Interpretable, predictive spatio-temporal models via enhanced pairwise directions estimation

Heng-Hui Luea and Sheng Li Tzengb

aDepartment of Statistics, Tunghai University, Taichung, Taiwan; bDepartment of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan

ABSTRACT
This article concerns predictive modeling for spatio-temporal data as well as model interpretation using data information in space and time. We develop a novel approach based on supervised dimension reduction for such data in order to capture nonlinear mean structures without requiring a prespecified parametric model. In addition to prediction as a common interest, this approach emphasizes the exploration of geometric information from the data. The method of Pairwise Directions Estimation (PDE) is implemented in our approach as a data-driven function searching for spatial patterns and temporal trends. The benefit of using geometric information from the method of PDE is highlighted, which aids effectively in exploring data structures. We further enhance PDE, referring to it as PDE+, by incorporating kriging to estimate the random effects not explained in the mean functions. Our proposal can not only increase prediction accuracy but also improve the interpretation for modeling. Two simulation examples are conducted and comparisons are made with several existing methods. The results demonstrate that the proposed PDE+ method is very useful for exploring and interpreting the patterns and trends for spatio-temporal data. Illustrative applications to two real datasets are also presented.

1. Introduction

Complicated phenomena in spatio-temporal data raise large challenges remaining to be overcome even in today’s new era of computing. Kriging, as a method widely used in modeling such data, typically contains two components; namely, a stationary Gaussian process and a mean function. Interestingly, most kriging approaches assume a very simple structure for the mean function. For example, data at n locations over T time points are assumed to be observed according to

\[ y(s, t) = \mu(s, t) + u(s, t) + \varepsilon(s, t), \]

where \( \mu(s, t) \) is a mean function, \( u(s, t) \) is a zero-mean Gaussian process with a covariance function \( C(s - s^*, t - t^*) \equiv \text{cov}(u(s, t), u(s^*, t^*)) \), and \( \varepsilon(s, t) \) is an additive white-noise. Usual choice of \( \mu(s, t) \) is either a linear combination of ‘known’ covariates (i.e. universal
kriging), or a constant not varying over space or time (i.e. ordinary kriging). It results in most analyses focusing on the covariance functions.

In Ref. [22], Martin and Simpson showed that ordinary kriging could obtain a poor prediction under the presence of strong trends. To explore the important spatial structures for spatio-temporal data, the renowned empirical orthogonal function (EOF) analysis is a commonly used method, which is based on principal component analysis; see Refs. [7,9] for review. Letting

\[ z(s, t) = \mu(s, t) + u(s, t), \]

EOFs formulate the spatio-temporal data approximately in the summation form as

\[ z(s, t) \approx \sum_{k=1}^{K} PC_k(t) EOF_k(s), \]

(see Refs. [3,26]). This linear combination of inner products of temporal and spatial functions in EOFs has close connection to many dimension reduction approaches for geostatistics in the literature. From the reduced rank perspective, a low-rank model for spatial data at a particular time point is considered in a form similar to (2) as

\[ z(s) \approx \sum_{k=1}^{K} \alpha_k \beta_k(s), \]

where \( \alpha_k \)'s are unknown scalars and \( \beta_k(s) \) is either a fully known function or a parametric basis function depending on a few parameters (e.g. [2,6,29]). When several time points are involved, \( \alpha_k \) in (3) can be further assumed as a time-varying random variables \( \alpha_k(t) \) (see Refs. [8,10,27,28]). Some of these works assumed \( (\alpha_1(t), \ldots, \alpha_K(t))' \) to be independent and identically distributed (i.i.d.) over time, while others linked \( \alpha_k(t-1) \) and \( \alpha_k(t) \) for equal time spaces in an autoregressive way. The \texttt{SpTimer} package developed by Bakar and Sahu [1] utilizing the low-rank framework is implemented to hierarchical Bayesian modeling for space-time data. In contrast to stochastic \( \alpha_k(t) \)'s and parametric \( \beta_k(s) \)'s above, additive models provide a non-parametric perspective on (3). They used spline functions to represent \( \alpha_k(t) \) and/or \( \beta_k(s) \) (e.g. [16,24]). The \texttt{mgcv} package based on Ref. [30] is a convenient tool for additive models.

As a motivating example for illustrating the intriguing discovery, the data of average temperatures for each daily of the year at 35 weather stations in Canada, shown in Figure 3(A,B), are considered. From Figure 3(A), a concave down pattern for daily temperatures is clearly observed. It seems to imply that a strong geometric information structure in the data. Two immediate questions are raised: are there any important time-related shapes hidden in those curves? Does the variation of curves relate to spatial patterns across the locations of stations? These questions stimulated us to investigate whether the inner products of spatial and temporal functions aid in exploring and interpreting the data. Our primary aim of the study is to gain an understanding of how temperature cycles at stations vary with location and/or with time. In general, this aim is not easy to accomplish. The aforementioned dimension reduction approaches turn the problem more tractable through different simplifications. For example, the EOFs impose the orthogonality on estimated components, the low-rank methods assume certain parametric function forms, and
additive models use a large collection of smooth functions and rely on coefficient shrinkage to eliminate unneeded functions. In contrast, we relax those simplifications through an extension of supervised dimension reduction approach into the analysis without requiring pre-specified parametric functions.

Instead of directly taking \( z(s,t) \) in (1), we assume \( z(s,t) = \sum_{k=1}^{K} \alpha_k(t) \beta_k(s) + u(s,t) \) with a substantially different mean distinct from typical kriging. The Pairwise Directions Estimation (PDE; [20]) is incorporated in our proposed method to estimate the \( \sum_{k=1}^{K} \alpha_k(t) \beta_k(s) \) in a data-driven way. In this article, we focus on explicitly finding both \( \alpha_k(t) \) and \( \beta_k(s) \) in order to detect potentially sophisticated mean structures in the spatio-temporal data. Despite recent flourish of research on supervised dimension reduction (e.g. [4, 17, 20]), such time-relevant high-dimensional data remains to be challenging. We shall propose a novel approach to adaptively capture important spatial structures and temporal patterns for spatio-temporal data. Our main idea is to combine PDE and kriging, two quite flexible spatio-temporal modeling methods, into an ensemble learner, which not only increases the prediction accuracy of PDE, but also gives more comprehensive interpretation than typical kriging. The numerical results and real applications show that our proposed method not only gives explainable mean trends but may also produce more accurate model prediction.

The remainder of the paper is organized as follows. Section 2 introduces our approach. We use two simulation examples to illustrate and evaluate our proposed method in Section 3. Section 4 applies our method to two real datasets, and Section 5 concludes with discussion on future work.

