Estimating basis functions in massive fields under the spatial mixed effects model

Karl Pazdernik1,2 | Ranjan Maitra3

1National Security Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA
2Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA
3Department of Statistics, Iowa State University, Ames, Iowa

Correspondence
Ranjan Maitra, Department of Statistics, Iowa State University, Ames, IA, USA. Email: maitra@iastate.edu

Funding information
National Institute of Food and Agriculture, Grant/Award Number: IOW03617; Office of Defense Nuclear Nonproliferation; Consortium for Nonproliferation Enabling Capabilities (CNEC)

Abstract
Spatial prediction is commonly achieved under the assumption of a Gaussian random field by obtaining maximum likelihood estimates of parameters, and then using the kriging equations to arrive at predicted values. For massive datasets, fixed rank kriging using the expectation–maximization algorithm for estimation has been proposed as an alternative to the usual but computationally prohibitive kriging method. The method reduces computation cost of estimation by redefining the spatial process as a linear combination of basis functions and spatial random effects. A disadvantage of this method is that it imposes constraints on the relationship between the observed locations and the knots. We develop an alternative method that utilizes the spatial mixed effects model, but allows for additional flexibility by estimating the range of the spatial dependence between the observations and the knots via an alternating expectation conditional maximization algorithm. Experiments show that our methodology improves estimation without sacrificing prediction accuracy while also minimizing the additional computational burden of extra parameter estimation. The methodology is applied to a temperature dataset archived by the United States National Climate Data Center, with improved results over previous methodology.

KEYWORDS
alternating expectation conditional maximization algorithm, bandwidth, basis functions, fixed rank kriging, maximum likelihood estimation, range parameter

1 INTRODUCTION

In geostatistics, spatial estimation and prediction are often the primary focus. It is assumed that nearby observations tend to be more similar than those far apart. Since observations are spatially correlated, modeling the dependence structure can provide insight into the spatial phenomena and can be used to improve prediction. To formalize ideas, let \( s = \{ s_1, s_2, \ldots, s_n \} \in D \) represent all observed locations and let \( s_0 = \{ s_{01}, s_{02}, \ldots, s_{0N} \} \in D \) represent the desired locations, where \( D \) is a spatial domain. The two sets of locations \( s \) and \( s_0 \) may (or may not) have common elements. Let \( y(\cdot) \) be the random process defined on \( D \). Generally, \( y(\cdot) \) is characterized as a linear combination of three main components: a mean structure \( \mu(\cdot) \), a zero-mean spatial process \( f(\cdot) \), and a zero-mean measurement error process \( \epsilon(\cdot) \), that is

\[
y(s) = \mu(s) + f(s) + \epsilon(s),
\]

where \( s \) is the location of interest.
where \( f(\cdot) \) and \( \epsilon(\cdot) \) are typically assumed to be Gaussian processes. Any likelihood-based estimation procedure involves evaluating the log-likelihood

\[
\ell(\theta) \propto -\frac{1}{2} \left[ y(s) - \mu(s) \right]' \left[ \text{Cov}[y(s), y(s)] \right]^{-1} \left[ y(s) - \mu(s) \right] - \frac{1}{2} \log(\det(\text{Cov}[y(s), y(s)])) \tag{2}
\]

Under squared error loss, the best linear unbiased predictions—equivalently, Matheron’s [36] kriged estimate—of \( y(\cdot) \) at the desired locations \( s_0 \) can be computed using the first two moments of \( y(\cdot) \) (see e.g., [10]) as:

\[
\hat{y}(s_0) = E[y(s_0)] + \text{Cov}[y(s_0), y(s)] \left[ \text{Cov}[y(s), y(s)] \right]^{-1} \times [y(s_0) - E[y(s_0)]] \tag{3}
\]

Both (2) and (3) involve inverting a covariance matrix of order \( n \times n \) which is an \( O(n^3) \) operation and thus computationally impractical for massive datasets in terms of both CPU time and memory. The need for such predictions in massive spatial datasets however exists, for example in satellite data where observations are recorded across the entire globe or in more localized spatial domains where the fine resolution results in a large \( n \). In such scenarios, it is computationally infeasible to obtain the kriged estimates provided by (3) even on modern computing hardware.

The development of efficient kriging methods has received substantial attention in the literature. Much of this development has focused either on approximating the kriging equations of (3) [4, 5, 15, 16, 22, 32, 41, 42, 45, 47, 51, 53] or on defining a covariance structure that allows for exact kriging, regardless of the size of the data [2, 11, 14, 23–25, 28, 35, 43].

In this paper, we develop methodology that is capable of modeling severe nonstationarity while providing useful parameter interpretation in an efficient manner and without loss of prediction accuracy. Our underlying framework uses the “spatial mixed effects” (SME) model defined by Cressie et al. [11] to obtain parameter estimates and spatial predictions, as it provides a good compromise between flexibility in the covariance structure and computational efficiency. See [6] for a more detailed comparison of some of the different approaches.

While Finley et al. [14] allow for parameter estimation for nonstationary random fields, this methodology assumes the dependence structure between observations is the same as between observations and “knots,” a lower-dimensional subspace. However, we desire a more flexible model that allows for cross-regional dependence, which can be achieved by the SME model. In addition, the computational cost is prohibitive for massive data, as pointed out by Bradley et al. [6]. Pazdernik et al. [46] provide an efficient means of parameter estimation for the SME model, but sacrifice a more complex covariance structure. Although Lindgren et al. [35] and Nychka et al. [43] are equally as efficient, again, the flexibility of the assumed covariance structure is more restrictive.

The use of so-called low-rank approaches, such as fixed rank kriging (FRK) through the SME model, has been called into question by Stein [50]. The author illustrates, through both theory and simulation, how prediction and model fit can actually be improved by assuming independent spatial block and using the corresponding methodology, such as [51, 53]. However, in this paper, our focus is on estimation and prediction of spatial models that may possess significant regional dependence, where the assumption of independent blocks is clearly inappropriate. Therefore, we now discuss the SME model in some detail.

### 1.1 SME model

Within the framework (1), the SME model uses a linear combination of fixed basis functions and a set of locations with reduced cardinality \( (m < n) \) known as knots to define the spatial process on the original set \( s \). A fine-scale variation process \( \delta(\cdot) \) is also added to the model. Let \( \eta(\cdot) \) represent a Gaussian random effect. Then define set of knots to be \( u = \{ u_1, \ldots, u_m \} \) and also let \( s_k(\cdot) \) define the basis function corresponding to the \( k \)th knot. Under the SME model, the usual spatial process \( f(\cdot) \) is then replaced by the linear combination of basis functions and the fine-scale variation process. Assuming a linear mean structure, the model for a single observation, \( s_i \), is

\[
y(s_i) = x(s_i)' \beta + \sum_{k=1}^{m} S_k(s_i) \eta(u_k) + \delta(s_i) + \epsilon(s_i), \tag{4}
\]

where \( x(\cdot) = [x_1(\cdot), \ldots, x_p(\cdot)]' \) is a vector of known covariates with (unknown) coefficients \( \beta = (\beta_1, \ldots, \beta_p)' \). Also, let \( \eta(u) \sim N(0, \mathbf{K}) \), \( \delta(s) \sim N(0, \sigma_{\delta}^2 \mathbf{V}_\delta) \), and \( \epsilon(s) \sim N(0, \sigma_{\epsilon}^2 \mathbf{V}_\epsilon) \) be mutually independent, and with \( \mathbf{V}_\delta \) and \( \mathbf{V}_\epsilon \) as known diagonal matrices having entries corresponding to the fine-scale and measurement error variances, respectively. These diagonal elements are realizations of known functions \( \upsilon_\delta(\cdot) \) and \( \upsilon_\epsilon(\cdot) \).

