JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–11: Theory and Methods
https://doi.org/10.1080/01621459.2022.2147531

Taylor & Francis
Taylor & Francis Group

Valid Model-Free Spatial Prediction

Huiying Mao[a], Ryan Martin[b], and Brian J. Reich[b]

[a]The Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC; [b]North Carolina State University, Raleigh, NC

ABSTRACT

Predicting the response at an unobserved location is a fundamental problem in spatial statistics. Given the difficulty in modeling spatial dependence, especially in nonstationary cases, model-based prediction intervals are at risk of misspecification bias that can negatively affect their validity. Here we present a new approach for model-free nonparametric spatial prediction based on the *conformal prediction* machinery. Our key observation is that spatial data can be treated as exactly or approximately exchangeable in a wide range of settings. In particular, under an infill asymptotic regime, we prove that the response values are, in a certain sense, locally approximately exchangeable for a broad class of spatial processes, and we develop a local spatial conformal prediction algorithm that yields valid prediction intervals without strong model assumptions like stationarity. Numerical examples with both real and simulated data confirm that the proposed conformal prediction intervals are valid and generally more efficient than existing model-based procedures for large datasets across a range of nonstationary and non-Gaussian settings.

ARTICLE HISTORY
Received September 2021
Accepted November 2022

KEYWORDS
Conformal prediction; Gaussian process; Kriging; Nonstationary; Plausibility

1. Introduction

Providing valid predictions of the response at an unobserved location is a fundamental problem in spatial statistics. For example, epidemiologists may wish to extrapolate air pollution concentrations from a network of stationary monitors to the residential locations of the study participants. There are a number of challenges one faces in carrying out valid prediction at a new spatial location, but one of the most pressing is that existing methods are model-based, so the reliability of the predictions depends crucially on the soundness of the posited model. For example, prediction intervals based on Kriging—see Cressie (1992) and Section 2.1—often rely on normality, and stationarity is often assumed to facilitate estimating the spatial covariance function required for Kriging. It is now common to perform geostatistical analysis for massive datasets collected over a vast and diverse spatial domain (Heaton et al. 2019). For complex processes observed over a large domain, the normality and stationarity assumptions can be questionable. Failing to account for nonstationarity can affect prediction accuracy, but typically has a larger effect on uncertainty quantification such as prediction intervals (Fuglstad et al. 2015). While there are now many methods available for dealing with nonstationary (Risser see 2016, for a recent review) and non-Gaussianity (Gelfand, Kottas, and MacEachern 2005; Duan, Guindani, and Gelfand 2007; Reich and Fuentes 2007; Rodriguez and Dunson 2011), these typically involve heavy computations. This exacerbates the already imposing computational challenges posed by massive datasets. Further, fitting the entire stochastic process may be unnecessary if only prediction intervals are desired. Nonparametric machine-learning methods can be used for prediction (Kim, Kwon Lee, and Mu Lee 2016a,b; Lim et al. 2017; Tai, Yang, and Liu 2017; Hengl et al. 2018; Franchi, Yao, and Kolb 2018; Wang, Guan, and Reich 2019; Li, Sun, and Reich 2020), but these methods typically focus on uncertainty estimation. In this article we propose a method with provably valid prediction intervals—exact in some cases, asymptotically approximate in others—for the response at a single location without requiring specification of a statistical model, and hence not inheriting the risk of model misspecification bias.

In recent years, the use of machine learning techniques in statistics has become increasingly more common. While there are numerous examples of this phenomenon, the one most relevant here is *conformal prediction*. This method originated in Vovk, Gammerman, and Shafer (2005) and the references therein (Shafer and Vovk see, also, 2008), but has appeared frequently in the recent statistics literature (Lei and Wasserman 2014; Lei et al. 2018; Guan 2019; Romano, Patterson, and Candes 2019; Tibshirani et al. 2019). What makes this method especially attractive is that it provides provably valid prediction intervals without specification of a statistical model. More precisely, the conformal prediction intervals achieve the nominal frequentist prediction coverage probability, uniformly over all data distributions; see Section 2.2. The crucial assumption behind the validity of conformal prediction is that the data are exchangeable.

Whether it is reasonable to assume exchangeability in a spatial application depends on how the data are sampled. On the one hand, if the locations are randomly sampled in the spatial domain, then exchangeability holds automatically; see Lemma 1. In such cases, standard conformal prediction can be used basically off the shelf. On the other hand, if the locations

CONTACT Brian J. Reich *brian_reich@ncsu.edu* North Carolina State University, 5212 SAS Hall, Box 8203, Raleigh, NC 27695-8203.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

are fixed in the spatial domain, then exchangeability does not hold in general. We show, however, that for a wide range of spatial processes, the response variables at tightly concentrated locations are approximately exchangeable; see Theorem 1. Therefore, a version of the basic conformal prediction method applied to these tightly concentrated observations ought to be approximately valid.

Using this insight about the connection between exchangeability and the sampling design, we propose two related spatial conformal prediction methods. The first, a so-called *global spatial conformal prediction* (GSCP) method, described in Section 3, is designed specifically for cases where the spatial locations are sampled at random. In particular, this global method produces a prediction interval which is marginally valid, i.e., valid on average with respect to the distribution of the target location at which prediction is desired; asymptotic efficiency of this global method is also investigated. The second, a *local spatial conformal prediction* (LSCP) method, described in Section 4, is designed specifically for the case when the spatial locations are fixed. Since our goal is to proceed without strong assumptions about the spatial dependence structure, it is only possible to establish approximate or local exchangeability. Therefore, the proposed local spatial conformal prediction method can only provide approximately valid predictions; see Theorem 2. But our goal in the fixed-location case resembles the "conditional validity" target in the conformal prediction literature (Barber et al. 2019; Chernozukov, Wüthrich, and Zhu 2021) so, given the impossibility theorems in the latter context, approximate validity is all that can be expected.

For both the global and local formulations, our proposed method is computationally feasible for large datasets and model-free in the sense that its validity does not depend on a correctly-specified model. In Sections 5 and 6, we show using real and simulated data that the proposed methods outperform both standard global Kriging and local approximate Gaussian process regression (Gramacy and Apley laGP; 2015) for nonstationary and non-Gaussian data. In addition to be useful for spatial applications, it is also an advancement in conformal prediction to the case of dependent data, and establishes the conditions on the spatial sampling design and data-generating mechanism that ensure (approximate) validity of the conformal prediction intervals.

The remainder of this article is organized as follows. Section 2 reviews spatial and conformal prediction. Sections 3 and 4 introduce the proposed methods, which are examined using simulations in Section 5 and a real data analysis in Section 6. Additional numerical and theoretical results, along with all proofs, are given in the Supplemental Materials.

2. Background

2.1. Spatial Prediction Methods

Let $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$, with $d \geq 1$, be the observable pairs at spatial location $s_i \in \mathcal{D} \subseteq \mathbb{R}^2$. Note that X_i can include covariates that are deterministic functions of the spatial location, such as elevation, genuinely stochastic covariates like wind speed, or even nonspatial covariates such as the smoking status of the resident at location. Write the data points as triples $Z_i = (s_i, X_i, Y_i)$, for $i = 1, 2, \ldots$. We assume only a single observation is made at each location and thus often adopt the notation $Y_i = Y(s_i)$ and $X_i = X(s_i)$.

