A smoothing spline approach for integrating spatially misaligned water quality data

by

Robert M. Sholl

B.S., Kansas State University, 2021

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Trevor J. Hefley

Copyright

Abstract

Contamination is a looming threat to U.S. water resources, but the inconsistency of data collection across studies and agencies has made it difficult to identify effective treatments or resolutions. In many areas, the available water quality datasets are compatible with standard regression analysis; however, they cannot be integrated with other data sources to directly investigate those mechanisms. Here we develop an approach for connecting data with mismatched spatial resolutions and leverage it to integrate decades of Kansas ambient water monitoring with best management practice data. Our method builds upon the existing krige-and-regress framework by fitting a model of similar structure to spatio-temporal kriging; cubic splines fit via gamma generalized additive models to adjust for the non-negative and nonlinear characteristics present in ambient monitoring data. Applying this model to Kansas monitoring and best management practice data, we identify several candidate practices associated with nutrient reduction. These findings demonstrate the potential of spatially connected water data in advancing understanding of nonpoint source pollution and informing targeted best management practices at the subwatershed level.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This project would have been possible without the guidance of Dr. Trevor Hefley, who served as my major professor and mentored me through my transition from the biological sciences to statistics. I would also like to thank my committee members, Dr. Josefina Lacasa and Dr. Xuan Xu; their advice, instruction, and feedback provided a persistent push forward for both my development as a statistician and my research. This project is the result of a collaboration between the Kansas Water Institute, the Kansas Department of Agriculture, and Statistical Analysis System. Beyond funding and data, those organizations gave their best to assist me: Dr. Susan Metzger, Julie Espy, and David Jones, who's joint expertise in environmental sciences and governance developed much of the theoretical backbone for this project. I would like to extend a special thanks to Elliot Inman, Jim Box, and Scott McClain, for making significant contributions to the ideation of the best management practice effect analysis.

I owe an unbelievable debt of gratitude to my colleague Aidan Kearns, who has contributed more to my education in statistics than any other individual. As well, I would like to thank to the entire Hefley lab team, Yoshimitsu Nagaoka, Abraham Arbelaez, and Jacob Spooner, who have each given everything from insights to laughter. I extend my thanks to everyone else in the department of statistics, staff, students, and faculty, as well as all of my students from STAT 225, 240, and 341; every day has been a great one because of the people I see whenever I walk into Dickens hall.

I am ever grateful to my friends for their support, my mother and father for their endless encouragement, and my wife, Sarah, who inspired me to pursue this education in the first place and continues to inspire me to better myself every day.

Introduction

Nutrient and sediment contamination are persistent threats to water quality across the U.S. Great Plains[1 2 3]. Much of this contamination originates from nonpoint source (NPS) pollution, chemical and physical pollutants transported from agricultural fields into surface waters through runoff and erosion[1]. Best management practices (BMPs) are interventions designed to reduce NPS pollution and control erosion; however, the direct effect of individual BMPs on NPS pollution loads is unclear[4]. The Kansas Department of Health and Environment (KDHE) has amassed 58 years (1967–2025) of ambient water monitoring data, which could uncover relationships between BMPs and NPS pollution if paired with the Kansas Department of Agriculture's (KDA) best management practice implementation records. Despite this opportunity, the two datasets cannot be directly integrated, because they were collected at incompatible spatial resolutions[5], with the ambient monitoring occurring at individual station locations and BMP implementation collected at the subwatershed level.

This difference in spatial resolution is known as spatial misalignment and is well documented in spatial statistics as part of the change-of-support problem[6]. To connect these misaligned data sets, they must be harmonized to a common spatial support. This is achieved by downscaling areal support to point support, upscaling point support to areal support, or interpolating areal data across mismatched boundaries[7]. The longevity of the Kansas ambient water data facilitates spatial interpolation of point observations and subsequent upscaling to the subwatershed level. The standard method for spatial interpolation of point data is kriging, which is the best linear unbiased predictor (BLUP) for unobserved spatial locations under the assumption of stationarity[6]. The predicted values (referred to as the kriged observations) can be aggregated to areal support and used in regression analysis, a process called krige-and-regress[8].

Under the assumptions of the general linear model, krige-and-regress provides lower bias coefficient estimation than the maximum likelihood estimator for spatially misaligned random variables[8]. The Kansas ambient monitoring data exhibits characteristics of non-normal and non-linear processes — violations of the assumptions behind the general linear model that existing krige-and-regress method cannot accommodate[8]. Beyond these constraints, kriging is computationally prohibitive given the sample size of the Kansas monitoring records, necessitating an approximation capable of handling non-Gaussian and nonlinear spatial data[7]. To address these limitations, we develop a framework based on cubic smoothing splines, fit via generalized additive models for distributional flexitivity, as a computationally efficient method of spatial interpolation similar in structure to spatio-temporal kriging. In the next section, we establish ontology and review existing methods which will become the foundation of our framework for our model we call spline-based psuedo-kriging. With this model as a substitute for kriging, we proceed with the krige-and-regress procedure to identify the most likely best management practices associated with total maximum daily load (TMDL) station delistings across the state of Kansas.

