Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Fixed Rank Filtering for Spatio-Temporal Data

Noel CRESSIE, Tao SHI, and Emily L. KANG

Datasets from remote-sensing platforms and sensor networks are often spatial, temporal, and very large. Processing massive amounts of data to provide current estimates of the (hidden) state from current and past data is challenging, even for the Kalman filter. A large number of spatial locations observed through time can quickly lead to an overwhelmingly high-dimensional statistical model. Dimension reduction without sacrificing complexity is our goal in this article. We demonstrate how a Spatio-Temporal Random Effects (STRE) component of a statistical model reduces the problem to one of fixed dimension with a very fast statistical solution, a methodology we call Fixed Rank Filtering (FRF). This is compared in a simulation experiment to successive, spatial-only predictions based on an analogous Spatial Random Effects (SRE) model, and the value of incorporating temporal dependence is quantified. A remote-sensing dataset of aerosol optical depth (AOD), from the Multi-angle Imaging SpectroRadiometer (MISR) instrument on the Terra satellite, is used to compare spatio-temporal FRF with spatial-only prediction. FRF achieves rapid production of optimally filtered AOD predictions, along with their prediction standard errors. In our case, over 100,000 spatio-temporal data were processed: Parameter estimation took 64.4 seconds and optimal predictions and their standard errors took 77.3 seconds to compute. Supplemental materials giving complete details on the design and analysis of a simulation experiment, the simulation code, and the MISR data used are available on-line.

Key Words: Aerosol optical depth (AOD); Fixed Rank Kriging (FRK); FRF; Spatial Random Effects (SRE) model; Spatio-Temporal Random Effects (STRE) model; Vector autoregressive (VAR) process.

1. INTRODUCTION

Many datasets not only contain attribute information, but also have spatial information (attribute data were collected somewhere) and temporal information (attribute data were collected at some time point). This spatial and temporal information can help to

Noel Cressie is Distinguished Professor (E-mail: *ncressie@stat.osu.edu*), Tao Shi is Assistant Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247. Emily L. Kang is Postdoctoral Fellow, SAMSI, Research Triangle Park, NC 27709-4006.

Journal of Computational and Graphical Statistics, Volume 19, Number 3, Pages 724–745
DOI: 10.1198/jcgs.2010.09051

separate out possible causative effects between two attributes from purely environmental (e.g., regional/seasonal) effects. Moreover, nearness of attributes in space or time is often manifested by a distinct lack of independence. In such circumstances, any model that involves a stochastic component should account for spatio-temporal dependence; in doing so, remarkable improvements can be made in filling in spatial and temporal gaps between observations, and filtering out unwanted measurement errors in the observations.

Statistics for spatio-temporal data have followed two paradigms. The first is to think of time as providing an extra dimension beyond the spatial dimensions (d, say), resulting in covariance models and analyses adapted to $(d+1)$-dimensional space. While there are many problems for which this descriptive statistical approach provides satisfactory solutions, it does not explicitly model the etiology of the phenomenon under study. In contrast, the dynamical-statistical approach models how the current state depends on previous states through dynamical relationships that are either mechanistic or probabilistic. The example best known in a purely temporal context is the standard Gaussian (Gau) autoregressive process of order 1 [AR(1)]: *Mechanistically*, the AR(1) process can be written as

$$Y_{t+1} = \alpha Y_t + \nu_{t+1}; \qquad t = 1, 2, \ldots,$$

where ν_{t+1} is independent of Y_t, $\{\nu_t\}$ are independent and identically distributed $\mathrm{Gau}(0, (1-\alpha^2)\sigma_\nu^2)$, and Y_1 is $\mathrm{Gau}(0, \sigma_\nu^2)$. This is a dynamical-statistical model, as is its *probabilistic* equivalent,

$$Y_{t+1} | Y_t, \ldots, Y_1 \sim \mathrm{Gau}(\alpha Y_t, (1-\alpha^2)\sigma_\nu^2); \qquad t = 1, 2, \ldots,$$

and Y_1 is $\mathrm{Gau}(0, \sigma_\nu^2)$. In contrast, a descriptive specification of the same AR(1) process is expressed through

$$\mathrm{cov}(Y_{t+k}, Y_t) = \alpha^k \sigma_\nu^2; \qquad k = 0, 1, 2, \ldots.$$

We prefer the dynamical (mechanistic or probabilistic) specification for several reasons. First, it is usually derived from scientific knowledge about the phenomenon under study. Second, one can derive covariance models (used in the descriptive specification) from the dynamical specification, and such models can be guaranteed to be valid (i.e., nonnegative-definite). Third, sequential updating allows rapid smoothing, filtering, and forecasting of the state from noisy and missing data observed at different times. These advantages have been well recognized in the signal-processing and time series literatures (Kalman 1960; Anderson 1984; Shumway and Stoffer 2006).

Statistics for *spatial* data also faces the problem of dealing with noisy and missing data, but there are no dynamics. In the geostatistics and, more generally, the spatial-statistics literature (e.g., Matheron 1963; Cressie 1993; Banerjee, Carlin, and Gelfand 2004), the predominant approach to spatial prediction is kriging. However, without a natural ordering in space, there is no obvious way to speed up kriging in the way that Kalman filtering does. Huang, Cressie, and Gabrosek (2002) ordered the data based on their resolutions, and Nychka et al. (1996) and Kammann and Wand (2003) used a space-filling sequence of spatial locations. The problem is fundamentally one of data-dimension reduction, to which one solution was recently given by Cressie and Johannesson (2006, 2008). They defined a Spatial Random Effects (SRE) model where the unknown random variables to be predicted are fixed in number and are coefficients of known (not necessarily orthogonal) spatial basis functions. This resulted in a spatial-prediction methodology they called *Fixed Rank Kriging* (FRK). Banerjee et al. (2008) developed a spatial model for large datasets, which then required an approximation to achieve dimension reduction.

Statistics for *spatio-temporal* data inherits a similar need for data-dimension reduction that we saw for spatial data, possibly more so since the data size quickly becomes massive as time progresses. Huang and Cressie (1996), Mardia et al. (1998), Wikle and Cressie (1999), among others, developed spatio-temporal Kalman filters (see Cressie and Wikle 2002, for a summary of that literature), although they did not address the massive-data problem. Johannesson, Cressie, and Huang (2007) proposed a spatio-temporal multiresolution approach to filtering, which from our perspective is restricted to "blocky" basis functions and coarse-resolution-only dynamics. Ghosh et al. (2010) proposed a Bayesian spatio-temporal model that decomposes variance matrices in terms of a lower-triangular matrix and a diagonal matrix, but those matrices are time-invariant and MCMC computations are needed. Lopes, Salazar, and Gamerman (2009) proposed a Bayesian spatial dynamic factor-analysis model that resembles our model, but its parameters are many, identifiability conditions need to be specified, and the use of MCMC leads to comparatively slow computations. A fully Bayesian spatio-temporal analysis that incorporates numerical-model output with irregularly sampled monitoring data was presented by Berrocal, Gelfand, and Holland (2010), again requiring MCMC computations. Generally speaking, for very large, streaming spatial data, Bayesian optimal filtering is unable to produce near-real-time predictions.

In this article, we build a Spatio-Temporal Random Effects (STRE) model that allows both dimension reduction (spatially) and rapid smoothing, filtering, or forecasting (temporally). Here we concentrate on filtering and develop a methodology we call *Fixed Rank Filtering* (FRF). Such an approach has obvious application to datasets from remote-sensing platforms and sensor networks (e.g., Kang 2009; Kang, Cressie, and Shi 2010).

The methodology we are proposing could be viewed as a type of data assimilation, which is a technique often used to re-initialize meteorological forecasts (e.g., Ghil and Malanotte-Rizzoli 1991; Talagrand 1997; Kalnay 2003). Meteorological data acquired rapidly can be processed rapidly using FRF and could then be used to re-initialize a numerical forecast. The dimension reduction and rapid computation of filtered values, along with their standard errors (a measure of uncertainty), should make our methodology of interest to the meteorology community. Data assimilation was reviewed from a hierarchical-statistical perspective in the recent article by Wikle and Berliner (2006), where they gave the working definition: "... (data assimilation) is an approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws; model output) to obtain an estimate of the distribution of the true state of a process." The model we present in Section 2 can be written hierarchically, and the "fusing" occurs after estimation of parameters.

Section 2 gives the necessary notation and terminology to formulate the problem of optimal filtering, smoothing, and forecasting from very large datasets; we develop FRF and give computational-complexity calculations that demonstrate the effect of its dimensionality reduction. The FRF algorithm is linearly scalable in the size of the spatio-temporal dataset. Section 3 summarizes a simulation study that illustrates the advantage of linking successive SRE models into a dynamic STRE model; for complete details, see the Appendix in the Supplemental Materials section. Section 4 shows the applicability of FRF to a remote-sensing dataset of AOD from the MISR (Multi-angle Imaging SpectroRadiometer) instrument; specifically, we investigate the advantage of the dynamical STRE model by deliberately removing data from a large region over North America and comparing the optimal spatio-temporal predictions with the corresponding optimal spatial-only predictions. We show that rapid production of statistically optimal, gap-filled, filtered AOD values, and their standard errors, is achievable from very large spatio-temporal datasets. Discussion and conclusions are given in Section 5.

2. FIXED RANK FILTERING (FRF) OF SPATIO-TEMPORAL DATA

The methodology at the core of this article is presented in this section. Our goal is fast statistical prediction of a hidden spatio-temporal process based on a potentially massive dataset. This is achieved through spatio-temporal models defined on a space of fixed dimension; the space is defined by the random coefficients of prespecified spatio-temporal basis functions, and the coefficients are assumed to evolve dynamically.

2.1 SPATIO-TEMPORAL RANDOM EFFECTS (STRE) MODEL

Consider a real-valued spatio-temporal process $\{Y(\mathbf{s};t):\mathbf{s}\in D\subset\mathbb{R}^d, t\in\{1,2,\ldots\}\}$, upon which we are interested in making inference based on data that have a component of measurement error. The domain D could be finite, countably infinite, or most generally have positive Lebesgue measure in $\mathbb{R}^d$. Observations and potential observations are given by the data process,

$$Z(\mathbf{s};t)=Y(\mathbf{s};t)+\varepsilon(\mathbf{s};t), \tag{2.1}$$

where $\{\varepsilon(\mathbf{s};t):\mathbf{s}\in D, t\in\{1,2,\ldots\}\}$ is a white-noise Gaussian process with mean zero and, for $\sigma_\varepsilon^2>0$, $\mathrm{var}(\varepsilon(\mathbf{s};t))=\sigma_\varepsilon^2 v_t(\mathbf{s})>0$. We assume further that $E(\varepsilon(\mathbf{s};t)\varepsilon(\mathbf{r};u))=0$, unless $\mathbf{s}=\mathbf{r}$ and $t=u$.