2. Proposed method

Consider a sequence of processes, \( \{z(s,t) : s \in D\} \) for \( t = 1, \ldots, T \), defined on a \( d \)-dimensional spatial domain \( D \subset \mathbb{R}^d \) with \( T \geq 1 \). The processes are assumed to have a stationary spatio-temporal covariance function \( \text{cov}(z(s,t), z(s^*, t^*)) \). The spatio-temporal random effects model of \( y(s,t) \) with a measurement error \( \varepsilon \) is considered in this article as

\[
y(s,t) = z(s,t) + \varepsilon(s,t),
\]

\[
z(s,t) = w(t)f(s) + u(s,t),
\]

for \( t = 1, \ldots, T \), where \( f(s) = (f_1(s), \ldots, f_\kappa(s))' \), \( w(t) = (w_1(t), \ldots, w_\kappa(t)) \), \( u(s,t) \) is a zero-mean random effect with a covariance function \( C(s-s^*, t-t^*) = \text{cov}(u(s,t), u(s^*, t^*)) \) for any pair of \( (s,t) \) and \( (s^*, t^*) \), and \( \varepsilon(s,t) \sim N(0, \sigma_\varepsilon^2(t)) \) is an additive white-noise uncorrelated with \( z(s,t) \). We assume that \( u(\cdot, t) \) and \( \varepsilon(\cdot, t) \) are mutually uncorrelated. More specifically, \( f_j(s) \) stands for \( f_j(\theta_j'x(s)) \), where \( x(s) \) is a spatial-only covariate vector, and we can represent the inner-product term in (5) in terms of the summation as

\[
w(t)f(s) = \sum_{j=1}^{\kappa} w_j(t)f_j(s).
\]

Here \( w_j(t), j = 1, \ldots, \kappa \), is an unknown basis function changing gradually with time \( t \), and \( f_j(s) \) is an unknown coefficient function standing for a certain spatial pattern. The function \( f_j(s) \) depends on a \( p \times 1 \) vector \( x(s) \) through an unknown vector \( \theta_j \) of weights. We refer
to $\theta'x(s)$ as a variate. Hence, determining $f_j(s)$ amounts to finding both $f_j$ and $\theta_j$, which is a typical theme in supervised dimension reduction via constructing new variate $\theta'x(s)$ instead of using $x(s)$ directly.

For covariates $x(s)$ over $\mathcal{H}^d$, $(s_1, \ldots, s_d, s_1^2, \ldots, s_d^2)'$ are considered in our simulation examples and real data applications. We certainly can incorporate more elaborate and domain-specific covariates into $x(s)$, which may be useful for comprehending the spatio-temporal phenomenon.

Note that in the semi-parametric models of (4) and (5), only $x(s)$ and $y(s, t)$ are observed, but other quantities are kept unknown. The aim of this study is to predict the process $z(s, t)$ by reconstructing those unknown terms $w(t)$, $f(s)$ and $u(s, t)$ based on the observed data $x(s_i)$ and $y(s_i, t)$, $i = 1, \ldots, n; t = 1, \ldots, T$, as parsimonious as possible. To estimate those terms in (5), we begin with finding $w(t)$ and $f(s)$ via PDE proposed by Lue [20], and then estimate $u(s, t)$ by applying the kriging method for the spatio-temporal data (e.g. [7,25]). We briefly introduce the two key building blocks, PDE and kriging, and then propose our estimation method.

### 2.1. Pairwise directions estimation

Suppose that $y(s, t)$ at $n$ distinct locations, $s_1, \ldots, s_n \in \mathcal{D}$, for $t = 1, \ldots, T$ with a covariate vector $x(s)$ are available. The original setting for time point $t$ is not necessarily equal-spaced, but for simplicity, we assume them to be equal-spaced. For ease of illustration, we denote $y(\cdot, t) = (y(s_1, t), \ldots, y(s_n, t))'$, $y(s, \cdot) = (y(s, 1), \ldots, y(s, T))'$ and $Y = (y(\cdot, 1), \ldots, y(\cdot, T))$. The rationale behind using PDE to approximate the inner-product term in (5) is tentatively considering

$$\phi_j'y(s, \cdot) = g_j(\theta_j'x(s)) + \hat{c}, \quad 1 \leq j \leq \kappa,$$

which originally deals with multivariate response data via a "pairwise" model. After some algebra, (7) leads to the following model

$$y(s, t) = \sum_{j=1}^{\kappa} \phi_j(t)g_j(\theta_j'x(s)) + e(s, t),$$

for $t = 1, \ldots, T$. In order to describe $y(s, t)$, we need to estimate a $T \times 1$ vector $\phi_j$, a $p \times 1$ vector $\theta_j$ and a link function $g_j$, which all are totally unknown. Under (7), it is natural to pairwise find the solutions of $\phi_j$ and $\theta_j$ via minimizing $\sum_i [\phi_j'y(s_i, \cdot) - g_j(\theta_j'x(s_i))]^2$. The estimation strategy is to find $\theta_j$ first and then estimate $\phi_j$ and $g_j$ in an iterative way. We allow $\phi_j, g_j$ and $\theta_j$ to be determined by the data with the number $\kappa$ as small as possible.

The initial estimate $\hat{\theta}_j$ of $\theta_j$, $j = 1, \ldots, \kappa$, can be found by two methods for efficiently implementing the PDE method. They are mrSIR and pe-mrPHD methods (for more details, see Refs. [18,19]). After initialization of $\theta_j$, we utilize an adaptive estimation of MAVE in Ref. [31] for a single-index case to obtain an initial estimate of $\phi_j$ by solving the minimization problem:

$$\min_{d_{j1}, d_{j2}, \phi_j} \sum_{i=1}^{n} \sum_{t=1}^{n} \left( \theta_j'x(s_i) - (d_{j1} + d_{j2}\phi_j y(s_i, \cdot) - y(s_i, \cdot)) \right)^2 \delta_{i\ell},$$

where $\delta_{i\ell}$ is a weight.
where $\delta_{i\ell} = K_{hy} \{ \phi_j'(y(s_i, \cdot) - y(s_\ell, \cdot)) / \sum_{i=1}^n K_{hy} \{ \phi_j'(y(s_i, \cdot) - y(s_\ell, \cdot)) \}$, $K$ is a kernel function and $h_Y$ is the bandwidth of $Y$.

The iterative algorithm for estimating $\phi_j$, $\theta_j$ and $g_j$ for $j = 1, \ldots, \kappa$ in (8) via PDE proceeds in the following steps:

1. Initialize $\hat{\theta}_{j(0)}$, $j = 1, \ldots, \kappa$, from mrSIR and/or pe-mrPHD, and then get $\hat{\phi}_{j(0)}$ as the minimizer of $\phi_j$ in (9). Compute the estimated basis functions $\hat{\phi}_{j(0)} = \hat{\Sigma}_Y \hat{\phi}_{j(0)}$, where $\hat{\Sigma}_Y$ is the sample covariance of $Y$ (see Lemma of [20]).

2. At the $\tau$-th iteration, fit a linear regression model to $y(s_i, \cdot)$ against $\{ \hat{\phi}_{j(\tau-1)} \}_{j=1}^\kappa$, i.e. assuming

$$ y(s_i, \cdot) = g_1 \theta_1 + \cdots + g_k \theta_k + \epsilon(s_i, \cdot), $$

for $i = 1, \ldots, n$, to obtain the estimated coefficients $\hat{g}_{ij(\tau)}$, $j = 1, \ldots, \kappa$, where the $(\tau)$ in the subscript denotes the iteration number.

3. Obtain the updated estimate $\hat{\phi}_{j(\tau)}$ from (9) by replacing $\theta_j'x(s_i)$ with $\hat{g}_{ij(\tau)}$ and then compute the updated basis functions $\hat{\phi}_{j(\tau)} = \hat{\Sigma}_Y \hat{\phi}_{j(\tau)}$, $j = 1, \ldots, \kappa$.

4. Repeat steps 2 and 3 until $||\hat{\phi}_{j(\tau)} - \hat{\phi}_{j(\tau-1)}|| < \Delta, j = 1, \ldots, \kappa$, for some given tolerance value $\Delta$ (e.g. $\Delta = 0.001$).