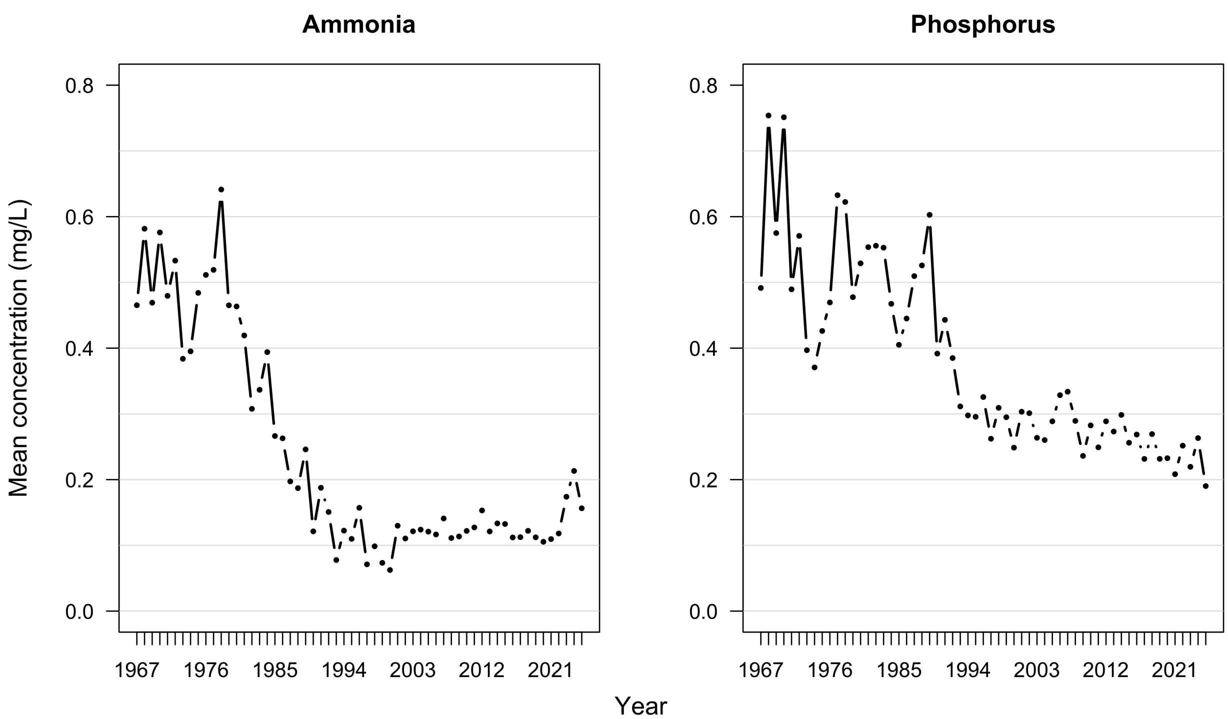


Figure 1.1: Ammonia (left) and phosphorous (right) are two of the many analytes measured through the Kansas Department of Health and Environment's ambient monitoring database that present characteristics of non-linear and non-normal data generating processes.

Methods

2.1 Data

The Kansas Department of Health and Environment (KDHE) has operated a state-wide ambient water monitoring program since 1967. The longevity and information density of this data is startling as it has been carefully maintained by KDHE throughout the programs existence. Wherever a monitoring station appears in the database one can see an image of over half a century of water quality trends, trials, and tribulations — although not always a clear picture due to the messy nature of highly tenured data collection programs. Additionally, the Kansas Department of Agriculture (KDA), Division of Conservation operates cost-share program aimed at natural resource protection, erosion control, and nonpoint source (NPS) pollution reduction. Since 2004, KDA has maintained records of all cost-shared BMP implementations within the Cost Share Information Management System (CSIMS). The database was developed with the intent to be used in detailed analyses for both state and academic applications, resulting in 1,404 unique covariates being recorded across 47,958 observations. Despite many of these covariates being tracked for brief one or two year periods, the mass of information within CSIMS presents numerous opportunities for post-hoc analyses of BMPs. KDA, KDHE, and the Kansas Water Institute made the decision in early 2025 to analyze these two datasets side-by-side in hopes to uncover the possible magnitude and direction of BMP effects on NPS pollution.

The ambient data was restricted to 16 analytes of concern for water quality and reduced to 538,813 complete observations after removal of erroneous entries (Figure 2.1). Yearly sampling frequency fluctuates substantially across the full monitoring record but stabilizes between 2004 and 2024, the period of record for the best management practice data and thus the interval selected for analysis (Figure 2.1). The spatial distribution of the samples is concentrated in eastern Kansas, following the major Kansas stream network (Figure 2.2). Gaps in spatial sampling east of the 98^{th} parallel are consistent with incomplete spatial coverage and may be considered missing completely at random (Figure 2.3). Missing samples west of the parallel appear associated with discontinuities in the stream network and potential resource constraints on the sampling effort; we treat these as missing at random and missing not at random respectively (Figure 2.4).

The data from CSIMS was extracted on June 13, 2025, and contains 21,044 complete observations after screening for missing values, with 108 unique BMPs recorded state-wide. The cleaning and manual validation of CSIMS is an ongoing effort that is outside the scope of this analysis, leading us to focus on addressing direct barriers to modeling. Some of the practices are duplicates with differing codes due to changes in naming convention or one-off implementations with similar impact to more frequently observed practices, creating instances of perfect collinearity and 0 degrees of freedom. To resolve these issues, many of these practices were consolidated in consultation with subject matter experts. CSIMS is recorded at the subwatershed level where practice implementation occurred, while the ambient monitoring stations are tracked as discrete points with no specified areal support. The temporal range of CSIMS spans from 2004 through June of 2025, slightly exceeding the upper bound of the ambient monitoring record, but as CSIMS is the target spatial support, this presents no temporal challenges. Given that each monitoring station falls within a subwatershed, but not every subwatershed contains a station, and spatial missingness is present throughout the study region, we require a method that addresses both spatial misalignment and interpolation.

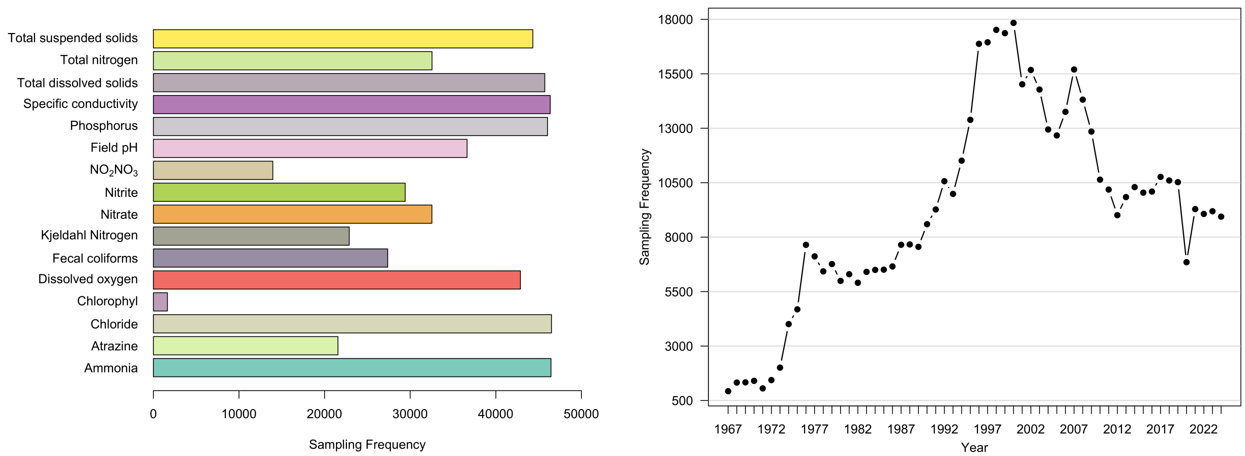


Figure 2.1: (Left) Sampling frequency by analyte. (Right) Sampling frequency by year.