Geostatistical analysis often assumes that the data follow a Gaussian process model, $Y_i = X_i^\top \beta + \theta_i + \varepsilon_i$, for $i = 1, \ldots, n$, where β is the vector of regression coefficients, $\varepsilon_1, \ldots, \varepsilon_n$ are independent Normal$(0, \tau^2)$ errors, $\theta_i = \theta(s_i)$, and θ is a mean-zero Gaussian process with isotropic covariance function $C(\theta_i, \theta_j) = \sigma^2 \rho(d_{ij})$, a function of the distance d_{ij} between locations s_i and s_j. A common example is the Matérn correlation function $\rho(d; \phi, \kappa)$ parameterized by correlation range ϕ and smoothness parameter κ. Denote the spatial covariance parameters as $\Theta = \{\sigma^2, \tau^2, \phi, \kappa\}$. The main assumptions of this model are that the data are Gaussian and the covariance function is stationary and isotropic, that is, it is a function only of the distance between spatial locations and is thus the same across the spatial domain.

Consider data $Z^{n+1} = (Z_1, \ldots, Z_n, Z_{n+1})$. In our applications, Z^n will be observed and (s_{n+1}, X_{n+1}) will be given, and the goal is to predict the corresponding Y_{n+1}. However, the ordering of the data is irrelevant so one can imagine different orderings that correspond to data point i in the last position, where $i = 1, \ldots, n, n+1$. That is, imagine we have the observed data $z_{(i)}^{n+1} = \{z_1, \ldots, z_{n+1}\} \setminus \{z_i\}$, along with (s_i, x_i) and parameter estimates $\widehat{\beta}$ and $\widehat{\Theta}$; then the predictive distribution of Y_i is normal with mean $\hat{\mu}_{n+1,i}(s_i, x_i)$ and variance $\sigma^2_{n+1,i}(s_i, x_i)$, where both the mean and variance depend on $\widehat{\Theta}$ and the configuration of the spatial locations; see the Supplementary Material for the specific expressions. The standardized residuals are

$$e_{n+1,i} = \frac{y_i - \hat{\mu}_{n+1,i}(s_i, x_i)}{\hat{\sigma}_{n+1,i}(s_i, x_i)}, \quad i = 1, \ldots, n, n+1, \qquad (1)$$

and the corresponding $100(1-\alpha)\%$ prediction interval for Y_i is $\hat{\mu}_{n+1,i}(s_i, x_i) \pm q_\alpha^\star \hat{\sigma}_{n+1,i}(s_i, x_i)$, where $q_\alpha^\star$ is the upper $\alpha/2$ quantile of a standard normal distribution.

2.2. Conformal Prediction

Here we take a step back and review conformal prediction for nonspatial problems; for a detailed treatment, see Vovk, Gammerman, and Shafer (2005) and Shafer and Vovk (2008). Suppose we have a data sequence $Z_1, \ldots, Z_n, Z_{n+1}, \ldots$, assumed to be exchangeable with joint distribution P, that is, $Z_1, Z_2, Z_3, \ldots$ and $Z_{\xi(1)}, Z_{\xi(2)}, Z_{\xi(3)}, \ldots$ have the same joint distribution for any permutation ξ defined on the positive integers. This data may be response-only, i.e., $Z_i = Y_i$, or may be response-covariate pairs, i.e., $Z_i = (X_i, Y_i)$; we will focus on the latter more general case. No assumptions about P are made here, beyond that it is exchangeable. We observe $Z^n = z^n$, and the goal is to predict Y_{n+1} at a new value X_{n+1} of the covariate. More specifically, we seek a procedure that returns, for any $\alpha \in (0, 1)$, a prediction interval $\Gamma^\alpha(Z^n; X_{n+1})$ that is *valid* in the sense that

$$\mathsf{P}^{n+1}\{\Gamma^\alpha(Z^n; X_{n+1}) \ni Y_{n+1}\} \geq 1 - \alpha, \quad \text{for all } (\alpha, n, \mathsf{P}), \qquad (2)$$

where P^{n+1} is the distribution of $(Z_1, \ldots, Z_n, Z_{n+1})$ under P. That we require the inequality (2) to hold for *all exchangeable distributions* P rules out the use of model-based procedures, such as likelihood or Bayesian methods.

The original conformal prediction method proceeds as follows. Define a *nonconformity measure* $\Delta(B, z)$, a function that takes two arguments: the first is a "bag" B that consists of a finite collection of data points; the second is a single data point z. Then $\Delta(B, z)$ measures how closely z represents the data points in bag B. For example, if π is a prediction rule and d is some measure of distance, then we might take $\Delta(B, z) = d(\pi_B, z)$, the distance between z and the value π_B returned by the prediction rule π applied to B. The choice of Δ depends on the context, though often there is a natural choice. Throughout this article we will assume Δ is symmetric in its first argument, so that shuffling the data in bag B does not change the value of $\Delta(B, z)$.

Given the nonconformity measure Δ, the next step is to appropriately transform the data via Δ. Specifically, augment the observed data $z^n = (z_1, \ldots, z_n)$ with a provisional value z_{n+1} of $Z_{n+1} = (X_{n+1}, Y_{n+1})$; this z_{n+1} value is generic and free to vary. Define

$$\delta_i = \Delta(z_{(i)}^{n+1}, z_i), \quad i = 1, \ldots, n, n+1.$$

Note that δ_{n+1} is special because it compares the actual observed data with this provisional value of the unobserved future observation. Next, compute the *plausibility* (Cella and Martin 2022) of z_{n+1} as a value for Z_{n+1} according to the formula

$$p(y_{n+1} \mid z^n, x_{n+1}) = \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{\delta_i \geq \delta_{n+1}\}, \tag{3}$$

where $1\{A\}$ denotes the indicator of event A. Note that this process can be carried out for any provisional value z_{n+1}, so the result is actually a mapping $\tilde{y} \mapsto p(\tilde{y} \mid z^n, \tilde{x})$, for a given $\tilde{x}$, which we will refer to as the *plausibility contour* returned by the conformal algorithm. This function can be plotted to visualize the uncertainty about Y_{n+1} based on the given $\tilde{x}$, data z^n, the choice of nonconformity measure, etc. Moreover, a prediction set $\Gamma^\alpha(z^n; \tilde{x})$ can be obtained as

$$\Gamma^\alpha(z^n; \tilde{x}) = \{\tilde{y} : p(\tilde{y} \mid z^n, \tilde{x}) > \alpha\}. \tag{4}$$

3. Global Spatial Conformal Prediction

3.1. GSCP Algorithm

The first approach we consider is a direct application of the original conformal algorithm to spatial prediction but with spatial dependence encoded in the nonconformity measure. In contrast to the local algorithm presented in Section 4, this method equally weights the nonconformity across all spatial locations in the plausibility contour evaluation. Therefore, we refer to this as global spatial conformal prediction, or GSCP for short.

From Section 2.1, recall that $z_i = (s_i, x_i, y_i)$ and $z^{n+1} = \{z_1, \ldots, z_{n+1}\}$; also, $z_{(i)}^{n+1}$ denotes $z^{n+1} \setminus \{z_i\}$, the full dataset with z_i excluded. Now define the nonconformity measure for the GSCP algorithm as

$$\delta_i = \Delta(z_{(i)}^{n+1}, z_i) = \left| \frac{y_i - \hat{\mu}_{n+1,i}(s_i, x_i)}{\hat{\sigma}_{n+1,i}(s_i, x_i)} \right|, \quad i = 1, \ldots, n+1, \tag{5}$$

where $\hat{\mu}_{n+1,i}(s_i, x_i)$ and $\hat{\sigma}_{n+1,i}(s_i, x_i)$ are, respectively, the mean response and standard error estimates at spatial location s_i with covariate x_i based on data in $z_{(i)}^{n+1}$. Then we define a plausibility contour exactly like in Section 2.2, with obvious notational changes. That is, for provisional values $(s_{n+1}, x_{n+1}, y_{n+1})$ of $(S_{n+1}, X_{n+1}, Y_{n+1})$, we have

$$p(y_{n+1} \mid z^n; s_{n+1}, x_{n+1}) = \frac{1}{n+1} \sum_{i=1}^{n} 1\{\delta_i \geq \delta_{n+1}\}. \tag{6}$$

The corresponding $100(1-\alpha)\%$ prediction interval for Y_{n+1}, denoted by $\Gamma^\alpha(z^n; s_{n+1}, x_{n+1})$, is just an upper level set of the plausibility contour, consisting of all those provisional y_{n+1} values with plausibility exceeding α, analogous to (4).

