.638

4 Application to PCTM

Atmospheric carbon dioxide (CO_2) is one of the most important greenhouse gases. Numerical models are typically used to study geophysical processes in carbon cycle, and to assess future climate change (e.g., Friedlingstein et al., 2006). Specifically, high-resolution atmospheric CO_2 fields over the globe produced by numerical models are often used to study atmospheric chemistry and dynamics. Global numerical models typically generate atmospheric CO_2 at relatively coarse-resolution, say, $100 \sim 500$ km grid cells. On the other hand, since surface-level emissions are more relevant to urban environments where the majority of people reside, high-resolution surface observing networks are required to monitor greenhouse gas emissions, and to devise mitigation strategies (e.g., Shusterman et al., 2016). OSSEs can be used to design these observing networks, and to evaluate data assimilation algorithms that combine their data with space-based observations like those from Japan’s Greenhouse Gases Observing Satellite (GOSAT) and NASA’s Orbiting Carbon Observatory-2 (OCO-2). In this section, we demonstrate how our downscaling framework can be used to construct high-resolution NR fields for OSSEs.

In our study, we downscale surface CO_2 concentrations generated by the PCTM/GEOS-4/CASA-GFED model. This numerical model has been widely used to study global CO_2 and to evaluate mapping algorithms (e.g., Parazoo et al., 2011; Hammerling et al., 2012; Zhang et al., 2014). We use PCTM to simulate global atmospheric CO_2 concentrations, with unit parts per million, at the surface on January 3, 2006. The spatial resolution of this output is 1° latitude by 1.25° longitude, which results in $M = 181 \times 288 = 52,128$ grid cells over the globe. Statistical downscaling is performed on this PCTM output to construct a global high-resolution surface CO_2 field on equal-area hexagonal grid cells, with 30 km intercell distances. These hexagonal grid cells are obtained from the Discrete Global Grid software (DGG, Sahr et al., 2003; Stough et al., 2014), and the hexagons are used as the BAUs in our study. The surface of the Earth is uniformly tiled by $N = 655,362$ BAUs. Exploratory analysis suggests a linear trend based on latitude, and nonstationary spatial dependence structure. In what follows, we describe how adaptive basis functions are selected for PCTM output with our forward selection algorithm, and present the downscaled high-resolution NRs based on both FRK and FGP.

To apply the forward basis selection algorithm, we begin with a set of 32 equally-spaced basis functions with radii 6241.1 km obtained from Cressie and Johannesson (2008). The centers of

these 36 basis functions are shown as blue asterisks in the left panels of Figure 7. We set the collection of candidate centers to the set of the centroids of all grid cells. In this application, pseudo-residuals are at the same coarse resolution as PCTM output, and local semivariogram analyse are carried out. We use a stopping criterion that requires the number of basis functions not exceed 450, and simultaneously, that the absolute change in the cutoff value, $L_0^{(i)}$, between two consecutive iterations not exceed 0.01. The forward selection algorithm stops after the ninth iteration, resulting in a total of $r = 431$ basis functions. The cutoff at the ninth iteration is $L_0^{(9)} = 1.61$, and the difference between the cutoffs at the eighth and ninth iterations is 0.17. Figure 8 plots the ELMSEs and the cutoffs, $L_0^{(i)}$, as a function of iteration, for $i = 1, \dots, 9$. The cutoff decreases as the number of iterations increases. The algorithm stops after the ninth iteration because of the upper limit on the total number of basis functions. Recall that inference, including parameter estimation and downscaling via conditional simulation, requires inverting $r \times r$ matrices and storing $N \times r$ matrices, thus large r is not desirable.

With these $r = 431$ adaptive basis functions, we implement the downscaling framework based on FRK and FGP. Figure 9 shows the PCTM output over the globe and the downscaled fields: the high-resolution NRs, from conditional simulation based on FRK and FGP, respectively. Although the spatial pattern in PCTM output is maintained by both high-resolution NRs, we judge the NR based on FGP to be superior, since the FRK NR presents a clear “salt-and-pepper” artifact, which is not realistic for atmospheric processes. Such “salt-and-pepper” artifacts appear more clearly when we zoom into a subregion, as shown in Figure 10. Specifically, note that when aggregated back to the $1^\circ \times 1.25^\circ$ resolution, the high-resolution NRs from both FRK and FGP match the PCTM output exactly. However, the FGP NR at high resolution does not present the salt-and-pepper artifacts. Using BIC to compare FRK and FGP with 431 adaptive basis functions. We found that $\text{BIC}_{FRK} = 22.65$ and $\text{BIC}_{FGP} = 20.37$, which also suggests that FGP performs better than FRK. This is also consistent with the empirical results in our simulation studies that show that adaptive FGP gives better model fit than adaptive FRK. Here, both FRK and FGP are run on an HP Intel Xeon E5-2690 machine with 12 GB memory and four cores at the Ohio Supercomputer Center (OSC, 1987). The adaptive basis function selection algorithm took about 15 minutes. The computation times for parameter estimation are about three minutes for FRK and ten hours in FGP. The latter takes more time because FGP requires additional calculations related to $N \times N$ sparse matrices and numerical optimization to update the spatial dependence