Assume that $Y(\mathbf{s};t)$ has the following structure:

$$Y(\mathbf{s};t)=\mu_t(\mathbf{s})+\nu(\mathbf{s};t), \tag{2.2}$$

where $\mu_t(\cdot)$ is a deterministic (spatio-temporal) mean function, or trend, modeling large-scale variation. For example, $\mu_t(\cdot)=\mathbf{X}_t(\cdot)'\boldsymbol{\beta}_t$ is a common choice, where $\mathbf{X}_t(\cdot)\equiv(X_{1,t}(\cdot),\ldots,X_{p,t}(\cdot))'$ represents a vector process of known covariates, and the coefficients $\boldsymbol{\beta}_t\equiv(\beta_{1,t},\ldots,\beta_{p,t})'$ are in general unknown.

The small-scale variation in (2.2) is modeled as a (spatio-temporal) Gaussian process. At any *fixed* time t, we assume that $\nu(\cdot;t)$ has zero mean and follows an SRE model (Cressie and Johannesson 2008):

$$\nu(\mathbf{s};t)=\mathbf{S}_t(\mathbf{s})'\boldsymbol{\eta}_t+\xi(\mathbf{s};t), \tag{2.3}$$

where $\mathbf{S}_t(\cdot) \equiv (S_{1,t}(\cdot), \ldots, S_{r,t}(\cdot))'$ represents a set of r known spatial basis functions, and $\boldsymbol{\eta}_t \equiv (\eta_{1,t}, \ldots, \eta_{r,t})'$ is a zero-mean Gaussian random vector with $r \times r$ covariance matrix given by K_t. The first term in (2.3) represents, for each t, smooth small-scale spatial variation, captured by the r-dimensional vector of basis functions $\mathbf{S}_t(\cdot)$. In general, the basis functions vary with time but, in our analyses in Sections 3 and 4, we have chosen them to be time invariant, which corresponds to $\nu(\cdot;\cdot)$ having scales of spatial variation that are relatively stable over time.

The second term in (2.3), $\xi(\cdot;\cdot)$, captures the fine-scale variability in exactly the same way that the so-called nugget effect in geostatistics reflects rapid transitions from gold to dross in geological applications (Matheron 1963). It is modeled as a white-noise Gaussian process in space and time with mean zero and variance σ_ξ^2, independent of $\boldsymbol{\eta}_t$. Further discussion of the fine-scale variation component can be found in Section 5.

Now let time progress. In all that follows, $\{\boldsymbol{\eta}_t\}$, $\xi(\cdot;\cdot)$, and $\varepsilon(\cdot;\cdot)$ are assumed to be independent of each other. The STRE model assumes that the component $\{\boldsymbol{\eta}_t : t = 1, 2, \ldots\}$ evolves according to the state equation,

$$\boldsymbol{\eta}_{t+1} = H_{t+1}\boldsymbol{\eta}_t + \boldsymbol{\zeta}_{t+1}; \qquad t = 1, 2, \ldots. \tag{2.4}$$

That is, (2.4) defines a vector autoregressive (VAR) process of order 1, where H_{t+1} is the $r \times r$ first-order autoregressive (or propagator) matrix, and the r-dimensional innovation vector $\boldsymbol{\zeta}_{t+1}$, which is independent of $\boldsymbol{\eta}_t$, has mean zero and innovation variance matrix, $\mathrm{var}(\boldsymbol{\zeta}_{t+1}) \equiv U_{t+1}$. Choosing a row of zeros in H_{t+1} means that the corresponding component of $\boldsymbol{\eta}_{t+1}$ does not evolve dynamically from $\boldsymbol{\eta}_t$, but it does have dependence on other components of $\boldsymbol{\eta}_{t+1}$ through $\boldsymbol{\zeta}_{t+1}$.

Combining (2.1), (2.2), and (2.3), the data process $Z(\cdot;\cdot)$ follows a Spatio-Temporal Mixed Effects (STME) model; that is,

$$Z(\mathbf{s};t) = \mu_t(\mathbf{s}) + \mathbf{S}_t(\mathbf{s})'\boldsymbol{\eta}_t + \xi(\mathbf{s};t) + \varepsilon(\mathbf{s};t); \qquad \mathbf{s} \in D, t = 1, 2, \ldots, \tag{2.5}$$

where $\mu_t(\cdot)$ is deterministic and the other components are stochastic. Wikle and Cressie (1999) formulated a spatio-temporal stochastic model whose form has a strong resemblance to (2.5), where their $\{\mathbf{S}_t(\cdot)\}$ were orthogonal and some components were specifically assumed not to evolve dynamically. The properties of the STRE model (2.3) and (2.4) will now be discussed.

Define the cross-covariances,

$$K_{t_1,t_2} \equiv \mathrm{cov}\big(\boldsymbol{\eta}_{t_1}, \boldsymbol{\eta}_{t_2}\big); \qquad t_1, t_2 = 1, 2, \ldots, \tag{2.6}$$

where we have already notated $K_{t,t}$ as simply K_t. From (2.4), for $t_1 < t_2$,

$$K_{t_1,t_2} = K_{t_1}\big(H_{t_2}H_{t_2-1}\cdots H_{t_1+1}\big)' \tag{2.7}$$

and

$$K_{t+1} = H_{t+1}K_t H_{t+1}' + U_{t+1}. \tag{2.8}$$

As a special case of (2.7), we have

$$L_{t+1} \equiv K_{t,t+1} = K_t H_{t+1}'; \qquad t = 1, 2, \ldots, \tag{2.9}$$

where the $r \times r$ matrix L_{t+1} captures the lag-1 cross-covariances in the STRE components $\{\boldsymbol{\eta}_t : t = 1, 2, \ldots\}$. Based on the propagator matrices $\{H_{t+1} : t = 1, 2, \ldots\}$, the innovation variance matrices $\{U_{t+1} : t = 1, 2, \ldots\}$, and K_1, we can calculate all variances and covariances of the random effects $\{\boldsymbol{\eta}_t : t = 1, 2, \ldots\}$ via (2.7) and (2.8).

In this article, we shall concentrate on predicting $Y(\cdot;\cdot)$ given by (2.2). We assume $\mu_t(\cdot)$ is known (in practice, through scientific models or estimated trends), and hence our effort here is devoted to predicting $\{\nu(\cdot;t) : t = 1, 2, \ldots\}$, the random-effects components of the model.

The data process $Z(\cdot;\cdot)$ is observed at a finite number, n_t, of locations $\{\mathbf{s}_{1,t}, \ldots, \mathbf{s}_{n_t,t}\}$ at time point t; then define the n_t-dimensional vector of data at time t to be

$$\mathbf{Z}(t) \equiv \big(Z(\mathbf{s}_{1,t};t), \ldots, Z\big(\mathbf{s}_{n_t,t};t\big)\big)'; \qquad t = 1, 2, \ldots.$$

Our interest is in inference on the process $Y(\cdot;\cdot)$: Given data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$, we wish to predict the random variable $Y(\mathbf{s}_0;t)$, regardless of whether there is a datum $Z(\mathbf{s}_0;t)$ available or not. This is known as *filtering*. Predicting the random variable $Y(\mathbf{s}_0;u)$, where $u \in \{1, 2, \ldots, t-1\}$, from data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$ is known as *smoothing*; and predicting the random variable $Y(\mathbf{s}_0;u)$, where $u \in \{t+1, t+2, \ldots\}$, from data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$ is known as *forecasting*.

Before we describe the inference methods, we specify the matrices and vectors used in the rest of the article: For $t = 1, 2, \ldots$, S_t is an $n_t \times r$ matrix whose (i, j) element is $S_{j,t}(\mathbf{s}_{i,t})$;

$$\mathrm{var}(\mathbf{Z}(t)) \equiv \Sigma_t = S_t K_t S_t' + \sigma_\xi^2 I_{n_t} + \sigma_\varepsilon^2 V_t \equiv S_t K_t S_t' + D_t \tag{2.10}$$

is an $n_t \times n_t$ positive-definite matrix; I_{n_t} is the $n_t \times n_t$ identity marix; and V_t is the $n_t \times n_t$ diagonal matrix, $\mathrm{diag}(\{v_t(\mathbf{s}_{i,t})\})$, consisting of the n_t measurement-error variances down the diagonal. Next, we need to evaluate $\mathrm{cov}(\mathbf{Z}(t), \xi(\mathbf{s}_0;t))$. Because of the independence of the processes in (2.2) and (2.3), the only contribution comes when $\mathbf{s}_0$ is an observation location. That is,

$$\mathbf{c}_t(\mathbf{s}_0) \equiv \mathrm{cov}(\mathbf{Z}(t), \xi(\mathbf{s}_0;t)) = \sigma_\xi^2\big(I(\mathbf{s}_0 = \mathbf{s}_{1,t}), \ldots, I\big(\mathbf{s}_0 = \mathbf{s}_{n_t,t}\big)\big)', \tag{2.11}$$

where $I(\cdot)$ is an indicator function. Hence,

$$\mathbf{k}_t(\mathbf{s}_0) \equiv \mathrm{cov}(\mathbf{Z}(t), Y(\mathbf{s}_0;t)) = S_t K_t \mathbf{S}_t(\mathbf{s}_0) + \mathbf{c}_t(\mathbf{s}_0). \tag{2.12}$$

Notice that the sizes of these matrices and vectors depend on n_t.

2.2 FIXED RANK KRIGING (FRK)

Fix time t; using only the *current* data $\mathbf{Z}(t)$ collected at time t and assuming the SRE model (2.3) with $\mu_t(\cdot)$ known, the FRK predictor for the process $Y(\mathbf{s}_0;t)$ is

$$\begin{aligned}\hat{Y}(\mathbf{s}_0;t)^{\mathrm{FRK}} &= E(Y(\mathbf{s}_0;t)|\mathbf{Z}(t))\\ &= \mu_t(\mathbf{s}_0) + \mathbf{k}_t(\mathbf{s}_0)'\Sigma_t^{-1}(\mathbf{Z}(t) - \boldsymbol{\mu}(t)); \qquad t = 1, 2, \ldots,\end{aligned} \tag{2.13}$$

where $\mathbf{k}_t(\mathbf{s}_0)$ is given by (2.12), and $\boldsymbol{\mu}(t) \equiv E(\mathbf{Z}(t)) = (\mu_t(\mathbf{s}_{1,t}), \ldots, \mu_t(\mathbf{s}_{n_t,t}))'$. As discussed by Cressie and Johannesson (2006, 2008) and Shi and Cressie (2007), due to the fixed rank r of K_t (much smaller than n_t) in (2.10), one may efficiently invert $\Sigma_t = S_t K_t S_t' + D_t$, by using a Sherman–Morrison–Woodbury formula (e.g., Henderson and Searle 1981). Specifically,

$$\Sigma_t^{-1} = D_t^{-1} - D_t^{-1} S_t \{K_t^{-1} + S_t' D_t^{-1} S_t\}^{-1} S_t' D_t^{-1}, \tag{2.14}$$

where from (2.10), D_t is an $n_t \times n_t$ diagonal matrix. Along with the predictor, $\hat{Y}(\mathbf{s}_0; t)^{\mathrm{FRK}}$, the FRK standard error can also be obtained; it is the root mean squared prediction error,

$$\begin{aligned}\sigma_t(\mathbf{s}_0)^{\mathrm{FRK}} &\equiv \left\{E(Y(\mathbf{s}_0; t) - \hat{Y}(\mathbf{s}_0; t)^{\mathrm{FRK}})^2\right\}^{1/2} \\ &= \{\mathbf{S}_t(\mathbf{s}_0)' K_t \mathbf{S}_t(\mathbf{s}_0) + \sigma_\xi^2 - \mathbf{k}_t(\mathbf{s}_0)' \Sigma_t^{-1} \mathbf{k}_t(\mathbf{s}_0)\}^{1/2}, \end{aligned} \tag{2.15}$$

where Σ_t^{-1} is given by (2.14). Notice that (2.13) is a simple-kriging version of FRK formulas given by Cressie and Johannesson (2006, 2008). The presence of the fine-scale variance, σ_ξ^2, was discussed by Cressie and Johannesson (2008) and subsequently included in FRK formulas given by Cressie and Kang (2010).