Any reasonable choice of $(\hat{\mu}, \hat{\sigma})$ estimates can serve the purpose here, including inverse distance weighting predictions (Henley 2012), deep learning predictions (Franchi, Yao, and Kolb 2018), Kriging predictions, etc. In our numerical results presented below, we use the Kriging estimates as defined in Section 2.1, so that δ_i is a standardized Kriging residual, $|e_i|$, from (1). Conformal prediction is invariant to monotone transformations of its δ_i's, and we found that similar results are obtained with other related measures, such as unstandardized Kriging residuals. A particular advantage of our recommended choice of δ_i's is that we can quickly compute the plausibility contour and prediction interval by exploiting the inherent quadratic structure of the Kriging-based nonconformity measure; see the Supplementary Materials. Moreover, note that validity of the GSCP-based prediction intervals does not require the Gaussian model associated with the Kriging method be correctly specified, nor does it depend on our choice of the δ_i's.

3.2. Theoretical Validity of GSCP

Given the importance of exchangeability to the validity of conformal prediction and the fact that the spatial dependence generally is incompatible with exchangeability, we might have some concerns about the validity of GSCP. However, there are practically relevant cases in which exchangeability does hold, in particular, when the spatial locations are sampled independently and identically distributed (iid). The following elementary lemma explains this.

Lemma 1. If the spatial locations $S_1, S_2, \ldots$ are iid, then $Z_1, Z_2, \ldots$, with $Z_i = (S_i, X(S_i), Y(S_i))$, is an exchangeable sequence.

Since randomly sampled spatial locations makes the data exchangeable, a validity property for GSCP follows immediately from the general theory in, for example, Shafer and Vovk (2008).

Theorem 1. Let (X, Y) be a stochastic process over $\mathcal{D}$ and let $S_1, S_2, \ldots$ be iid draws in $\mathcal{D}$. Let $Z_i = (S_i, X(S_i), Y(S_i))$ for $i = 1, 2, \ldots$, and define the coverage probability function

$$c(\alpha, n, \mathsf{P}) = \mathsf{P}^{n+1}\{\Gamma^\alpha(Z^n; X_{n+1}, S_{n+1}) \ni Y(S_{n+1})\}.$$

Then the proposed GSCP is valid in the sense that

$$c(\alpha, n, \mathsf{P}) \geq 1 - \alpha, \quad \text{for all } (\alpha, n, \mathsf{P}), \tag{7}$$

where P^{n+1} is the joint distribution of $Z_1, \ldots, Z_n, Z_{n+1}$ under P. Moreover, if $\delta_1, \ldots, \delta_{n+1}$ in (5) have a continuous distribution, then

$$c(\alpha, n, \mathsf{P}) \leq 1 - \alpha + (n+1)^{-1}, \quad \text{for all } (\alpha, n, \mathsf{P}). \tag{8}$$

The upper bound in (8), which follows from the same arguments as in Lei et al. (2018), implies that the GSCP method is not only valid but also efficient in the sense that the coverage probability is not too much larger than the nominal level. That is, the coverage condition is not being achieved simply giving excessively wide intervals. Some further details on the efficiency of the global spatial conformal prediction procedure are investigated in the Supplementary Materials. Note, also, that Theorem 1 makes no assumptions about the distribution of (X, Y), so it surely covers non-Gaussian and nonstationary processes.

Theorem 1 gives a marginal validity result in the sense that it accurately predicts the response $Y(S_{n+1})$ at $X(S_{n+1})$, for a randomly sampled spatial location S_{n+1}. However, it does not ensure conditional validity, that is, the case where $S_{n+1} = s^\star$ with $s^\star$ being a fixed spatial location. There are negative results in the literature (Lei and Wasserman 2014) which state that strong conditional validity—for all P and almost all targets $s^\star$—is impossible with conformal prediction. Considerable effort has been expended recently trying to achieve "approximate" conditional validity in some sense; see, for example, Lei and Wasserman (2014), Tibshirani et al. (2019), Barber et al. (2019), and Chernozukov, Wüthrich, and Zhu (2021). Remarkably, there is at least one scenario in which a strong conditional validity result can be achieved in our context. In particular, as we show in the Appendix, the GSCP-based intervals are both marginally and conditionally valid for the special case of an isotropic process sampled uniformly on a sphere. Admittedly, these are rather strong conditions, so one would hope for (approximately) valid prediction under much less. In Section 4, we show that asymptotically valid prediction intervals at a fixed location can be obtained under only mild conditions on the sampling scheme and the unknown process.

4. Local Spatial Conformal Prediction

4.1. LSCP Algorithm

For valid prediction at a fixed location $s^\star$, we propose a local spatial conformal prediction (LSCP) approach that is based on only those data points in the neighborhood of $s^\star$. Fix an integer $m > 0$ and select a neighborhood around $s^\star$ that contains m many locations, s_{i_j}, for $j = 1, \ldots, m$. Note that $\{s_{i_j} : j = 1, \ldots, m\}$ is a subset of the full set of spatial locations $s_1, \ldots, s_n$. Without structural assumptions about the response process, such as stationarity, the data at locations far from $s^\star$ are not obviously relevant to prediction at $s^\star$, so removing—or down-weighting (Section 4.3)—them from the local analysis is reasonable. Plus, in applications where the infill asymptotic regime is appropriate, there are many observations nearby $s^\star$, so m could be taken to be large.

From here, we can proceed very much like in Section 3. For notational simplicity, assume that indices $i = 1, \ldots, m$ correspond to those m spatial locations closest to $s^\star$. Now let $Z_i = (s_i, X_i, Y_i)$, for $i = 1, \ldots, m$, denote the observations at these m closest locations to $s^\star$. With a slight abuse of that notation, set $s_{m+1} = s^\star$ and (x_{m+1}, y_{m+1}) as the provisional values of X and Y at $s^\star$. Then define the nonconformity scores exactly as before:

$$\delta_i = \Delta(z_{(i)}^{m+1}, z_i), \quad i = 1, \ldots, m+1.$$

With this, we can readily obtain the plausibility contour function:

$$p(y_{m+1} \mid z^m, s^\star, x_{m+1}) = \frac{1}{m+1} \sum_{i=1}^{m+1} 1\{\delta_i \geq \delta_{m+1}\}. \quad (9)$$

Specific details are presented in Algorithm 2 in the Supplementary Materials. The output of this algorithm is a $100(1-\alpha)\%$ prediction interval for $Y_{m+1} = Y(s^\star)$, depending on Z^m and the observed $X_{m+1} = X(s^\star)$, which we denote by $\Gamma_{s^\star}^{\alpha}(Z^m; X_{m+1})$.