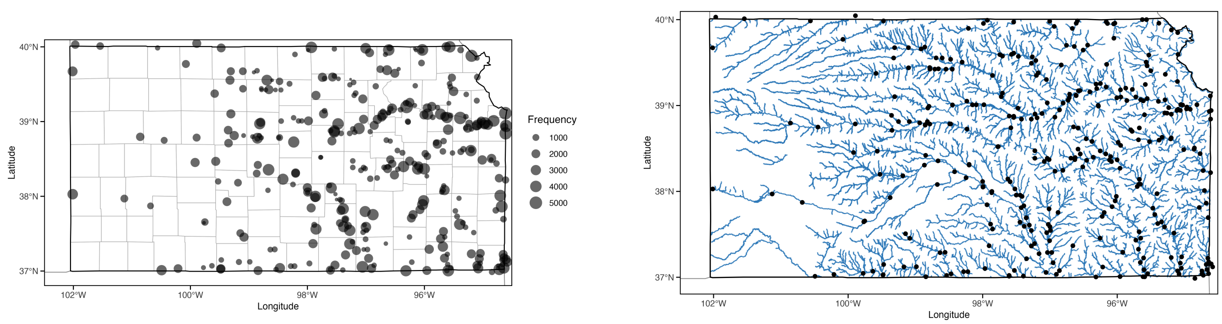


Figure 2.2: (Left) The spatial frequency distribution of ambient monitoring samples across the full study duration (1967–2025). (Right) The distribution of monitoring stations closely follows the structure of the Kansas major stream network.

2.2 Kriging

The method of kriging produces optimal predictions for a given random variable at unobserved locations by estimating the spatial covariance structure for observations within the same domain. Generally, kriging assumes spatial observations arise from a constant mean Gaussian process and uses a variogram to estimate the covariance. Kriging is, at its core, the BLUP for any given spatial covariance structure under the assumption of stationarity and thus exactly comparable to any other method producing the same estimate[6]. This characteristic has led to many discussions, simulations, and proofs of Gaussian process regression

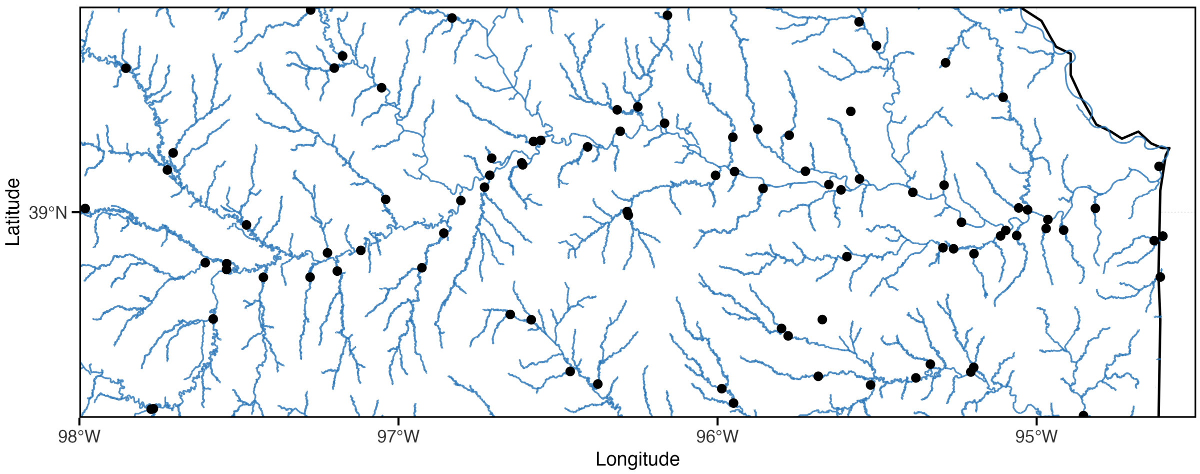
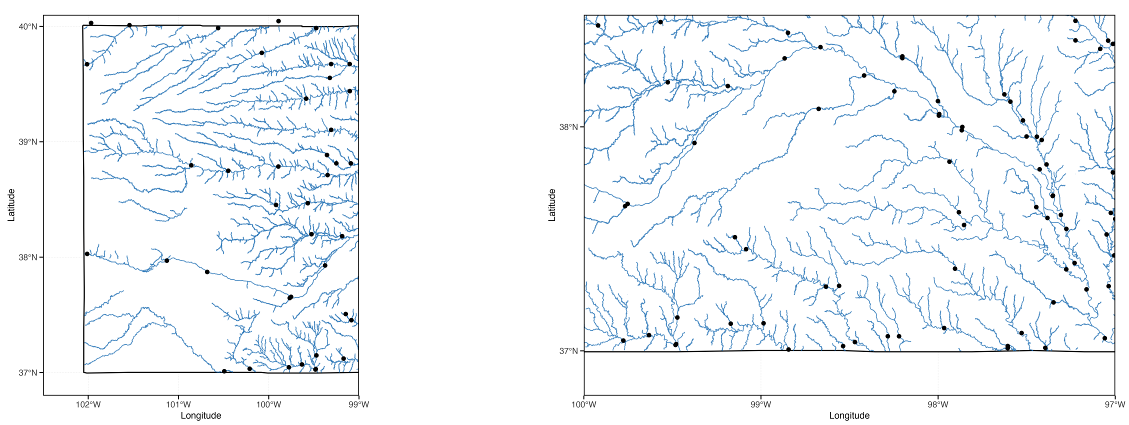


Figure 2.3: Western Kansas (west of 99°W). Spatial missingness in this area may be considered to follow a completely random process since, across the steam network, the gaps in sampling appear somewhat uniformly.



(a) **(b)**

Figure 2.4: (a) East-central Kansas (east of 98°W, 38.5°N–39.5°N). (b) South-central Kansas (100°W–97°W, south of 38.5°N). Some of the regions appear to be missing sampling locations due to random processes like the distance between streams or non-random processes such as insufficient resources to build monitoring stations.