Define the dispersion matrix of \( y(\cdot) \) by \( \Sigma = \mathbf{S} \mathbf{K} \mathbf{S}' + \sigma_{\delta}^2 \mathbf{V}_\delta + \sigma_{\epsilon}^2 \mathbf{V}_\epsilon \) with \( \mathbf{S} \) denoting the full \( n \times m \)-matrix of basis functions with \( k \)th column given by \( \mathbf{S}_k(s) \). Defining \( \mathbf{X} \) as the design matrix for the mean structure, the log-likelihood for \( y(\cdot) \) in simplified matrix notation is then
the usual
\[ \ell(K, \sigma_k^2, \sigma_c^2; \beta; y) \propto -\frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta) - \frac{1}{2} \log(|\Sigma|). \] (5)

Spatial predictions require the covariance between the set of observations and the desired locations, \( \text{Cov}[y(s_0), y(s)] \). Let \( A = [S_1(s_0), \ldots, S_m(s_0)]' \) be the \( N \times m \) matrix of basis functions relating the knots to the desired locations and define \( I_t \) as an \( N \times n \) matrix with entries equal to \( I_t(i, j) = I[s_0 = s_j] \). We then define the covariance,
\[ C(s_0) = \text{Cov}[y(s_0), y(s)] = AKS' + \sigma_c^2 V_{s_0} I_s \] (6)
and the kriging estimates—see Section 3.4.5 of [10]—as
\[ \hat{y}(s_0) = X(s_0)\hat{\beta} + C(s_0)\Sigma^{-1}(y - X\hat{\beta}). \] (7)
where \( \hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \). The kriging standard error (KSE) is given by \( \hat{\sigma}_k(S_0) = \{AKA' + \sigma_v^2 V_{s_0} - C(s_0) \Sigma^{-1}C(s_0)' + [X(s_0) - X'\Sigma^{-1}C(s_0)]'(X'\Sigma^{-1}X)^{-1}[X(s_0) - X'\Sigma^{-1}C(s_0)]\}^{1/2} \) [11] where \( V_{s_0} \) is a diagonal matrix with diagonal entries given by the function \( v_{s_0}(-) \) evaluated at \( s_0 \). By defining \( D = \sigma_v^2 V_{s_0} + \sigma_c^2 V_{c} \), the computational burden of matrix inversion is reduced using the identity
\[ \Sigma^{-1} = D^{-1} - D^{-1}S(K^{-1} + S'D^{-1}S)^{-1}S'D^{-1}. \] (8)
which follows from the Sherman–Morrison–Woodbury formula [20]. Note that the inversion of the \( n \times n \) matrix \( \Sigma^{-1} \) is replaced by that of two (smaller) \( m \times m \) matrices and the diagonal matrix, \( D \). Cressie and Johannesson [11] recommended the method of moments estimator for the parameters in order to guarantee positive definiteness of \( \hat{K} \), that is the estimate of \( K \).

Katzfuss and Cressie [30] provided an alternative approach to maximum likelihood estimates (MLEs) that uses the expectation–maximization (EM) algorithm [13] with \( y \) treated as the observed data and \( q \) and \( e \) as missing observations. In the absence of independent information regarding the measurement error of the process, Kang et al. [27] provide a method of estimating \( \sigma_v^2 \) that involves extrapolating the semivariogram back to the origin. Therefore, Katzfuss and Cressie [30] assume that \( \sigma_v^2 \) is known or can be obtained independently. We adopt the same approach in this paper. Starting from initial values, the variance parameter estimates are obtained upon updating
\[ K^{(t+1)} = K^{(t)} - K^{(t)}S'^{-1}(y - X\hat{\beta}) \]
\[ + \left[ K^{(t)}S'^{-1}(y - X\hat{\beta}) \right]' \left[ K^{(t)}S'^{-1}(y - X\hat{\beta}) \right], \]
from the \( t \)th to the \((t+1)\)th iteration, and proceeding till convergence. The estimate of \( \beta \) (denoted as \( \hat{\beta} \)) is obtained using generalized least squares (GLS) as in (7).

The SME methodology assumes that the basis functions are fully specified smoothing functions. Cressie and Johannesson [11] and Katzfuss and Cressie [30] use the local bisquare function. We provide further details in Section 2.1 but note that this function requires a tuning parameter, which the authors set at a constant 1.5 times the minimum distance between the inter-knot distance. In the context of spatial statistics, this tuning parameter has interpretation similar to the range parameter and so estimation provides insight into the rate of decay of spatial dependence. Furthermore, inaccurate specification of this parameter has the potential to also result in poor prediction, therefore our methodology details an extension to the EM algorithm that allows for estimation of a “range” parameter within \( S \). To be explicit, the major contributions and benefits of this work are the following:

- Computationally efficient estimation of the tuning, range parameter \( (b) \) found in the basis functions of FRK is made possible via the alternating expectation conditional maximization (AECM) algorithm.
- Full flexibility of the underlying spatial process at the knot locations is maintained to allow for a nonstationary covariance \( (K) \).
- We also find anecdotal evidence in our experiments and application that the added flexibility achieved by estimating this range parameter can help mitigate the negative effects of model misspecification in the mean structure of the spatial process.

This paper is organized as follows. In Section 2, we propose a combination of concepts from the EM approach to FRK and the AECM [7, 37] that allows for computationally practical estimation of a continuous parameter within the basis functions, in a similar manner as the ECM is used for spatio-temporal problems by Xu and Wikle [55]. Results of our extensive simulation-based evaluations of our methodology in Section 3 show that estimation of the basis functions can improve prediction under certain situations. We next demonstrate the applicability of our methodology to predict temperatures across the United States and show its potential to help mitigate the negative effects of some cases of model misspecifications. We conclude with some discussion and pointers to future work.
We also have a Supplementary Material detailing additional evaluations and investigations: sections, figures and tables in the supplement are prefixed with “S-”.

2 | METHODOLOGY

2.1 | Choice of basis functions

Fixed rank kriging and the corresponding EM estimation methodology recommend the use of multi-resolution basis functions in order to capture different scales of spatial variation. In particular, they use the local bisquare function which defines the basis function at the \( l \)-th resolution as

\[
S_{l}(x) = \Psi \left( \frac{|x - u_{l}|}{r_{l}} \right) \quad \forall x \in D, \quad (10)
\]

where \( u_{l} \) is a knot assigned to the \( l \)-th resolution, with \( l = \{1, \ldots, L\} \). \( \Psi(d) \) is the local bisquare function defined as

\[
\Psi(d) = \begin{cases} (1-d^{2})^{2}, & 0 \leq d \leq 1 \\ 0, & d > 1 \end{cases}, \quad (11)
\]

where \( r_{l} = b \min\{ \|u_{l(i)} - u_{l(j)}\| : j \neq i, 1 \leq i, j \leq m \} \) and \( b \) is some constant. We follow the recommendation and notation of other authors by defining our matrix of basis functions \( S \) by (10) and (11). This form is particularly useful in that (11) sets any value to zero where the distance between the location and the knot is greater than \( r_{l} \), the “bandwidth.” This allows the opportunity to utilize matrix operations and algorithms (see e.g., [12]) that have been specifically designed to exploit sparsity. The distances are often defined in terms of the Euclidean norm; however, this is not a necessary condition, as shown in our application to temperature data.

In practice, the number of knots used in prediction will be a function of the computational resources available and the desired prediction accuracy. Consequently, this input parameter may be considered to be known. From (10), it is clear that the remaining unknowns are the resolution \( l \) and the bandwidth constant \( b \). The optimal resolution is one of a finite, and generally small, set of positive integers. Given a finite set, estimation and prediction with varying resolutions remain easily parallelizable processes and so a model selection approach comparing performance of varying levels of resolution can be implemented. The domain of \( b \) is \( \mathbb{R}^{+} \), so a model selection approach requires a discretized domain. This approach can provide an estimate for \( b \) that will produce reasonable predictions, however numerous estimation chains may be necessary and if the set of a priori selected possible values for \( b \) does not encompass the true value, then the estimate will be biased. Maintaining a model selection approach to determining the optimal resolution, \( l \), we now direct our focus to improved estimation of \( b \).

2.2 | Restricted maximum likelihood estimation of bandwidth

This section develops methodology for obtaining a restricted maximum likelihood (REML) estimator of \( b \) along with those for \( K \) and \( \sigma_{\delta}^{2} \). As noted by Katzfuss and Cressie [29], an analytical solution to ML estimation of the parameters in the SME model does not appear to exist and direct numerical ML estimation is also challenging in that it requires maintaining a positive-definite covariance matrix \( K \). For these reasons, the EM algorithm is used for parameter estimation.

When implementing the EM algorithm to estimate \( b \), we wish to keep the advantageous structure of \( S \) intact. To avoid singularity issues during estimation, we also assume that \( \sigma_{\delta}^{2} \) is known and that GLS is used to estimate \( \beta \), as before. Thus, we follow [30] and the remaining two random variables play the role of “missing values”: \( \eta \) and \( \delta \). Assuming independence of \( \eta \sim N(0, K) \) and \( \delta \sim N(0, \sigma_{\delta}^{2}V_{\delta}) \), we denote \( \theta^{[l]} = \{K^{[l]}, \sigma_{\delta}^{2[0]}, b^{[l]}\} \) as the parameter values at the \( t \)-th iteration. The analytical solution to the M-step of the EM algorithm then requires maximizing

\[
Q(\theta; \theta^{[l]}) = -\frac{1}{2} \{ \log |K| + tr(K^{-1}E_{\theta^{[l]}}[\eta \eta']|y|) + \log(\sigma_{\delta}^{2[0]}|V_{\delta}|) \\
+ \frac{1}{\sigma_{\delta}^{2}} tr(V_{\delta}^{-1}E_{\theta^{[l]}}[\delta \delta'|y]) + \frac{1}{\sigma_{\delta}^{2}} tr(V_{\epsilon}^{-1}|y - X\beta - SE_{\theta^{[l]}}[\eta|y]| - E_{\theta^{[l]}}[\delta|y]|y - X\beta - SE_{\theta^{[l]}}[\eta|y]| - E_{\theta^{[l]}}[\delta|y]|y]) \}. \quad (12)
\]

The derivations for \( Q(\theta; \theta^{[l]}) \) are provided in Section S-1.1 of the Supplementary Material.