4.2. Theoretical Validity of LSCP

Our theoretical results hinge on a definition of local exchangeability. Let $\mathcal{D} \subset \mathbb{R}^2$ be a compact spatial domain, for example, $[0,1]^2$. For a generic $\mathbb{R}^d$-valued stochastic process T defined on $\mathcal{D}$, with $d \geq 1$, define the *localized version* of T, relative to a location $s^\star \in \mathcal{D}$, as

$$\widetilde{T}_r(u) = T(s^\star + ru), \quad u \in \mathcal{U} = \{u \in \mathbb{R}^2 : \|u\| \leq 1\}, \quad (10)$$

indexed by the unit disk $\mathcal{U}$ and the radius $r > 0$. Now suppose that T can be decomposed as

$$T(s) = \psi\left(L(s), E(s)\right), \quad s \in \mathcal{D}, \quad (11)$$

where L and E are independent $\mathbb{R}^d$-valued stochastic process, L is a continuous spatial process, E is a nonspatial process, and ψ is a deterministic, continuous, $\mathbb{R}^d$-valued function. More specifically, suppose that L and E, respectively, satisfy the following conditions:

- L is L_2-*continuous* at $s^\star$ in the sense that its localized version $\widetilde{L}_r$ satisfies $\mathsf{E}\|\widetilde{L}_r(u) - \widetilde{L}_0(u)\|^2 \to 0$ as $r \to 0$ for any $u \in \mathcal{U}$;
- E is *locally iid* at $s^\star$, that is, its localized version $\widetilde{E}_r$ converges in distribution to an iid process as $r \to 0$.

This formulation is too abstract to be useful, but formulating a general result here is appropriate. Appendix A.2 describes several common spatial models that satisfy (11), and generalizes the above formulation to the case where the covariates are also considered stochastic processes.

These assumptions yield a certain kind of *local exchangeability* which will be used below to show the LSCP algorithm achieves a desired validity property. We first establish this local exchangeability result, which may be of independent interest.

Proposition 1. Suppose that T can be decomposed as in (11), where L is L_2-continuous at $s^\star$, E is locally iid at $s^\star$, and L and E are independent. Then the localized process $\widetilde{T}_r$ in (10) converges in distribution as $r \to 0$, and the limit is an exchangeable process in the sense that its finite-dimensional distributions are exchangeable.

Using Proposition 1, we can establish the (asymptotically approximate) theoretical validity of the LSCP method. To set the scene, those m spatial locations closest to $s^\star$ fall in a neighborhood of some radius r. As is common in the spatial statistics literature (Stein 1990; Cressie 1992), we adopt an *infill asymptotic*

regime in which the region $\mathcal{D}$ remains fixed while the number of observations n goes to infinity, hence filling the space. The relevant point for our analysis is that under this regime the number of observations made in any neighborhood of $s^\star \in \mathcal{D}$ will go to infinity. Such a regime is natural—and necessary—in cases without structural assumptions about Y, e.g., stationarity, where it is simply not possible to learn the local features of a process at $s^\star$ if data are not concentrated in a neighborhood around $s^\star$. Under the infill asymptotic framework , if m is fixed and the number of locations n is increasing to fill the bounded space $\mathcal{D}$, then the radius of the neighborhood in which those m points fall is vanishing. For example, if the spatial locations are (roughly) uniformly distributed in $\mathcal{D}$, then the number of points in a neighborhood of radius r would be proportional to nr^2; setting this equal to m gives $r = r_n = (m/n)^{1/2} \to 0$ as $n \to \infty$.

It follows from Proposition 1 that the joint distribution of the response Y at these m-many spatial locations around $s^\star$ (corresponding to m-many vectors in $\mathcal{U}$) would be approximately exchangeable and, consequently, a conformal prediction algorithm that creates nonconformity scores using only these m observations would be valid for predicting $Y(s^\star)$.

Theorem 2. Consider an infill asymptotic regime with n spatial locations in the bounded domain $\mathcal{D}$, with $n \to \infty$. Fix an integer $m > 0$ and let $r = r_n \to 0$ be such that the m closest locations to $s^\star$ fall in a neighborhood of radius r. Under the assumptions of Proposition 1, the nonconformity measure Δ is a continuous function of its inputs, and if the limiting distribution in Proposition 1 is continuous, then the LSCP prediction intervals are asymptotically valid at $s^\star$ in the sense that

$$\lim_{n\to\infty} \mathsf{P}^{m+1}\{\Gamma^\alpha_{s^\star}(Z^m; X_{m+1}) \ni Y_{m+1}\} = 1 - \alpha + O(m^{-1}),$$

where P^{m+1} is the joint distribution of $Z_1, \ldots, Z_m, Z_{m+1}$ at the m spatial locations and at $s^\star$.

4.3. The Smoothed LSCP Algorithm

Theorem 2 implies that the local spatial conformal prediction with m nearest neighbors is approximately valid under the infill asymptotic regime. However, in practice completely disregarding the contribution of the observations outside the m nearest neighbors may be unsatisfactory, so we propose a smoothed version of the LSCP algorithm (sLSCP).

The GSCP algorithm weights all $n + 1$ nonconformity measures δ_i equally in the plausibility contour computation in (3), but this is questionable for nonstationary processes with stochastic properties that vary throughout the spatial domain. To allow for nonstationarity, the sLSCP algorithm weights the nonconformity measures δ_i by how far the corresponding s_i is from the prediction location. Let f be a nonincreasing function, and define weights

$$w_i \propto f(d_i), \quad i = 1, \ldots, n+1,$$

where $d_i = \|s_i - s^\star\|$, $d_{n+1} \equiv 0$, and the proportionality constant ensures that $\sum_{i=1}^{n+1} w_i = 1$. Different f functions can be applied, but we recommend the normalized Gaussian kernel function with bandwidth η,

$$w_i = \frac{\exp(-d_i^2/2\eta^2)}{1 + \sum_{j=1}^{n} \exp(-d_j^2/2\eta^2)}, \quad i = 1, \ldots, n+1. \tag{12}$$

Note that, if $\eta \to \infty$, then $w_i \to (n+1)^{-1}$ for each i, which corresponds to the GSCP algorithm. Finally, with these new weights, the plausibility contour at a provisional value $(s_{n+1}, x_{n+1}, y_{n+1})$ of $Y(s^\star)$ is given by

$$p_w(y_{n+1} \mid Z^{n+1}_{(n+1)}, s_{n+1}, x_{n+1}) = \sum_{i=1}^{n+1} w_i 1\{\delta_i \geq \delta_{n+1}\}, \tag{13}$$

As before, we recommend the Kriging-based strategy with δ_i the standardized residual in (1).

Since we are interested only in the local structure of Y, it is natural that locations far from $s^\star$ have negligible weight, as in (12). But including all n observations requires some nontrivial calculations, e.g., inverting a large $n \times n$ covariance matrix. Therefore, to avoid cumbersome and ultimately irrelevant computation, we recommend using only the $M \ll n$ observations closest to $s^\star$ for both the Kriging predictions that determine δ_i and in the plausibility scores in (13). The resulting method is both locally adaptive and computationally efficient even for large data sets.

The tuning parameter η can be selected using cross validation, as illustrated in Sections 5 and 6. The value of M is determined by the bandwidth η so that all observations with substantial w_i are included, as are observations that are required for the Kriging prediction of these observations. Typically the number of nearby observations required to approximate the Kriging prediction is a small subset of the total number of observations (Stein 2002). As a rule of thumb, M could be selected to roughly include all observations within $2\eta + r^\star$ radius of $s^\star$, where 2η captures observations with substantial weights, and $r^\star$ is selected so that all the M observations include the nearest 15 neighbors of the observation within 2η of $s^\star$. We summarize the details of Algorithm sLSCP in Algorithm 1. For simplicity, we use sLSCP and LSCP indistinguishably.