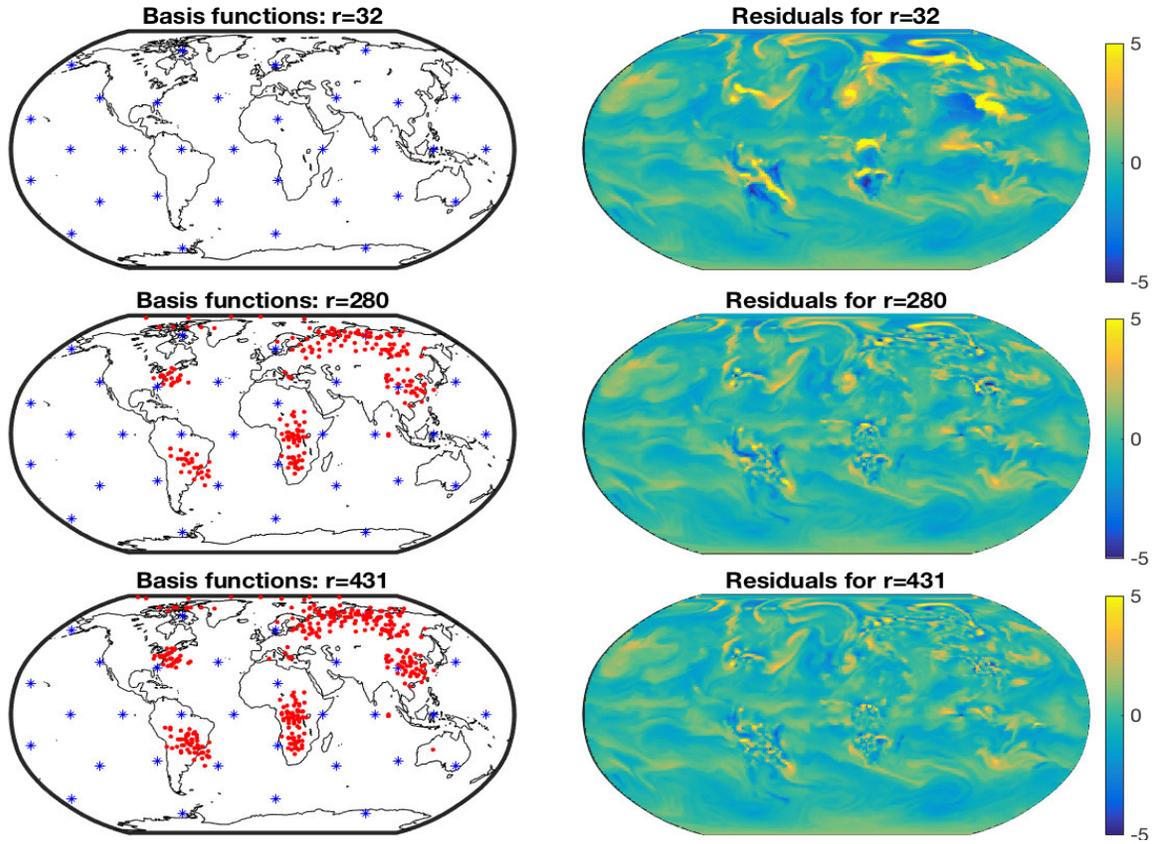


Figure 7. Adaptive basis functions and corresponding residuals for PCTM output. The asterisks * in three left panels represent 32 initial basis centers. The dots • represent new basis centers added using the forward selection algorithm. The three right panels show the residuals in the forward selection algorithm with different number of basis functions $r = 32, 280,$ and $431,$ respectively.

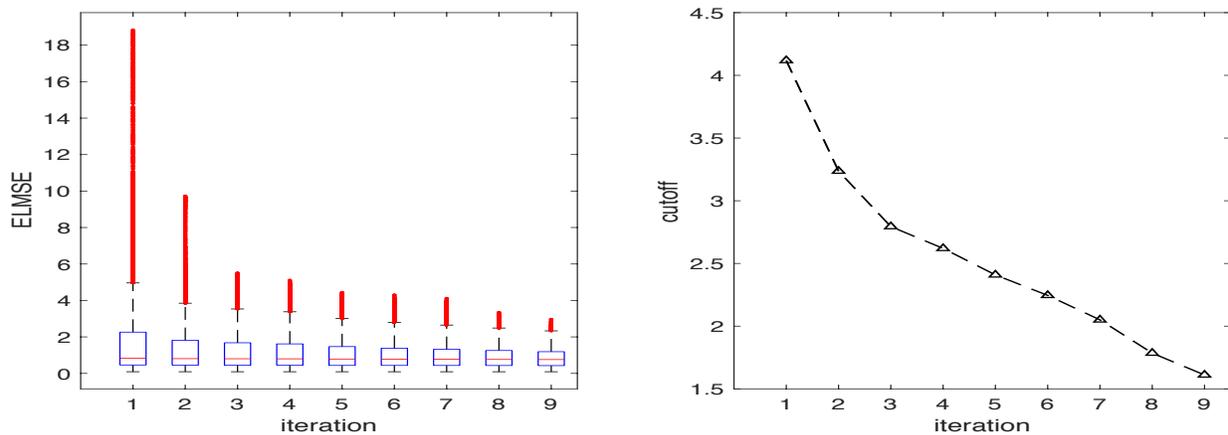


Figure 8. Diagnostics of the forward selection algorithm at each iteration for PCTM outputs. The left panel shows the boxplot of ELMSEs at each iteration of the forward selection algorithm. In the right panel, the triangles \triangle show the 90th percentile of ELSMEs, i.e., cutoff, at each iteration of the forward selection algorithm.

parameter in the CAR part of the model within each iteration of the EM algorithm. To generate one downscaled NR, FRK took about one minute, and FGP about four minutes. Therefore, we are able to produce ensembles of high-resolution NRs efficiently for both FRK and FGP. We have tried implementing local kriging in parallel on the HP Intel Xeon E5-2690 machine with 12 GB memory and 12 cores. Note that local kriging does not define a valid joint predictive distribution for all BAUs. Therefore, the downscaled field from local kriging fails to satisfy the important aggregation requirement. Furthermore, due to the large size of $N = 655,362$, it took about 20 hours to run local kriging at all N BAUs. Computation time required for local kriging could be shortened when more computing cores (and thus more extensive parallelization) are available.

5 Conclusions and Discussion

We have presented a unified model-based statistical spatial downscaling framework that can be used in OSSEs to construct realizations of high-resolution NRs. Our model explicitly handles change-of-support that occurs because of the gap in spatial resolutions of the numerical model outputs and that of the desired NRs. Our downscaling framework differs from that in previous studies in two important ways. First, we only utilize numerical model outputs and do not require physical observations, which makes our method suitable in the context of OSSEs, in particular. Second, our downscaled NRs match the coarse-resolution outputs exactly when they are aggregated back to the resolution of the numerical model in order to preserve physical relationships embodied in the numerical model.