5. Use the final estimate $\hat{\phi}_j$ in step 4 to obtain the final estimates $\hat{\theta}_j$ and $\hat{g}_j$, $j = 1, \ldots, \kappa$, by solving the minimization problem:

$$ \min_{d_{j_1}, d_{j_2}, \theta_j} \sum_{i=1}^n \sum_{\ell=1}^n \left( \phi_j'(y(s_i, \cdot)) - \{ d_{j_1} + d_{j_2} \theta_j' x(s_i) - x(s_\ell) \} \right)^2 \delta_{i\ell}, $$

where $\delta_{i\ell} = K_{hx} \{ \theta_j'(x(s_i) - x(s_\ell)) / \sum_{i=1}^n K_{hx}(\theta_j'(x(s_i) - x(s_\ell))$ and $h_x$ is the bandwidth of $x(s)$. We use $\hat{d}_{j_1}, \ell = 1, \ldots, n$, to be the estimate of $g_j(\cdot)$ in (8).

In step 2 of the algorithm, $\hat{g}_{ij} \equiv g_j(\theta_j'x(s_i))$, $j = 1, \ldots, \kappa$ obtained from the regression coefficients in (10), is a working estimate of $\hat{g}_j$ at $s_i$.

### 2.2. Function scaling estimation

The PDE captures the inner-product term, $w(t)f(s)$, of nonlinear mean structures in (5) without requiring a presupposed parametric model; namely, $\hat{\phi}_j$ and $\hat{g}_j$ found by PDE in step 5 are used to respectively estimate $w_j$ and $f_j$ in (6) for $j = 1, \ldots, \kappa$. In order to address the identifiability issue of functional estimation, we let the estimate $\hat{\omega}_j = (\hat{\omega}_j(1), \ldots, \hat{\omega}_j(T))'$ be a normalized vector of unit norm, and then a scaling factor $\hat{b}_j(s)$ is introduced for prediction purpose. Under (5), we use $\sum_{j=1}^{\kappa} \hat{\omega}_j(t) \hat{b}_j(s) \hat{f}_j(s)$ to approximate $w(t)f(s)$ via a linear fit. More specifically, $\hat{b}_j(s_i) = (\hat{b}_1(s_i), \ldots, \hat{b}_\kappa(s_i))'$ are the estimated coefficients obtained from the linear regression of $y(s_i, t)$ against $\{ \hat{f}_j(s_i) \hat{\omega}_j(t) \}_{j=1}^{\kappa}$ for each $i = 1, \ldots, n$. Then we refer to the $j$th scaled coefficient function $\hat{f}_j(s)$ as the estimates $(\hat{b}_j(s_1) \hat{f}_j(s_1), \ldots, \hat{b}_j(s_n) \hat{f}_j(s_n))'$, $j = 1, \ldots, \kappa$. Based on the scaling procedure, it would not only improve the prediction accuracy, but also give appropriate magnitude for the inner products. In calculating $\hat{\omega}(s_0, \cdot)$
at a predicted location \( s_0 \), the scaled coefficients \( \hat{b}_j(s_0) \), \( j = 1, \ldots, \kappa \), can be obtained by using the average of \( k \) nearest neighbors around \( \hat{\theta}_jX(s_0) \) in the observed data (e.g. \( k = 3 \)).

### 2.3. Incorporation of kriging

Since PDE does not take the random effects in (5) into account, we incorporate kriging to estimate this term for model prediction. The kriging method is dedicated to modeling the covariance function \( C(s - s^*, t - t^*) \equiv \text{cov}(u(s, t), u(s^*, t^*)) \) that measures the strength of dependency between any pairs of variables at \( (s, t) \) and \( (s^*, t^*) \). In (5), \( w(t) \) and \( f(s) \) are unknown deterministic functions. When measurement errors are absent with \( \epsilon(\cdot, \cdot) = 0 \), an equivalence exists between the variogram \( \nu(s - s^*, t - t^*) \equiv \text{var}((y(s, t) - y(s^*, t^*)) = E[(u(s, t) - u(s^*, t^*))^2] \) and the covariance function \( C(\cdot, \cdot) \) under the second-order stationary assumption of \( u(s, t) \), i.e. \( E(u(s, t)) \) is constant (assumed zero here) over \( s \) and \( t \), and \( \text{cov}(u(s, t), u(s^*, t^*)) \) only depends on \( s - s^* \) and \( t - t^* \); see Ref. [5] or [23]. Then the minimum-mean-squared-error linear prediction for \( u(s, t) \) at \( (s_0, t_0) \) of interest can be found by \( E[u(s_0, t_0)|Y] \) under the joint normality and assuming the parameters to be known; see Ref. [5].

We consider the isotropic product-sum model in this article with covariance in the form of

\[
C(h_s, h_t) = C_s(h_s; \xi_1) + C_t(h_t; \xi_2) + k_0 C_s(h_s; \xi_1)C_t(h_t; \xi_2)
\]

(11)

for any pair of space-time points \((s, t)\) and \((s + hs, t + ht)\) with valid covariance functions \( C_s(h_s; \xi_1) \) and \( C_t(h_t; \xi_2) \), respectively. We use exponential covariance functions for \( C_s \) and \( C_t \) in what follows, where both \( \xi_1 \) and \( \xi_2 \) contain two parameters for the marginal variance and the scale of \( C_s \) or \( C_t \), and \( k_0 \) is an additional parameter for the spatio-temporal interaction. When the first two terms on the right-hand side in (11) vanish, the covariance function becomes a separable model. Although other non-separable covariance functions can be applied in this framework, we mainly focus on (11) due to its ease of understanding and acceptable flexibility.

In the presence of measurement errors as (4), the variation of \( y(\cdot, \cdot) \) comes from both \( u(\cdot, \cdot) \) and \( \epsilon(\cdot, \cdot) \), and hence \( \nu(s - s^*, t - t^*) = E[(y(s, t) - y(s^*, t^*))^2] + E[(\epsilon(s, t) - \epsilon(s^*, t^*))^2] \). We shall instead consider a variogram model corresponding to the following covariance function between \((s, t)\) and \((s + hs, t + ht)\):

\[
C(h_s, h_t) + \sigma^2 \cdot \mathbb{I}(h_s = 0) \cdot \mathbb{I}(h_t = 0),
\]

where \( \mathbb{I} \) is the indicator function, and \( \sigma^2 = \text{var}(\epsilon(s, t)) \). For approximation of \( u(s, t) \), we need to calculate the empirical variogram, namely, the sample version of \( \nu(s - s^*, t - t^*) \), for the detrended process \( y(s, t) - \hat{w}(t)f(s) \). The estimation of parameters \( \sigma^2, k_0, \xi_1 \) and \( \xi_2 \) is based on least-squares variogram fitting, which is the default method implemented in the package \texttt{gstat}. After obtaining \( \hat{w}(t)f(s) \) by PDE in Section 2.1, we apply kriging to estimate \( u(s, t) \) based on the residuals \( y(s, t) - \hat{w}(t)f(s) \) through using

\[
E \left[ u(s_0, t_0) \mid \left\{ y(s_i, t) - \sum_{j=1}^k \hat{w}_j(t)f_j(s); i = 1, \ldots, n, t = 1, \ldots, T \right\} \right]
\]
with plugging-in estimated parameters under the joint normality assumptions of \(u(\cdot, \cdot)\) and \(\varepsilon(\cdot, \cdot)\) in (4) and (5). Here \(\hat{w}_j(t)\) and \(\hat{f}_j(s)\) are replaced respectively with \(\hat{\Phi}_j\) and \(\hat{g}_j\) found by PDE in its steps 4 and 5 for \(j = 1, \ldots, \kappa\). We perform a one-more-step iteration to obtain the final estimates for each component in (5), whose algorithm is given in Section 2.4.