Maximizing (12) involves taking partial derivatives with respect to each parameter. For \( K \) and \( \sigma_{\delta}^{2} \), this is fairly straightforward as these parameters appear separately within the summation. Thus, the resulting updating scheme includes the equations in (9). Partial differentiation with respect to \( b \), however, involves taking the derivative of a quartic term nested within the matrix of multi-resolution basis functions. An analytic solution does not exist, and the usual EM algorithm does not appear to show much promise. So we develop an AECM algorithm to estimate \( b \).
2.3 Estimating bandwidth through AECM

In order to exploit the analytical iterative updating scheme for $K$ and $\sigma^2$, we partition the parameters as $\theta_1 = (K, \sigma^2)$ and $\theta_2 = \{b\}$. The CM-step of the AECM algorithm alternates between maximizing the likelihood function (12) with respect to $\theta_1$ (keeping $\theta_2$ fixed at its current value) and $\theta_2$ (by holding $\theta_1$ fixed at its current value). Therefore, $\theta_1$ is updated, as before, by (9). An estimate for $\theta_2$, however, cannot be calculated analytically nor numerically, so in lieu of maximizing (12), we maximize the restricted log-likelihood function,

$$
\ell(K, \sigma^2, b; y, \sigma^2_c) \propto -\frac{1}{2} \left( y - X\hat{b} \right)' \Sigma^{-1} \left( y - X\hat{b} \right) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} \log \left( |X' \Sigma^{-1} X| \right). 
$$

(13)

When restricted to be a function of $\theta_2$, (13) involves a covariance matrix which is in a form that allows for the use of dimension reduction equalities in the matrix calculations. As noted in [11], the Sherman–Morrison–Woodbury formula can be used to obtain a computationally efficient form of $\Sigma^{-1}$. To avoid the burdensome computation of the determinant, we also invoke Sylvester’s determinant theorem [11, 18], which is an analogue to the Sherman–Morrison–Woodbury formula. Let $X$ be an invertible $n \times n$ matrix, $A$ be an $n \times m$ matrix, $B$ an $m \times n$ matrix, and $I_m$ denote the $m \times m$ identity matrix. Then $\det(X + AB) = \det(X) \det(I_m + BX^{-1}A)$. Also, let $C$ be the Cholesky decomposition of $K$ ($K = CC'$). This theorem, along with the fact that $K$ is positive-definite and some algebra, yields the reduction

$$
\det(D + SKS') = \det(D) \det(C)^2 \det(K^{-1} + S'D^{-1}S) 
$$

(14)

which provides us with the advantage that it requires computing the inverse of an $m \times m$ matrix and a determinant of an $m \times m$ matrix instead of the determinant of an $n \times n$ matrix. Also, a Cholesky decomposition of the two $m \times m$ matrices, $K$ and $K^{-1} + S'D^{-1}S$, can be used in both the calculation of the determinant using (14) and the inverse covariance in the log-likelihood that uses (8), providing an additional reduction in computation. Finally, since $D$ is a diagonal matrix and $C$ is a lower-triangular matrix, the determinant of each is simply (and efficiently) computed as the product of the diagonal elements of $D$ and $C$. Let $C_2$ be the Cholesky decomposition of $(K^{-1} + S'D^{-1}S)$ and $C_3$ be the Cholesky decomposition of $X' \Sigma^{-1} X$. The final form of the restricted log-likelihood is then

$$
\ell(K, \sigma^2, b; y, \sigma^2_c) \propto -\frac{1}{2} \left( y - X\hat{b} \right)' \Sigma^{-1} \left( y - X\hat{b} \right) - \frac{1}{2} \log(|D|) - \log(|C_2|) - \log(|C_3|),
$$

(15)

where $\Sigma^{-1}$ is defined as in (8). Parameter estimates using our AECM algorithm are obtained by computing $\theta_1$ from (9) given $\theta_2^{[t]}$, then maximizing (15) with respect to $\theta_2$; given $\theta_1^{[t]}$, and repeating this process until convergence.

Our maximization step requires solving at most an $m \times m$ linear system of equations; however, to perform a full maximization of $\theta_2$ between each update of $\theta_1^{[t]}$ would be immensely inefficient. Numerical maximization techniques for $\theta_2$ may also fail when the log-likelihood function is ill-behaved. However, exploratory results suggest that the function remains locally quadratic in $b$ near its MLE, so we use a search algorithm to explore the log-likelihood for its maximum with respect to $b$.

Given a quadratic function and a desired maximum, we can use the quadratic search algorithm [39] that requires only three evaluations of the likelihood. In order to improve initial starting values, we propose using the golden search algorithm [1, 33] to provide a “burn-in” phase. Also, since the overall discrepancy between the starting values and the final estimates of $K$ could be very large, it is best to obtain a reasonable estimate of $\theta_1$ before any update of $\theta_2$. So we also suggest a burn-in phase of the EM algorithm at a few values of $b$. The full updating algorithm is therefore as follows:

1. Obtain $\theta_1$ by (9) until weak convergence.
2. (a) Set $\theta_2^{[1]} = \{b_1, b_2, b_3, b_4\}$ based on the Golden Search algorithm.
   (b) Set $\theta_2^{[1]} = \{b_1, b_2, b_3\}$ based on the Quadratic Search algorithm.
3. Obtain $\theta_1$ by (9) for each $\theta_2^{[1]}$.
4. Evaluate (15) for each set of $\theta_1^{[1]}$ and $\theta_2^{[1]}$.
5. Set $\theta_1^{[t+1]}$ to the set of parameter values that maximize (15).
6. Repeat 2(a)–5 until weak convergence.
7. Repeat 2(b)–5 until convergence.

The convergence criteria for the initial burn-in phase of the EM algorithm need not be very strict, as the log-likelihood function typically plateaus after only a few iterations. However, obtaining better initial values for $\theta_1$ is still advisable to avoid the problem of the evaluated log-likelihood getting trapped in a local maxima. Also, Cressie and Johansson [11] suggest using $b = 1.5$ in the bisquare basis functions, so we will use a set of initial values for $b$ that encompasses 1.5 for the EM algorithm.
burn-in phase. In the next section, we investigate, through simulation, the impact of estimating the bandwidth constant, $b$, of the local bisquare basis functions.

3 | EXPERIMENTAL EVALUATIONS

3.1 | Experimental setup

The previous section outlined an AECM algorithm for efficiently computing an MLE for $b$ in addition to that for $K$ and $\sigma_s^2$. Here, we evaluate performance of our method through a series of simulation experiments using the local bisquare basis function of (10). Our experimental design was similar to that in [30] and which used (4) as the generative model. While this does not allow for clear relation to the data-generating process, it makes direct assessment and comparison of the different methodologies more straightforward.

For our experiments, we chose a one-dimensional spatial domain, $D = \{1, \ldots, 256\}$, with $n = 64$ observed locations selected either completely randomly or following a “clustered” random sample. The clusters were created by splitting $D$ into intervals of 32 locations and omitting every other interval. To elucidate, a clustered random sample is a subset of $\{1, \ldots, 32\} \cap \{64, \ldots, 96\} \cap \{128, \ldots, 160\} \cap \{192, \ldots, 224\}$. Five knots, centered at $(0.5, 64.5, 128.5, 192.5, 256.5)$, were used at a single resolution.

The main focus of our simulation was to understand the effects on estimation and prediction when aspects of the covariance structure are varied. To that end, we used a simple linear mean structure ($\beta_0 = 5$ and $\beta_1 = 0.08$) for all simulations. To cover the wide variety of covariance matrices $K$ allowed by FRK, we simulated covariance matrices to have stationary, positively correlated nonstationary and unrestricted nonstationary structures. The stationary covariance was modeled after a Matérn covariance matrix [49] which, for locations $s_i$ and $s_j$, is defined by

$$\text{Cov}[\mathbf{f}(s_i), \mathbf{f}(s_j)] = \frac{\rho}{2^{\frac{\nu}{\theta}} \Gamma(\nu)} \left( \frac{||s_i - s_j||}{\theta} \right)^\nu K_\nu \left( \frac{||s_i - s_j||}{\theta} \right),$$

where $K_\nu$ is the modified Bessel function of the second kind of order $\nu$, $\nu$ is the smoothness parameter, $\rho$ is the sill or scale parameter, and $\theta$ is the range parameter, with $\nu, \theta \in (0, \infty)$ and $\rho \in [0, \infty)$. For our simulation setup, we held the parameters fixed at $\rho = 9$, $\theta = 96$, and $\nu = 1$. The unrestricted nonstationary structure was created through the matrix product of a realization from the Wishart [54] distribution ($\mathbf{W} \sim \mathcal{W}(2I_5, 10)$, where $I_5$ is the $5 \times 5$ identity matrix) sandwiched in between diagonal matrices with elements increasing from 1 to 5. This simulation procedure created significant nonstationarity with expected value given by a diagonal matrix with elements $\{2, 8, 18, 32, 50\}$. The positively correlated nonstationary structure was simply the absolute value of the unrestricted nonstationary matrix described above.