2.3 Fixed Rank Filtering (FRF)

Spatio-temporal processes are important components of many environmental models. Prediction of the unknown state of the environment at a given location and time, along with a measure of uncertainty of that prediction, is a fundamental problem. In this section, we shall present a spatio-temporal Kalman filter based on the model (2.1), (2.2), (2.3), and (2.4), where the dimension reduction in (2.3) allows extraordinarily fast computation times (Sections 2.4 and 4.2).

Suppose that data are observed at more time points than just a single t. Then, by taking into account the temporal dynamics in the process $\{\boldsymbol{\eta}_t : t = 1, 2, \ldots\}$ and all available observations $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$, we are able to improve our prediction precision over the purely spatial FRK given in Section 2.2. Assuming the STRE model (2.3) and (2.4), the optimal predictor of $\boldsymbol{\eta}_t$, given observations up to and including t, is expressed recursively in terms of a Kalman filter (Kalman 1960; Shumway and Stoffer 2006, sec. 6.2). Assuming initial values, $\hat{\boldsymbol{\eta}}_{0|0}$ and $P_{0|0}$, we obtain

$$\begin{aligned}\hat{\boldsymbol{\eta}}_{t|t} &\equiv E(\boldsymbol{\eta}_t | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) \\ &= \hat{\boldsymbol{\eta}}_{t|t-1} + G_t\left\{\mathbf{Z}(t) - \boldsymbol{\mu}(t) - S_t \hat{\boldsymbol{\eta}}_{t|t-1}\right\}; \qquad t = 1, 2, \ldots, \end{aligned} \tag{2.16}$$

with $r \times r$ mean-squared-prediction-error matrix

$$P_{t|t} \equiv E\left[\left(\hat{\boldsymbol{\eta}}_{t|t} - \boldsymbol{\eta}_t\right)\left(\hat{\boldsymbol{\eta}}_{t|t} - \boldsymbol{\eta}_t\right)'\right] = P_{t|t-1} - G_t S_t P_{t|t-1}, \tag{2.17}$$

where the definitions of $\hat{\boldsymbol{\eta}}_{t|t-1}$ and $P_{t|t-1}$ are given below; the $r \times r$ Kalman gain matrix G_t is given by

$$\begin{aligned}G_t &= P_{t|t-1} S_t'\left(S_t P_{t|t-1} S_t' + D_t\right)^{-1} \\ &= P_{t|t-1} S_t'\left(D_t^{-1} - D_t^{-1} S_t\left\{P_{t|t-1}^{-1} + S_t' D_t^{-1} S_t\right\}^{-1} S_t' D_t^{-1}\right); \end{aligned} \tag{2.18}$$

and recall from (2.10) that $D_t \equiv \sigma_\xi^2 I_{n_t} + \sigma_\varepsilon^2 V_t$ and $V_t \equiv \text{diag}(v_t(\mathbf{s}_{t,1}), \ldots, v_t(\mathbf{s}_{t,n_t}))$. To obtain (2.18), we rely again on a Sherman–Morrison–Woodbury formula, in a like manner to (2.14). The one-step-ahead-forecast quantities are given by

$$\hat{\boldsymbol{\eta}}_{t|t-1} = H_t \hat{\boldsymbol{\eta}}_{t-1|t-1}, \tag{2.19}$$

$$P_{t|t-1} = H_t P_{t-1|t-1} H_t' + U_t, \tag{2.20}$$

where recall that the $r \times r$ matrix $U_t \equiv \text{var}(\boldsymbol{\zeta}_t)$ in (2.4).

Our goal is optimal prediction of $Y(\mathbf{s}_0; t)$ based on the data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$. The optimal filter is

$$\hat{Y}(\mathbf{s}_0; t|t) = \mu_t(\mathbf{s}_0) + \mathbf{S}_t(\mathbf{s}_0)' E(\boldsymbol{\eta}_t | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) + E(\xi(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)).$$

The second term was derived just above in (2.16). The third term was assumed by Wikle and Cressie (1999) to depend only on $\mathbf{Z}(t)$, but in fact it needs to be filtered like the second term.

Define the filter, $\hat{\xi}_{t|t}(\mathbf{s}_0) \equiv E(\xi(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$; prediction variance, $Q_{t|t}(\mathbf{s}_0) \equiv \text{var}(\xi(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$; and prediction covariance, $\mathbf{R}_{t|t}(\mathbf{s}_0) \equiv \text{cov}(\boldsymbol{\eta}_t, \xi(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$. After computing the conditional distribution of $(\mathbf{Z}(t), \boldsymbol{\eta}_t, \xi(\mathbf{s}_0; t))'$ given $\mathbf{Z}(1), \ldots, \mathbf{Z}(t-1)$, which is Gaussian, it can be shown that the conditional distribution of $(\boldsymbol{\eta}_t, \xi(\mathbf{s}_0; t))'$ given $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$ is also Gaussian with mean components $\hat{\boldsymbol{\eta}}_{t|t}$ given by (2.16) and

$$\hat{\xi}_{t|t}(\mathbf{s}_0) = \mathbf{c}_t(\mathbf{s}_0)' \big(S_t P_{t|t-1} S_t' + D_t\big)^{-1} \big(\mathbf{Z}(t) - \boldsymbol{\mu}(t) - S_t \hat{\boldsymbol{\eta}}_{t|t-1}\big). \tag{2.21}$$

The variance–covariance components are $P_{t|t}$ given by (2.17),

$$Q_{t|t}(\mathbf{s}_0) = \sigma_\xi^2 - \mathbf{c}_t(\mathbf{s}_0)' \big(S_t P_{t|t-1} S_t' + D_t\big)^{-1} \mathbf{c}_t(\mathbf{s}_0), \tag{2.22}$$

$$\mathbf{R}_{t|t}(\mathbf{s}_0) = -G_t \mathbf{c}_t(\mathbf{s}_0), \tag{2.23}$$

where recall that the gain matrix G_t is given by (2.18).

Hence, the optimal filter of $Y(\mathbf{s}_0; t)$ follows straightforwardly as

$$\hat{Y}(\mathbf{s}_0; t|t) = \mu_t(\mathbf{s}_0) + \mathbf{S}_t(\mathbf{s}_0)' \hat{\boldsymbol{\eta}}_{t|t} + \hat{\xi}_{t|t}(\mathbf{s}_0), \tag{2.24}$$

where $\hat{\boldsymbol{\eta}}_{t|t}$ is given by (2.16), and $\hat{\xi}_{t|t}(\mathbf{s}_0)$ is given by (2.21). Following the nomenclature FRK, described in Section 2.2, we call (2.24) the *FRF* (*Fixed Rank Filtering*) predictor and notate it as

$$\hat{Y}(\mathbf{s}_0; t)^{\text{FRF}} \equiv \hat{Y}(\mathbf{s}_0; t|t); \qquad \mathbf{s}_0 \in D. \tag{2.25}$$

Its root mean squared prediction error (to be compared to the FRK standard error (2.15)) is

$$\begin{aligned}
\sigma(\mathbf{s}_0; t)^{\text{FRF}} &\equiv \big\{E(Y(\mathbf{s}_0; t) - \hat{Y}(\mathbf{s}_0; t)^{\text{FRF}})^2\big\}^{1/2} \\
&= \big\{E\big(\mathbf{S}_t(\mathbf{s}_0)'\big(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{t|t}\big) + \xi(\mathbf{s}_0; t) - \hat{\xi}_{t|t}(\mathbf{s}_0)\big)^2\big\}^{1/2} \\
&= \big\{\mathbf{S}_t(\mathbf{s}_0)' P_{t|t} \mathbf{S}_t(\mathbf{s}_0) + \sigma_\xi^2 - \mathbf{c}_t(\mathbf{s}_0)' \big(S_t P_{t|t-1} S_t' + D_t\big)^{-1} \mathbf{c}_t(\mathbf{s}_0) \\
&\quad - 2\mathbf{S}_t(\mathbf{s}_0)' G_t \mathbf{c}_t(\mathbf{s}_0)\big\}^{1/2},
\end{aligned} \tag{2.26}$$

which is derived from (2.17), (2.22), and (2.23). We call (2.26) the FRF standard error.

It should be noticed that when $\mathbf{s}_0 \notin \{\mathbf{s}_{1,t}, \ldots, \mathbf{s}_{n_t,t}\}$, $\mathbf{c}_t(\mathbf{s}_0) = \mathbf{0}$. In this case, the effect of the fine-scale variability term, $\xi(\cdot; t)$, is only seen in the mean squared prediction error. That is, when $\mathbf{s}_0$ is not the location of a datum, $\hat{Y}(\mathbf{s}_0; t)^{\mathrm{FRF}} = \mu_t(\mathbf{s}_0) + \mathbf{S}_t(\mathbf{s}_0)'\hat{\boldsymbol{\eta}}_{t|t}$, and $\sigma(\mathbf{s}_0; t)^{\mathrm{FRF}} = \{\mathbf{S}_t(\mathbf{s}_0)' P_{t|t} \mathbf{S}_t(\mathbf{s}_0) + \sigma_\xi^2\}^{1/2}$.