Algorithm 1: Smoothed local spatial conformal prediction (sLSCP).

Input: observations $z_i = (s_i, x_i, y_i)$, $i = 1, \ldots, n$; predict location $s^\star$; nonconformity measure Δ; significance level α; and a fine grid of candidate response values weight parameter $\eta \in (0, \infty)$; number of neighbors to consider $M \leq n$

Output: $(1-\alpha)100\%$ prediction interval, Γ^α, for $Y(s^\star)$

determine M through η if not given;
form $z_i, i = 1, \ldots, M$, based on M locations closest to $s^\star$;
$s_{M+1} \leftarrow s^\star$;
calculate weights $w_i, i = 1, \ldots, M+1$ as in (12);
for y_{M+1} *in the specified grid* **do**
  for *i = 1 to M+1* **do**
    define $z^{M+1}_{(i)}$ by removing y_i from z^{M+1};
    $\delta_i \leftarrow \Delta(z^{M+1}_{(i)}, y_i)$;
  end
  compute plausibility for y_{M+1} as $p_w(y_{M+1} \mid \cdots)$ in (13);
  include y_{M+1} in Γ^α if $p_w(y_{M+1} \mid \cdots) \geq t_M(\alpha)$;
end
return Γ^α.

It is important for our proposed methods to be computationally feasible for large datasets. Conformal prediction itself is relatively expensive since it requires fitting the underlying model once for each held-out data point being predicted. In particular, the Kriging residuals and the associated nonconformity score computations require us to compute $\hat{\mu}_{n+1,i}(s_i, x_i)$ and $\hat{\sigma}_{n+1,i}(s_i, x_i)$ for each $i = 1, \ldots, n$, which involves n many evaluations of $(\widehat{\beta}, \widehat{\Theta})$. To overcome this computational bottleneck, various adjustments have been considered in the literature. One is the *split conformal prediction* strategy—also called *inductive conformal prediction* in Vovk, Gammerman, and Shafer (2005)—which is common; see, for example, Lei et al. (2018, Sec. 2.2). The idea is to split the data into two parts: one for fitting the underlying model and the other for running conformal prediction with the fitted model from the first part fixed. The theoretical validity of split conformal prediction is now well-known, for example, Section 3 of the Supplementary Materials. Alternatively, as is common in parametric Kriging, one could use the entire dataset to estimate $(\widehat{\beta}, \widehat{\Theta})$ and then use the entire dataset again for prediction with the parameter estimates plugged in as if they were the "true values." Given that the number of parameters in the working spatial model is relatively small, both approaches should perform well for moderate to large n. The simulation results presented in the Supplemental Materials suggest that this plug-in conformal is more efficient than split conformal in terms of width of the corresponding prediction intervals (or, more precisely, in terms of the interval score as defined in Section 5.2), so the numerical results in Sections 5 and 6 below are based on plug-in versions of the proposed GSCP and LSCP algorithms.

5. Simulation Study

5.1. Data Generation

We consider one mean-zero Gaussian stationary process (Scenario 1) and seven non-Gaussian and/or nonstationary data-generating scenarios (Scenarios 2–8). Data are generated based on transformations of a latent Gaussian process $Z(s)$ and a white noise process $E(s)$ with standard normal distribution, where $s = (s_x, s_y) \in [0, 1]^2$. The mean-zero stationary Gaussian process process $Z(s)$ has a Matérn covariance function with variance $\sigma^2 = 3$, range $\phi = 0.1$, and smoothness $\kappa = 0.7$. Data are sampled on the $N \times N$ grid of $n = N^2$ points in the unit square, $s \in \{N^{-1}, 2N^{-1}, \ldots, 1\}^2$, with $N = 20$ or $N = 40$. The scenarios are:

1. $Y(s) = Z(s) + E(s)$;
2. $Y(s) = Z(s)^3 + E(s)$;
3. $Y(s) = q[\Phi\{Z(s)/\sqrt{3}\}] + E(s)$ where Φ is the standard normal distribution function and q is the Gamma$(1, 3^{-1/2})$ quantile function;
4. $Y(s) = \sqrt{3}Z(s)|E(s)|$;
5. $Y(s) = \text{sign}\{Z(s)\}|Z(s)|^{s_x+1} + E(s)$;
6. $Y(s) = \sqrt{\omega(s)/3}\, Z(s) + \sqrt{1-\omega(s)}\, E(s)$ where $\omega(s) = \Phi(\frac{s_x-0.5}{0.1})$;
7. $Y(s) = Z(s) + s_x E(s)$;
8. $Y(s) = Z(s) + 10\exp(-50\|s - c\|^2)$ where $c = (0.5, 0.5)$;

Scenario 1 is Gaussian and stationary, Scenarios 2–4 are stationary but non-Gaussian, and Scenarios 5–8 are nonstationary either in the spatial variance (Scenarios 5 and 6), error variance (Scenario 7), or mean (Scenario 8). Scenario 3 generates skewed data to assess the method's performance when the symmetry of the base Kriging model is violated.

5.2. Prediction Methods and Metrics

For each dataset we apply the global and local (with $\eta = 0.1$) conformal spatial prediction algorithms. For the parametric Kriging method and the initial Kriging predictions of our proposed conformal prediction, we estimate the spatial covariance parameters using empirical variogram methods (Cressie 1992). The empirical variograms are calculated using the `variog` function in the R package `geoR`, and the covariance parameters are chosen to minimize the weighted (by number of observations) squared error between the empirical and model-based variograms.

We compare the proposed conformal prediction methods with standard global Kriging prediction and the local Kriging (laGP) method of Gramacy and Apley (2015) that dynamically defines the support of a Gaussian process predictor based on a local subset of the data. For laGP, we use the function provided by the `laGP` package in R the local sequential design scheme starting from 6 points to 50 points through an empirical Bayes mean-square prediction error criterion.

Methods are trained using a completely randomset of 90% of the observations and tested on the remaining 10%. Each scenario is repeated 100 times, and performance is evaluated using average coverage of $(1 - \alpha)100\%$ prediction intervals, average interval width, and average interval score (Gneiting and Raftery 2007), defined as

$$S_\alpha(I; y^n) = \frac{1}{n}\sum_{i=1}^{n}\big\{(I_u - I_l) + \tfrac{2}{\alpha}(I_l - y_i)_+ + \tfrac{2}{\alpha}(y_i - I_u)_+\big\},$$

where $I = [I_l, I_u]$ is the $100(1 - \alpha)\%$ prediction interval, y^n contains the observations $y_1, \ldots, y_n$, and $z_+ = z \vee 0$ denotes the "positive part." A smaller interval score is desirable as this rewards both high coverage and narrow intervals. We use $\alpha = 0.1$ in this simulation study.

5.3. Results

We present results averaged over data sets and all spatial locations in Table 1. For the nonstationary scenarios varying across s_x (Scenarios 5–7), we present the results by the first spatial coordinate (s_x) averaged over the datasets and the second coordinate (s_y) in Figure 1, e.g., the value of coverage plotted at $s_x = N^{-1}$ is the average of the coverage over the N points of the form (N^{-1}, s_y) for $s_y \in \{N^{-1}, 2N^{-1}, \ldots, 1\}$.