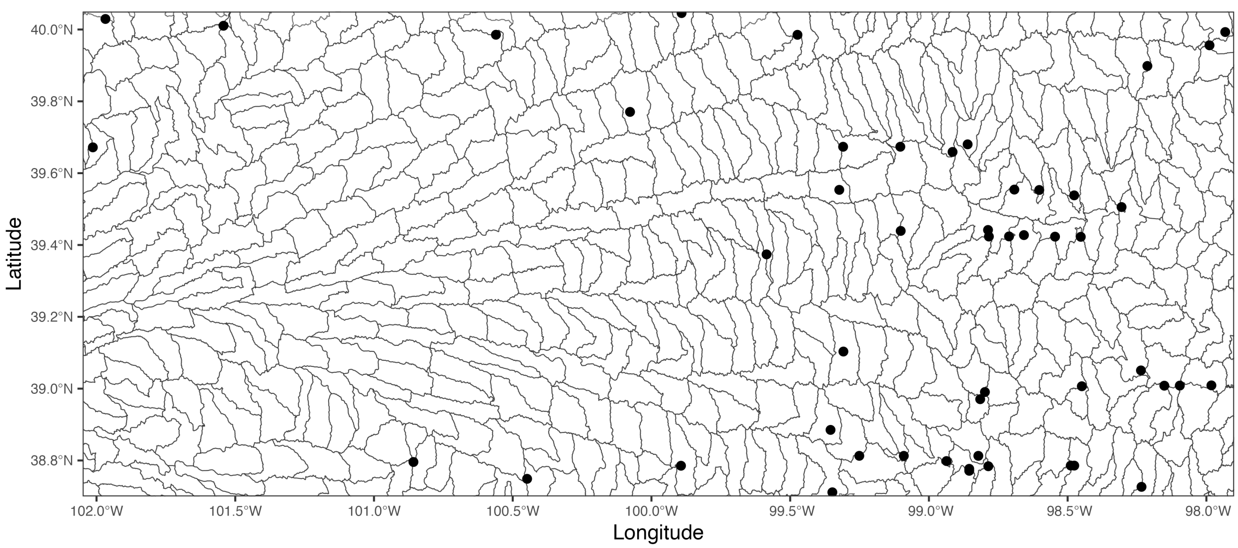


Figure 2.5: Sampled points for ambient water monitoring in the Republican and Solomon river basins, with subwatershed geometries overlayed. The points fall within subwatershed boundaries but are incompatible for direct matching to the areas for inferential purposes. Additionally, not all subwatershed areas have points within them to be matched with.

(GPR) being identical to the method of kriging[9]. With this in mind, we proceed with our discussion over kriging using the language of GPR and make all ontological distinctions between methods assuming their equivalent implementations through GPR. Kriging methods are differentiated by their treatment of the mean; whether it is known and constant (simple kriging), unknown and constant (ordinary kriging), or unknown and variable (universal kriging). Universal kriging is particularly useful for representing complex processes, such as hydrologic flow, where variability may be spatially dominant, but the mean effect is dependent on non-spatial covariates in the local environment[10]. These models can be further characterized by the usage of external covariates to account for cross-correlation (cokriging) or the addition of temporal variables to account for joint covariance in space-time data (spatio-temporal kriging)[7]. Spatio-temporal universal kriging retains the benefit of a parametric mean function while also accounting for non-constant temporal processes by treating time as an additional dimension, rather than a separate process[7]. Additionally, while lag effects are not directly modeled in spatio-temporal kriging models, universal kriging allows their incorporation via the mean function[7]. This framework produces the BLUP at unobserved locations for spatio-temporal data where the mean effect of the response variable is best informed using non-spatial covariates and temporal lag effects, which addresses many of the challenges posed by the Kansas ambient monitoring data.

Krige-and-regress, as described by Madsen et al. (2008), provides a framework for applying kriging to address spatial misalignment problems, assuming the misaligned data arise from the same spatial process. This can be extended to account for spatial misalignment across temporal records, wherein a spatio-temporal kriging model is used to perform spatial interpolation at each time step. This allows regression analysis to be performed as if both variables had been observed at the same resolution and temporal reference. The purely spatial krige-and-regress has been shown to produce, for spatially misaligned data, an estimator with lower bias than the maximum likelihood estimate; however, the results from a spatio-temporal extension have not been formally established[8]. The assumptions of the linear model (linearity, homoskedasticity, residual independence, and normality) must be approximately met to obtain the krige-and-regress estimator and realize its inferential benefits[8]. These assumptions are seemingly violated in the Kansas ambient monitoring data, with the strong presence of non-linear and non-normal characteristics throughout. By relaxing those assumptions to accommodate non-normal and non-linear processes, the spatio-temporal krige-and-regress addresses the challenge of connecting Kansas ambient monitoring data to BMP implementations.

While the problem of spatial alignment is addressed by krige-and-regress, of equal concern is the computational cost incurred by the method of kriging. The best case for simple kriging is cubic time scaling with sample size, but for spatio-temporal kriging this is multiplied by cubic scaling of temporal record length (Table 2.1)[7]. This bottleneck is the primary motivation behind the widespread application of GPR and low-rank Gaussian process methods, which offer variable computational discounts to achieve the same kriging predictor. The problems presented thus far, non-normality, non-linearity, and immense time complexity, can be addressed simultaneously by using generalized additive models to fit cubic smoothing splines to create a model similar to spatio-temporal kriging. With this modeling framework we develop a non-normal, non-linear, and computationally efficient spline model, which we refer to as spline-based psuedo-kriging (SBK), to act as a substitute for spatio-temporal kriging in the spatio-temporal krige-and-regress framework.

	Simple kriging $O(n^3)$		Spatio-temporal kriging $O(n^3t^3)$	
n	Operations	Est. Time	Operations	Est. Time
100	1.00×10^6	333 µs	1.25×10^8	41.7 ms
200	8.00×10^6	2.7 ms	1.00×10^9	333 ms
1,000	1.00×10^9	333 ms	1.25×10^{11}	41.67 s
2,000	8.00×10^9	2.7 s	1.00×10^{12}	5.56 min
10,000	1.00×10^{12}	5.6 min	1.25×10^{14}	11.57 hr
20,000	8.00×10^{12}	44.4 min	1.00×10^{15}	92.59 hr

Table 2.1: Computational complexity and estimated wall-clock time for simple kriging $O(n^3)$ and Spatio-Temporal Kriging $O(n^3t^3)$ as a function of the number of spatial observations n, with $t = 5$ time steps. Wall-clock estimates apply the leading constant of Cholesky decomposition ($\frac{1}{3}n^3$) and assume one nanosecond per floating-point operation.