We further proposed a data-driven algorithm to sequentially add basis functions to the low-rank component of the model to learn nonstationary spatial structures and localized features from data. This forward selection algorithm specifies both the centers and bandwidths of basis functions, adaptively. When the number of observation locations is extremely large, we suggest that our algorithm be combined with spatial clustering methods such as Marchetti et al. (2017) to further improve computational efficiency. Our current algorithm selects the centers and bandwidths of basis functions in a forward fashion. One interesting extension would be a modification of such procedure to eliminate some basis centers by incorporating cross-validation steps in the algorithm. LASSO-type variable selection methods (e.g. Tibshirani, 1996; Bondell et al., 2010) can also be potentially developed for basis function selection, though more work is needed to

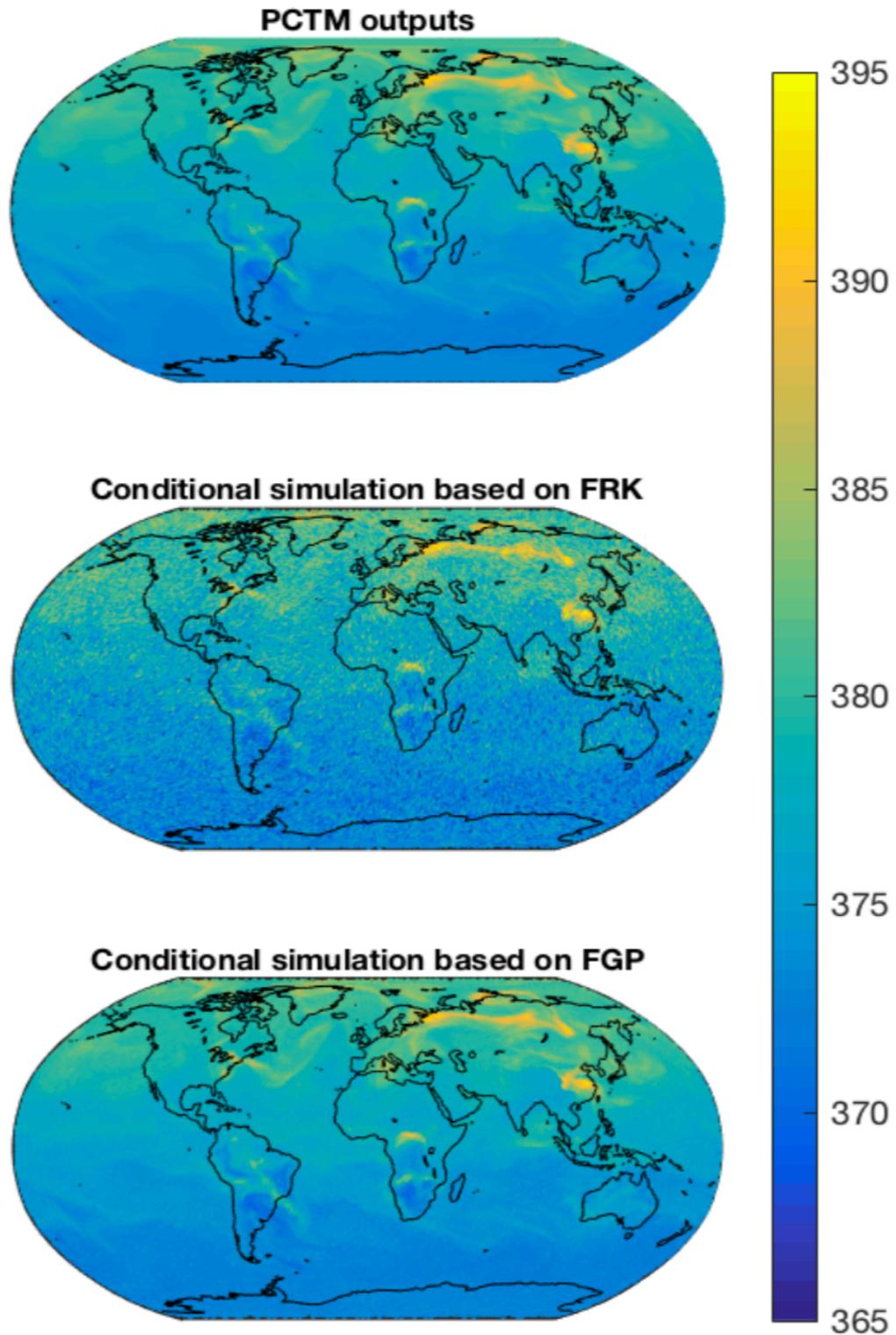


Figure 9. Global map of PCTM outputs and downscaled fields. The top panel shows the surface CO₂ concentrations with unit parts per million (ppm) from PCTM at $1^\circ \times 1.25^\circ$ latitude and longitude; middle and bottom panels show the downscaled fields for surface CO₂ based on FRK and FGP at 30 km spatial resolution.