**2.4. Enhanced PDE**

We propose a new enhanced method of PDE, called PDE+, through incorporating the PDE with kriging for modeling spatio-temporal data. In what follows, \(\varepsilon(s, \cdot), \hat{\varepsilon}(s, \cdot), u(s, \cdot), \) and \(\hat{u}(s, \cdot)\) are defined in a similar way as \(y(s, \cdot)\) with respect to \(\{y(s, t) ; \ t = 1, \ldots, T\}\). The algorithm for PDE+ proceeds in the following two steps:

(I) Apply PDE to the data \(\{x(s), y(s, t)\}\) for obtaining \(\hat{f}(s)\) and \(\hat{w}(t)\), and then get the residuals \(\varepsilon(s_i, \cdot)\) from a linear fit of \(y(s_i, \cdot)\) against \(\{\hat{w}_j\}_{j=1}^\kappa\) for \(i = 1, \ldots, n\). Use kriging in Section 2.3 on \(\varepsilon(s, \cdot)\) to get an initial estimate \(\hat{u}(s, \cdot)\) of \(u(s, \cdot)\).

(II) Let \(\hat{\varepsilon}(s, \cdot)\) be the difference between \(y(s, \cdot)\) and \(\hat{u}(s, \cdot)\) from step I, and then repeat step I once by treating the observed data as \(\{x(s), \hat{\varepsilon}(s, t)\}\).

The final predictive model of PDE+ is defined as the one in step II. The reason for taking once iteration on PDE+ is for computational efficiency. For prediction, we incorporate function scaling estimation procedure in Section 2.2 for measuring the prediction error based on the testing data. In this article, the dominant eigenvalue sequence and the scatterplots of \(g_j\) on \(\theta'x(s)\) are effectively used to determine the parameter \(\kappa\) in (7) in practice.

**3. Numeric results**

**3.1. Simulation setups**

We shall apply our proposed method for studying predictive modeling on spatio-temporal data through two simulated examples. Suppose that we observe data \(y(s, t)\) at location \(s\) and time \(t\). Let the spatial-only covariate vector be \(x(s) = (s_1, s_2, s_1^2, s_2^2)'\), where \(s = (s_1, s_2)'\), generated from given spatial locations and \(f_j(\cdot)\) be the coefficient function for unknown spatial structure shared across \(t\). Under (5), we think of elements of \(w(t)\) and \(f(s)\) as unknown deterministic functions. In all simulations and real applications, we have conducted the sample splitting procedure by removing some spatial locations and some time points (20% removal). To make comparisons with our proposal, we consider several existing methods. To evaluate the performance of prediction accuracy, we compute four indices, including two versions of prediction error criteria, namely root integrated mean squared error (RIMSE) and rooted prediction mean squared error (RPMSE), defined by

\[
\text{RIMSE} = \frac{1}{n_t} \sum_{i=1}^{n_t} ||y(s_i, \cdot) - \hat{z}(s_i, \cdot)||,
\]

\[
\text{RPMSE} = \left( \frac{1}{n_t T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} (y(s_i, t) - \hat{z}(s_i, t))^2 \right)^{1/2},
\]

(12)
where \( n_t \) is the size of locations in the testing data, \( \hat{z}(s_i, t) \) denotes the estimate of \( z(s_i, t) \), and \( \hat{z}(s_i, \cdot) = (\hat{z}(s_i, 1), \ldots, \hat{z}(s_i, T))' \). We also assess the model prediction based on continuous ranked probability score (CRPS) and interval score (Iscore) discussed in Ref. [12]. For a specific real-value \( y \), CRPS and Iscore are respectively defined as:

\[
\text{CRPS}(F, y) = -\int_{-\infty}^{\infty} (F(v) - \mathbb{I}(v \geq y))^2 \, dv, \tag{13}
\]

\[
\text{Iscore}(\alpha_0, y) = (I_U - I_L) + \frac{2}{\alpha_0} (I_L - y) \mathbb{I}(y < I_L) + \frac{2}{\alpha_0} (y - I_U) \mathbb{I}(y > I_U), \tag{14}
\]

where \( F(\cdot) \) is the cumulative distribution of a probabilistic forecasting model, \([I_L, I_U]\) is its \((1 - \alpha_0)\) prediction interval. We shall report the averaged CRPS and Iscore over all observed \( y \) values for \( \alpha_0 = 10\% \) in what follows. According to four indices above, the smaller the index levels, the higher is the prediction accuracy.

The cross-validation estimates of RIMSE, RPMSE, CRPS and Iscore are collected over 100 replications for each example. Seven other methods to be compared, referred to as naive, SpTimer, kriging-P, kriging-N, mgcv, OKFD and WMWC, are briefly described below.

The naive method takes the simple predictor \( \hat{z}(s, t) = \frac{1}{n_t} \sum_{i=1}^{n_t} y(s_i, t) \), i.e. the sample mean of all locations at time \( t \). The SpTimer method uses a hierarchical Bayesian approach for space-time data. Its model is given by

\[
y(s, t) = \alpha + R(s, t) + \omega(s, t),
\]

\[
R(s, t) = \rho R(s, t - 1) + \varsigma(s, t),
\]

where \( \varsigma(s, t) \)'s are zero-mean spatial random effects independent over time with an isotropic stationary Matérn covariance function. This method has been implemented in an R package named SpTimer. The kriging method typically considers the term \( w(t)f(s) \) in (5) as a constant \( \mu \). The ordinary kriging estimates \( \mu + u(s_0, t_0) \) by \( \sum_{i=1}^{n_t} \sum_{t=1}^{T} P_{it}(s_0, t_0)y(s_i, t) \), where \( \{P_{it}(s_0, t_0)\} \) depend on the variogram \( v(\cdot, \cdot) \) under the constraint of \( \sum_{i} \sum_{t} P_{it}(s_0, t_0) = 1 \), with plugging all estimated covariance parameters. We refer to this typical kriging as parametric kriging, or kriging-P for short. Instead of using parametric form, [32] recently proposed a non-parametric method to estimate the spatio-temporal mean function and covariance function. We refer to it as non-parametric kriging (kriging-N). If the assumptions of stationarity, constant mean, or isotropic structure lack to hold, such a non-parametric method expectedly performs better than kriging-P.

The mgcv method exploits additive models with tensor products. It assumes

\[
E[y(s, t)] = q_1(s_1) + q_2(s_2) + q_3(t) + q_4(s_1, s_2) + q_5(s_1, t) + q_6(s_2, t) + q_7(s_1, s_2, t), \tag{15}
\]

where \( q_\cdot \)'s are constructed by superimposing several known basis functions and its noises are assumed to follow \( N(0, \sigma^2) \) independently over space and time. With these higher-order interactions, the model tends to overfit, and hence some penalized approach is necessary. The famous R package mgcv provides convenient ways for determining an appropriate penalty.

There have been some research treating spatial-temporal data as a set of functional data, i.e. curves, at multiple locations, by specifying spatial correlations among
curves [11,13–15]. The OKFD method used in Refs. [11,14] estimates a mean curve \( \mu(s, t) = \mu(t) \) and a variance curve \( \text{var}(y(s, t)) = \sigma^2(t) \) for all \( s \), followed by modeling the \( \text{cov}(y(s, t), y(s^*, t)) \) based on a spatially stationary covariance model \( C_s(||s - s^*||, t) \). The OKFD was implemented in the R package \texttt{geoR} and \texttt{fd}. Besides, the literature [13] and [15] discussed an expansion of \( y(s, t) = \mu(t) + \sum_{i=1}^{K} \alpha_k(t) \beta_k(s) \), where the mean curve \( \mu(t) \) and the components of \( \alpha_k(t) \) and \( \beta_k(s) \) are estimated based on weighted averages of quantities from all observed locations, without assuming spatial stationarity. This method is referred to as weighted-mean-weighted-component (WMWC).