In the rest of this subsection, we give the equations for spatio-temporal random effects smoothing and forecasting for the STRE model, although our simulation in Section 3 and the application in Section 4 are concerned only with filtering. *Smoothing* involves optimal prediction of $Y(\mathbf{s}_0; u)$, where $u \in \{1, \ldots, t-1\}$, from data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$. Assuming the STRE model (2.3) and (2.4), the optimal predictor of $\boldsymbol{\eta}_u$ given observations up to and including t ($> u$) can be expressed recursively (e.g., Shumway and Stoffer 2006). Recall that $\hat{\boldsymbol{\eta}}_{u|u}$ and $P_{u|u}$, for $u \in \{0, 1, \ldots, t\}$, are available from the Kalman filter (2.16) and (2.17). Then the optimal smoother is

$$\hat{\boldsymbol{\eta}}_{u|t} \equiv E(\boldsymbol{\eta}_u | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) = \hat{\boldsymbol{\eta}}_{u|u} + J_u \{\hat{\boldsymbol{\eta}}_{u+1|t} - \hat{\boldsymbol{\eta}}_{u+1|u}\}, \tag{2.27}$$

with $r \times r$ mean-squared-prediction-error matrix

$$\begin{aligned} P_{u|t} &\equiv E[(\hat{\boldsymbol{\eta}}_{u|t} - \boldsymbol{\eta}_u)(\hat{\boldsymbol{\eta}}_{u|t} - \boldsymbol{\eta}_u)'] \\ &= P_{u|u} + J_u (P_{u+1|t} - P_{u+1|u}) J_u'; \qquad u = 1, \ldots, t-1. \end{aligned} \tag{2.28}$$

In (2.27) and (2.28), we use (2.19) and (2.20), the one-step-ahead forecast equations that define $\hat{\boldsymbol{\eta}}_{u+1|u}$ and $P_{u+1|u}$, respectively, and

$$J_u \equiv P_{u|u} H_{u+1}' P_{u+1|u}^{-1}, \tag{2.29}$$

where recall that H_{u+1} is the $r \times r$ propagator matrix defined by (2.4).

The optimal smoother of $Y(\mathbf{s}_0; u)$; $u \in \{1, \ldots, t-1\}$, is

$$\hat{Y}(\mathbf{s}_0; u|t) = \mu_u(\mathbf{s}_0) + \mathbf{S}_u(\mathbf{s}_0)' E(\boldsymbol{\eta}_u | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) + E(\xi(\mathbf{s}_0; u) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)).$$

The second term was derived just above in (2.27); for the third term, we obtain

$$\begin{aligned} E(\xi(\mathbf{s}_0; u) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) &\equiv \hat{\xi}_{u|t}(\mathbf{s}_0) = \hat{\xi}_{u|u}(\mathbf{s}_0) + \mathbf{M}_u(\mathbf{s}_0)'(\hat{\boldsymbol{\eta}}_{u+1|t} - \hat{\boldsymbol{\eta}}_{u+1|u}), \\ \mathrm{var}(\xi(\mathbf{s}_0; u) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) &\equiv Q_{u|t}(\mathbf{s}_0) \\ &= Q_{u|u}(\mathbf{s}) + \mathbf{M}_u(\mathbf{s}_0)'(P_{u+1|t} - P_{u+1|u})\mathbf{M}_u(\mathbf{s}_0), \\ \mathrm{cov}(\boldsymbol{\eta}_u, \xi(\mathbf{s}_0; u) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) &\equiv \mathbf{R}_{u|t}(\mathbf{s}_0) = \mathbf{R}_{u|u}(\mathbf{s}_0) + J_u (P_{u+1|t} - P_{u+1|u}) \mathbf{M}_u(\mathbf{s}_0), \end{aligned}$$

where $\mathbf{M}_u(\mathbf{s}_0) \equiv P_{u+1|u}^{-1} H_{u+1} \mathbf{R}_{u|u}(\mathbf{s}_0)$, and $\mathbf{R}_{u|u}(\mathbf{s}_0)$ is given by (2.23).

Hence, the optimal smoother of $Y(\mathbf{s}_0; t)$ is

$$\hat{Y}(\mathbf{s}_0; u|t) = \mu_u(\mathbf{s}_0) + \mathbf{S}_u(\mathbf{s}_0)'\hat{\boldsymbol{\eta}}_{u|t} + \hat{\xi}_{u|t}(\mathbf{s}_0); \qquad u = 1, \ldots, t-1. \tag{2.30}$$

Using a derivation like that of (2.26), the root mean squared prediction error of $\hat{Y}(\mathbf{s}_0; u|t)$ is

$$\sigma(\mathbf{s}_0; u|t) = \{E(\mathbf{S}_u(\mathbf{s}_0)'(\boldsymbol{\eta}_u - \hat{\boldsymbol{\eta}}_{u|t}) + \xi(\mathbf{s}_0; u) - \hat{\xi}_{u|t}(\mathbf{s}_0))^2\}^{1/2}$$

$$= \left\{\mathbf{S}_u(\mathbf{s}_0)' P_{u|t} \mathbf{S}_u(\mathbf{s}_0) + Q_{u|t}(\mathbf{s}_0) \right.$$
$$\left. + 2\mathbf{S}_u(\mathbf{s}_0)' \mathbf{R}_{u|t}(\mathbf{s}_0)\right\}^{1/2}; \qquad u = 1, \ldots, t-1. \tag{2.31}$$

Once again, notice that when $\mathbf{s}_0 \notin \{\mathbf{s}_{1,u}, \ldots, \mathbf{s}_{n_u,u}\}$, $\mathbf{c}_u(\mathbf{s}_0) = \mathbf{0}$. In this case, $\hat{Y}(\mathbf{s}_0; u|t) = \mu_u(\mathbf{s}_0) + \mathbf{S}_u(\mathbf{s}_0)' \hat{\boldsymbol{\eta}}_{u|t}$, and $\sigma(\mathbf{s}_0; u|t) = \{\mathbf{S}_u(\mathbf{s}_0)' P_{u|t} \mathbf{S}_u(\mathbf{s}_0) + \sigma_\xi^2\}^{1/2}$.

Forecasting involves optimal prediction of $Y(\mathbf{s}_0; u)$, where $u \in \{t+1, t+2, \ldots\}$, from data $\mathbf{Z}(1), \ldots, \mathbf{Z}(t)$. Assuming the STRE model (2.3) and (2.4), the optimal predictor of $\boldsymbol{\eta}_u$ is (e.g., Shumway and Stoffer 2006), for $u = t+1, t+2, \ldots,$

$$\hat{\boldsymbol{\eta}}_{u|t} \equiv E(\boldsymbol{\eta}_u | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) = \left(\prod_{i=t+1}^{u} H_i\right) \hat{\boldsymbol{\eta}}_{t|t}, \tag{2.32}$$

with $r \times r$ mean-squared-prediction-error matrix

$$P_{u|t} = \left(\prod_{i=t+1}^{u} H_i\right) P_{t|t} \left(\prod_{i=t+1}^{u} H_i\right)' + U_u$$
$$+ \sum_{i=t+1}^{u-1} \left\{\left(\prod_{j=i+1}^{u} H_j\right) U_i \left(\prod_{j=i+1}^{u} H_j\right)'\right\}, \tag{2.33}$$

where it is understood that the last term in (2.33) is zero if $u = t+1$.

The optimal forecast of $Y(\mathbf{s}_0; u)$; $u > t$, is

$$\hat{Y}(\mathbf{s}_0; u|t) = \mu_u(\mathbf{s}_0) + \mathbf{S}_u(\mathbf{s}_0)' E(\boldsymbol{\eta}_u | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)) + E(\xi(\mathbf{s}_0; u) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t)).$$

The second term was derived just above in (2.32); the third term is trivially zero, since $\xi(\cdot; u)$ is independent of $Z(\cdot; 1), \ldots, Z(\cdot; t)$ when $u > t$.

Similar considerations to those for filtering and smoothing lead to $\hat{\xi}_{u|t}(\mathbf{s}_0) \equiv 0$, $Q_{u|t}(\mathbf{s}_0) \equiv \sigma_\xi^2$, and $\mathbf{R}_{u|t}(\mathbf{s}_0) \equiv \mathbf{0}$. Hence, the optimal forecast of $Y(\mathbf{s}_0; u)$ follows straightforwardly as

$$\hat{Y}(\mathbf{s}_0; u|t) = \mu_u(\mathbf{s}_0) + \mathbf{S}_u(\mathbf{s}_0)' \hat{\boldsymbol{\eta}}_{u|t}; \qquad u = t+1, t+2, \ldots, \tag{2.34}$$

and its root mean squared prediction error is

$$\sigma(\mathbf{s}_0; u|t) = \left\{\mathbf{S}_u(\mathbf{s}_0)' P_{u|t} \mathbf{S}_u(\mathbf{s}_0) + \sigma_\xi^2\right\}^{1/2}; \qquad u = t+1, t+2, \ldots. \tag{2.35}$$

2.4 COMPUTATIONAL COMPLEXITY OF FRF

To illustrate the computational advantage of FRF based on the STRE model (2.3) and (2.4), we compare the computational complexity of FRF to that of an evaluation of $E(Y(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$ based on a generic joint Gaussian distribution of $(Y(\mathbf{s}_0; t), \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$. The computational complexity is calculated in terms of the number of observed time units t, the number of observed spatial locations n_u at time u; $u = 1, \ldots, t$, and the dimension r of the hidden VAR process $\{\boldsymbol{\eta}_t\}$. Define $n_{+t} \equiv \sum_{u=1}^{t} n_u$ and $n_{\cdot t} \equiv n_{+t}/t$.

Given observed data $\{\mathbf{Z}(1), \ldots, \mathbf{Z}(t)\}$, the calculation of $E(Y(\mathbf{s}_0; t) | \mathbf{Z}(1), \ldots, \mathbf{Z}(t))$ requires inverting the covariance matrix, $\text{cov}((\mathbf{Z}(1)', \ldots, \mathbf{Z}(t)')')$, of size $n_{+t} \times n_{+t}$. Thus, the computational complexity of an evaluation of $E(Y(\mathbf{s}_0;t)|\mathbf{Z}(1),\ldots,\mathbf{Z}(t))$ based on a generic joint Gaussian distribution is $O(n_{\cdot t}^3 t^3)$, which results in a computation that will fail if $n_{\cdot t}$ or t is too large. One way to reduce the computational complexity is to implement vector-based Kalman filtering in the time dimension. This results in a reduction from $O(n_{\cdot t}^3 t^3)$ to $O(n_{\cdot t}^3 t)$, but this computation will also fail if $n_{\cdot t}$ is too large.

Under the STRE model (2.3) and (2.4), FRF computes the prediction $\hat{Y}(\mathbf{s}_0;t)^{\mathrm{FRF}}$, using a Sherman–Morrison–Woodbury formula for the matrix inverses that reduces the computational complexity from $O(n_{\cdot t}^3 t^3)$ to $O(n_{\cdot t} t)$, as we now demonstrate. From Section 2.3, the computation of $\hat{Y}(\mathbf{s}_0;t)^{\mathrm{FRF}}$; $t > 1$, is carried out sequentially. From the predictor $\hat{\boldsymbol{\eta}}_{t-1|t-1}$ and its mean squared prediction error $P_{t-1|t-1}$, at time $t-1$, we first obtain the one-step-ahead predictor $\hat{\boldsymbol{\eta}}_{t|t-1}$ and its corresponding mean-squared-prediction-error matrix $P_{t|t-1}$ from (2.19) and (2.20), respectively; the computational costs are $O(r^3)$ for both steps. Due to a Sherman–Morrison–Woodbury formula like (2.14), the computational complexity of the Kalman gain G_t in (2.18) is $O(r^3 n_t)$, reduced from $O(r^2 n_t^3)$ if one were to carry out a direct evaluation of $P_{t|t-1}S_t'(S_t P_{t|t-1}S_t' + D_t)^{-1}$. Hence, the computational complexity of $P_{t|t}$ in (2.17) and $\hat{\boldsymbol{\eta}}_{t|t}$ in (2.16) is $O(r^3 n_t)$. A similar $O(r^3 n_t)$ complexity calculation can be obtained for $\hat{\xi}_{t|t}(\mathbf{s}_0)$, $Q_{t|t}(\mathbf{s}_0)$, and $\mathbf{R}_{t|t}(\mathbf{s}_0)$. Finally, the computational complexities of $\hat{Y}(\mathbf{s}_0;t)^{\mathrm{FRF}}$ and $\sigma(\mathbf{s}_0;t)^{\mathrm{FRF}}$ are both $O(r^3 n_{+t}) = O(r^3 n_{\cdot t}t) = O(n_{\cdot t}t)$, since r is fixed.