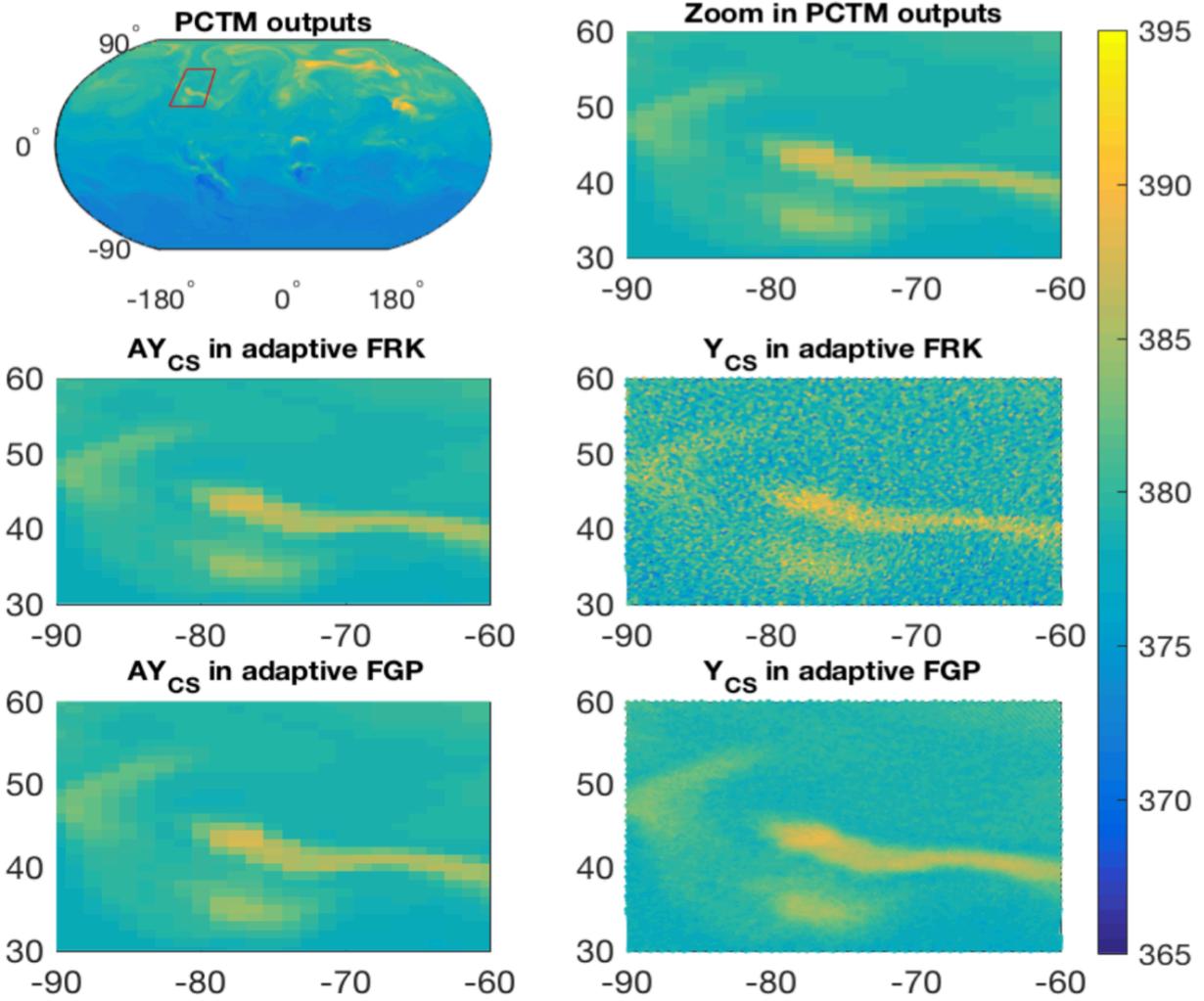


Figure 10. PCTM outputs and aggregated downscaled fields at $1^\circ \times 1.25^\circ$ resolution, and downscaled fields based on FRK and FGP at 30 km spatial resolution (in ppm). Top-left panel shows the global PCTM outputs with zoom-in region identified by the black rectangle. The top-right panel shows the PCTM outputs for the zoom-in region at 30 km resolution. Middle-left panel shows the aggregated conditional simulated values based on FRK; middle-right panel shows the conditional simulated values based on FRK. Bottom-left panel shows the aggregated conditional simulated values based on FGP; bottom-right panel shows the conditional simulated values based on FGP.

make the classical regularization terms computationally practical for large spatial datasets.

In this article, we also show that by combining the low-rank and the CAR components together in the FGP model, the resulting downscaled field and spatial predictions are improved substantially. As in Ma and Kang (2018a), alternative models that take into account the modeling and computational complexity trade-off can be adopted to describe spatial dependence in ξ , such

as the Gaussian Markov random field in Lindgren et al. (2011). Expert knowledge of the likely behavior of the fine-scale process, if available, can also be incorporated into the model for ξ .

Our downscaling framework is designed to produce NRs at very fine spatial resolution. Ideally, we would like NRs not only at fine spatial resolution in a plane (horizontal resolution), but also at fine resolution in the vertical and temporal dimensions. As our downscaling framework is model-based, it can be extended to allow multiple input dimensions (longitude, latitude, height, time). The tensor basis functions in Nguyen et al. (2017) provide a way to describe both vertical and horizontal dependence. The spatio-temporal version of FGP model (Ma and Kang, 2018b) can be used to generate NRs across time. We will consider these extensions in future work.

The current downscaling framework can be used as a building block in hierarchical models for non-Gaussian distributions or nonlinear constraints. Compared to heuristic methods, our downscaling framework provides a coherent and rigorous way to propagate physical relationships at coarse resolutions down to fine resolutions. Our current downscaling framework can also be extended to the multivariate downscaling framework to generate NRs for multiple geophysical processes. This is closely related to the theme of super-resolution imaging (Tian and Ma, 2011), which requires jointly downscaling multiple coarse-resolution images to obtain a single fine-resolution image.

Finally, our downscaling model is able to generate whole ensembles of high-resolution spatial fields through conditional simulation. These ensembles can facilitate probabilistic uncertainty quantification in observation system design and data assimilation algorithm evaluation at the fine resolutions at which those systems and algorithms are intended to operate. Meanwhile, our methods can be further extended to handle both numerical model output and physical observations, and thus may be useful in statistical emulation and uncertainty quantification for multi-fidelity computer models. It is also of interest to use the spatial modeling strategy in this article together with methods for computer model calibration to address a broad range of problems including data assimilation and retrievals in remote sensing and atmospheric sciences. These topics will be investigated in future research.

Acknowledgements

The research was partially carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation. This research was Ma’s Ph.D. dissertation and was partially supported by the Charles Phelps Taft Dissertation Fellowship at the University of Cincinnati. Kang’s research was partially supported by the Simons Foundation’s Collaboration Award (#317298) and the Taft Research Center at the University of Cincinnati. We wish to acknowledge Dr. Noel Cressie, Dr. Matthias Katzfuss, and Dr. Vineet Yadav for valuable discussions and suggestions on this work. We thank the Editor, Associate Editor and two anonymous referees for constructive comments and suggestions.

Appendix

A The EM Algorithm for Downscaling Models with FRK and FGP

In this section, we will give complete derivation of EM algorithms for FRK and FGP under downscaling framework in detail. Recall that the complete data log-likelihood function is given in Eq. (2.16). The EM algorithm consists of two steps: E-step and M-step, and these two steps are run iteratively starting with initial values until the EM algorithm converges. Given parameter estimates $\theta_{[\ell]}$ in the ℓ -th iteration of the EM algorithm, the conditional distribution of η given $\tilde{\mathbf{Y}}$ is multivariate normal with mean $\mu_{\eta|\tilde{\mathbf{Y}},\theta_{[\ell]}}$ and covariance matrix $\Sigma_{\eta|\tilde{\mathbf{Y}},\theta_{[\ell]}}$, which are

$$\mu_{\eta|\tilde{\mathbf{Y}},\theta_{[\ell]}} = \mathbf{K}_{[\ell]}(\mathbf{A}\mathbf{S})^T(\mathbf{A}\Sigma_{[\ell]}\mathbf{A}^T)^{-1}[\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\beta_{[\ell]}], \quad (\text{A.1})$$