Moreover, the original PDE method is also included in the comparisons. We can think of PDE+ as an ensemble learner combining both PDE and kriging. The comparisons among PDE, kriging, and PDE+ aid in numerically verifying that the proposed integration leads to better prediction of the data.

According to the algorithms of the above methods, the default values are used to set the turning parameters. We emphasize the accuracy of prediction as the indices in (12)–(14) and the findings of interesting pattern structures. The strategy for determining the structural dimension \( \kappa \) in model (7) is based on the maximal eigenvalue ratio criterion (MERC) proposed by Luo et al. [21]. More specifically, they suggested to select \( \kappa \) by \( \hat{\kappa} = \arg \max_{i=1,\ldots,p-1} \frac{\hat{\lambda}_i}{\hat{\lambda}_{i+1}} \), where \( \hat{\lambda}_i \) is the estimated eigenvalue of the eigenvalue decomposition for mrSIR (or pe-mrPHD). For the two real data applications in the next section, both of the MERC are two. From simulations and practical applications in the next two sections, PDE successfully estimates the effective dimension reduction directions.

### 3.2. Comparison of simulation results

In each simulation run, we set \( T = 20 \) and generate \( y(s, \cdot) \) at \( n \) locations which consist of a learning set with size \( n_l (= 80\% \times n) \) and a testing set with size \( n_t (= 20\% \times n) \). We also exclude learning data at randomly sampled 4 (= 20\% \times T) time points for validation purposes. The location set \( \{s_1, \ldots, s_n\} \) is drawn from \( D = [-1, 1]^2 \) using simple random sampling. We set the number of slices for mrSIR to be 10 for all runs. We denote the absolute value of cosine angle between the estimate \( \hat{\theta}_j \) and the true vector \( \theta_j \) as \( |\cos(\hat{\theta}_j)| \), \( j = 1, \ldots, \kappa \), for evaluating the accuracy of directional estimation.

**Example 3.1:** Consider a process, \( \{z(s, t) : s \in [-1, 1]^2\} \), generated according to (5) with trigonometric and quadratic structures by setting

\[
\begin{align*}
f_1(s) &= \cos(0.5\pi||s - (-0.5, -0.5)'||^2), \\
f_2(s) &= \sin(0.5\pi||s - (0.5, 0.5)'||^2), \\
w_1(t) &= (0.5 t - 5)^2, \\
w_2(t) &= 5 \sin(0.1\pi t),
\end{align*}
\]

for \( t = 1, \ldots, T \), and \( u(s, t) \) being a zero-mean random effect with the covariance

\[
\text{cov}(u(s, t), u(s^*, t^*)) = \begin{cases} 
\exp(-0.5||s - s^*||) & \text{if } t = t^*; \\
0 & \text{otherwise}.
\end{cases}
\]
Using (4), the data \{y(s, t)\} with \(\varepsilon(s, t)\) i.i.d. \(N(0, 0.25)\) over time and space are generated. With letting \(\theta_1 = (0.5, 0.5, 0.5, 0.5)\)' and \(\theta_2 = (-0.5, -0.5, 0.5, 0.5)'\), the coefficient functions \(f_j\)'s can be reformulated as \(f_1(s) = \cos(0.5\pi(2\theta_1^T x(s) + 0.5))\) and \(f_2(s) = \sin(0.5\pi(2\theta_1^T x(s) + 0.5))\), where \(x(s) = (s_1, s_2, s_1^2, s_2^2)'\). Both unknown vectors, \(\theta_1\) and \(\theta_2\), are sufficient dimension reduction directions needed to be estimated.

To illustrate the application of PDE+ via data visualization, a single run is taken. The data with \(n = 100\) are generated according to models (16) and (17). We separately conduct the eigenvalue decomposition of mrSIR and pe-mrPHD for the learning data to obtain initial estimates \(\hat{\theta}_j\)'s. It turns out that the modified eigenvalues \((0.73, 0.31, 0, 0)\) found by mrSIR suggest one \(x(s)\) variate. One significant direction is also found by pe-mrPHD with the eigenvalues \((2.24, 1.83, 0.68, 0.29)\). By choosing \(h_Y = 3, h_x = 0.5\) and taking the two initial estimates \(\hat{\theta}_j, j = 1, 2\), we proceed with the proposed algorithm to dimension reduction for \(x(s_i)\) and \(y(s_i, t), i = 1, \ldots, n; t = 1, \ldots, T\). After attaining the iterative convergence, two leading directions, \((0.501, 0.506, 0.490, 0.502)'\) and \((-0.496, -0.496, 0.515, 0.492)'\), for \(x(s)\) variates are found along with \(|\cos(\hat{\theta}_1)| = 0.999\) and \(|\cos(\hat{\theta}_2)| = 0.999\), which are consistently close to the theoretical vectors \(\theta_j, j = 1, 2\). The scatterplot of the first estimated basis function, shown in Figure 1(A), reveals a noticeable quadratic pattern. Figure 1(B) shows a clear sine pattern for the second estimated basis function. Figure 1(C,D) displays coefficient functions which are very close to the true trigonometric patterns.

In reference to the sampling performance, the results based on 100 simulated replicates according to (16) and (17) are summarized in Table 1. It reports the mean and standard deviation for four indices obtained by all methods. PDE+ does outperform with the smallest averaged values of indices for this model and with the best prediction interval because of its capture of clear signal for functional patterns. For example, the mean of RIMSE obtained by PDE+ is about 5.18, which is about 27.6\% (= 6.61/5.18 - 1) improvement over kriging-P; while the improvement with respect to RPMSE is more significant about 31.8\% (= 1.74/1.32 - 1). Not surprisingly, PDE uses locally linear information via smoothing techniques to capture the function forms of \(w_j\) and \(f_j, j = 1, \ldots, k\), which make the rest of processes more well-suited for typical kriging models. Though the original PDE and kriging-P obtain similar results in this setup, the effectiveness of prediction is validated via combining the strengths of PDE and kriging into PDE+.

Both kriging-N and WMWC via removing a non-constant mean function prior to modeling the complicated covariance structure indeed pay some price in estimation. It is not appropriate for OKFD to ignore the correlation interactions between space and time in estimation. In contrast, the naive and SpTimer produce considerable error in prediction, however.

**Example 3.2:** Generate a process, \(\{z(s, t) : s \in [-1, 1]^2\}\), from (5) with setting

\[
\begin{align*}
  f_1(s) &= 15/\{-0.75 + \exp(||s - (-0.5, -0.5)'||^2)\}, \\
  f_2(s) &= 1.5(-2 + ||s - (0.5, 0.5)'||^2)^2, \\
  w_1(t) &= \arctan(0.1\pi t), \quad w_2(t) = 2\log(0.75 + (0.1t - 1)^2),
\end{align*}
\]
for $t = 1, \ldots, T$, and $u(s, t)$ being a zero-mean random effect with the covariance

$$
cov(u(s, t), u(s^*, t^*)) = 0.25 \exp(-0.5||s - s^*||) \exp(-0.8||t - t^*||)
+ \exp(-0.5||s - s^*||) + 0.5 \exp(-0.8||t - t^*||).
$$

We also generate $\varepsilon(s, t)$ i.i.d. $N(0, 0.5)$ over time and space. We need to estimate two sufficient dimension reduction vectors, $\theta_1 = (0.5, 0.5, 0.5, 0.5)'$ and $\theta_2 = (-0.5, -0.5, 0.5, 0.5)'$, as previously mentioned.