2.3 Spline approximation

Smoothing splines are estimates of functions separated into intervals, often defined by observed data. Spline interpolation optimizes the placement of interval boundaries, referred to as knots, to model complex processes by balancing proximity to observations and smoothness of the estimated function[11]. The result is an estimate analogous to polynomial regression with improved flexibility to local variability[12]. Just as with polynomial regression, individual splines each form a basis function: mathematical functions that transform covariates[13]. Basis functions map inputs to a target function space, thus when applied to a covariate they are a method for taking a variable from one domain to another (e.g. projecting geographic coordinates onto a spline surface)[13]. Cubic splines have been shown to be effective approximators for the various kriging methods due to this property of forming basis functions and mapping covariates to new domains, although at the direct cost of kriging variance[7;14]. The cubic spline can be thought of as a piecewise function composed from third order polynomials joined to form a smooth curve[15].

Cubic splines are limited to one-dimensional predictors, a problem for the two-dimensional spatial coordinates found in the Kansas datasets[16]. Thin-plate splines, a multi-dimensional extension of cubic splines, correct for this limitation while retaining all of the benefits of cubic splines described above[12]. The use of cubic splines with temporal data runs the risk of introducing spurious or artificial seasonal patterns[17]. These distortions in temporal signal can be prevented by constraining knots to capture periodicity[17]. Placing this constraint on a cubic spline creates what is known as a cyclic cubic spline, which serves as an estimator of periodic means in temporal data[17]. In this framework, the thin-plate spline and cyclic cubic spline form separate spatial and temporal bases respectively. To construct a spatio-temporal basis function, the thin-plate and cyclic cubic splines are joined with a Kronecker (tensor) product[18]. Spatio-temporal basis functions are often formed via these tensor products due to their convenient representations of spatio-temporal covariance matrices, ability to scale mismatched units to the same joint process, and computational efficiency[15;18].

A natural framework for fitting smoothing splines to data is to incorporate them as the functional bases in additive models (Eq. 2.1), a special case of the linear model that includes smooth terms as linear predictors[15].

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I}). \tag{2.1}$$

Similarly, generalized additive models insert smooth terms into the generalized linear model architecture, expanding on the additive model by relaxing the distributional assumption on the data generating process to account for any exponential family distribution with expected value μ and scale parameter ϕ (Eq. 2.2)[15]. The expected value of the response, y, is modeled as the linear combination of parametric fixed effects and the summation of p smooth functions of predictors, with the canonical link, $g(\cdot)$, used to map the model to the scale of μ (Eq. 2.3)[15].

$$y_i \sim \mathrm{EF}(\mu_i, \phi), \tag{2.2}$$

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{l} \beta_k x_{ik} + \sum_{j=1}^{p} f_j(x_{ij}). \tag{2.3}$$

Generalized additive models, when fit with smoothing spline bases, account for non-linear characteristics present in the Kansas ambient monitoring data, although the smoothing spline coefficients lack any immediate parametric interpretation. Together these components yield a method for developing a spatio-temporal covariance structure, analogous to spatio-temporal kriging, from a set of computationally efficient basis functions. Additionally, the ability to incorporate any exponential family distribution in the assumed data generating process addresses a wide range of response distributions encountered in water quality monitoring. We now present the SBK modeling framework, which accounts for non-normal and non-linear processes in spatio-temporal krige-and-regress problems with substantially reduced computational expense.

2.4 Spline-based psuedo-kriging

To account for strict non-negativity, our model assumes a gamma-distributed response, (Eq. 2.4).

$$f(y_i;\,\lambda_i,\alpha) = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\lambda_i}\right)^{\alpha} y_i^{\alpha-1}\exp\!\left(-\frac{\alpha y_i}{\lambda_i}\right), \qquad y_i > 0. \tag{2.4}$$

The expected value is modeled as a linear combination of the parametric components, and the tensor product of the smoothing spline bases for spatial and temporal effects to form an approximate spatio-temporal covariance structure (Eq. 2.5). We choose a thin-plate spline and cyclic cubic spline basis for the spatial and temporal variables respectively, (Eq. 2.6).

$$\lambda_i = \exp\left[\beta_0 + \sum_{j=2}^{J}\beta_1\gamma_{ij} + g(s_{1i}, s_{2i}, \delta_i)\right], \tag{2.5}$$

$$g(s_1, s_2, \delta_i) = \sum_{k=1}^{K_1}\sum_{m=1}^{K_2}\eta_{km}\, t_k(s_1, s_2)\, c_m(\delta_i), \tag{2.6}$$

$$\hat{g} = \arg\min_{g}\left[-\,\ell(\lambda_i;\, y_i) + \frac{1}{2}\left(\nu_1\, S_1(g) + \nu_2\, S_2(g)\right)\right]. \tag{2.7}$$

$$S_1(g) = \iint\left[\left(\frac{\partial^2 g}{\partial s_1^2}\right)^2 + 2\left(\frac{\partial^2 g}{\partial s_1\,\partial s_2}\right)^2 + \left(\frac{\partial^2 g}{\partial s_2^2}\right)^2\right] ds_1\, ds_2, \tag{2.8}$$

$$S_2(g) = \int\left(\frac{\partial^2 g}{\partial\,\delta^2}\right)^2 d\,\delta. \tag{2.9}$$

Spline coefficients were estimated via penalized regression (Eq. 2.8), quantity and location of knots for the smoothing splines were determined by generalized cross validation. All analyses were implemented using R[19]; these estimation procedures are available through the 'mgcv' package[20].

2.5 Model fitting and evaluation

In addition to SBK, we fit a multiple linear regression (MLR) and spatio-temporal kriging model using the same covariates. The models were compared using holdout data partitioned to match the spatial missingness structure (Figure 2.6). Two analytes, total phosphorous and total suspended solids, were chosen as responses for model fitting to maintain consistency with our final analysis of BMP effects. Due to the substantial computational cost of kriging, the models were initially fit using a subset of the total study area, the Solomon River basin. The MLR and SBK models were then fit to the full study area using the same partitioning approach for a large-scale comparison. Prediction intervals over the holdout data were obtained via posterior predictive simulation and used to estimate mean coverage.