In Scenario 1, the Gaussian and stationary process, the performance of GSCP, LSCP, and Kriging are comparable (Table 1). Kriging performs well in this case since the data generating mechanism aligns with its underlying assumption, but the conformal methods are competitive with the parametric model in terms of both coverage and interval width. In Scenarios 2, 3, and 4, the non-Gaussian but stationary processes, GSCP, LSCP, and Kriging perform more or less the same in terms of interval score and outperform laGP. However, the coverage of the conformal

Table 1. Performance comparison for simulation scenarios ("Scen") without a covariate. The metrics are the empirical coverage of 90% prediction intervals ("Cov90"), the width of prediction intervals ("Width") and the interval score ("IntScore"), each averaged over location and dataset. The methods are global (GSCP) and local (LSCP) conformal prediction, stationary and Gaussian Kriging ("Kriging") and local approximate Gaussian process ("laGP") regression.

		$N = 20$			$N = 40$		
Scen	Method	Cov90	Width	IntScore	Cov90	Width	IntScore
	GSCP	0.906	4.67	5.78	0.897	4.00	5.06
	LSCP	0.890	4.57	5.95	0.891	3.99	5.12
	Kriging	0.912	4.73	5.78	0.888	3.90	5.07
	LaGP	0.877	4.78	6.50	0.879	4.27	5.63
2	GSCP	0.895	33.62	67.16	0.897	22.12	43.80
	LSCP	0.896	31.05	58.37	0.910	21.74	36.47
	Kriging	0.931	44.07	69.38	0.924	27.75	44.80
	LaGP	0.913	40.57	69.38	0.928	31.28	47.24
3	GSCP	0.908	4.79	6.32	0.895	4.05	5.27
	LSCP	0.893	4.65	6.27	0.893	4.03	5.24
	Kriging	0.919	4.95	6.33	0.887	3.96	5.28
	LaGP	0.883	4.92	6.83	0.880	4.29	5.77
4	GSCP	0.902	7.06	11.06	0.895	6.25	10.25
	LSCP	0.892	6.84	11.23	0.895	6.08	9.74
	Kriging	0.918	7.70	11.12	0.908	6.71	10.26
	LaGP	0.898	7.40	11.83	0.901	6.68	10.51
5	GSCP	0.900	6.51	9.40	0.898	5.03	7.11
	LSCP	0.887	6.36	9.06	0.892	5.05	6.75
	Kriging	0.924	7.18	9.52	0.897	5.11	7.15
	LaGP	0.887	7.20	10.42	0.891	5.98	8.08
6	GSCP	0.894	2.78	3.69	0.896	2.63	3.60
	LSCP	0.878	2.66	3.47	0.895	2.38	3.06
	Kriging	0.897	2.78	3.69	0.888	2.54	3.61
	LaGP	0.865	2.58	3.60	0.869	2.32	3.21
7	GSCP	0.905	3.63	4.55	0.896	2.77	3.71
	LSCP	0.888	3.53	4.59	0.896	2.70	3.46
	Kriging	0.915	3.77	4.55	0.889	2.70	3.72
	LaGP	0.869	3.92	5.31	0.881	3.17	4.14
8	GSCP	0.906	3.04	3.74	0.899	1.93	2.45
	LSCP	0.880	3.00	3.91	0.895	1.92	2.46
	Kriging	0.928	3.25	3.78	0.915	2.05	2.48
	LaGP	0.863	3.39	4.64	0.871	2.41	3.20

methods, especially the GSCP algorithm, is closer to the nominal level than Kriging and the Kriging intervals are generally wider than the conformal intervals.

Figure 1 shows the results for nonstationary Scenarios 5 and 6 when $N = 40$. LSCP performs the best among the four methods for these nonstationary scenarios. For Scenario 5, the process is Gaussian to the west and non-Gaussian to the east. The global prediction methods GSCP and Kriging generate prediction intervals with similar width for all s_x (ignoring edge effects), while LSCP and laGP provide wider intervals on the east (s_x near 1) than the west (s_x near 0). LSCP has coverage around 90% for all s_x, and the lowest interval scores, especially in the east where the process is more non-Gaussian. Similarly, in Scenario 6, the correlation is stronger in the east than the west, and the LSCP performs the best by providing adaptive prediction interval width and valid coverage across space.

We also conducted a simulation study when spatial locations are sampled uniformly on $[0, 1]^2$. The performance is very similar to that when locations are fixed at equally-spaced grid points, so we only show the latter in the article. Additional results for the scenarios with covariates (thus a comparison with universal Kriging) and a sensitivity analysis confirming our method's robustness to the estimates of the spatial covariance parameters are included in the Supplemental Materials.

Table 2. Performance comparison for the canopy height data. The metrics are the width, coverage ("Cov90") and interval score ("IntScore") of 90% prediction intervals, each averaged over 10,000 randomly chosen test locations. The methods are local conformal prediction (LSCP), stationary and Gaussian Kriging ("Kriging") and local approximate Gaussian process ("laGP") regression.

	Width	Cov90	IntScore
LSCP	2.87	87.9%	4.33
Kriging	5.44	96.6%	6.81
laGP	5.04	91.1%	6.63

6. Real Data Analysis

This section demonstrates the performance of conformal prediction method using the canopy height data in Figure 2a. The data were originally presented in Cook et al. (2013) and were analyzed using a nearest-neighbor Gaussian process model in Datta et al. (2016). The data are available in the R package `spNNGP` (Finley, Datta, and Banerjee 2017). There are $n = 1,723,137$ observations and clear nonstationarity and non-normality. For example, there are several heterogeneous areas with small height canopies around the location with longitude and latitude being 729,000 and 470,000, respectively.

We compare methods using 90% prediction intervals for 10,000 test locations chosen completely at random from the full dataset. Since the data clearly exhibit nonstationarity we do not apply GSCP. We select the kernel function and bandwidth parameter using cross-validation over the validation locations. The average interval score is consistently smaller for the Gaussian kernel (ranging between 4.3 and 4.4 by η) than the uniform kernel (ranging between 4.9 and 5.2 by η) and minimized by the Gaussian kernel with $\eta = 6 \times 10^{-4}$. Table 2 compares the performance of LSCP on the 10,000 test locations with Kriging and laGP. LSCP outperforms the other methods as the empirical coverage of LSCP is the closet to the desired 90% and the LSCP minimizes the interval score.

Figure 2 plots the interval widths for each method. Unlike Kriging, the LSCP and laGP interval widths are locally adaptive with wider intervals in heterogeneous areas and more narrow intervals in homogeneous areas. Comparing LSCP and laGP, LSCP generally provides narrower intervals than laGP, which means the proposed method is more efficient than laGP. In addition, the locations of the observations that fall outside the prediction intervals are uniformly distributed for LSCP and laGP, but clustered in the heterogeneous areas for Kriging. In short, the proposed spatial conformal prediction algorithm shows its superiority in this real data analysis.

7. Discussion

In this article, we proposed a spatial conformal prediction algorithm to provide valid, robust, and model-free prediction intervals by combining spatial methods and the classical conformal prediction. We provided both global and local versions to accommodate different stationarity cases and sampling designs.