3. THE ROLE OF TEMPORAL DEPENDENCE: A SIMULATION STUDY

In this section, we quantify the possible gains to be made when temporal dependence is incorporated into an optimal statistical-prediction procedure that predicts the current spatial state given all the spatio-temporal data. When we only use the current spatial data in an optimal spatial-statistical predictor (i.e., kriging), the prediction algorithm is much simpler (although the computational complexity is comparable). This section addresses the question of when it is worth the trouble of building a STRE model and implementing FRF (Section 2.3), in comparison to implementing purely spatial FRK based on an SRE model (Section 2.2). We answer the question through a simulation study and show that, when the temporal dependence is strong, we can obtain statistical efficiency gains of FRF over FRK of up to 450%. Theoretically, FRF will never do worse than FRK, but the simulation study shows that FRF's gains can be substantial.

We now give a summary of the simulation study; for more details, see the Appendix in the Supplemental Materials section. The spatial domain considered in this study is one-dimensional, namely, $D = \{s : s = 1,\ldots,256\}$, and the time dimension is discretized as $t = 1,2,\ldots,50$. We simulate the processes $\{Y(s;t)\}$ and $\{Z(s;t)\}$ according to (2.2), (2.4), and (2.5), where we assume that: $\mu_t(s) = 0$ for $s \in D$; $t = 1,2,\ldots,50$; and the spatial basis functions $\mathbf{S}_t(\cdot) \equiv \mathbf{S}(\cdot)$ do not depend on t and are made up of 30 W-wavelets (e.g., Kwong and Tang 1994; Nychka, Wikle, and Royle 2002; Shi and Cressie 2007) from the first four resolutions. The STRE process $\{\boldsymbol{\eta}_t\}$ used to generate the realizations of the simulation is assumed to be stationary with $K_t \equiv K$, $H_{t+1} \equiv H$, and $U_{t+1} \equiv U$ not dependent on t, and hence $K = HKH' + U$. The covariance matrix K is chosen such that $\|SKS' - \Sigma^0\|^2$ is minimized, where $(\Sigma^0)_{ij} \equiv \exp\{-|i-j|/\theta\}$ with $\theta = 25$ and $\|\cdot\|$ denotes the Frobenius norm (e.g., Hastie 1996; Donoho, Mallet, and von Sachs 1998; Cressie and Johannesson 2008):

$$\|A-B\|^2 \equiv \mathrm{tr}((A-B)'(A-B)) = \sum_{j,k}(A_{j,k}-B_{j,k})^2.$$

The role of the matrix Σ^0 is to calibrate the spatial dependence in the STRE model; here it is quite strong with an equivalent range of $3\theta = 75$. The other parameters $\{H, U, \sigma_\xi^2, \sigma_\varepsilon^2\}$ are chosen in ways that relate to factors in the experiment and will be described in Section 3.1.

3.1 FACTORS OF THE SIMULATION EXPERIMENT

We compare FRF and FRK under a variety of conditions. We denote these prediction methods as the factor PM. At time t, we assume observations are *missing* over a specified region $D^M \subset D$, and they are present at the remaining locations, namely $D^O \equiv D \setminus D^M$. While FRK makes spatial predictions of Y at time t given only the incomplete current data, FRF combines the incomplete current data *and* the (possibly incomplete) past data. In our experiment, FRF and FRK are compared over the *missing region* D^M and over the *observed region* D^O, separately. As well as PM, six other factors are included in the experiment. The details of *all* factors are now presented.

Prediction methods (PM). Two prediction methods, FRF and FRK, are considered. The formulas for both methods are presented in Section 2; see (2.13), (2.15), (2.24), and (2.26).

Temporal dependence (TD). The temporal dependence is specified through the propagator matrix H. We first define a target matrix T^0 through

$$(T^0)_{i,j} \equiv \rho \exp\{-|i-j|/\theta\}; \qquad i, j = 1, \dots, 256,$$

where recall that $\theta = 25$, and we call ρ the temporal-dependence parameter. This specification of T^0 is motivated by the multivariate structure discussed by Ver Hoef and Cressie (1993), where there ρ indicated the strength of cross-dependence. Then H is obtained by minimizing the Frobenius norm between $SHKS'$ and $\mathrm{diag}(SKS')^{1/2}\, T^0\, \mathrm{diag}(SKS')^{1/2}$, and U is obtained from the relationship between K, H, and U: $U = K - HKH'$, where we always check that U is positive-definite. We use ρ to define the levels of temporal dependence. Four levels of TD are considered: $\rho = 0.975$, 0.7, 0.5, and 0.1.

Fine-scale variation (FV). The proportion of fine-scale variation to the total variability of $Y(\cdot; t)$ is defined as

$$\mathrm{FVP} \equiv \frac{n\sigma_\xi^2}{\mathrm{tr}(SKS') + n\sigma_\xi^2},$$

where recall that $n = 256$, $r = 30$, and S is the $n \times r$ matrix where (i, j) element is $S_j(\mathbf{s}_i)$. Two levels of FV are considered: 0 and 0.05.

Signal-to-noise ratio (SN). We define the signal-to-noise ratio as

$$\mathrm{SNR} \equiv \frac{\mathrm{tr}(SKS') + n\sigma_\xi^2}{n\sigma_\varepsilon^2}. \tag{3.1}$$

Two levels of SN are considered: SNR = 10 and 1; σ_ε^2 then depends on the SNR chosen.

Prediction time point (PT). Predictions are made at times $t = 1$ through $t = 50$, but we do not look at results for all these times. We consider prediction time point, PT, as a possible factor at levels $t = 10$, 15, 25, and 45.

Missing-data width (MW) *and missing-data location* (ML). We use these two factors to define the missing region D^M. Missing-data width specifies the number of contiguous locations with missing data (a swath, in the terminology of remote sensing), notated by w. Missing-data location specifies where the swath of missing data is located in the space, either b or c:

- $b \equiv$ missing in the beginning: $D^M = \{1, \dots, w\}$;
- $c \equiv$ missing in the center: $D^M = \{120 - \frac{w-1}{2}, \dots, 120 + \frac{w-1}{2}\}$,

for $w = 25, 51$, and 103. That is, three levels of MW are considered: $\omega = 25, 51$, and 103; and two levels of ML are considered: b and c.

3.2 Results of the Simulation Experiment

Figure 1 illustrates what *one* replicate of the simulation looks like for a certain combination of the factors, along with the corresponding behavior of FRF and FRK. The upper panel shows the observations $\mathbf{Z}(t)$ at $t = 9$ and $t = 10$, simulated from $\rho = 0.975$, FVP $= 0.05$, and SNR $= 10$, with the superimposed intervals illustrating the missing locations D^M. The lower panel shows $\mathbf{Y}(10)$, $\hat{\mathbf{Y}}(10)^{\mathrm{FRF}}$, and $\hat{\mathbf{Y}}(10)^{\mathrm{FRK}}$, in both the missing region D^M and in the observed region D^O. FRF is clearly superior to FRK in D^M and much less so in D^O.

We define a response variable based on the empirical mean squared prediction error (MSPE) in each simulation run: let D^* denote either D^O or D^M, and let $\hat{Y}_l(\cdot; t)$ denote a generic predictor of $Y_l(\cdot; t)$, for the lth realization of the simulation. Then define

$$\mathrm{MSPE}(D^*)_l \equiv \frac{1}{|D^*|} \sum_{s \in D^*} (\hat{Y}_l(s; t) - Y_l(s; t))^2; \qquad l = 1, \dots, L, \tag{3.2}$$

where $|D^*|$ is the number of spatial locations in D^* and $L = 625$ is the total number of simulation runs in this study; this choice of L is discussed in the Appendix.

We use an analysis of variance (ANOVA) to study which factors are important and under which scenarios FRF provides substantial improvement over FRK. In the Appendix, we justify a fourth-root transformation of the MSPE:

$$A(D^*) \equiv \mathrm{ave}\big\{\mathrm{MSPE}(D^*)_l{}^{1/4}\big\}; \qquad D^* = D^O \text{ or } D^M, \tag{3.3}$$

where the average is taken over the $L = 625$ simulation runs.

The response upon which the ANOVA is based is

$$A^{\mathrm{FRK}}(D^M) - A^{\mathrm{FRF}}(D^M),$$

and the ANOVA table showing up to two-way interactions can be found in the Appendix.

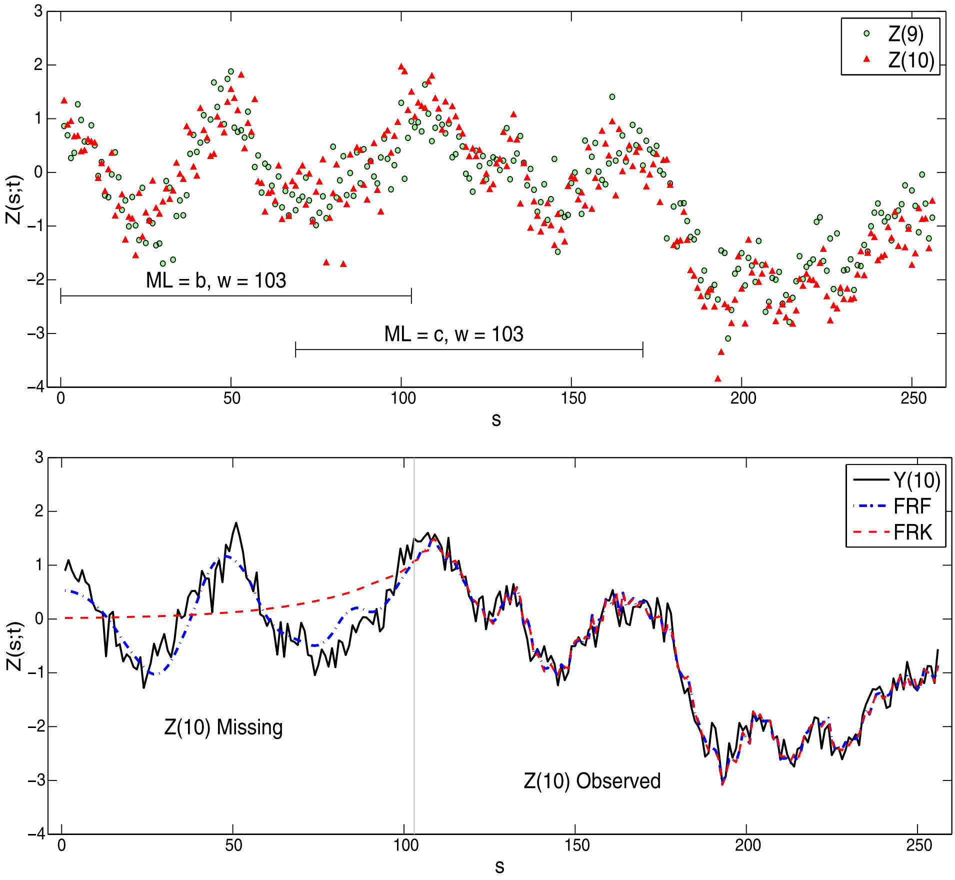

Figure 1. One example of simulated data and predictions. *Upper panel*: $\mathbf{Z}(9)$ (circles), $\mathbf{Z}(10)$ (triangles); some of these data will be declared missing, according to the intervals shown in the upper panel. Data are simulated with $\rho = 0.975$, SNR $= 10$, and FVP $= 0.05$. *Lower panel*: Conditional on level b of ML and level $w = 103$ of MW, the plot shows $\mathbf{Y}(10)$ (solid line), $\hat{\mathbf{Y}}(10)^{\text{FRF}}$ (dot-dashed line), and $\hat{\mathbf{Y}}(10)^{\text{FRK}}$ (dashed line). The online version of this figure is in color.