To compare our method with the krige-and-regress framework, the SBK model and a universal kriging model were fit to the unpartitioned Solomon River basin data. To emulate a crude implementation of the krige-and-regress model in the presence of temporal data, the universal kriging model was fit to each year individually with the model equation in 2.10.

$$\mathbb{E}[y_i] = \beta_0 + s_{1i} + s_{2i} + \delta. \tag{2.10}$$

Total phosphorous and total suspended solids were used as the response variable for modeling in order to focus inference on a string of station delistings in the Solomon River basin for those two analytes during 2022. A prediction surface was obtained for each model from 2020 to 2022, classified to subwatersheds by majority rule, and averaged within each subwatershed. The predictions for the universal kriging model were fit to the linear mixed model with a cyclic cubic spline to account for day-to-day effects in (2.11).

$$\begin{gathered} y_i = \beta_0 + b_{0l} + \beta_1 c_{ij} + \beta_2 \gamma_{ik} + \eta_m c_m(\delta_i) + \epsilon_{ijk}, \\ \epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2), \quad b_{0l} \sim N(0, \sigma_b^2). \end{gathered} \tag{2.11}$$

For the SBK predictions, we fit a gamma distributed generalized linear mixed model which is formed by applying the likelihood in (2.4) to the model equation of (2.11). Least squares (LS) means were computed for each BMP coefficient in the two mixed models, along with 80% confidence intervals.

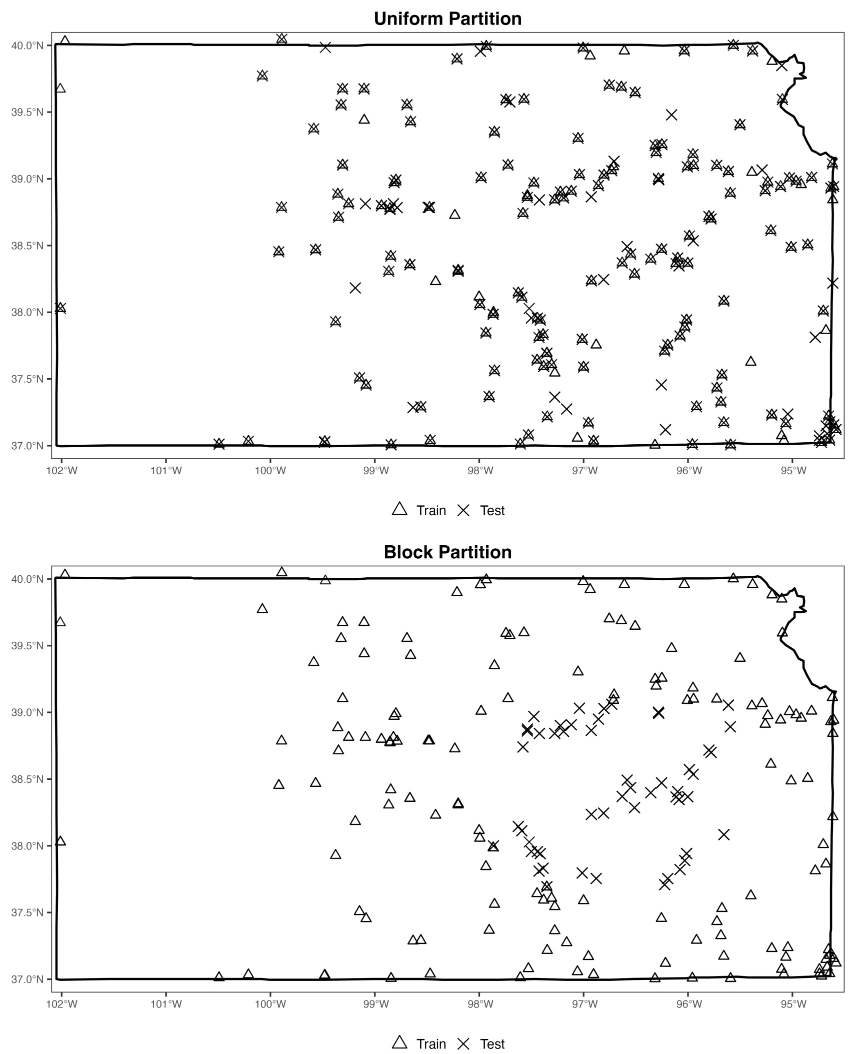


Figure 2.6: Out-of-sample validation was performed by using the a temporal split (not pictures) and two spatial partitions above to check the capability of models to maintain callibration in the presence of the missing data geometries naturally present in the Kansas data. The uniform partition was built with a 50% removal of all spatial locations uniformly spread across the state to emulate data missing completely at random. The block partition removed the central 30% of locations to emulate data missing not at random.

Results

Table 3.1 provides an analysis of the SBK model's callibration when compared to MLR and spatio-temporal kriging. The SBK model showed the lowest mean coverage for all partitions, as well as the least absolute distance from the prediction interval width (0.95) for the temporal (Solomon: 0.960, Kansas: 0.948) and grid (Solomon: 0.948, Kansas: 0.958) paritions. For the central partition it maintained the lowest absolute distance Notably, the MLR and spatio-temporal kriging models performed similarly on average across each partition within the Solomon River basin.

	Solomon River Basin			**Full Kansas Data**	
Partition	MLR	ST-Kriging	SBK	MLR	SBK
Total Phosphorus					
Temporal	0.972	0.972	0.960	0.972	0.957
Grid	0.966	0.967	0.948	0.978	0.958
Central	0.974	0.972	0.948	0.988	0.912
Total Suspended Solids					
Temporal	0.979	0.978	0.966	0.973	0.949
Grid	0.970	0.970	0.952	0.980	0.961
Central	0.973	0.973	0.954	0.985	0.960

Table 3.1: Out-of-sample mean prediction interval coverage, by partition and water-quality constituent, for the reduced (Solomon River basin) and full (Kansas) study areas.

The LS mean intervals for BMP effects on phosphorus in figure 3.1 shows high uncertainty propogating throughout the mean estimates. One candidate practice for phosphorous, which we defined as neither interval bounds crossing zero, was identified for krige-and-regress (On farm trial, $\hat{\mu} = 0.00855$, 80%CI $[0.00265, 0.01444]$) and two were found for SBK (On farm trial, $\hat{\mu} = -0.14026$, 80%CI $[-0.24154, -0.03898]$; irrigation group

$\hat{\mu} = 0.06555$, 80%CI $[0.00151, 0.12959]$). For total suspended solids the mean intervals for krige-and-regress appear to have the same high uncertainty as with phosphorous; by comparison, the SBK intervals are shifted further away from zero (Figure 3.2). One TSS practice for krige-and-regress (On farm trial, $\hat{\mu} = 0.53448$, 80%CI $[0.16593, 0.90303]$) was beyond zero, while three candidates for SBK (Onsite wastewater system, $\hat{\mu} = 0.05683$, 80%CI $[0.00324, 0.11041]$, sediment reduction group, $\hat{\mu} = 0.07971$, 80%CI $[0.02776, 0.13166]$, waterway group, $\hat{\mu} = 0.04505$, 80%CI $[0.00416, 0.08593]$) were identified. Table 3.2 summarises those BMP 80% confidence intervals isolated for their signal clarity.