$$\Sigma_{\eta|\tilde{\mathbf{Y}},\theta_{[\ell]}} = \mathbf{K}_{[\ell]} - \mathbf{K}_{[\ell]}(\mathbf{A}\mathbf{S})^T(\mathbf{A}\Sigma_{[\ell]}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{S})\mathbf{K}_{[\ell]}^T, \quad (\text{A.2})$$

where the subscript “[ℓ]” indicates that the quantity is evaluated with parameters $\beta_{[\ell]}$, $\mathbf{K}_{[\ell]}$, $\sigma_{\xi,[\ell]}^2$ in FRK model, and with parameters $\beta_{[\ell]}$, $\mathbf{K}_{[\ell]}$, $\tau_{[\ell]}^2$, $\gamma_{[\ell]}$ in FGP model. In E-step, taking conditional

expectation of the complete data log-likelihood w.r.t. $\boldsymbol{\eta}$ given $\tilde{\mathbf{Y}}$ with parameters $\boldsymbol{\theta}_{[\ell]}$ will give the $Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{[\ell]})$ function. The twice-negative Q function is

$$\begin{aligned}
-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{[\ell]}) &= E_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}[-2 \ln L(\boldsymbol{\eta}, \tilde{\mathbf{Y}})] \\
&= \ln |\mathbf{K}| + \ln |\mathbf{D}^{-1}| + [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}]^T \mathbf{D} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}] \\
&\quad - 2[\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}]^T \mathbf{D} (\mathbf{A}\mathbf{S}) \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} + \text{tr}\{[(\mathbf{A}\mathbf{S})^T \mathbf{D} (\mathbf{A}\mathbf{S}) + \mathbf{K}^{-1}] \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}\} \\
&\quad + \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}^T \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}.
\end{aligned}$$

In M-step, the Q function is maximized w.r.t. parameters $\boldsymbol{\theta}$ to obtain updated parameters $\boldsymbol{\theta}_{[\ell+1]}$. As the formulas for FRK and FGP are slightly different, we first give formulas for parameter updates in FRK. In the downscaling model based on FRK, taking derivative of $-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{[\ell]})$ w.r.t. $\boldsymbol{\beta}$, \mathbf{K} , σ_ξ^2 and setting it to zero will give

$$\boldsymbol{\beta}_{[\ell+1]} = [(\mathbf{A}\mathbf{X})^T (\mathbf{A}\mathbf{A}^T)^{-1} (\mathbf{A}\mathbf{X})]^{-1} (\mathbf{A}\mathbf{X})^T (\mathbf{A}\mathbf{A}^T)^{-1} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{S}) \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}], \quad (\text{A.3})$$

$$\mathbf{K}_{[\ell+1]} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} + \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}^T, \quad (\text{A.4})$$

$$\begin{aligned}
\sigma_{\xi, [\ell+1]}^2 &= \{[\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}_{[\ell+1]}]^T (\mathbf{A}\mathbf{A}^T)^{-1} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}_{[\ell+1]} - 2(\mathbf{A}\mathbf{S}) \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}] \\
&\quad + \text{tr}[(\mathbf{A}\mathbf{S})^T (\mathbf{A}\mathbf{A}^T)^{-1} (\mathbf{A}\mathbf{S}) \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}]\} / M - \sigma_\epsilon^2.
\end{aligned} \quad (\text{A.5})$$

In the downscaling model based on FGP, taking derivative of $-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{[\ell]})$ w.r.t. $\boldsymbol{\beta}$, \mathbf{K} , and setting it to zero will give

$$\hat{\boldsymbol{\beta}} = [(\mathbf{A}\mathbf{X})^T \mathbf{D} (\mathbf{A}\mathbf{X})]^{-1} (\mathbf{A}\mathbf{X})^T \mathbf{D} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{S}) \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}], \quad (\text{A.6})$$

$$\mathbf{K}_{[\ell+1]} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} + \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}^T, \quad (\text{A.7})$$

where $\mathbf{K}_{[\ell+1]}$ is updated explicitly, but $\hat{\boldsymbol{\beta}}$ depends on values of τ^2 and γ . To get parameters updates $\tau_{[\ell+1]}^2$ and $\gamma_{[\ell+1]}$, the following function needs to be minimized w.r.t. τ^2, γ :

$$\begin{aligned}
f(\tau^2, \gamma) &= \ln |\mathbf{D}| + [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\hat{\boldsymbol{\beta}}]^T \mathbf{D} [\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\hat{\boldsymbol{\beta}}] \\
&\quad - 2[\tilde{\mathbf{Y}} - (\mathbf{A}\mathbf{X})\boldsymbol{\beta}]^T \mathbf{D} (\mathbf{A}\mathbf{S}) \boldsymbol{\mu}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}} + \text{tr}[(\mathbf{A}\mathbf{S})^T \mathbf{D} (\mathbf{A}\mathbf{S}) \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\tilde{\mathbf{Y}}, \boldsymbol{\theta}_{[\ell]}}],
\end{aligned} \quad (\text{A.8})$$

where $\mathbf{D} = [\mathbf{A}(\mathbf{I} - \gamma\mathbf{H})^{-1}\mathbf{A}^T/\tau^2 + \mathbf{A}\mathbf{A}^T/\sigma_\epsilon^2]^{-1}$. By plugging in the function $f(\tau^2, \gamma)$ with $\hat{\boldsymbol{\beta}}$ in Eq. (A.6), numerical optimization such as interior-point or active-set algorithm can be used to obtain optimal values for τ^2 and γ . The optimal values $\tau_{[\ell+1]}^2$ and $\gamma_{[\ell+1]}$ are then plugged in Eq. (A.6) to obtain parameter updates $\boldsymbol{\beta}_{[\ell+1]}$. The function $f(\tau^2, \gamma)$ can be evaluated efficiently

as it has same computational cost to evaluate the twice negative marginal log-likelihood function (2.10). To accelerate the EM algorithm in FGP, we use Akein's acceleration scheme to update EM algorithm, which is called SQUAREM algorithm in Berline and Roland (2007) and Varadhan and Roland (2008).