Figure 2 highlights the effects of TD, MW, and their interaction. Observe that $A^{\text{FRK}}(D^M) - A^{\text{FRF}}(D^M)$ is generally above zero and decreases as the temporal dependence becomes weaker. For each ρ, the response, $A^{\text{FRK}}(D^M) - A^{\text{FRF}}(D^M)$, increases as the missing-data width w increases, although for $\rho = 0.1$, the differences are very small.

We define the relative efficiency of FRF to FRK at a given location s, by first averaging over replications of the simulation, as well as the prediction times (see the Appendix), and then taking the ratio of their respective empirical mean squared prediction errors:

$$E(s) \equiv \frac{\text{ave}\{(\hat{Y}_l(s;t)^{\text{FRK}} - Y_l(s;t))^2\}}{\text{ave}\{(\hat{Y}_l(s;t)^{\text{FRF}} - Y_l(s;t))^2\}} \times 100\%; \qquad s \in D.$$

The locations s where $E(s) \geq 100\%$ represent regions of the space where FRF is preferred to FRK.

In Figure 3, we show the relative efficiencies, $\{E(\mathbf{s}) : \mathbf{s} \in D\}$, for different factor combinations. From the upper panels, we see that FRF always outperforms FRK when predicting

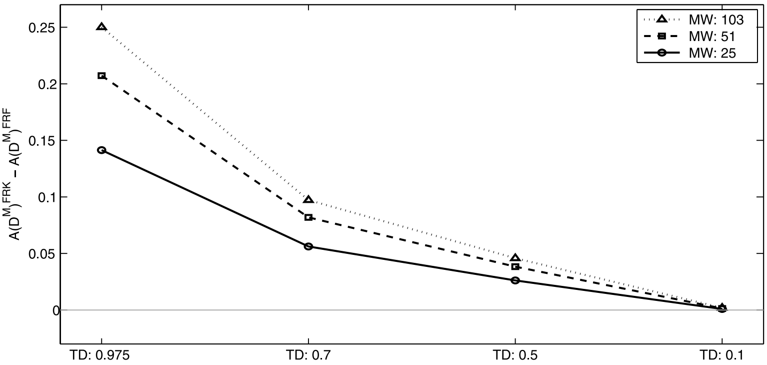

Figure 2. Plots showing interaction between TD and MW in the ANOVA of the response, $A^{\mathrm{FRK}}(D^M) - A^{\mathrm{FRF}}(D^M)$.

locations in D^M, sometimes remarkably so. Relative efficiencies of around 450% are possible when $\rho = 0.975$, which drop to around 125% when $\rho = 0.5$. From the lower panels, we see that the wider the missing region, the higher the relative efficiency of FRF to FRK in D^M. As the prediction location s gets closer to D^O, the efficiency drops sharply and once in D^O maintains a level marginally above 100%.

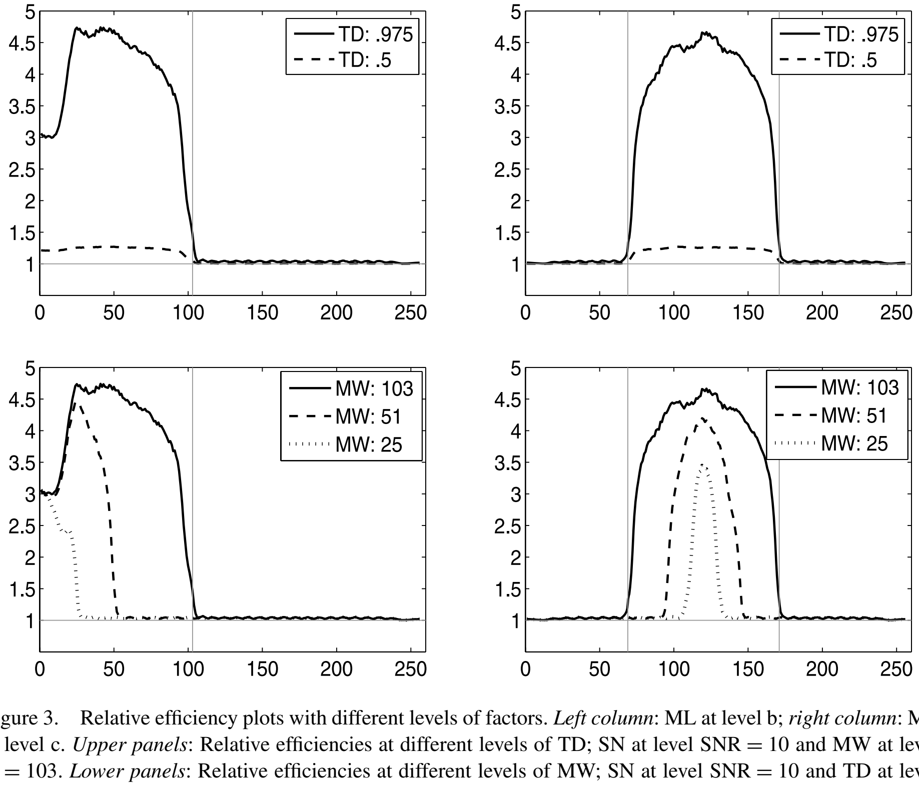

Figure 3. Relative efficiency plots with different levels of factors. *Left column*: ML at level b; *right column*: ML at level c. *Upper panels*: Relative efficiencies at different levels of TD; SN at level SNR = 10 and MW at level $w = 103$. *Lower panels*: Relative efficiencies at different levels of MW; SN at level SNR = 10 and TD at level $\rho = 0.975$.

4. SPATIO-TEMPORAL FILTERING OF AEROSOL OPTICAL DEPTH

We now implement FRF to process a very large spatio-temporal dataset collected by MISR, one of NASA's key instruments measuring and monitoring global aerosol distributions (Diner et al. 1999; Kaufman et al. 2000). MISR's cameras cover a swath at the Earth's surface that is approximately 360 km wide and extends across the daylight side of the Earth from the Arctic down to Antarctica. There are 233 geographically distinct swaths (also called paths) that are visited on a repeat cycle of 16 days; that is, MISR collects data on the exact same path every 16 days. Since the satellite is polar-orbiting, the paths can overlap, particularly at higher latitudes. The spatial resolution of MISR level-2 aerosol data is 17.6 × 17.6 km; these level-2 data are then converted to level-3 data at a much lower spatial (0.5° × 0.5°) and temporal (1-day) resolution by averaging the observations that fall in these lower-resolution pixels on a given day. The streaming nature of MISR data makes the spatio-temporal dataset quickly massive; in this section, we take advantage of the temporal dependence in the level-3 AOD data by implementing the FRF methodology given in Section 2.

We use MISR level-3 data collected between July 1 and August 9, 2001. As illustrated in the article by Shi and Cressie (2007), in which MISR level-3 AOD were processed using FRK, the distribution of AOD is heavily right-skewed. Hence, we take log(AOD) as the observation. The time unit we have chosen is eight days, and we obtain an individual datum by taking a weighted average of daily level-3 log(AOD) values in a given 8-day period, where the weight is defined by the number of level-2 observations $N_d(\mathbf{s})$ in each level-3 pixel $\mathbf{s}$ on day d. Time unit 1 corresponds to July 1–8, time unit 2 corresponds to July 9–16, ..., and time unit 5 corresponds to August 2–9, 2001.

We apply both FRF and FRK to data collected in a rectangular region D between longitudes $-125°$ and $+3°$ and between latitudes $-20°$ and $+44°$, which is shown in the upper-left panel of Figure 4. The study region covers North and South America, the western part of the Sahara Desert in Africa, the Iberian Peninsula in Europe, and parts of the Atlantic and Pacific Oceans. We use this region because dust from the Sahara Desert is typically transported across the Atlantic Ocean to North America. The number of level-3 pixels in the region is $128 \times 256 = 32{,}768$, and the number of data for each time unit is on the order of $20{,}000$. Our example demonstrates that prediction based on the STRE model can process very large spatio-temporal datasets, and the resulting FRF is beneficial when current data have large parts where there are no data.

4.1 STME MODEL SPECIFICATION

In our analysis of the spatio-temporal MISR data, log(AOD), we fit the STME model given by (2.5). Based on our exploratory data analysis, the trend term $\mu_t(\mathbf{s})$ is modeled as $\mathbf{X}(\mathbf{s})'\boldsymbol{\beta}$, where $\mathbf{X}(\mathbf{s}) = (1, 1_O(\mathbf{s}), 1_A(\mathbf{s}))'$; here, $1_O(\cdot)$ denotes the indicator function for the oceans and $1_A(\cdot)$ denotes the same but for the Americas. Hence, the trend does not depend on t. The coefficient $\boldsymbol{\beta}$ is fitted by ordinary least squares (OLS) based on all the