Best Management Practice	Analyte	Estimate	SE	Lower CI	Upper CI
Krige-and-regress					
On Farm Trial	Phos.	0.00855	0.00231	0.00265	0.01444
On Farm Trial	TSS	0.53448	0.14451	0.16593	0.90303
Spline-based psuedo-kriging					
On Farm Trial	Phos.	-0.14026	0.03449	-0.24154	-0.03898
Irrigation Group	Phos.	0.06555	0.02679	0.00151	0.12959
OnSite Wastewater System	TSS	0.05683	0.02222	0.00324	0.11041
Sediment Reduction Group	TSS	0.07971	0.02203	0.02776	0.13166
Waterway Group	TSS	0.04505	0.01675	0.00416	0.08593

Table 3.2: Least squares means for the effect of best management practices on nonpoint source pollution with 80% confidence interval upper bounds below 0 or lower bounds above 0. Spline-based psuedo-kriging estimates were log transformed to match scale with the krige-and-regress means.

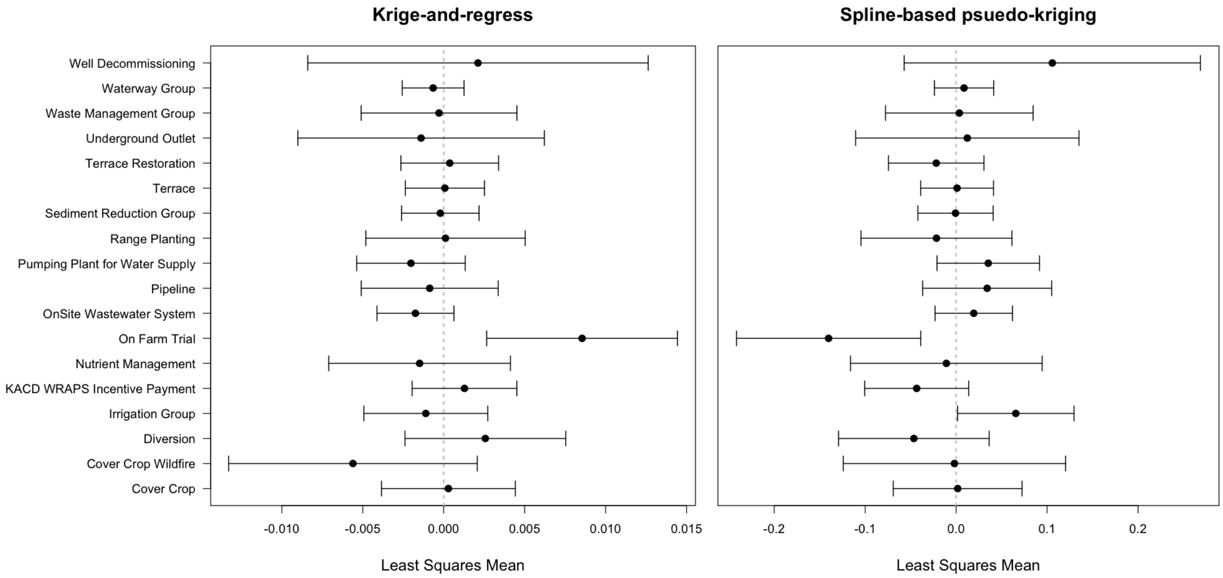


Figure 3.1: Least squares means with 80% confidence intervals for the effect of best management practices on phosphorus loads from 2020–2022 in the Solomon River basin.

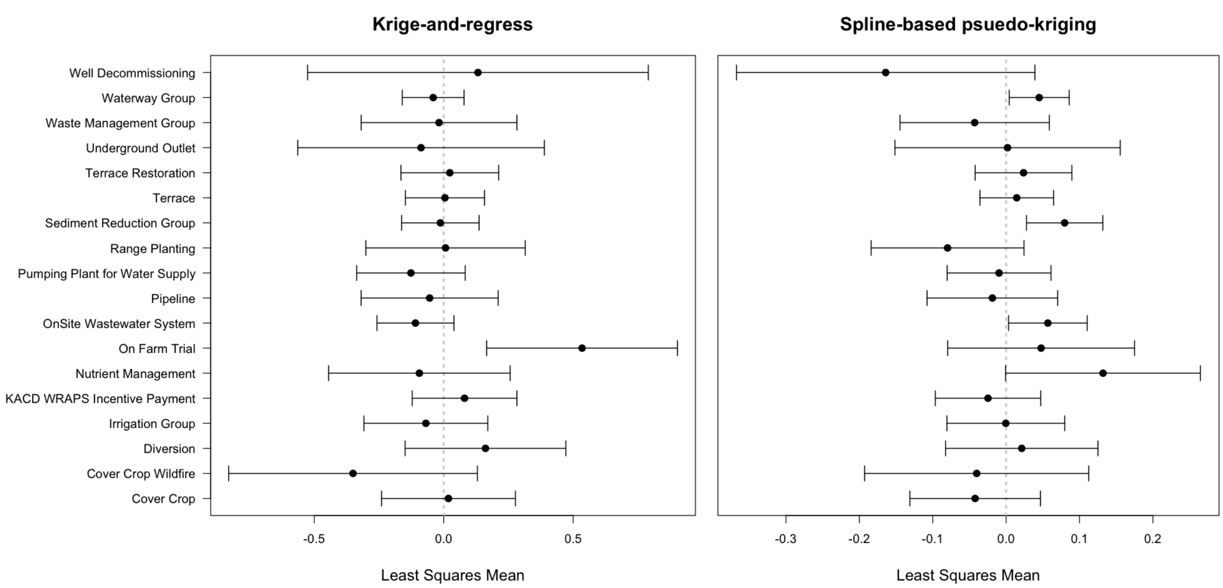


Figure 3.2: Least squares means with 80% confidence intervals for the effect of best management practices on total suspended solid loads from 2020–2022 in the Solomon River basin.