The initial value for β in the EM algorithms of FRK and FGP can be set as the ordinary least square estimate $\hat{\beta}_{\text{ols}} = [(\mathbf{A}\mathbf{X})^T(\mathbf{A}\mathbf{X})]^{-1}(\mathbf{A}\mathbf{X})^T\tilde{\mathbf{Y}}$. The initial value for \mathbf{K} can be set to $0.9\hat{\sigma}_{\tilde{\mathbf{Y}}}^2\mathbf{I}_r$, where $\hat{\sigma}_{\tilde{\mathbf{Y}}}^2$ is the empirical variance of $\tilde{\mathbf{Y}}$. The initial value for σ_{ξ}^2 is $0.1\hat{\sigma}_{\tilde{\mathbf{Y}}}^2$ in FRK. The initial value for τ^2 can be set to $0.1\hat{\sigma}_{\tilde{\mathbf{Y}}}^2$, and γ is constrained in the interval $(1/\lambda_1, 1/\lambda_N)$, where λ_1, λ_N are the smallest and largest eigenvalues for the proximity matrix \mathbf{H} . The EM algorithm starts with the initial values $\theta_{[\ell]}$ at $\ell = 0$, and then the E-step and M-step are carried out iteratively with new initial values from previous M-step until certain convergence criterion is satisfied, e.g., the difference of the parameters θ at two consecutive iterations is less than a threshold. The convergence of the EM algorithms is monitored by the twice-negative-marginal-log-likelihood function (2.10).

B Technical Proofs

Proof of Proposition 1:

- (1) Recall the definition of \mathbf{Y}_{CS} in Algorithm 1, $\mathbf{Y}_{\text{CS}} = \mathbf{Y}_{\text{NS}} + \Sigma\mathbf{A}^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\mathbf{Y}_{\text{NS}})$. It follows immediately that $\mathbf{A}\mathbf{Y}_{\text{CS}} = \mathbf{A}\mathbf{Y}_{\text{NS}} + \mathbf{A}\Sigma\mathbf{A}^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\mathbf{Y}_{\text{NS}}) = \mathbf{A}\mathbf{Y}$. Thus, conditional on $\mathbf{A}\mathbf{Y} = \tilde{\mathbf{Y}}$, we have $\mathbf{A}\mathbf{Y}_{\text{CS}} = \tilde{\mathbf{Y}}$.
- (2) Let $\mathbf{H} \equiv \Sigma\mathbf{A}^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}$. Then $\mathbf{Y}_{\text{CS}} = \mathbf{Y}_{\text{NS}} + \mathbf{H}(\mathbf{A}\mathbf{Y} - \mathbf{A}\mathbf{Y}_{\text{NS}})$. The expectation of \mathbf{Y}_{CS} given $\mathbf{A}\mathbf{Y}$ is

$$\begin{aligned}
E[\mathbf{Y}_{\text{CS}} | \mathbf{A}\mathbf{Y}] &= E(\mathbf{Y}_{\text{NS}}) + \mathbf{H}[\mathbf{A}\mathbf{Y} - \mathbf{A}E(\mathbf{Y}_{\text{NS}})] \\
&= \boldsymbol{\mu} + \mathbf{H}(\mathbf{A}\mathbf{Y} - \mathbf{A}\boldsymbol{\mu}), \\
&= \boldsymbol{\mu} + \Sigma\mathbf{A}^T(\mathbf{A}\Sigma\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{Y} - \mathbf{A}\boldsymbol{\mu}).
\end{aligned}$$

The covariance matrix of \mathbf{Y}_{CS} given $\mathbf{A}\mathbf{Y}$ is

$$\begin{aligned}
\text{cov}(\mathbf{Y}_{\text{CS}} \mid \mathbf{A}\mathbf{Y}) &= \text{cov}(\mathbf{Y}_{\text{NS}} - \mathbf{H}\mathbf{A}\mathbf{Y}_{\text{NS}} \mid \mathbf{A}\mathbf{Y}) = \text{cov}[(\mathbf{I} - \mathbf{H}\mathbf{A})\mathbf{Y}_{\text{NS}} \mid \mathbf{A}\mathbf{Y}] \\
&= (\mathbf{I} - \mathbf{H}\mathbf{A})\text{cov}(\mathbf{Y}_{\text{NS}})(\mathbf{I} - \mathbf{H}\mathbf{A})^T = (\mathbf{I} - \mathbf{H}\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{H}\mathbf{A})^T \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T\mathbf{H}^T - \mathbf{H}\mathbf{A}\boldsymbol{\Sigma} + \mathbf{H}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{H}^T \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}.
\end{aligned}$$

Since \mathbf{Y}_{NS} follows the multivariate normal distribution and \mathbf{Y}_{CS} is a linear transformation of \mathbf{Y}_{NS} (conditional on $\mathbf{A}\mathbf{Y}$), it is easy to verify that $\mathbf{Y}_{\text{CS}} \mid \mathbf{A}\mathbf{Y}$ follows a multivariate normal distribution with mean and covariance given above.

- (3) It is obvious that \mathbf{Y}_{CS} follows a multivariate normal distribution since it is a linear combination of multivariate normal vectors \mathbf{Y} and \mathbf{Y}_{NS} . Therefore, it suffices to show that \mathbf{Y}_{CS} has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It follows that

$$E(\mathbf{Y}_{\text{CS}}) = E[E(\mathbf{Y}_{\text{CS}} \mid \mathbf{A}\mathbf{Y})] = \boldsymbol{\mu} + \mathbf{H}[\mathbf{A}E(\mathbf{Y}) - \mathbf{A}\boldsymbol{\mu}] = \boldsymbol{\mu}.$$

The covariance matrix of \mathbf{Y}_{CS} is

$$\begin{aligned}
\text{cov}(\mathbf{Y}_{\text{CS}}) &= \text{cov}(\mathbf{Y}_{\text{NS}} - \mathbf{H}\mathbf{A}\mathbf{Y}_{\text{NS}} + \mathbf{H}\mathbf{A}\mathbf{Y}) = \text{cov}[(\mathbf{I} - \mathbf{H}\mathbf{A})\mathbf{Y}_{\text{NS}}] + \text{cov}(\mathbf{H}\mathbf{A}\mathbf{Y}) \\
&= (\mathbf{I} - \mathbf{H}\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{H}\mathbf{A})^T + \mathbf{H}\mathbf{A}\boldsymbol{\Sigma}(\mathbf{H}\mathbf{A})^T \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma} \\
&= \boldsymbol{\Sigma}.
\end{aligned}$$

- (4) It follows that $E(\mathbf{Y}_{\text{CS}} - \mathbf{Y})^2 = \text{cov}(\mathbf{Y}_{\text{CS}} - \mathbf{Y}) = \text{cov}(\mathbf{Y}_{\text{CS}}) - \text{cov}(\mathbf{Y}_{\text{CS}}, \mathbf{Y}) - \text{cov}(\mathbf{Y}, \mathbf{Y}_{\text{CS}}) + \text{cov}(\mathbf{Y})$, which is $2[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A}\boldsymbol{\Sigma}]$.