Modeling the covariance as spatial in nature is only useful when significant spatial dependence exists in the data. To quantify the magnitude of spatial structure, we compare two ratios of variances: (i) $\text{tr}(SKS' + \sigma_b^2V_b)/\text{tr}(\sigma_e^2V_e)$, the signal-to-noise ratio, and (ii) $\text{tr}(\sigma_b^2V_b)/\text{tr}(SKS' + \sigma_b^2V_b)$, the fine-scale variation proportion. Testing the limits of our methodology, we varied the fine-scale variation parameter $\sigma_e^2$ to be within $\{0.01, 0.1, 1\}$ and the measurement error $\sigma_b^2$ to lie within $\{1, 10, 100\}$. This resulted in the ratio (i) to range between $\{0.037, 30.96\}$ and for ratio (ii) to range between $\{0.0003, 0.215\}$. Our values for $\sigma_e^2$ are chosen to represent situations of low, medium, and high noise. Since we are comparing methods that estimate spatial covariance components, we expect both methods to perform poorly in the high-noise scenario. However, with the extra model complexity introduced in the AECM approach, the high-noise results are likely to be skewed to favor the EM approach since it should include less potential for overfitting.

In the interest of simplicity, the weighting functions were set to one ($w_{ij} = 1 \forall i, j$), creating identity matrices for $V_b$ and $V_e$, and the number of resolutions was held constant at $l = 1$. As the primary parameter under evaluation, the levels of the bandwidth constant, $b$, are fairly important. First, since $b = 1.5$ is recommended and used in the EM approach, we included this to quantify the additional performance achieved when the ideal bandwidth constant is known. Next, we varied $b$ adequately enough in either direction to create realistic patterns in the overlap of the basis function. Specifically, we chose $b = \{1, 2\}$ to provide intuitive “low” and “high” overlap scenarios. We considered increasing $b$ beyond 2 but noticed that the effect of the overlap becomes severe far too quickly and loses spatial structure to aggregation over a large subset of the data. The lower limit for $b$ is set at 0.5 because any lower value will potentially completely ignore observations. However, even at the limit and when assuming uniform knot placement, any observation that is located at the exact halfway point between knots will not contribute to estimation. This is an unusual scenario, unlikely to be seen in practice. However, we still included this limit in our investigations in order to fully assess the limits of each methodology. So, the bandwidth constant values selected for our simulation were ($b = \{0.5, 1, 1.5, 2\}$).
The R [48] package fields [44] was used to simulate 1000 Gaussian random fields for each combination of covariance type, parameter value, and sampling design. For each simulated field, we obtained MLEs using both our AECM algorithm (so, estimating \( b \)) and the EM algorithm (leaving \( b = 1.5 \) fixed). From these estimates, we compared model fits quantified by Kullback and Leibler’s [34] K–L divergence and calculated the corresponding kriging estimates and standard errors at all locations within \( D \).

3.2 | Results

Computational cost is crucial for any spatial prediction method intended for massive datasets. The simulation results, however, were generated using a relatively small sample size in order to permit evaluation of the methodology in a wide range of settings. Without implementing parallel processing techniques, the computational complexity of our method is a known \( O(3nr^2) \) versus \( O(nr^2) \) for the EM approach, after burn-in. A thorough comparison of computational cost is, therefore, saved for the application section where the sample size is larger.

The other relevant points of comparison in our simulation experiments are parameter estimation error, model fit, and performance in prediction. We assess error in parameter estimation by calculating the median absolute deviation (MAD) given the true parameter values. However, to properly interpret these results, we need to first reiterate that not every scenario is equal. For example, when the true \( b = 1.5 \), the EM by design should perform better across all metrics since it is given the true parameter value. The simulations when the true \( b = 0.5 \) represent an unusual scenario, as previously mentioned. Finally, the high-variance examples, where \( \sigma_e = 100 \) contain such a large amount of noise that any spatial dependence is almost impossible to detect. That being the case, both the EM and AECM methods are likely to exhibit patterns typical of overfitting the data. Figure 1 displays the results: we see that parameters that are part of the mean structure exhibit similar patterns. The error in slope and intercept estimation is decreased by the AECM algorithm in the presence of low measurement variance (\( \sigma^2 = 1 \)) and increased given high measurement variance (\( \sigma^2 = 100 \)). This can be seen by a positive slope relative to the diagonal line in Figure 1A,B. The magnitude of deviation is also an increasing function of measurement error for either approach. The AECM algorithm decreases the error in estimating \( \beta_0 \) when the true value of \( b \) is not equal to 1.5, provided the measurement error is low (most circles that are not colored blue fall below the diagonal line). When \( \sigma^2 = 10 \), both approaches perform similarly and when \( \sigma^2 = 100 \), again, the AECM algorithm exhibits signs of overfitting.

This pattern is also present in the estimation of \( \beta_1 \), however the relative improvement by the AECM algorithm in low-to-medium noise simulations (excluding \( b = 1.5 \)) is larger. When \( \sigma^2 \) is less than 100 and the covariance of \( K \) is nonstationary, the AECM algorithm consistently improves estimation of \( \beta_1 \). These results hold even when we vary covariance type, fine-scale variation, or sampling design. Figure 1C shows the MAD for all unique elements of \( \hat{K} \). These results are more difficult to interpret and mostly appear similar between the AECM and EM approaches, but there are some patterns that can be observed. Specifically, in the presence of large measurement error (displayed by squares) or when the true \( b = 1.5 \) or greater (blue or green symbols), the EM algorithm consistently reduces the MAD. However, as previously discussed, these are largely scenarios where we would expect the EM approach to perform better. When \( b = 1 \), the pattern more closely resembles similar conclusions drawn from Figure 1A,B.

In summary, there is a preference for the AECM method when measurement error is low (\( \sigma^2 = 1 \)) and for the EM method when measurement error is high (\( \sigma^2 = 100 \)). It is not obvious if covariance type and sampling design exhibit significant differences in performance between the two methods.

Accurate estimation of variance parameters in spatial fields can often be challenging. In particular, the fine-scale variation parameter in the SME model is often underestimated. However, this is where the AECM appears to be most advantageous. Allowing a more flexible spatial structure by estimating the range parameter with respect to the knots (\( b \)) provides fairly consistent reduction in the MAD for \( \sigma^2_b \). This improvement is especially noticeable when \( b = 0.5 \) and \( \sigma^2_b \) (gray circles) where the MAD for the EM method is often observed at over twice that of the AECM approach. Only when both independent variance components are equal (\( \sigma^2_b = \sigma^2_\epsilon \)) and occasionally when the measurement error is high does the AECM method fail to improve estimation.

Clearly, when observations contain extreme levels of measurement error, parameter estimation given a SME model is problematic. Assuming any spatial model when nonspatial variance components dominate the total variance can have ill-effects on estimation accuracy. This is no more evident than when exploring distributions of \( \hat{b} \) obtained through simulation, provided in Figure 2. Distributions are further segregated and color-coded between those that showed significant spatial autocorrelation (\( \epsilon < .05 \)) as measured by Moran’s I [38] and those that did not. (Some distributions of \( \hat{b} \) are truncated in order to highlight the range of highest density within each distribution; however, these estimates account for less than 0.17% of the total simulations.) We see in the figure that there is a clear distinction between distributions with and without
significant spatial autocorrelation. When $\sigma^2 = 100$, nearly all simulated fields do not exhibit enough spatial dependence to compensate for the large measurement error. Without significant spatial dependence, $\hat{b}$’s tend to be negatively biased with a large percentage of the distribution residing between 0 and 1, made worse when $K$ has a Matérn structure. This pattern persists for all levels of measurement error, however the percentage of fields with insignificant spatial dependence decreases with decreasing $\sigma^2$. Estimation of $b$ performs fairly well for true values of $b \leq 1.5$, however when $b = 2$, there is again a tendency for $\hat{b}$ to be underestimated.