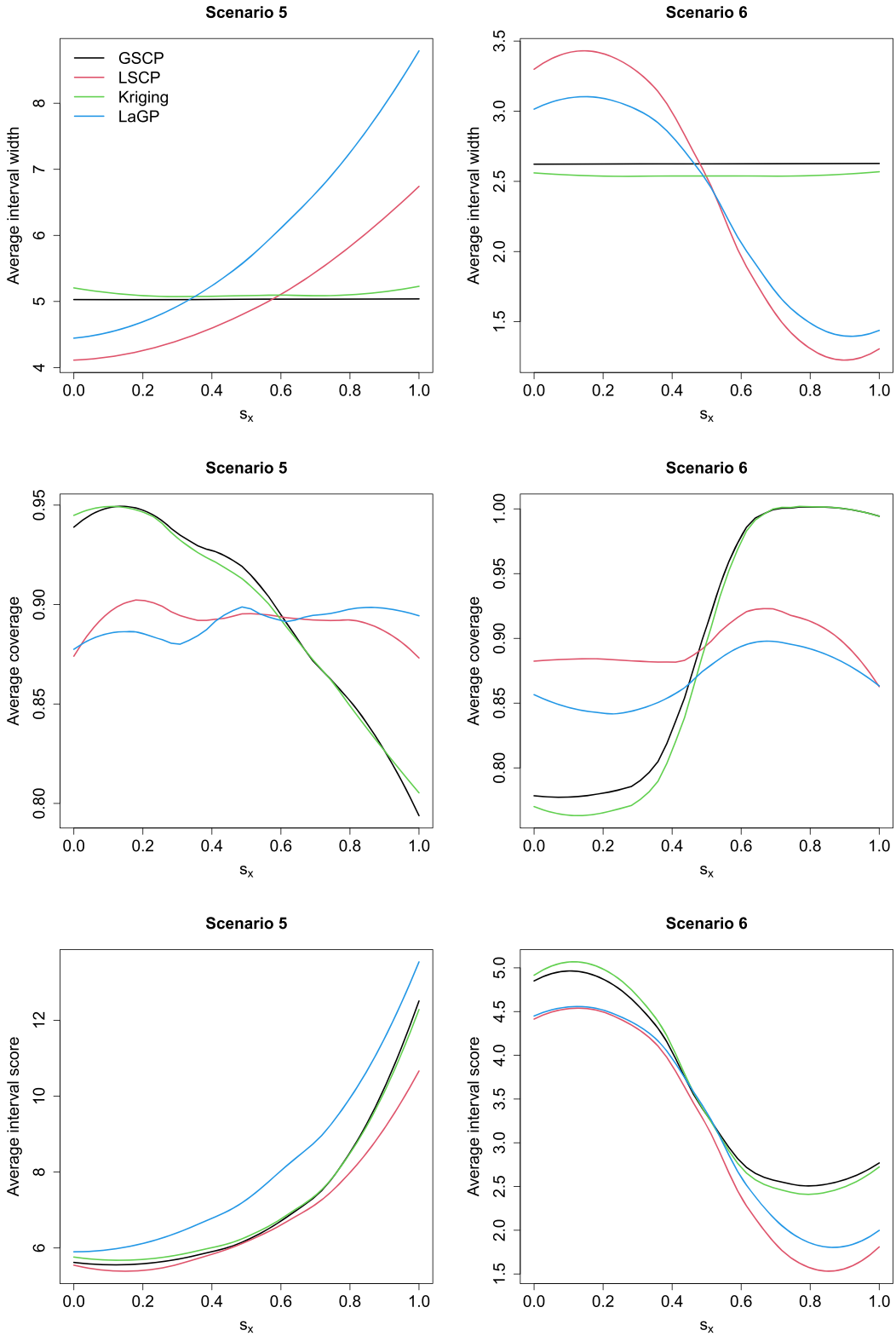

Figure 1. Performance comparison by s_x for Scenario 5: $Y(s) = \text{sign}\{Z(s)\} \cdot |Z(s)|^{s_x+1} + E(s)$ and Scenario 6: $Y(s) = \sqrt{\omega(s)/3} \cdot Z(s) + \sqrt{1-\omega(s)} \cdot E(s)$ where $\omega(s) = \Phi(\frac{s_x-0.5}{0.1})$ when $N = 40$ (results are smoothed over s_x for clarity).

We proved their validity under various sampling designs and data-generating mechanisms. To the authors' knowledge, this work is among the first in making the classical conformal algorithm work for nonexchangeable data. Our simulation studies and real data analyses demonstrate the advantage of the proposed spatial conformal prediction algorithms. We also developed an R package entitled `scp` (*https://github.com/mhuiying/scp*) to compute the plausibility contours and generate spatial prediction intervals using either Kriging residual or any other user-defined nonconformity measure.

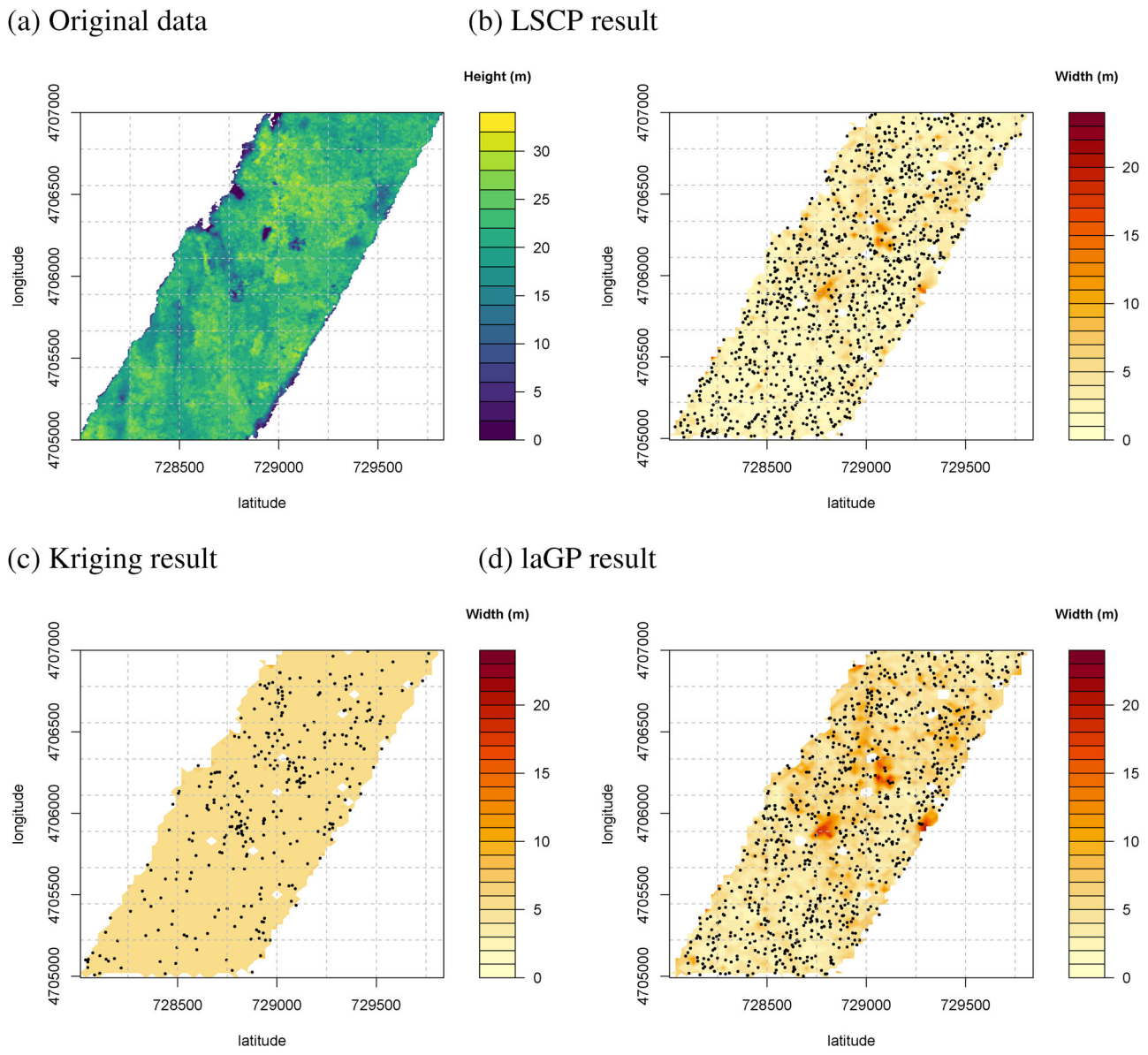

Figure 2. (a) Heatmap of the canopy height data; (b)–(d) Prediction interval width (color) and locations not covered (points) for LSCP, Kriging, and laGP. Longitude and latitude are in UTM Zone 18.