References

- Abida, R., Attié, J.-L., El Amraoui, L., Ricaud, P., Lahoz, W., Eskes, H., Segers, A., Curier, L., de Haan, J., Kujanpää, J., Nijhuis, A. O., Tamminen, J., Timmermans, R., and Veeffkind, P. (2017). Impact of spaceborne carbon monoxide observations from the S-5P platform on tropospheric composition analyses and forecasts. *Atmospheric Chemistry and Physics*, 17(2):1081–1103.
- Akritas, A. G., Akritas, E. K., and Malaschonok, G. I. (1996). Various proofs of sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, 42(4):585 – 593. Sybolic Computation, New Trends and Developments.

- Atkinson, P. (2001). Geostatistical regularization in remote sensing. In Tate, N. and Atkinson, P., editors, *Modelling Scale in Geographical Information Science*, pages 237–260. John Wiley and Sons Ltd.
- Atkinson, P. M. (2013). Downscaling in remote sensing. *Int. J. Applied Earth Observation and Geoinformation*, 22:106–114.
- Atlas, R., Hoffman, R. N., Ma, Z., Emmitt, G. D., Jr., S. A. W., Greco, S., Tucker, S., Bucci, L., Annane, B., Hardesty, R. M., and Murillo, S. (2015). Observing system simulation experiments (OSSEs) to evaluate the potential impact of an optical autocovariance wind lidar (OAWL) on numerical weather prediction. *Journal of Atmospheric and Oceanic Technology*, 32(9):1593–1613.
- Ba, S. and Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838–1860.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press.
- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.
- Berlinet, A. and Roland, C. (2007). Acceleration schemes with application to the EM algorithm. *Computational Statistics and Data Analysis*, 51(8):3689–3702.
- Berrocal, V. J., Craigmire, P. F., and Guttorp, P. (2012). Regional climate model assessment using statistical upscaling and downscaling techniques. *Environmetrics*, 23(5):482–492.
- Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A bivariate spacetime downscaler under space and time misalignment. *Ann. Appl. Stat.*, 4(4):1942–1975.
- Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, 66(4):1069–1077.
- Brasseur, G. P. and Jacob, D. J. (2017). *Modeling of Atmospheric Chemistry*. Cambridge University Press, Cambridge.
- Cracknell, A. P. (1998). Review article synergy in remote sensing-what’s in a pixel? *International Journal of Remote Sensing*, 19(11):2025–2047.
- Craigmire, P. F., Calder, C. A., Li, H., Paul, R., and Cressie, N. (2009). Hierarchical model building, fitting, and checking: a behind-the-scenes look at a Bayesian analysis of arsenic exposure pathways. *Bayesian Anal.*, 4(1):1–35.
- Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17(5):563–586.
- Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition.
- Cressie, N. and Johannesson, G. (2006). Spatial prediction for massive datasets. In *Mastering the Data Explosion in the Earth and Environmental Sciences*, pages 1–11. Canberra, Australia: Australian Academy of Science.

- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.
- Cressie, N. and Kang, E. L. (2010). High-resolution digital soil mapping: Kriging for very large datasets. In *Proximal Soil Sensing*, pages 49–63. Springer.
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.
- Edwards, D. P., Arellano, A. F., and Deeter, M. N. (2009). A satellite observation system simulation experiment for carbon monoxide in the lowermost troposphere. *Journal of Geophysical Research: Atmospheres*, 114, D14304. DOI:<http://dx.doi.org/10.1029/2008JD011375>.
- Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.
- Errico, R. M., Yang, R., Priv, N. C., Tai, K.-S., Todling, R., Sienkiewicz, M. E., and Guo, J. (2013). Development and validation of observing-system simulation experiments at NASA’s global modeling and assimilation office. *Quarterly Journal of the Royal Meteorological Society*, 139(674):1162–1178.
- Eskes, H. J., Velthoven, P. F. J. V., Valks, P. J. M., and Kelder, H. M. (2003). Assimilation of GOME total-ozone satellite observations in a three-dimensional tracer-transport model. *Quarterly Journal of the Royal Meteorological Society*, 129(590):1663–1681.
- Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873 – 2884.
- Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K. G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N. (2006). Climate Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison. *Journal of Climate*, 19(14):3337–3353.
- Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45.
- Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.
- Glotter, M., Elliott, J., McInerney, D., Best, N., Foster, I., and Moyer, E. J. (2014). Evaluating the utility of dynamical downscaling in agricultural impacts projections. *Proceedings of the National Academy of Sciences*, 111(24):8776–8781.
- Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.
- Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large com-

- puter experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.
- Guillas, S., Bao, J., Choi, Y., and Wang, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmospheric Environment*, 42(6):1338 – 1348.
- Gutmann, E. D., Rasmussen, R. M., Liu, C., Ikeda, K., Gochis, D. J., Clark, M. P., Dudhia, J., and Thompson, G. (2012). A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *Journal of Climate*, 25(1):262–281.
- Haas, T. C. (1990). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759–1769.
- Hammerling, D. M., Michalak, A. M., and Kawa, S. R. (2012). Mapping of CO₂ at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2. *Journal of Geophysical Research: Atmospheres*, 117, D06306.
- Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60.
- Higdon, D., Kennedy, M., Cavendish, J. C., Cafoe, J. A., and Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM J. Scientific Computing*.
- Hoffman, R. N. and Atlas, R. (2016). Future observing system simulation experiments. *Bulletin of the American Meteorological Society*, 97(9):1601–1616.
- Kang, E. L., Cressie, N., and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38(2):271–289.
- Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24(3):189–200.
- Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.
- Katzfuss, M. and Cressie, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4):430–446.
- Kawa, S. R., Erickson, D. J., Pawson, S., and Zhu, Z. (2004). Global CO₂ transport simulations using meteorological data from the NASA data assimilation system. *Journal of Geophysical Research: Atmospheres*, 109, D18312.
- Kawa, S. R., MAO, J., ABSHIRE, J. B., COLLATZ, G. J., SUN, X., and WEAVER, C. J. (2010). Simulation studies for a space-based CO₂ lidar mission. *Tellus B*, 62(5):759–769.
- Kennedy, M. C. and O’Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13.
- Kennedy, M. C. and O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.
- Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press.
- Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J., and Schwartz, J. (2011). Assessing temporally