The difference between the estimation methods clearly varies across parameters. To summarize the fit of all parameter estimates and, consequently, the overall model fit, we compared Kullback–Leibler (KL) divergence between the two methods. KL divergence quantifies the amount of information lost when an estimated distribution, $Q$, is used in place of the true distribution, $P$. The KL divergence [50] for two multivariate normal distributions, $P \equiv N(\mu_P, \Sigma_P)$ and $Q \equiv N(\mu_Q, \Sigma_Q)$, on $\mathbb{R}^n$ simplifies to

$$2\text{KL}(P, Q) = \text{tr} \left( \Sigma_P^{-1} \Sigma_Q \right) + (\mu_Q - \mu_P)' \Sigma_Q^{-1} (\mu_Q - \mu_P) - \log |\Sigma_P| + \log |\Sigma_Q| - n.$$  

Figure 3 provides the proportion out of 1000 simulations where the KL divergence is less for AECM than it is for EM.
Thus, values over 0.5 indicate an improvement in overall model fit when estimating $b$. Interestingly, although individual parameter estimates almost unanimously improved using AECM when $b \leq 1$ and $\sigma^2 = 1$, the opposite holds when comparing KL divergence. In these cases, overall model fit is better assuming a fixed $b = 1.5$. When $b \geq 1.5$, on the contrary, nearly every combination of simulation input resulted in a reduced KL divergence using the AECM algorithm. In the presence of high measurement errors, results were similar between estimating $b$ and leaving it fixed at 1.5, regardless of the true value of $b$. When $\sigma^2 = 10$, the AECM algorithm decreases KL divergence the most consistently, reaching a proportion as high as 0.86.

Our final performance comparisons were in the context of prediction. To that end, we summarized our results by three metrics: mean square prediction error (MSPE), ratio of KSE, and prediction interval coverage (PIC). MSPE is the usual statistical summary which calculates the average squared loss between the kriging predictions and the true value at unobserved locations. Since this is a simulation setting, we can define the true value as the uncorrupted spatial process, $\{y(s_0) - e(s_0)\}$. The median KSE at each spatial location is computed given estimated parameters by either method. The ratio KSE (rKSE) is the ratio between this median KSE and the KSE computed given the true parameter values at similar simulation input.
values. Ideally, the rKSE is equal to 1. PIC is the percentage of prediction intervals that contain the true value, averaged over all locations. Since we assumed 95% prediction intervals, we expect the PIC to be roughly 0.95 for each case. For each summary statistic, the median was computed over all values of $\sigma^2$ and sampling designs. Results are provided in Table 1. Medians for every unique combination of input values are provided in Section S-2.1 of the Supplementary Material. Clearly, knowing the true parameter values improves prediction, as its MSPE is always the lowest. Likewise, when the true $b = 1.5$, the EM approach is generally best in all metrics, since it is provided one of the true parameter values. If the measurement error is small enough and the true $b \leq 1$, the AECM algorithm improves MSPE compared to the EM approach. However, for $\sigma^2 \geq 10$, holding $b = 1.5$ fixed provides the most accurate prediction. The KSE is often underestimated, with the AECM method exacerbating the issue. However, when the true $b = 0.5$ and $\sigma^2 = 1$, the EM approach severely overestimates KSE. The lower KSE also has the consequence of lowering the PIC, as narrower confidence bands will inevitably miss the true values at a higher rate. However, the anticipated improved performance using AECM when $b = 0.5$ and $\sigma^2 = 1$ is absent. To elucidate this pattern, a single realization from our simulation with prediction regions is provided in Figure 4, with the PIC at each location included to facilitate comparison. The PIC is exclusively closer to 0.95 for the EM method across locations (top plot), but as a plot of the actual process (bottom) accentuates, the EM method misses clear signals in the data by over-smoothing predictions. Overestimating KSE, in this case, allows the prediction intervals to include the true values, however, it is an inaccurate representation of the process.
| K   | $\sigma^2_i$ | $b$  | MSPE | rKSE | PIC |
|-----|------------|------|------|------|-----|
|     |            |      | True | AECM | EM  |
| M   | 1          | 0.5  | 0.271| 0.508| 1.689|
| M   | 1          | 1    | 0.216| 0.359| 0.369|
| M   | 1          | 1.5  | 0.194| 0.331| 0.258|
| M   | 1          | 2    | 0.165| 0.330| 0.291|
| M   | 10         | 0.5  | 0.814| 1.734| 2.292|
| M   | 10         | 1    | 0.913| 1.705| 1.326|
| M   | 10         | 1.5  | 0.829| 1.699| 1.256|
| M   | 100        | 0.5  | 1.865| 9.446| 6.951|
| M   | 100        | 1    | 2.626| 9.383| 5.902|
| M   | 100        | 1.5  | 2.829| 9.414| 5.891|
| M   | 100        | 2    | 2.764| 9.505| 6.011|
| WP  | 1          | 0.5  | 0.653| 1.055| 2.413|
| WP  | 1          | 1    | 0.222| 0.345| 0.865|
| WP  | 1          | 1.5  | 0.219| 0.329| 0.269|
| WP  | 1          | 2    | 0.170| 0.333| 0.325|
| WP  | 10         | 0.5  | 1.215| 2.300| 3.183|
| WP  | 10         | 1    | 1.138| 1.919| 1.706|
| WP  | 10         | 1.5  | 1.039| 1.812| 1.346|
| WP  | 10         | 2    | 0.806| 1.866| 1.363|
| WP  | 100        | 0.5  | 2.782| 12.810| 3.126|
| WP  | 100        | 1    | 3.630| 12.343| 7.638|
| WP  | 100        | 1.5  | 3.380| 11.372| 6.994|
| WP  | 100        | 2    | 2.806| 11.668| 6.899|
| WN  | 1          | 0.5  | 0.545| 1.002| 2.316|
| WN  | 1          | 1    | 0.221| 0.338| 0.881|
| WN  | 1          | 1.5  | 0.218| 0.318| 0.260|
| WN  | 1          | 2    | 0.167| 0.331| 0.316|
| WN  | 10         | 0.5  | 1.215| 2.329| 3.126|
| WN  | 10         | 1    | 1.127| 1.930| 1.738|
| WN  | 10         | 1.5  | 1.021| 1.838| 1.362|
| WN  | 10         | 2    | 0.774| 1.854| 1.358|
| WN  | 100        | 0.5  | 2.841| 12.964| 8.895|
| WN  | 100        | 1    | 3.650| 12.359| 7.629|
| WN  | 100        | 1.5  | 3.373| 12.036| 6.960|
| WN  | 100        | 2    | 2.935| 11.847| 6.957|
As an example with $b = 0.5$, Figure 4 also highlights how unusual it is to have nonoverlapping basis functions. The coverage proportion (PIC) even when provided the true parameter values (orange line in the top plot of Figure 4) shows that correct coverage cannot be attained uniformly throughout the spatial domain when the range of the basis functions are so restricted. The data itself (bottom plot of Figure 4) show that nonoverlapping basis functions can also create artifacts that resemble mean structure. However, as seen by the predicted fit, allowing a more flexible bandwidth parameter may be able to account for an underspecified mean structure. We explore this idea further in the application of our methodology in the next section.

In summary, the results of our experiments show that estimating the bandwidth constant in the local bisquare basis functions by means of the AECM algorithm can be advantageous to both individual parameter estimation and prediction, but only under certain circumstances. When the true $b$ is small, and the measurement error is also low, the AECM method is generally beneficial. However, the improved parameter estimation and prediction are combined with an underestimate of the overall variance in the process. In the extreme case ($b = 0.5$), this combination produces obvious large-scale trends in the data that is not captured by the mean structure. Clearly, fitting a more complex model for the mean, such as a higher order polynomial, would also allow the flexibility necessary to fit such data. The decision of whether or not to estimate $b$ is then analogous to a common statistical concern in mixed models of whether to treat an effect as fixed (mean structure) or random (spatial process), which depends on the context and desired interpretation. As expected, our results as $b$ approaches 1.5 confirm that the extra parameter estimation is unnecessary, however the potential negative effects on prediction accuracy, at least when measurement error is low, are fairly minimal. In terms of overall model fit, as measured by KL divergence, estimating $b$ by means of the AECM algorithm is equivalent or better than the standard approach of fixing $b = 1.5$ in most situations that we explored, excluding the cases with low measurement error ($\sigma_\epsilon = 1$) and relatively low bandwidth constant ($b \leq 1$). We believe that the unusual covariance structure caused by limited overlap in the basis functions ($b \leq 1$) may underestimate the variability in the observed data while creating artifacts in the mean structure. Therefore, an especially low estimate for the bandwidth could be used as an indicator that the mean structure should be revisited and that quantification of overall uncertainty may be underestimated.
Having gained these insights from our simulation study, we now apply our methodology to the National Climate Data Center (NCDC) data of monthly temperatures recorded across the continental United States of America.

4 | APPLICATION: PREDICTING TEMPERATURES OVER THE CONTIGUOUS UNITED STATES

We applied our methodology to temperature data recorded by the Cooperative Observer Program (COOP). Established in 1890, the COOP is the largest and oldest weather and climate observation network [40], recording weather information every 24 h from over 11,700 volunteer citizens and institutions. The data are available online at http://www.image.ucar.edu/Data/US.monthly.met/. There is considerable variability in the number of weather stations across the contiguous United States for which measurements are available at any point of time as well as in the number of variables recorded (monthly maximum temperature, minimum temperature, precipitation, etc.) at these time points.