The data with $n = 150$ are generated according to models (18), (19) and (4) for a single run. Using the eigenvalue decomposition, the modified eigenvalues (1.50, 0.35, 0, 0) found by mrSIR suggest one $x(s)$ variate and another significant direction is also found by pe-mrPHD with the eigenvalues (2.50, 1.93, 0.65, 0.50). Two leading directions for $x(s)$ variates, $(0.520, 0.480, 0.530, 0.467)'$ and $(-0.510, -0.488, 0.517, 0.483)'$, are found by taking $h_Y = 10, h_x = 0.5$, after attaining the iterative convergence. The estimates along with $|\cos(\hat{\theta}_1)| = 0.998$ and $|\cos(\hat{\theta}_2)| = 0.999$ are consistently close to the theoretical vectors.
Table 1. Averaged values of four indices in the testing locations obtained by methods for data generated from models in (16) and (17) in 100 replicates (values given in parentheses are the corresponding standard errors).

| Method       | PDE+ | PDE  | naive | SpTimer | kriging-P | kriging-N | mgcv | OKFD | WMWC |
|--------------|------|------|-------|---------|-----------|-----------|------|------|------|
| RIMSE        | 5.18 | 6.91 | 37.17 | 32.74   | 6.61      | 13.30     | 6.93 | 8.39 | 13.45 |
| (1.21)       | (2.67)| (3.33)| (3.57) | (1.31)  | (2.03)    | (1.13)    | (8.50)| (4.28)|      |
| RPMSE        | 1.32 | 1.79 | 8.91  | 8.79    | 1.74      | 3.27      | 1.71 | 2.07 | 3.11 |
| (0.46)       | (0.85)| (0.63)| (0.83) | (0.44)  | (0.54)    | (0.40)    | (2.03)| (0.95)|      |
| CRPS         | 0.05 | 0.10 | 5.57  | 5.43    | 0.17      | 1.06      | 0.25 | 0.25 | 0.22 |
| (0.06)       | (0.12)| (0.57)| (0.83) | (0.38)  | (0.78)    | (0.37)    | (0.31)| (0.48)|      |
| Iscore       | 0.19 | 0.25 | 4.35  | 4.31    | 0.22      | 0.37      | 0.25 | 0.33 | 0.48 |
| (0.27)       | (0.37)| (5.34)| (5.30) | (0.45)  | (0.17)    | (0.38)    | (0.43)| (0.82)|      |

Figure 2. Plots of (A) the first and (B) the second basis functions with (C) the first and (D) the second scaled coefficient functions for a randomly selected simulation from model in (18) and (19).

\( \theta_j, j = 1, 2 \). The first two estimated basis functions, shown in Figure 2(A,B), reveal clear arctangent and quadratic patterns. Figure 2(C,D) displays estimated coefficient functions which are very close to the true views.

The results based on 100 simulation runs according to (18) and (19) are summarized in Table 2. As regards the predictive accuracy, PDE+ does outperform with the smallest
Table 2. Averaged values of four indices in the testing locations obtained by methods for data generated from models in (18) and (19) in 100 replicates (values given in parentheses are the corresponding standard errors).

|          | PDE+ | PDE naive | SpTimer | kriging-P | kriging-N | mgcv | OKFD | WMWC |
|----------|------|-----------|---------|-----------|-----------|------|------|------|
| RIMSE    | 5.83 | 6.43      | 65.38   | 61.05     | 10.44     | 17.57| 13.55| 10.44|
|          | (1.35)| (1.59)    | (8.66)  | (7.29)    | (5.05)    | (4.14)| (2.61)| (19.86)|
| RPMSE    | 1.65 | 1.81      | 17.53   | 16.35     | 3.57      | 5.85 | 4.12 | 2.85 |
|          | (0.53)| (0.63)    | (2.49)  | (2.43)    | (1.91)    | (1.67)| (0.78)| (5.24)|
| CRPS     | 0.36 | 0.41      | 11.5    | 10.41     | 0.91      | 2.11 | 0.45 | 0.45 |
|          | (0.30)| (0.44)    | (10.11) | (9.28)    | (1.66)    | (3.40)| (0.60)| (0.64)|
| Iscore   | 0.23 | 0.25      | 7.70    | 6.78      | 0.46      | 0.99 | 0.48 | 0.67 |
|          | (0.15)| (0.26)    | (10.39) | (9.52)    | (0.89)    | (2.75)| (0.44)| (0.52)|

averaged indices compared to other methods. The original PDE significantly outperforms kriging-P and kriging-N in this example. In contrast, other methods perform worse due to lacking the capture of the substantial mean structure. The mean of RIMSE obtained by PDE+ is about 5.83, which is about 79% (= 10.44/5.83 − 1) improvement over kriging-P; while the improvement is more significant about 116% (= 3.57/1.65 − 1) with respect to RPMSE, for example. In contrast, the naive and SpTimer produce considerable error in prediction, however.

4. Applications to real data

To illustrate the application of our approach to predictive modeling for empirical studies, we analyze two real datasets. In the following datasets, the longitudes and latitudes are available for given spatial locations. Additional information about elevation in the second dataset is usable. We set the respective spatial-only covariate vector to be \( x(s) = (s_1, s_2, s_1^2, s_2^2)' \) and \( x(s) = (s_1, s_2, s_1^2, s_2^2, m)' \), where \( s_1 \) is the longitude, \( s_2 \) is the latitude, and \( m \) is the elevation.

4.1. Canadian weather data

We applied the proposed method to a dataset of average temperatures over the years 1960–1994 for each day of the year at 35 weather stations in Canada. The data are available in the \texttt{fda} package on Comprehensive R Archive Network (CRAN). The observations \( \{y(s, t)\} \) of interest are the average of temperatures measured on 35 location sites for 365 days.

The annual temperature cycle at 35 stations shown in Figure 3(A) indicates a clear concave-down pattern. The map of observation locations is shown in Figure 3(B). For illustration, we first standardize \( x(s) \) for being centered and removing scale. We randomly partition \( \{y(s_i, \cdot); i = 1, \ldots, 35\} \) into a learning set with size 30 and a testing set with size 5 for further comparison analysis. We exclude learning data at randomly sampled 73 (\( = 20\% \times 365 \)) days for validation purpose. Details in one single analysis run are demonstrated. We conduct the eigenvalue decomposition for the learning data to obtain initial estimates \( \hat{\theta}'s \). Two significant directions are found by pe-mrPHD with the eigenvalues (2.71, 2.35, 1.20, 0.81). By choosing the bandwidths, \( h_Y = 3.5 \) and \( h_X = 1 \), we proceed with the proposed algorithm to dimension reduction for \( x(s_i) \) and \( y(s_i, t), i = 1, \ldots, 30; t = \)
Figure 3. Plots of (A) the average daily temperatures over individual locations, (B) the spatial distribution of locations, (C) the first and (D) the second basis functions, and (E) the first and (F) the second scaled coefficient functions for Canadian Weather data.
After convergence, two leading directions are \((-0.109, -0.950, 0.217, -0.198)\) and \((0.314, 0.110, 0.900, 0.281)\). It shows that there exist respective dominant effects, \textit{latitude} and \textit{squared longitude}, for the first two variates.