Discussion

Spatial misalignment between water monitoring and BMP data is a fundamental hazard of their data generating process; it isn't a case were better data collection is a reasonably achievable solution. Our efforts were focused on developing an extension to the existing krige-and-regress framework so that these misaligned data sources can be connected post-hoc. The callibration seen in the SBK model affirms the model's ability to consider the data generating process in the presence of missing data without overestimating variance. The drop in mean coverage seen in the full Kansas data central partition can be attributed to the tendency of the model to catastrophically fail when faced with extreme information gaps. This behavior was also observed with the south-western corner of the SBK prediction surfaces, where there was no ambient montioring data collected and the surrounding points contained high outlier values. The MLR and spatio-temporal kriging models having similar out-of-sample callibration appears to be primarily the by-product of their shared distributional assumption and strictly linear parameters. Despite both models equally over-estimating the variance present in the data partitions there's no evidence that MLR and spatio-temporal kriging are inferentially analogous, they simply share predictive capacity in this specific case. We recommend our SBK model in any instance where both of these assumptions are violated, however it may be worthwhile for researchers to consider a normal nonlinear model or generalized linear model if only one assumption is clearly violated.

Results from our regression analysis present a clear view of the outcome in replacing kriging with the SBK model in krige-and-regress. The standard krige-and-regress framework provides highly uncertain results when non-linear and non-mormal characteristics are present in the data. Recommending on farm trial as a candidate for *increasing* phosphorus pollution does not significantly challenge intuition but is questionable when considering the low prevalence of the practice in the Solomon River basin (1.67% of all practices implemented between 2020 and 2022). The SBK and regress framework (SBK-R) presents more scientifically consistent results with less obvious reliance on practice sample size (11.22% for irrigation group, 10.50% for onsite wastewater system, 7.64% for sediment reduction, and 21.24% for waterway group). On farm trial changing signs based on model is a result we lack a formal explanation for at this time, but it is suspected that this is a by-product of the extreme uncertainty due to low sample size. Fitting both frameworks with total suspended solids as a response was meant to investigate the nature of the both models in the presence of delusive discrete data. The SBK-R candidate practices are both rational and certain enough to consider for future investigation, which is how all candidate practices identified should be treated. The unvertainty in all of these estimates and intervals is far too inconsistent to treat any relationships identified as causal or even highly correlative.

The choice of 80% confidence intervals was meant to increase the potential pool of candidate practice given the low cost to type I error in selecting practice for further investigation. Pushing the cutpoint for the dichotimization of candidate practices provided a striking result, changing the uninformative results of the krige-and-regress framework to vastly overinformative. SBK-R showed a near resistance to this change, which could be attributable to the differing response scales for the two methods, but is indicative of the LS means crossing zero having no reasonable effect. The high uncertainty propogated throughout LS mean estimates is very likely associated with the incomplete grouping of low sample size practices, lack of auxillary data for large regions of missing data, as well as gaps in spatial records for private landowners and facilities. These intervals could be tightened significantly and used for stronger decision ruling if the previously discussed variance inflation problems are addressed.

The spline-based psuedo-kriging model and subsequent SBK and regress framework present promise in resolving the problem of spatial misalignment in water quality data. Our results show that, in the presence of non-linearity and non-normality the SBK model is a more effective method for interpolating misaligned space than kriging or spatio-temporal kriging. When using the predicted surface of SBK as the response for a regression model with shared distribution, we have shown a more informative estimate of variance for identifying candidate BMPs for change in NPS pollution. We thus recommend the use of our SBK-R framework in all instances where the krige-and-regress framework would be appropriate but the assumptions of linearity and normality have been clearly violated.

Bibliography

[1] Jabbar, F. K., and Grote, K. (2019). Statistical assessment of nonpoint source pollution in agricultural watersheds in the Lower Grand River watershed, MO, USA. *Environmental Science and Pollution Research*, 26(2), 1487–1506.

[2] Falcone, J. A., Murphy, J. C., and Sprague, L. A. (2018). Regional patterns of anthropogenic influences on streams and rivers in the conterminous United States, from the early 1970s to 2012. *Journal of Land Use Science*, 13(6), 585–614.

[3] Deng, Q., Sharretts, T., Ali, T., Ao, Y. Z., Chiarelli, D. D., Demeke, B., Marston, L., Mehta, P., Mekonnen, M., Rulli, M. C., Tuninetti, M., Xie, W., and Davis, K. F. (2025). Deepening water scarcity in breadbasket nations. *Nature Communications*, 16(1), 1110.

[4] Easton, Z., Stephenson, K., Benham, B., Böhlke, J. K., Buda, A., Collick, A., Fowler, L., Gilinsky, E., Miller, A., Noe, G., Palm-Forster, L. H., Shabman, L., and Wynn-Thompson, T. (2025). The nonpoint source challenge: Obstacles and opportunities for meeting nutrient reduction goals in the Chesapeake Bay watershed. *JAWRA Journal of the American Water Resources Association*, 61(3), e70034.

[5] Pacifici, K., Reich, B. J., Miller, D. A. W., and Pease, B. S. (2019). Resolving misaligned spatial data with integrated species distribution models. *Ecology*, 100(6), e02709.

[6] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York, NY.

[7] Cressie, N., and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

[8] Madsen, L., Ruppert, D., and Altman, N. S. (2008). Regression with spatially misaligned data. *Environmetrics*, 19(5), 453–467.

[9] Suryasentana, S. K., and Sheil, B. B. (2023). Demystifying the connections between Gaussian process regression and kriging theories. In *Proceedings of the 9th International Conference on Offshore Site Investigation and Geotechnics (OSIG 2023)*. Society for Underwater Technology, London, UK.

[10] Daniel, E. B. (2011). Watershed modeling and its applications: A state-of-the-art review. *The Open Hydrology Journal*, 5(1), 26–50.

[11] Storath, M., and Weinmann, A. (2024). Smoothing splines for discontinuous signals. *Journal of Computational and Graphical Statistics*, 33(2), 651–664.

[12] Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 65(1), 95–114.

[13] Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3), 632–646.

[14] Mardia, K. V., Kent, J. T., Goodall, C. R., and Little, J. A. (1996). Kriging and splines with derivative information. *Biometrika*, 83(1), 207–221.

[15] Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC, New York, NY.

[16] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer, New York, NY.

[17] Lusa, L., and Ahlin, Č. (2020). Restricted cubic splines for modelling periodic data. *PLOS ONE*, 15(10), e0241364.

[18] Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman and Hall/CRC, Boca Raton, FL.

[19] R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[20] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(1), 3–36.