The mean temperature recorded over a regular grid is an important summary in climate science [26], as it is used to benchmark climate-derived models. Inevitably, temperatures are not recorded for all locations on the predetermined regular grid. Thus, kriging is necessary to estimate temperatures at the remaining unobserved locations. Also, since spatial dependence typically exhibits a high degree of smoothness with respect to temperature, this response should help determine if \( b = 1.5 \) is a good default smoothing parameter value.

As in Pazdernik et al. [46], we restrict our attention to mean daily temperature readings for April 1990 in this paper. The COOP had daily minimum and maximum temperatures recorded at 5030 locations across the contiguous United States—these were summarized into a single univariate measurement at each of these locations in terms of an average of the mean monthly minimum and maximum temperatures. In this setup, note that performing kriging requires iteratively solving a 5030 × 5030 linear system, a task which would be computationally challenging in terms of CPU time and memory. Thus, efficient forms of kriging such as fixed rank kriging, with parameter estimation achieved through either an EM or AECM approach are helpful. Also, this dataset is used here simply for illustration, our method can easily be applied to larger sample sizes, as it is scalable with computational cost growing at a linear rate with the sample size.

4.1 | Known covariates

Temperature is known to be affected by geographic location and elevation. Since temperature cannot reasonably be modeled with a constant mean structure in the presence of these factors, a set of covariates is necessary to reduce the remaining structure to a zero-mean spatial process and zero-mean measurement error. Assuming linearity in the effects of our covariates, we can model the nonzero mean structure with the additive model \( \mathbf{x}(s, h_s) \beta = \beta_0 + g_{\text{Elev}}(h_s) \beta_{\text{Elev}} + g_{\text{Lat}}(s) \beta_{\text{Lat}} + g_{\text{Lon}}(s) \beta_{\text{Lon}} \) where \( h_s \) is the elevation at location \( s \).

The exact functional relationship between temperature and our covariates is unknown, however we expect temperature to decrease with increasing elevation and latitude. One option is to assume a simple linear mean structure for all covariates. We will refer to this framework as the “reduced model.” However, this is likely an over-simplification and, hence, smoothing regression splines fits based on AIC were also used to model the mean structure. Specifically, \( g_{\text{Elev}}(h_s) \) required a cubic regression spline with nine degrees of freedom. \( g_{\text{Lat}}(s) \) was modeled with a quadratic regression spline with four degrees of freedom, and \( g_{\text{Lon}}(s) \) fit best with a cubic regression spline with nine degrees of freedom. We will refer to this scenario as the “full model.” Since the AECM method showed the ability to accommodate irregular patterns in the data not captured by the mean structure, we are interested in a comparison between these reduced and full models. The model fits and their resulting residuals are shown in Figures 5 and 6, respectively. Although the relationship between temperature and each covariate appears decidedly nonlinear, validation of model assumptions, by way of a residual plot, does not feature this violation. Visual assessment suggests that both mean structures are reasonable.

4.2 | Results and analysis

REML estimation was carried out by maximizing (15) by our AECM algorithm and by the iterative updating scheme of (9) obtained by the EM algorithm. Following the strategy proposed by Kang et al. [27], \( \sigma_2^2 = 2.059 \) was determined separately. We selected \( m = 369 \) knots representing 7.3% of the original dataset on a regular triangular grid within a space-filling framework and tested our methodology on up to four levels of resolution. Distance matrices were obtained by using the \texttt{rdist.earth} function in the \texttt{fields} package in \texttt{R} [48], which computes the Great circle distance between any two locations on the Earth.
The size of this dataset, despite making use of the fixed rank kriging equations, also warranted the use of more efficient software. Functions were translated to MATLAB [52], which is optimized for matrix and vector algebra and can be easily parallelized, however an R implementation also exists [56]. We note that a major burden of the AECM method is that every iteration during parameter estimation costs three to four times that of the EM method and that subsequent iterations require previous results, so care had to be taken in parallelization. However, despite the additional burden, recall that the main benefit of our AECM algorithm is additional parameter estimation.

We compared model performance in terms of (a) prediction by MSPE through five-fold cross-validation and (b) efficiency, summarized by the elapsed time for estimation, in minutes (see Table 2). All reported computations were performed using a standard desktop computer with 16GB of RAM and a 64-bit Intel quad-core i7 processor. Also included in the table are parameter estimates for $b$, the range of $K$, and $\sigma^2_\delta$ and the corresponding evaluated log-likelihood.

The performance measures show that, as expected, predictions stemming from the full model provide lower MSPEs. For comparison within a model, we select the optimal number of resolutions based on MSPE and compare those results, hence those results are in bold font. When using the full model, the setup with two resolutions minimized the MSPE for both methods, with the AECM method barely outperforming the EM method. Three- and one-resolution settings were optimal when using the reduced linear model for the EM and AECM methods, respectively. Here again, AECM produced lower MSPE and also produced the lowest MSPE across all options, albeit often with fairly negligible improvements. The largest discrepancy within the same model and resolution pair occurred for the simplest model, reduced with only one resolution, where the AECM approach substantially decreased MSPE, relative to the EM. Although the AECM algorithm is not optimal across the board for every number of resolutions, these results also indicate that even when suboptimal resolutions are chosen, prediction error is not substantially increased.

The EM method is, by design, a more efficient algorithm than AECM for this problem and the timing results substantiate that fact. Although the AECM method requires additional computation time, the increase observed was only between 50% and 79% within a resolution/model combination. Considering the amount of communication required between workers (necessary after each iteration), MATLAB manages this extra burden surprisingly well. Also note that the relative increase was lower across all resolutions when assuming the reduced model.

For the full model, parameter estimates for $\sigma^2_\delta$ were essentially identical between the two methods and estimation of $b$ validated the reasonable choice of bandwidth.
**Table 2**  Prediction and estimation results from both the EM and AECM algorithms, using 1 to 4 resolutions and the full or reduced model on the US temperature dataset

| Model     | Res | Method | MSPE  | Minutes | $\hat{b}$ | Range ($\hat{K}$) | $\hat{\sigma}^2$ | $\hat{\varepsilon}(\hat{b})$ |
|-----------|-----|--------|-------|---------|----------|-------------------|----------------|----------------|
| Full      | 1   | EM     | 0.7706| 14.6    | NA       | 158.64           | 0.0046         | -7215.72       |
| Full      | 1   | AECM   | 0.7716| 26.2    | 1.508    | 155.26           | 0.0046         | -7215.76       |
| Full      | 2   | EM     | 0.7672| 16.8    | NA       | 210.11           | 0.0047         | -7213.01       |
| Full      | 2   | AECM   | 0.7670| 27.3    | 1.468    | 203.06           | 0.0047         | -7212.60       |
| Full      | 3   | EM     | 0.7680| 19.0    | NA       | 210.11           | 0.0047         | -7214.88       |
| Full      | 3   | AECM   | 0.7691| 28.3    | 1.473    | 318.34           | 0.0046         | -7216.26       |
| Full      | 4   | EM     | 0.7684| 19.0    | NA       | 326.18           | 0.0046         | -7214.88       |
| Full      | 4   | AECM   | 0.7704| 28.3    | 1.473    | 326.32           | 0.0046         | -7216.26       |
| Reduced   | 1   | EM     | 0.8050| 23.0    | NA       | 537.85           | 0.0030         | -7257.36       |
| Reduced   | 1   | AECM   | 0.7896| 37.0    | 1.890    | 297.90           | 0.0034         | -7244.87       |
| Reduced   | 2   | EM     | 0.7946| 25.8    | NA       | 507.49           | 0.0031         | -7245.30       |
| Reduced   | 2   | AECM   | 0.7918| 39.7    | 1.546    | 603.01           | 0.0033         | -7242.13       |
| Reduced   | 3   | EM     | 0.7929| 30.3    | NA       | 575.08           | 0.0030         | -7244.78       |
| Reduced   | 3   | AECM   | 0.7954| 47.8    | 1.271    | 372.84           | 0.0033         | -7245.35       |
| Reduced   | 4   | EM     | 0.7944| 27.4    | NA       | 617.49           | 0.0030         | -7245.56       |
| Reduced   | 4   | AECM   | 0.7978| 41.1    | 1.281    | 354.57           | 0.0032         | -7245.11       |

Note: The MSPE values in bold font represent the best performing resolutions for both the EM and AECM methods, with corresponding time taken and parameter estimation also in bold font.