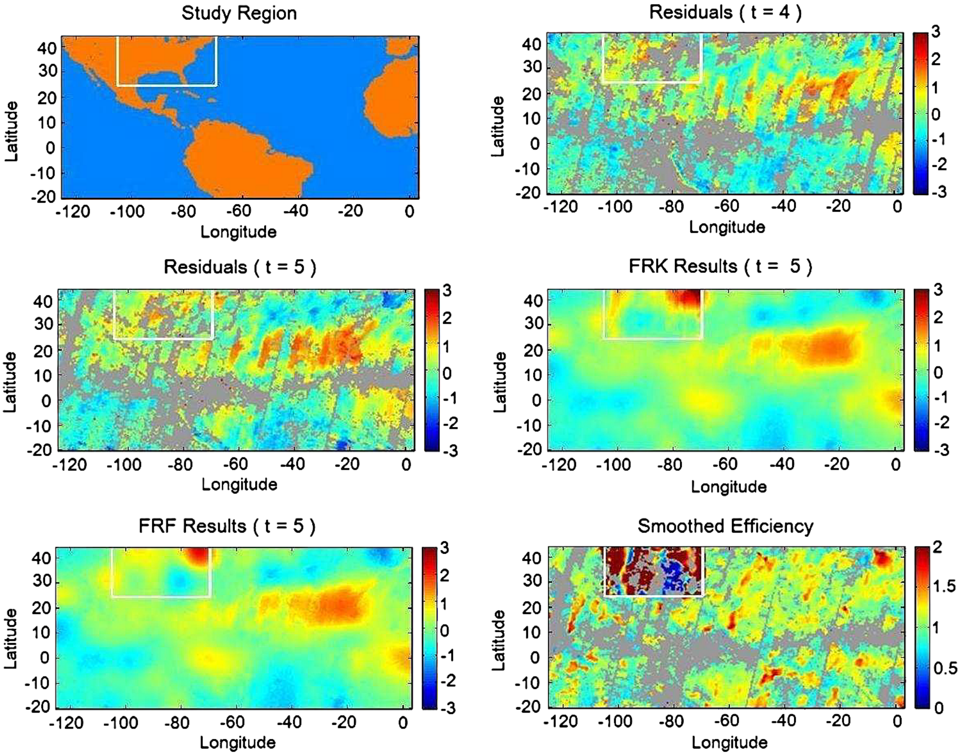

Figure 4. Aerosol optical depth (AOD). *Upper-left panel*: Study region with validation region for $t = 5$ highlighted in the white box. *Upper-right* and *middle-left panels*: Detail residuals of log(AOD) for $t = 4$ and $t = 5$, respectively, with validation region for $t = 5$ highlighted in the white box. Pixels with no data are colored gray. *Middle-right panel*: FRK-prediction map. *Lower-left panel*: FRF-prediction map. *Lower-right panel*: Smoothed empirical-relative-efficiency map of FRF relative to FRK. On the color scale, dark red/brown denotes 200% relative efficiency and green denotes 100% relative efficiency. Pixels with no data and hence no smoothed empirical relative efficiency are colored gray.

observed data $\{\mathbf{Z}(t) : t = 1, \ldots, 5\}$. The resulting estimator, $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, is used to obtain the detail residuals:

$$R(\mathbf{s}_i; t) \equiv Z(\mathbf{s}_i; t) - \mathbf{X}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}, \tag{4.1}$$

from which the parameters related to the random processes are estimated. The detail-residual maps at times $t = 4$ and $t = 5$ are shown in the upper right and the middle left panels, respectively, of Figure 4. In this section, we concentrate on comparing the performance of FRF and FRK in predicting the detrended process.

To model the process $\nu(\mathbf{s}; t)$ given by (2.3), we chose $\mathbf{S}(\cdot)$ from multiresolution W-wavelet basis functions using the strategy given by Shi and Cressie (2007). This resulted in $r = 32 + 62 = 94$ basis functions with all 32 W-wavelets from the first scale and the 62 W-wavelets with "large" absolute coefficients from the second scale. The fine-scale variation term, $\xi(\cdot; \cdot)$, is modeled as Gaussian white noise with mean zero and variance σ_ξ^2.

The error term $\varepsilon(\mathbf{s}; t)$ in (2.5) is modeled as independent $\mathrm{Gau}(0, \sigma_\varepsilon^2 v_t(\mathbf{s}))$. Recall that the datasets were generated by averaging level-2 MISR data, so it is reasonable to assume in (2.1) that $v_t(\mathbf{s}) = 1/\widetilde{N}_t(\mathbf{s})$, where $\widetilde{N}_t(\mathbf{s})$ is the number of level-2 observations in the average. Based on our exploratory data analysis, σ_ε^2 does not depend on t.

The unknown parameters are $\{\sigma_\xi^2, \sigma_\varepsilon^2\}$, K_1, and $\{H_{t+1}, U_{t+1} : t = 1, \ldots, 4\}$, which we estimate. This article concentrates principally on the FRF methodology presented in Section 2; for details of the parameter-estimation procedure and parameter estimates for this dataset, we refer readers to the work of Kang, Cressie, and Shi (2010).

4.2 COMPARISON OF FRF AND FRK

We conducted a validation experiment by leaving out all those data at $t = 5$ in a rectangular region that covers the United States lying east of Denver, the northeastern corner of Mexico, and the southern part of Canada in the Great Lakes area (between longitudes −105° and −69.5° and between latitudes +24.5° and +44°). We call this validation region D^V, delineated by the white box in the upper-left panel of Figure 4, and we call the complement $D^C \equiv D \setminus D^V$. Let D^O be the set of all observation locations at $t = 5$. Thus, $D^{VO} \equiv D^V \cap D^O$ consists of all pixels in the validation region D^V, where there are data available for validation at $t = 5$; we emphasize that those data are not used for prediction. Similarly, $D^{CO} \equiv D^C \cap D^O$ consists of all pixels in the complementary region D^C, where there are data available for prediction at $t = 5$.

After substituting the estimates $\{\hat{\sigma}_\xi^2, \hat{\sigma}_\varepsilon^2\}$, $\hat{K}_1$, and $\{\hat{H}_{t+1}, \hat{U}_{t+1} : t = 1, \ldots, 4\}$ into the FRF and FRK equations (2.13), (2.15), (2.24), and (2.26) with $\mu_t(\cdot) = 0$, we obtained $\hat{\nu}(\mathbf{s}; t)^{\mathrm{FRK}}$ and $\hat{\nu}(\mathbf{s}; t)^{\mathrm{FRF}}$, for $t = 2, \ldots, 5$. (Details on the estimates are given in the work of Kang, Cressie, and Shi (2010).) The middle-right and lower-left panels of Figure 4 show the FRK- and FRF-prediction maps, respectively, at time $t = 5$. It is clear that $\hat{\nu}(\mathbf{s}; t)^{\mathrm{FRF}}$ and $\hat{\nu}(\mathbf{s}; t)^{\mathrm{FRK}}$ do not differ much at locations in D^C. However, the two methods predict quite differently in D^V, where FRF takes advantage of the extra information provided by $\mathbf{Z}(4)$ at the previous time.

To compare FRF and FRK, we evaluate the prediction error, $\hat{\nu}(\mathbf{s}; 5) - R(\mathbf{s}; 5)$, for each $\mathbf{s}_{i,5} \in D$, and define the average efficiency as

$$\bar{E}(D^*) = \frac{\sum_{\mathbf{s} \in D^*} (\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRK}} - R(\mathbf{s}; 5))^2}{\sum_{\mathbf{s} \in D^*} (\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRF}} - R(\mathbf{s}; 5))^2} \times 100\%; \qquad D^* = D^{VO}, D^{CO}.$$

We calculated $\bar{E}(D^{VO}) = 177\%$ and $\bar{E}(D^{CO}) = 108\%$, demonstrating that FRF outperforms FRK substantially in regions with a lot of missing data, which is consistent with the results of the simulation experiment in Section 3.

In addition to reporting the average efficiencies, it is informative to illustrate the spatial relationship between the efficiency and the missing-data locations. We define the efficiency at each $\mathbf{s} \in D^O$ at time 5 as

$$\widetilde{E}(\mathbf{s}) \equiv \frac{\mathrm{SPE}(\mathbf{s})^{\mathrm{FRK}}}{\mathrm{SPE}(\mathbf{s})^{\mathrm{FRF}}}; \qquad \mathbf{s} \in D^O = D^{VO} \cup D^{CO},$$

where the *smoothed prediction errors*, $\mathrm{SPE}(\mathbf{s})^{\mathrm{FRF}}$ and $\mathrm{SPE}(\mathbf{s})^{\mathrm{FRK}}$, are computed by spatially smoothing the observed $\{(\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRF}} - R(\mathbf{s}; 5))^2 : \mathbf{s} \in D^O\}$ and $\{(\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRK}} - R(\mathbf{s}; 5))^2 : \mathbf{s} \in D^O\}$, respectively, using a Gaussian kernel with bandwidth 0.5°. The map of $\{\widetilde{E}(\mathbf{s}) : \mathbf{s} \in D^O\}$ is shown on the lower-right panel of Figure 4. It is clear that $\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRF}}$ and $\hat{\nu}(\mathbf{s}; 5)^{\mathrm{FRK}}$ do not differ much at locations in D^{CO}, but predict quite different values in D^{VO}. The efficiency map indicates that the prediction error of FRF is typically much lower in D^{VO} (since it exploits the temporal dependence between $R(\cdot; 4)$ and $R(\cdot; 5)$). We do see parts of D^O where the reverse is true, just as in Figure 1 where FRK sometimes predicts the true process better than FRF. Overall, though, FRF is superior for predicting log(AOD) in the validation region. Recall that $\overline{E}(D^{VO}) = 177\%$.

To evaluate the strength of the temporal dependence in the data, we compare the diagonal elements of the fitted time-lag-1 covariance matrix to those of the fitted spatial covariance matrix. Define

$$\hat{\rho}(\mathbf{s}) \equiv \frac{\mathbf{S}(\mathbf{s})' \hat{L}_5 \mathbf{S}(\mathbf{s})}{\sqrt{\mathbf{S}(\mathbf{s})' \hat{K}_5 \mathbf{S}(\mathbf{s})} \sqrt{\mathbf{S}(\mathbf{s})' \hat{K}_4 \mathbf{S}(\mathbf{s})}}; \qquad \mathbf{s} \in D.$$

This corresponds to the temporal-dependence level ρ given in Section 3. For the MISR data, the median of $\{\hat{\rho}(\mathbf{s}) : \mathbf{s} \in D\}$ is 0.77, which is relatively strong. Hence, the MISR data taken in 8-day time units have relatively strong lag-1 correlations between $t = 4$ and $t = 5$ and are highly suitable for FRF.

Along with the comparison of precisions of FRF and FRK, we also compared their computation times. All computations were carried out in Matlab on a Windows laptop with a dual core 2.0 GHz processor and 3 GB memory. The computations related to FRK took 13.3 sec to fit the parameter model and 21.8 sec to compute the predictions and the associated standard errors for the 32,768 pixels in the study region for $t = 5$. The recursion in the Kalman filter was started with $\hat{\boldsymbol{\eta}}_{1|1} \equiv E(\boldsymbol{\eta}_1 | \mathbf{Z}(1))$ and $P_{1|1} \equiv \mathrm{var}(\boldsymbol{\eta}_1 | \mathbf{Z}(1))$, where K_1 was estimated from $\mathbf{Z}(1)$ and substituted into these expressions. By comparison to FRK, the computations related to FRF took 64.4 sec to fit the model parameters (including binning) and 77.3 sec to compute the predictions and standard errors up to and including time $t = 5$. In practice, it is not necessary to restart the filtering procedure at each time point. The information needed from the past is summarized in the form of an $r \times r$ matrix $P_{t|t}$ and an r-dimensional vector $\boldsymbol{\eta}_{t|t}$. When the new data arrive in at time $t + 1$, we fit the model parameters for time $t + 1$ and then compute the predicted values and the prediction standard errors by filtering. The *incremental time* for FRF to compute the predictions at time $t = 5$, given the filtering results at time $t = 4$, was only 19.7 sec, which was about the same as the computing time for FRK. Additionally, notice that $n_1 = 20{,}970$, $n_2 = 19{,}398$, $n_3 = 20{,}819$, $n_4 = 20{,}167$, $n_5 = 21{,}759$, but we only need to save a 94×94 matrix $P_{t|t}$ and a 94-dimensional vector $\hat{\boldsymbol{\eta}}_{t|t}$, where $94 \ll n_t$, for $t = 1, 2, \ldots, 5$. So, the required storage space for FRF is quite small, due to the fixed rank r.