- and spatially resolved PM_{2.5} exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, 45(35):6267 – 6275.
- Konomi, B. A., Sang, H., and Mallick, B. K. (2014). Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 23(3):802–829.
- Lenderink, G., van Ulden, A., van den Hurk, B., and Keller, F. (2007). A study on combining global and regional climate model results for generating climate scenarios of temperature and precipitation for the Netherlands. *Climate Dynamics*, 29(2):157–176.
- 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(4):423–498.
- Liu, X., Mizzi, A. P., Anderson, J. L., Fung, I. Y., and Cohen, R. C. (2017). Assimilation of satellite NO₂ observations at high spatial resolution using OSSEs. *Atmospheric Chemistry and Physics*, 17(11):7067–7081.
- Ma, P. and Kang, E. L. (2018a). Fused Gaussian process for very large spatial data. *Journal of Computational and Graphical Statistics*. Under revision.
- Ma, P. and Kang, E. L. (2018b). Spatio-temporal data fusion for massive sea surface temperature data from MODIS and AMSR-E instruments. *In preparation*.
- Marchetti, Y., Nguyen, H., Braverman, A., and Cressie, N. (2017). Spatial data compression via adaptive dispersion clustering. *Computational Statistics & Data Analysis*. DOI:<http://dx.doi.org/10.1016/j.csda.2017.08.004>.
- Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, 107(499):1004–1018.
- Nguyen, H., Cressie, N., and Braverman, A. (2017). Multivariate spatial data fusion for very large remote sensing datasets. *Remote Sensing*, 9(2):142. DOI:10.3390/rs9020142.
- Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185.
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599.
- OSC (1987). Ohio Supercomputer Center. Columbus, OH: Ohio Supercomputer Center. <http://osc.edu/ark:/19495/f5s1ph73>.
- Parazoo, N. C., Denning, A. S., Berry, J. A., Wolf, A., Randall, D. A., Kawa, S. R., Pauluis, O., and Doney, S. C. (2011). Moist synoptic transport of CO₂ along the mid-latitude storm track. *Geophysical Research Letters*, 38, L09804.
- Putman, W. M. and Suarez, M. (2011). Cloud-system resolving simulations with the NASA Goddard Earth Observing System global atmospheric model (GEOS-5). *Geophysical Research Letters*, 38, L16809. DOI:10.1029/2011GL048438.
- Ravishanker, N. and Dey, D. (2002). *A First Course in Linear Model Theory*. Chapman and

Hall/CRC, Cova Raton, FL.

- Reich, B. J., Chang, H. H., and Foley, K. M. (2014). A spectral method for spatial downscaling. *Biometrics*, 70(4):932–942.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall.
- Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.
- Sahr, K., White, D., and Kimerling, A. J. (2003). Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134.
- Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.
- Shi, H. and Kang, E. L. (2017). Spatial data fusion for large non-Gaussian remote sensing datasets. *Stat*, 6(1):390–404.
- Shi, T. and Cressie, N. (2007). Global statistical analysis of MISR aerosol data: a massive data product from NASA’s Terra satellite. *Environmetrics*, 18(7):665–680.
- Shusterman, A. A., Teige, V. E., Turner, A. J., Newman, C., Kim, J., and Cohen, R. C. (2016). The BERkeley atmospheric CO₂ observation network: initial evaluation. *Atmospheric Chemistry and Physics*, 16(21):13449–13463.
- Stough, T., Braverman, A., Cressie, N., Kang, E. L., Michalak, A. M., Nguyen, H., and Sahr, K. (2014). Visualizing massive spatial datasets using multi-resolution global grids. Technical Report 05-14 (24 pp.), NIASRA Working Paper.
- 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(10):3311–3319.
- Tian, J. and Ma, K.-K. (2011). A survey on super-resolution imaging. *Signal, Image and Video Processing*, 5(3):329–342.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Tsai, C.-C., Yang, S.-C., and Liou, Y.-C. (2014). Improving quantitative precipitation nowcasting with a local ensemble transform Kalman filter radar data assimilation system: observing system simulation experiments. *Tellus A: Dynamic Meteorology and Oceanography*, 66(1):21804.
- Tzeng, S. and Huang, H.-C. (2017). Resolution adaptive fixed rank kriging. *Technometrics*. DOI:10.1080/00401706.2017.1345701.
- Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.
- Wakefield, J. and Lyons, H. (2017). Spatial Aggregation and the Ecological Fallacy. In *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 541–558. CRC Press.

- Webster, W. P. and Duffy, D. Q. (2015). High resolution nature runs and the big data challenge. *International Computing in Atmospheric Sciences Workshop*.
- Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396.
- Wikle, C. K. and Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2):1–16.
- Wilby, R. L. and Wigley, T. M. L. (1997). Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548.
- Zammit-Mangion, A. and Cressie, N. (2017). FRK: An R package for spatial and spatio-temporal prediction with large datasets. <https://arxiv.org/abs/1705.08105>.
- Zhang, X., Gurney, K. R., Rayner, P., Liu, Y., and Asefi-Najafabady, S. (2014). Sensitivity of simulated CO₂ concentration to regridding of global fossil fuel CO₂ emissions. *Geoscientific Model Development*, 7(6):2867–2874.
- Zhou, J., Fuentes, M., and Davis, J. (2011). Calibration of numerical model output using non-parametric spatial density functions. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(4):531–553.
- Zhou, Y. and Michalak, A. M. (2009). Characterizing Attribute Distributions in Water Sediments by Geostatistical Downscaling. *Environmental Science & Technology*, 43(24):9267–9273.
- Zhu, Y., Kang, E. L., Bo, Y., Tang, Q., Cheng, J., and He, Y. (2015). A robust fixed rank kriging method for improving the spatial completeness and accuracy of satellite SST products. *IEEE Transactions on Geoscience and Remote Sensing*, 53(9):5021–5035.
- Zoogman, P., Jacob, D. J., Chance, K., Zhang, L., Sager, P. L., Fiore, A. M., Eldering, A., Liu, X., Natraj, V., and Kulawik, S. S. (2011). Ozone air quality measurement requirements for a geostationary satellite mission. *Atmospheric Environment*, 45(39):7143 – 7150.