Abbreviations: AECM, alternating expectation conditional maximization; EM, expectation–maximization.

constant being 1.5, as mentioned by Cressie and Johannesson [11]. However, the resolution that both maximized the evaluated log-likelihood and minimized the MSPE also provided an estimate of $b$ with the largest deviation from 1.5. The reduced model produced estimates of $b$ as high as 1.890 for optimal resolution as measured by MSPE and at 1.546 for the optimal resolution measured by evaluated log-likelihood. All values of $\hat{\sigma}^2$ for the reduced model are lower relative to the full model, but the decrease is consistently less for the AECM method. If we assume that the full model is a more appropriate description of our data, then this provides evidence that the AECM method counteracts the effects of model misspecification by allowing the range of the basis functions, or bandwidth, to vary. The range of $\hat{K}$ using the AECM method also shows greater agreement between analogous models for almost every resolution, further corroborating this hypothesis. Complete image plots of $\hat{K}$ are provided in Section S-2.2 of the Supplementary Material.

For the remainder of this section, we assume that the full model with two resolutions is our best model and, consequently, we focus on those results. Plots of the KSEs using both EM and AECM methods are provided in Figure 7. Both algorithms produce extremely similar results, however the scales are slightly shifted. The range of KSE values for the EM method is between 0.0688 and 0.1640, whereas the AECM method produces values between 0.0689 and 0.1669. Considering the evidence we observed that the fixed rank kriging approach underestimates KSE when $b \approx 1.5$, we are inclined to believe that the estimates of KSE resulting from the AECM method are likely closer to the truth. Other than this difference, the pattern of KSE is fairly consistent between methods across the entire contiguous United States, with pockets of higher variability existing in the western states, particularly along the southern coast of California. The pattern along the California coastline is particularly interesting because, at first glance, it resembles edge effects often seen in spatial prediction. However, knot locations extend well beyond the boundary of the contiguous United States. Also, when comparing Figures 7 and 8, we see that this increased KSE coincides with the negative effect contributed by the spatial dependence portion of the prediction. This suggests that the mean structure was not as predictive of temperature in these regions, which leads to increased uncertainty in the overall predictions.

Figure 8 displays the data with superimposed knots (Figure 8A) for two resolutions and the corresponding spatially predicted field using the AECM method. The predictions are decomposed into three parts: the mean
structure ($X(s_0)\hat{\beta}$), the spatial dependence ($C(s_0)\Sigma^{-1}(y - X\hat{\beta})$), and the complete additive kriging estimates that are given by $\hat{y}(s_0) = X(s_0)\hat{\beta} + C(s_0)\Sigma^{-1}(y - X\hat{\beta})$. It is clear from Figure 8B that the covariates are critical to temperature prediction, since the mean structure reflects the general trends of the temperature in April in the contiguous United States. In particular, the image depicts the negative relationship that exists between elevation/latitude and temperature. However, Figure 8C shows that certain geographically related aspects of temperature are not easily described by a linear function of location. Mirroring patterns of high KSE, the spatially dependent components show warmer than expected areas in Washington, Oregon, California, and along most of the Nevada
border. Cooler than expected regions occur mainly along the southern coastline of California and in northern Montana. The cooler California coast relative to location can be explained by the California Current which transports northern waters southward, cooling the ocean temperature near California and, consequently, its land temperature as well [21]. The resulting complete kriging estimates are found in Figure 8D. The linear combination of the mean structure and the spatial dependence provides predictions with anticipated patterns, the mild climate relative to latitude that exists for the Pacific states being most evident. Thus, the results of applying our methodology on this dataset are encouragingly explained by climate science.

5 | DISCUSSION

This paper develops methodology to estimate a continuous tuning parameter in the SME model, analogous to the range parameter. The methodology maintains the ability to model severe nonstationarity in an efficient manner, without loss of prediction accuracy, thus making it scalable to massive datasets. This is accomplished using the SME model and fixed rank formulas developed by Cressie and Johannesson [11] and the EM algorithm to obtain ML estimates of the parameters in the variance structure, suggested by Katzfuss and Cressie [30]. In addition to previous work estimating variance components associated with the knots and fine-scale variation ($K$ and $\sigma^2$), we estimate the range parameter in the basis functions by using the AECM algorithm. By making use of simple parallel processing options in MATLAB, we minimize the computational burden of additional parameter estimation.

The estimation of $b$ is valuable because knowledge of the range parameter in a spatial process can be of scientific interest. Our argument is that, in addition to knowledge gained about the spatial process, estimating $b$ can improve estimation of other parameters. In so doing, we infuse flexibility into the model that allows the basis functions to describe the optimal range of spatial dependence. This flexibility has even been shown to reduce MSPE under certain situations when compared to its fixed $S$ counterpart. In particular, the AECM approach shows improvement in parameter estimation and prediction performance when the signal-to-noise ratio is high (i.e., low $\sigma^2$) and the spatial range is low (i.e., true value of $b$ is less than 1.5). A low signal-to-noise ratio will result in a spatial process that is overwhelmed by noise and a true value of $b$ that is large implies that the spatial process is relatively smooth given the significant range and overlap between basis functions. In either case, the added model flexibility will be unnecessary and possibly detrimental. That said, there may be value in using this methodology for any new dataset with significant spatial dependence where the underlying mean structure is unknown. As was observed in the temperature application, when the relationship between the covariates and the response is over-simplified or the number of knot resolutions is sub-optimal, estimating $b$ can provide indication of the model misspecification. As a mixed effects model, the total variability in the data is described by a combination of the deterministic and stochastic components. Therefore, we hypothesize that in the absence of a fully specified mean structure, the extra variance can be captured by the extra flexibility in the variance components provided by estimating $b$. And, in any case, when the resulting estimate of $b$ is near 1.5, the performance of the model will be similar, if not better.

Although $K$ is completely flexible, outside of standard requirements for covariance matrices, the locations of knots and the choice of basis function impose certain assumptions about the spatial dependence. The nonstationarity of $K$ is unconstrained, however, the pattern of nonstationarity in the full covariance $\Sigma$ is dependent on knot location. While a uniform or nearly uniform space-filling design is common, scenarios with irregularly located observations may benefit from irregular knot location guided by optimal space-filling algorithms such as [17] or [3]. In this case, the scaling function, $\eta$, should be reconsidered since the minimum distance between knots may be an outlier, resulting in an overly sparse and poorly-performing $S$ matrix. A more representative summary statistic, such the mean or the median, may adequately mitigate this effect.

Also, in choosing the local bisquare basis function, we are assuming that the area of spatial dependence is circular, with diameter controlled by $b$. However, if an anisotropic pattern exists in the data, it would be straightforward to replace the local bisquare basis function with something more appropriate. The current restriction is that the direction and relative scale of the anisotropic behavior would need to be known or estimated outside of the AECM algorithm. Interesting directions for future work, thus, would include optimal knot location and an estimation procedure that can support multiple parameters in the basis functions. This would allow for more complex structures including, but not limited to, the estimation of the direction and scale of anisotropy and a varying range for each resolution ($b_i$). While added parameter estimation would add computational cost, a clear concern in large spatial fields, the impact may be dampened using parallel programs, borrowing ideas from, for instance, the expectation gathering maximization algorithm for large spatio-temporal data [8].

Also worth noting is that, while this methodology allows for more flexible covariance structure and can
mitigate the effects of model misspecification, it does not improve uncertainty quantification. A case study comparing efficient spatial prediction techniques found that while FRK may perform comparably in terms of prediction error, it lags behind the current state-of-the-art in terms of uncertainty quantification [19]. Since a significant contributor to the FRK methodology has recently switched to scaled Vecchia approximations to approximate Gaussian processes [31], it is possible that the FRK framework may simply be unsuitable for proper uncertainty quantification. Thus, we see that while we have addressed the issue of computationally practical estimation and prediction in massive fields under the SME model, several questions remain that are worthy of further attention.

ACKNOWLEDGMENTS

The authors sincerely thank the Editor, Associate Editor, and three reviewers whose helpful and insightful comments on an earlier version of this article greatly improved its content. We are also very grateful to Joseph Guinness, Brian Reich, and Jonathan Stallrich for their valuable input and suggestions. K.T. Pazdernik received funding from the Consortium for Nonproliferation Enabling Capabilities (CNEC), which is sponsored by the National Nuclear Security Administration (NNSA) Office of Nuclear Nonproliferation of the Department of Energy. R. Maitra acknowledges support from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project IOW03617. The content of this paper however is solely the responsibility of the authors and does not represent the official views of the NIFA or the USDA. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data are publicly available at http://www.image.ucar.edu/Data/US.monthly.met/.