5. DISCUSSION AND CONCLUSIONS

This article establishes that FRF is able to use past data as well as current data to great effect when estimating a process from a noisy, incomplete, and very large spatio-temporal dataset. The gains can be substantial when the temporal dependence is strong and there are past data at or near locations where the current data have gaps.

The potential of this technology in remote sensing is apparent; data gaps caused by the geometry of the paths and their relation to the geoid could be predicted with help from nearby current observations, as well as nearby observations taken in the recent past. In Section 4, we demonstrate that large gains are made when the gaps are large. The spatio-temporal statistical model "borrows strength" from regions of the geoid where data are available. This assumes that the underlying process in unobserved regions is statistically dependent on other regions *according to the fitted statistical model*; model diagnostics should provide a check of such an assumption.

The STRE model that underlies FRF also allows Fixed Rank *Smoothing* (Section 2.3), and this may be where the model will be the most useful. Based on data from a whole repeat cycle, spatial maps for smaller time units can be made by exploiting the temporal dependence.

The spatial basis functions, $\mathbf{S}_t(\cdot)$, for each time point t are not restricted to be orthogonal, although empirical orthogonal functions are a natural choice. In our experience (Cressie and Johannesson 2006, 2008; Shi and Cressie 2007), it is very important to choose them to be *multiresolutional* to capture different scales of variability. When those scales do not vary much over time, then neither will $\mathbf{S}_t(\cdot)$. It is currently an open problem to optimize the choice (type and number) of basis functions in space and time.

The fine-scale variation has a simple covariance structure, but it could be modified to a diagonal matrix, or a finer-scale spatial random effects model, $\mathbf{T}(\cdot)'\boldsymbol{\xi}$, where the components of $\mathbf{T}(\cdot)$ fluctuate at high spatial frequencies. Fixed-rank models have a great deal of flexibility to capture multiscale variability.

We have chosen to work with detrended data and *estimate* parameters of the model based on the method of moments. This may prove practically difficult if the data are spatially sparse. A fully Bayesian approach where a prior distribution is put on trend parameters $\{\boldsymbol{\beta}_t\}$, the spatial-variability parameters $\{K_t\}$, and the propagator matrices $\{H_t\}$, would avoid this practical difficulty. The dimension-reduction referred to earlier would greatly speed up a Markov chain Monte Carlo algorithm needed to produce optimal predictions and the associated uncertainty (posterior standard errors), but the computations will be much slower than what we achieve in this article (where we estimate and "plug in" parameters). One test of whether a statistical methodology will be used by a remote-sensing instrument team is whether a day's worth of data can be processed in (much) less than a day. Estimation of parameters rather than putting priors on them allows us to pass this test.

In conclusion, we propose a STRE model for very large spatio-temporal datasets, which in effect moves all the calculations onto a space of fixed dimension r. It allows optimal smoothing, filtering, or forecasting. We demonstrate through filtering both simulated and real (remotely sensed aerosol) data that the method is efficient and extremely rapid. (Over 100,000 spatio-temporal aerosol observations required on the order of a minute to estimate all parameters and a minute to filter at all 163,840 pixels.) The rapidity of the filtering, once the parameters are estimated, indicates that almost-real-time processing of sensor-network data is possible using our approach.

SUPPLEMENTAL MATERIALS

Appendix: Design and analysis of a simulation study to compare FRF with FRK. (JCGS-simu-study.pdf)

Matlab code for simulation: Matlab code to perform the simulation study reported in Section 3. (JCGS-simu-code.zip)

Datasets for MISR application: MISR data used in Section 4, and a Matlab routine for loading the data. (MISR-AOD.zip)

ACKNOWLEDGMENTS

We thank the referees and an associate editor for their helpful comments. We are grateful for comments on Section 3 from Anna Michalak; that part of the research was presented at the Fifth International Workshop on Greenhouse Gas Measurements from Space in Pasadena, CA, June 2008. Matthias Katzfuss and Dorit Hammerling gave helpful comments on an earlier version of this article. Cressie's research was supported by the National Science Foundation under awards DMS-0706997 and DMS-0707009, the Office of Naval Research under grant N00014-08-1-0464, and the National Aeronautics and Space Admininstration (NASA) under grant NNX08AJ92G issued through the ROSES Carbon Cycle Science Program and grant NNH08ZDA001N issued through the Advanced Information Systems Technology ROSES 2008 Solicitation. Shi's research was partially supported by NASA grant NNG06GD31G. MISR data were obtained by courtesy of the NASA Langley Research Atmospheric Science Data Center.

[Received March 2009. Revised April 2010.]

REFERENCES

Anderson, B. D. O. (1984), *Adaptive Control*, Oxford: Pergamon Press. [725]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman & Hall/CRC. [725]

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, Ser. B*, 70, 825–848. [726]

Berrocal, V., Gelfand, A. E., and Holland, D. M. (2010), "A Spatio-Temporal Downscaler for Output From Numerical Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 176–197. [726]

Cressie, N. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley. [725]

Cressie, N., and Johannesson, G. (2006), "Spatial Prediction of Massive Datasets," in *Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, Canberra: Australian Academy of Science, pp. 1–11. [725,729,730,743]

——— (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society, Ser. B*, 70, 209–226. [725,727,729,730,735,743]

Cressie, N., and Kang, E. L. (2010), "High-Resolution Digital Soil Mapping: Kriging for Very Large Datasets," in *Proximal Soil Sensing*, eds. R. Viscarra-Rossel, A. B. McBratney, and B. Minasny, Amsterdam: Elsevier. [730]

Cressie, N., and Wikle, C. K. (2002), "Space-Time Kalman Filter," in *Encyclopedia of Environmetrics*, eds. A. H. El-Shaarawi and W. W. Piegorsch, New York: Wiley, pp. 2045–2049. [726]

Diner, D. J., Asner, G. P., Davies, R., Knyazikhin, Y., Muller, J., Nolin, A. W., Pinty, B., Schaaf, C. B., and Stroeve, J. (1999), "New Directions in Earth Observing Scientific Applications of Multiangle Remote Sensing," *Bulletin of the American Meteorological Society*, 80, 2209–2228. [739]

Donoho, D. L., Mallet, S., and von Sachs, R. (1998), "Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods," Technical Report 517, Stanford University, Stanford. [735]

Ghil, M., and Malanotte-Rizzoli, P. (1991), "Data Assimilation in Meteorology and Oceanography," in *Advances in Geophysics*, Vol. 33, New York: Academic Press, pp. 141–266. [726]

Ghosh, S. K., Bhave, P. V., Davis, J. M., and Lee, H. (2010), "Spatio-Temporal Analysis of Total Nitrate Concentrations Using Dynamic Statistical Models," *Journal of the American Statistical Association*, 105, 538–551. [726]

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society, Ser. B*, 58, 379–396. [735]

Henderson, H. V., and Searle, S. R. (1981), "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53–60. [730]

Huang, H.-C., and Cressie, N. (1996), "Spatio-Temporal Prediction of Snow Water Equivalent Using the Kalman Filter," *Computational Statistics and Data Analysis*, 22, 159–175. [726]

Huang, H.-C., Cressie, N., and Gabrosek, J. (2002), "Fast, Resolution-Consistent Spatial Prediction of Global Processes From Satellite Data," *Journal of Computational and Graphical Statistics*, 11, 63–88. [725]

Johannesson, G., Cressie, N., and Huang, H.-C. (2007), "Dynamic Multi-Resolution Spatial Models," *Environmental and Ecological Statistics*, 14, 5–25. [726]

Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME—Journal of Basic Engineering, Ser. D*, 82, 35–45. [725,730]

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge: Cambridge University Press. [726]

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Applied Statistics*, 52, 1–18. [725]

Kang, E. L. (2009), "Reduced-Dimension Hierarchical Statistical Models for Spatial and Spatio-Temporal Data," Ph.D. dissertation, The Ohio State University, Dept. of Statistics. [726]

Kang, E. L., Cressie, N., and Shi, T. (2010), "Using Temporal Variability to Improve Spatial Mapping of Satellite Data," *Canadian Journal of Statistics*, 38, to appear. [726,741]

Kaufman, Y. J., Holben, B. N., Tanre, D., Slutsker, I., and Smirnov, A. (2000), "Will Aerosol Measurements From Terra and Aqua Polar Orbiting Satellites Represent the Daily Aerosol Abundance and Properties?" *Geophysical Research Letters*, 27, 3861–3864. [739]

Kwong, M. K., and Tang, P. T. P. (1994), "W-Matrices, Nonorthogonal Multiresolution Analysis, and Finite Signals of Arbitrary Length," Technical Report MCS-P449-0794, Argonne National Laboratory. [734]

Lopes, H. F., Salazar, E., and Gamerman, D. (2009), "Spatial Dynamic Factor Analysis," *Bayesian Analysis*, 3, 759–792. [726]

Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998), "The Kriged Kalman Filter," *Test*, 7, 217–285. [726]

Matheron, G. (1963), "Principles of Geostatistics," *Economic Geology*, 58, 1246–1266. [725,728]

Nychka, D., Bailey, B., Ellner, S., Haaland, P., and O'Connell, M. (1996), *FUNFITS: Data Analysis and Statistical Tools for Estimating Functions*, Raleigh: North Carolina State University. [725]

Nychka, D., Wikle, C., and Royle, J. A. (2002), "Multiresolution Models for Nonstationary Spatial Covariance Functions," *Statistical Modeling*, 2, 315–331. [734]

Shi, T., and Cressie, N. (2007), "Global Statistical Analysis of MISR Aerosol Data: A Massive Data Product From NASA's Terra Satellite," *Environmetrics*, 18, 665–680. [729,734,739,740,743]

Shumway, R. H., and Stoffer, D. S. (2006), *Time Series Analysis and Its Applications, With R Examples* (2nd ed.), New York: Springer. [725,730,732,733]

Talagrand, O. (1997), "Assimilation of Observations, an Introduction," *Journal of the Meteorological Society of Japan*, 75, 191–209. [726]

Ver Hoef, J. M., and Cressie, N. (1993), "Multivariable Spatial Prediction," *Mathematical Geology*, 25, 219–240; errata: 26 (1994), 273–275. [735]

Wikle, C. K., and Berliner, L. M. (2006), "A Bayesian Tutorial for Data Assimilation," *Physica D*, 230, 1–16. [726]

Wikle, C. K., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Filtering," *Biometrika*, 86, 815–829.