According to (5), both basis functions \(w_j(t)\) and coefficient functions \(f_j(s)\) affect \(z(s, t)\) through the inner product form. Figure 3(C,D) displays the estimated basis functions which reveal similar concave patterns. Figure 3(C) displays that \(\hat{w}_1\) is positive and has much higher values at the left and right sides along the horizontal axis (Day) (indicating a winter season), but lower values in the middle (indicating a summer season). The levels of function \(\hat{f}_1(s)\) over spatial locations \(s\) shown in Figure 3(E) reveal a decreasing trend as the latitude increases. This figure also shows a large portion of locations with negative values for \(\hat{f}_1(s)\). As a result, the influence of the first dominant component \(\hat{w}_1(t)\hat{f}_1(s)\) just goes consistently with the common sense that the larger latitude is, the lower temperature is in winter; the lower latitude is, the higher temperature is in summer. The shape of Figure 3(D) roughly contrasts the temperature difference between summer and winter, due to the two seasons with the opposite signs of the function value for \(\hat{w}_2\). Figure 3(F) displays the levels of \(\hat{f}_2(s)\) over locations, indicating the maximum of the absolute value of \(\hat{f}_2(s)\) occurred nearly around the continental areas. Note that \(\hat{f}_2\) is negative for all areas. Figure 3(F) indicates that the bright-red (continental) regions have greater temperature difference compared to the faint-red (coastal) regions. Consequently, the secondary component \(\hat{w}_2(t)\hat{f}_2(s)\) implies that continental areas tend to have lower temperatures in winter, and higher temperatures in summer compared to the coastal regions. The colors in Figure 3(A,B) are based on the corresponding temperature difference. The pattern in Figure 3(F) can be also found in Figure 3(B). More specifically, Figure 3(A) implies the temperature differences are greater in brown curves (continental areas) than in pink curves (coastal areas).

In reference to the sampling performance, the results based on 100 sets of random data partitions, as previously mentioned are summarized in Table 3. With including a three-term interaction in (15) (i.e. the \(q_7\) term), \textit{mgcv} produces a considerable error with the respective mean values of 2487.9, 235.8, 7.3, and 14.7 for RIMSE, RPMSE, CRPS and Iscore. The larger values of indices found by \textit{mgcv} indicate its heavy overfitting. It should be aware of the constraint impact on model complexity, even though \textit{mgcv} provides its automatic penalty determination. The results for \textit{mgcv} shown in Table 3 are obtained through excluding the \(q_7\) term. Interestingly, \textit{PDE} captures the sharp functional patterns for \(w_j(t)\) and \(f_j(s)\), \(j = 1, 2\), useful for improving model prediction. Not unexpectedly, \textit{PDE+} does outperform with the smallest averaged indices through integrating \textit{PDE} for the dominant mean structure with kriging for estimating the random effect. From the outputs of Table 3 and Figure 3, \textit{PDE+} indeed shows an interpretable predictive model in addition to improving the accuracy of model prediction.

### 4.2. German PM\(_{10}\) data

We applied the proposed method to a dataset of air quality obtained from European Commission's Airbase and air quality e-reporting repositories (AQER). The monthly averages of PM\(_{10}\) concentrations within Germany between January 2016 and June 2018 are considered. The package \texttt{saggetr} on CRAN was used to extract hourly PM\(_{10}\) averages, and
Table 3. Averaged values of four indices in the testing locations obtained by methods for Canadian Weatherdata in 100 validations (values given in parentheses are the corresponding standard errors).

| Method       | RIMSE   | RPMSE  | CRPS   | Iscore  |
|--------------|---------|--------|--------|---------|
| PDE+         | 50.21   | 3.01   | 0.72   | 0.47    |
| PDE naive    | 101.93  | 5.79   | 2.24   | 1.11    |
| SpTimer      | 112.67  | 6.68   | 2.73   | 1.78    |
| kriging-P    | 96.64   | 6.47   | 2.61   | 1.72    |
| kriging-N    | 55.58   | 3.32   | 0.76   | 0.50    |
| mgcv         | 54.72   | 3.14   | 1.11   | 0.47    |
| OKFD         | 174.77  | 1.76   | 1.84   | 1.95    |
| WMWC         | 238.75  | 7.06   | 10.16  | 2.72    |

an aggregation to obtain monthly averages was performed at each of background stations. Totally 221 stations have complete monthly records over the 30 months considered. The monthly average concentrations for 221 stations shown in Figure 4(A) indicate several peaks occurring particularly at the beginning of years.

In addition to location information, this dataset also contains elevation values. We will demonstrate that spatial covariates show two dominate effects in the PDE+ model. The map of station locations is shown in Figure 4(B), where the color gradients represent the average concentrations of PM$_{10}$ values over the 30 months of individual sites. For illustration, we first standardize $x(s)$ for being centered and removing scale for further analysis. We randomly partition $\{y(s_i, \cdot); i = 1, \ldots, 221\}$ into a learning data with size 177 and a testing data with size 44. We also exclude learning data at randomly sampled 6 $(= 20\% \times 30)$ months for validation purpose.

In one single analysis run, the eigenvalue decomposition is conducted for the learning data to obtain initial estimates $\hat{\theta}_j$’s. Two significant directions are found by pemrPHD with the eigenvalues $(0.68, 0.60, 0.21, 0.17, 0.07)$. After the convergence of PDE algorithm with choosing the bandwidths $h_Y = 6$ and $h_X = 1$, two leading directions, $(0.28, 0.17, 0.21, 0.08, -0.91)'$ and $(0.71, 0.60, 0.06, -0.08, 0.35)'$, for $x(s)$ variates are obtained. It reveals that there exist respective dominant effects; namely, elevation (with negative association) for the first variate, and a linear combination of longitude and latitude (with approximately equal weight) for the second variate.

Recall that both $w_j(t)$ and $f_j(s)$ affect $z(s, t)$ through the inner product form. Figure 4(C) displays the first estimated basis function with a conceivable ‘W’ pattern reflecting the peaks of the data. It indicates that German PM$_{10}$ concentrations are higher in winter than in summer. Figure 4(E) shows a linearly increasing trend of $\hat{f}_1$ associated with a large negative elevation effect in $\hat{\theta}_1$. As a result, the influence of the first dominant component $\hat{w}_1(t)\hat{f}_1(s)$ shows that the larger (smaller) the elevation levels, the smaller (larger) is the value of PM$_{10}$. Figure 4(D) displays a near equal-partitioned pattern fluctuating around zero in the vertical axis. Figure 4(F) exhibits a linearly increasing pattern along $\hat{\theta}_2'x(s)$, roughly standing for a direction from northeast to southwest, due to the equal-weight effect of longitude and latitude in $\hat{\theta}_2$. Consequently, the influence of the second component $\hat{w}_2(t)\hat{f}_2(s)$ illustrates that the air pollution of PM$_{10}$ in Germany widely spreads over in the areas from northeast to southwest, similar to the colored pattern shown in Figure 4(B). Compared to Figure 4(E), the second component may serve as the complement part of influence on PM$_{10}$ due to the magnitudes.
Figure 4. Plots of (A) the monthly averages of PM$_{10}$ concentrations over individual locations, (B) the spatial distribution of locations, (C) the first and (D) the second basis functions, and (E) the first and (F) the second scaled coefficient functions for German PM$_{10}$ data.
Table 4. Averaged values of four indices in the testing locations obtained by methods for German PM$_{10}$ data in 100 validations (values given in parentheses are the corresponding standard errors).