ORCID

Karl Pazdernik https://orcid.org/0000-0003-1410-4006
Ranjan Maitra https://orcid.org/0000-0002-3515-8532

REFERENCES

1. M. Avriel and D. Wilde, Optimality proof for the symmetric Fibonacci search technique, Fibonacci Quart. 4(3) (1966), 265–269.
2. S. Banerjee, A. Gelfand, A. Finley, and H. Sang, Gaussian predictive process models for large spatial data sets, J. R. Stat. Soc. Ser. B 70(4) (2008), 825–848.
3. M. Bia and P. Van Kerm, Space-filling location selection, Stata J. 14(3) (2014), 605–622.
4. S. D. Billings, R. K. Beatson, and G. N. Newsam, Interpolation of geophysical data using continuous global surfaces, Geophysics 67(6) (2002), 1810–1822.
5. S. D. Billings, R. K. Beatson, and G. N. Newsam, Smooth fitting of geophysical data using continuous global surfaces, Geophysics 67(6) (2002), 1823–1834.
6. J. Bradley, N. Cressie, and T. Shi, A comparison of spatial predictors when datasets could be very large, Stat. Surv. 10 (2016), 100–131.
7. W. Chen and R. Maitra, Model-based clustering of regression time series data via APECM – An ECM algorithm sung to an even faster beat, Stat. Anal. Data Min. 4 (2011), 567–578.
8. W.-C. Chen, G. Ostroouchov, D. Pugmire, Prabhut, and M. Wehner, A parallel EM algorithm for model-based clustering applied to the exploration of large spatio-temporal data, Technometrics 55(4) (2013), 513–523.
9. Code by Richard A. Becker, All R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P. Minka, and A. Deckmyn. maps: Draw geographical maps. R package version 3.3.0, 2018, available at https://CRAN.R-project.org/package=maps.
10. N. Cressie, Statistics for spatial data, Revised ed., Wiley Series in Probability and Mathematical Statistics, New York, 1993.
11. N. Cressie and G. Johannesson, Fixed rank kriging for very large spatial data sets, J. R. Stat. Soc. B 70(1) (2008), 209–226.
12. T. Davis, Direct methods for sparse linear systems, Fundamentals of Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, 2006.
13. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood for incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B 39 (1977), 1–38.
14. A. Finley, H. Sang, S. Banerjee, and A. Gelfand, Improving the performance of predictive process modeling for large datasets, Comput. Stat. Data Anal. 53(8) (2008), 2873–2884.
15. M. Fuentes, Spectral methods for nonstationary spatial processes, Biometrika 89(1) (2002), 197–210.
16. R. Furrer, M. Genton, and D. Nychka, Covariance tapering for interpolation of large spatial datasets, J. Comput. Graph. Stat. 15(3) (2006), 502–523.
17. A. E. Gelfand, S. Banerjee, and A. Finley, Spatial design for knot selection in knot-based dimension reduction models, in Spatio-Temporal Design: Advances in efficient Data Acquisition, Wiley, Chichester, 2013, 142–169.
18. D. Harville, Matrix algebra from a statistician’s perspective, Springer, New York, 2008.
19. M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hpgg, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion, A case study competition among methods for analyzing large spatial data, J. Agric. Biol. Environ. Stat. 24(3) (2019), 398–425.
20. H. Henderson and S. Searle, On deriving the inverse of a sum of matrices, Soc. Ind. Appl. Math. Rev. 23(1) (1981), 53–60.
21. B. Hickey, The California current system – Hypotheses and facts, Prog. Oceanogr. 8(4) (1979), 191–279.
22. D. Higdon, A process-convolution approach to modelling temperatures in the North Atlantic Ocean, Environ. Ecol. Stat. 5(2) (1998), 173–190.
23. H. C. Huang, N. Cressie, and J. Gabrosek, Fast, resolution-consistent spatial prediction of global processes from satellite data, J. Comput. Graph. Stat. 11 (2002), 63–88.

24. G. Johannesson and N. Cressie, Variance–covariance modeling and estimation for multi-resolution spatial models, in geoENV IV–Geostatistics for Environmental Applications, X. Sanchez-Vila, J. Carrera, and J. Gomez-Hernandez, Eds., Kluwer, Dordrecht, 2004, 319–330.

25. G. Johannesson, N. Cressie, and H. C. Huang, Dynamic multi-resolution spatial models, Environ. Ecol. Stat. 14 (2007), 5–25.

26. C. Johns, D. Nychka, T. Kittel, and C. Daly, Infilling sparse records of spatial fields, J. Am. Stat. Assoc. 98(464) (2003), 796–806.

27. E. Kang, N. Cressie, and T. Shi, Using temporal variability to improve spatial mapping with application to satellite data, Can. J. Stat. 38 (2010), 271–289.

28. M. Katzfuss, A multi-resolution approximation for massive spatial datasets, J. Am. Stat. Assoc. 112(517) (2016), 201–214.

29. M. Katzfuss and N. Cressie. Maximum likelihood estimation of covariance parameters in the spatial-random-effects model. Proceedings of the 2009 Joint Statistical Meetings, American Statistical Association Alexandria, VA, pp. 3378–3390, 2009.

30. M. Katzfuss and N. Cressie, Spatio-temporal smoothing and em estimation for massive remote-sensing data sets, J. Time Ser. Anal. 32 (2011), 430–446.

31. M. Katzfuss, J. Guinness, and E. Lawrence. Scaled Vecchia approximation for fast computer-model emulation, 2020. arXiv preprint arXiv:2005.00386.

32. C. Kaufman, M. Schervish, and D. Nychka, Covariance tapering for likelihood-based estimation in large spatial data sets, J. Am. Stat. Assoc. 103(484) (2008), 1545–1555.

33. J. Kiefer, Sequential minimax search for a maximum, Proc. Am. Math. Soc. 4(3) (1953), 502–506.

34. S. Kullback and R. Leibler, On information and sufficiency, Ann. Math. Stat. 22(1) (1951), 79–86.

35. F. Lindgren, H. Rue, and J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, J. R. Stat. Soc. Ser. B: Stat. Methodol. 73(4) (2011), 423–498.

36. G. Matheron, Traite de Geostatistique Applique, Vol 1, Technip, Paris, 1962.

37. X. Meng and D. van Dyk, The EM algorithm – An old folk-song sung to a fast new tune, J. R. Stat. Soc. 59(3) (1997), 511–567.

38. P. Moran, Notes on continuous stochastic phenomena, Biometrika 37(1) (1950), 17–23.

39. L. Muu and N. Quy, A global optimization method for solving convex quadratic bilevel programming problems, J. Glob. Optim. 26(2) (2003), 199–219.

40. National Weather Service. Cooperative Observer Program (COOP), 2000.

41. D. Nychka, Spatial-process estimates as smoothers, in Smoothing and Regression: Approaches, Computation, and Application, M. G. Schimek, Ed., Wiley, New York, 2000, 393–424.

42. D. Nychka, B. Bailey, S. Ellner, P. Haaland, and M. O’Connell, FUNFITS: Data analysis and statistical tools for estimating functions, North Carolina State University, Raleigh, 1996.

43. D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain, A multiresolution Gaussian process model for the analysis of large spatial datasets, J. Comput. Graph. Stat. 24(2) (2015), 579–599.

44. D. Nychka, R. Furrer, J. Paige, and S. Sain, fields: Tools for spatial data. R package version 8.3-5, 2015, available at https://CRAN.R-project.org/package=fields.

45. D. Nychka, C. Wikle, and J. A. Royle, Multiresolution models for nonstationary spatial covariance functions, Stat. Model. 2(4) (2002), 315–331.

46. K. T. Pazdernik, R. Maitra, D. Nychka, and S. R. Sain, Reduced basis kriging for big spatial fields, Sankhyā 80(2) (2018), 280–300. https://doi.org/10.1007/s13171-018-0129-7

47. J. Quiñero Candela and C. Rasmussen, A unifying view of sparse approximate Gaussian process regression, J. Mach. Learn. Res. 6 (2005), 1939–1959.

48. R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2015. https://www.R-project.org/

49. M. Stein, Interpolation of spatial data, Springer, New York, 1999.

50. M. Stein, Limitations on low rank approximations for covariance matrices of spatial data, Spatial Stat. 8 (2014), 1–19.

51. M. Stein, Z. Chi, and L. Welty, Approximating likelihoods for large spatial data sets, J. R. Stat. Soc. Ser. B: Stat. Methodol. 66(2) (2004), 275–296.

52. The MathWorks, Inc, MATLAB and statistics toolbox release 2014a, The MathWorks, Inc., Natick, MA, 2014.

53. A. Vecchia, Estimation and model identification for continuous spatial processes, J. R. Stat. Soc. Ser. B: Methodol. 50 (1988), 297–312.

54. J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika 20A(1/2) (1928), 32–52.

55. K. Xu and C. K. Wikle, Estimation of parameterized spatio-temporal dynamic models, J. Stat. Plan. Inference 137 (2007), 567–588.

56. A. Zammit-Mangion and N. Cressie, Frk: An r package for spatial and spatio-temporal prediction with large datasets, 2017. arXiv preprint arXiv:1705.08105.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** K. Pazdernik, and R. Maitra, Estimating basis functions in massive fields under the spatial mixed effects model, Stat. Anal. Data Min.: ASA Data Sci. J. 14 (2021), 430–448. https://doi.org/10.1002/sam.11537