|          | PDE+ | PDE naive | SpTimer | kriging-P | kriging-N | mgcv | OKFD | WMWC |
|----------|------|-----------|---------|-----------|-----------|------|------|------|
| RIMSE    | 17.28| 27.64     | 19.17   | 18.32     | 17.68     | 24.78| 18.95| 90.19| 90.79|
| (1.98)   | (1.95)| (1.19)    | (0.98)  | (3.49)    | (1.40)    | (1.07)| (2.79)| (9.90) |
| RPMSE    | 3.29 | 5.12      | 3.68    | 3.47      | 3.43      | 4.68 | 3.56 | 16.71| 16.84|
| (0.38)   | (0.37)| (0.21)    | (0.24)  | (0.7)     | (0.25)    | (0.20)| (0.49)| (1.78) |
| CRPS     | 0.86 | 1.66      | 1.22    | 1.18      | 1.01      | 2.46 | 1.18 | 14.51| 13.32|
| (0.55)   | (0.7) | (1.31)    | (1.24)  | (0.88)    | (2.67)    | (0.99)| (2.68)| (5.32) |
| Iscore   | 0.48 | 0.76      | 0.65    | 0.57      | 0.49      | 1.52 | 0.55 | 11.26| 10.49|
| (0.29)   | (0.39)| (1.13)    | (1.09)  | (0.55)    | (2.38)    | (0.66)| (2.91)| (5.34) |

The results based on 100 random data partitions for sampling performance are summarized in Table 4. To make a compatible comparison, kriging-P uses universal kriging with elevation being the covariate in the mean part. For methods without a natural way to include the covariates (e.g. OKFD and naive), we simply subtract the mean function obtained by a linear regression of $y(s, t)$ on elevation from $y(s, t)$. Again, PDE indeed captures the interesting functional patterns for $w_j(t)$ and $f_j(s)$, $j = 1, 2$. In the aspect of predictive accuracy, PDE+ does outperform with the smallest averaged indices, as shown in Table 4. More importantly, PDE+ can greatly help the explanation of the spatio-temporal trends, which is difficult for all other methods.

5. Discussion

Exploring the mean structure for spatio-temporal data via dimension reduction is studied in this article. Effective dimension reduction is a much more challenging statistical issue on spatio-temporal data, which was less paid attention in the past decades. We focus on dimension reduction for the spatial covariates $x(s)$ to increase the interpretability of the predictive model. Searching the optimum of possible linear combinations for $x(s)$ is crucial for solving such a problem. Another important issue is to discover the connection of linear or nonlinear association with the spatio-temporal variables of interest. The PDE+ can successfully achieve the above goals. In simulations, the performance of PDE is comparable to kriging, while in real applications, kriging works better than PDE. Overall, PDE+, combining PDE and kriging, outperforms kriging in all situations. The proposed method based on dimension reduction effectively demonstrates its interpretable and accurate prediction from all the numerical results.

In contrast, few methods utilize a data-driven approach for providing the estimation of mean structures in the literature. Two exceptions investigated in the numerical examples are kriging-N and WMWC, which may trade stability off for flexibility. The number of unknown parameters in the data-driven estimation of $w_j(t)f_j(s)$ in (6) is relatively small compared to the available $n \times T$ data points. For example, $w_j(t)$ and $\theta_j$ in the $j$th component have respective $T$ and $p$ parameters, while $f_j(s) = f_j(\theta'_j x(s))$ is merely a univariate smoother for the variate $\theta'_j x(s_i)$. Besides, the number of parameters in $C(s - s^*, t - t^*)$ for $u(s, t)$ is usually less than a dozen. Using the smaller number of parameters in PDE+ not only prevent the modeling from overfitting, but also allow it to be more easily interpretable. Hence, we can think of PDE+ as a parsimonious ensemble learner with great prediction ability as typical ensemble methods, but not being a hard-to-understand black box.
Acknowledgments

The authors thank Professor Hsin-Cheng Huang for useful comments greatly improving the article. We also would like to thank the associated editor and the two anonymous reviewers for their helpful and thorough suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Lue’s and Tzeng’s research works were supported in part by grants from the Ministry of Science and Technology of Taiwan [grant numbers MOST107-2118-M-029-001 and MOST 107-2118-M-110-004-MY3], respectively.

References

[1] K.S. Bakar and S.K. Sahu, sptimer: spatio-temporal bayesian modelling using R, J. Stat. Softw. 63 (2015), pp. 1–32.
[2] S. Banerjee, A.E. Gelfand, A.O. Finley, and H. Sang, Gaussian predictive process models for large spatial data sets, J. R. Stat. Soc. Series B. Stat. Methodol. 70 (2008), pp. 825–848.
[3] I. Braud and C. Obled, On the use of empirical orthogonal function (EOF) analysis in the simulation of random fields, Stoch. Hydrol. Hydraul. 5 (1991), pp. 125–134.
[4] R. Coudret, S. Girard, and J. Saracco, A new sliced inverse regression method for multivariate response, Comput. Stat. Data. Anal. 77 (2014), pp. 285–299.
[5] N. Cressie, Statistics for Spatial Data, John Wiley & Sons, Hoboken, 2015.
[6] N. Cressie and G. Johannesson, Fixed rank kriging for very large spatial data sets, J. R. Stat. Soc. Series B. Stat. Methodol. 70 (2008), pp. 209–226.
[7] N. Cressie and C.K. Wikle, Statistics for Spatio-Temporal Data, John Wiley & Sons, Boca Raton, 2011.
[8] N. Cressie, T. Shi, and E.L. Kang, Fixed rank filtering for spatio-temporal data, J. Comput. Graph. Stat. 19 (2010), pp. 724–745.
[9] U. Demšar, P. Harris, C. Brunsdon, A.S. Fotheringham, and S. McLoone, Principal component analysis on spatial data: an overview, Ann. Assoc. Am. Geogr. 103 (2013), pp. 106–128.
[10] A. Fassò and M. Cameletti, A unified statistical approach for simulation, modeling, analysis and mapping of environmental data, Simulation 86 (2010), pp. 139–153.
[11] R. Giraldo, P. Delicado, and J. Mateu, geofd: Spatial Prediction for Function Value Data (2020). Available at https://CRAN.R-project.org/package=geofd, R package version 2.0.
[12] T. Gneiting and A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. 102 (2007), pp. 359–378.
[13] O. Gromenko, P. Kokoszka, L. Zhu, and J. Sojka, Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends, Ann. Appl. Stat. 6 (2012), pp. 669–696.
[14] R. Ignaccolo, J. Mateu, and R. Giraldo, Kriging with external drift for functional data for air quality monitoring, Stoch. Environ. Res. Risk. Assess. 28 (2014), pp. 1171–1186.
[15] P. Kokoszka and M. Reimherr, Some recent developments in inference for geostatistical functional data, Rev. Colomb. Estad. 42 (2019), pp. 101–122.
[16] D.J. Lee and M. Durbán, P-spline anova-type interaction models for spatio-temporal smoothing, Stat. Model. 11 (2011), pp. 49–69.
[17] K.C. Li, Y. Aragon, K. Shedden, and C. Thomas Agnan, Dimension reduction for multivariate response data, J. Am. Stat. Assoc. 98 (2003), pp. 99–109.
[18] H.H. Lue, Sliced inverse regression for multivariate response regression, J. Stat. Plan. Inference 139 (2009), pp. 2656–2664.
Appendix

The codes for implementing the PDE+ method of this article are available at https://github.com/hhlue/PDEplus. We also illustrate how to extract the components of a PDE+ model, and how to predict based on PDE+, through a simulation in Example 3.1. In addition, the datasets of monthly German PM$_{10}$ concentrations in Section 4.2 can also be found from the same webpage.