An attractive feature of the proposed algorithms is that they are model-free in the sense that their theoretical validity does not depend on correct specification of a model. In our implementations we use the squared residuals from a simple parametric model to define the nonconformity terms. However, as anticipated by our theoretical results, our simulation study shows that the methods work well even if the parametric family or the mean and covariance functions are misspecified, and that the results are insensitive to inaccurate estimation of the parameters in the parametric model. This robustness allows the methods to be applied broadly and with confidence.

The local conformal prediction method relies on a dense grid of points around each prediction location and thus a large dataset. Computation time often prohibits application of spatial methods to large datasets. Fortunately, we are able to apply our method to large datasets by exploiting local algorithms and an explicit formula for the plausibility contour using the Kriging-based nonconformity score. Similar derivations are needed for prediction procedures other than Kriging in order to maintain computational efficiency.

Future research directions include extending the work to spatial processes with discrete observations, e.g., when $Y(s)$ is binary or a count. Generalized spatial linear models are compatible with our current framework (see the Appendix), but continuity in the distribution function is required. Therefore, further studies would be required to establish the validity of conformal prediction for discrete data. Another limitation of the proposed algorithms is that they only produce intervals for a single location. Generalizing the algorithms to produce joint intervals for multiple locations would be useful in some applications. One option is to use a Bonferroni correction; of course, this may be inefficient for many simultaneous predictions, but greater efficiency would require model assumptions to link the multiple locations and facilitate information sharing. It would also be of interest to extend the proposed spatial conformal prediction methods to spatiotemporal data, perhaps building on

recent work for time series data (Xu and Xie 2020; Zaffran et al. 2022).

Appendix

A.1: Conditional validity on a sphere

An obstacle that prevents a conditional validity result in the existing literature is an "edge effect." That is, conditional validity is typically achieved at targets in the middle of the domain, but fails at targets in the extremes; see Figure 1(b) in Lei and Wasserman (2014). So if it were possible to eliminate the edge effect—even if in a trivial way, by eliminating the edge itself—then there is hope for establishing a conditional validity result. In our spatial context, but perhaps not in other cases, it may not be unreasonable to assume that the spatial locations are sampled iid from a uniform distribution on a sphere. Since the sphere has no edges and a uniform distribution has no extremes, there is no "edge effect" preventing conditional validity. Some additional structure in the (X, Y) process is also needed here, in particular, it should be isotropic in the sense that the correlation structure only depends on the distance between spatial locations. Note that, if the mean of the parametric base model is correctly specified, so that the conditional distribution of $Y - X\beta$, given X, is free of X, then the stationarity assumption about X can be removed.

To our knowledge, Proposition 2 below gives the first finite-sample conditional validity result for conformal prediction in the literature, albeit under rather strong conditions.

Proposition 2. Let (X, Y) be an isotropic stationary process over the sphere $\mathcal{D} = \{s \in \mathbb{R}^3 : \|s\| = 1\}$, and suppose that the locations $S_1, \ldots, S_n, S_{n+1}$ are independent and uniformly distributed on $\mathcal{D}$. For Γ^α as described above, define the conditional coverage probability function

$$c(s^\star \mid \alpha, n, \mathsf{P}) = \mathsf{P}^{n+1}\{\Gamma^\alpha(Z^n; S_{n+1}, X_{n+1}) \ni Y_{n+1} \mid S_{n+1} = s^\star\}.$$

Then the GSCP-based predictions are conditionally valid, that is, $c(s^\star \mid \alpha, n, \mathsf{P}) \geq 1 - \alpha$ for all $(\alpha, n, s^\star)$ and all P under which (X, Y) is stationary and isotropic and S are iid uniform on $\mathcal{D}$.

A.2: Locally-exchangeable processes

To better understand how the locally-exchangeable processes in Section 4.2 relate to our prediction problem, consider a simple case with no covariates, where $\{Y(s) : s \in \mathcal{D}\}$ is the only random process under consideration. In that case, we want to show that $T(s) = Y(s)$ has this local exchangeability property. Then the sufficient condition (11) above amounts to assuming there exists a suitable real-valued function ψ_Y, along with appropriate processes L_Y and E_Y, such that

$$Y(s) = \psi_Y\left(L_Y(s), E_Y(s)\right), \quad s \in \mathcal{D}.$$

There are a number of common models for continuous responses that meet this condition, including the additive model in Section 2.1, certain generalized spatial linear models Diggle, Tawn, and Moyeed (1998), spatial copula models (Krupskii and Genton 2019), and max-stable processes (Reich and Shaby 2012). For example, in a generalized spatial linear model, with suitable spatial process L_Y and Gaussian white noise E_Y, take

$$\psi_Y(\ell, e) = H^{-1}_{g(\ell)}\left(\Phi^{-1}(e)\right), \tag{A.1}$$

where H_ξ is the distribution function for an exponential family with natural parameter ξ, g is the link function, and Φ is the standard normal distribution function. Of course, to meet the continuity requirement, the exponential family must have a density with respect to Lebesgue measure.

For the practically relevant case with both a response Y and covariate X process, the idea is similar but the notation is more complicated. The goal is to find conditions under which the joint process $T(s) = (X(s), Y(s))$ has a representation as in (11). Admittedly, it is challenging to consider the joint process directly, which is why we aim to give a simpler sufficient condition based on the marginal distribution of X and the conditional distribution of Y, given X. Consider the following decomposition:

$$\begin{aligned} X(s) &= \psi_X\left(L_X(s), E_X(s)\right) \\ Y(s) &= \psi_Y\left(L_Y(s), E_Y(s) \mid X(s)\right). \end{aligned} \tag{A.2}$$

Roughly, this amounts to assuming that each of X and Y has a decomposition like that described above for Y alone. Individual assessments of the distributional properties of X and Y are more manageable than directly considering their joint distribution. And as the following simple lemma states, separate considerations of its marginal and conditional structure suffice to establish a decomposition of the joint structure.

Lemma 2. If (X, Y) can be decomposed as in (A.2), if the pairs of processes (L_X, L_Y) and (E_X, E_Y) in are individually L_2-continuous and locally iid, respectively, and if $(\ell, e) \mapsto \psi_X(\ell, e)$ and $(\ell, e, x) \mapsto \psi_Y(\ell, e \mid x)$ are both continuous, then $T(s) = (X(s), Y(s))$ satisfies the conditions of Proposition 1.

As we discussed above, the decomposition of the X marginal as in (A.2) is quite flexible. For example, it is quite common that X could be expressed as $X = L_X + E_X$ for a spatial and non-spatial components, L_X and E_X, respectively, but the additive form is not necessary. And just like in our discussion above Proposition 2 above, if the mean of the parametric base model is correctly specified, then these assumptions about X here can be dropped. Similarly, if the decomposition of Y in the response-only model was flexible, then the corresponding conditional decomposition in (A.2) must be equally flexible. For example, the same generalized spatial linear model can be considered, but now it is allowed to depend smoothly on the covariate.

Acknowledgments

The authors thank the reviewers for their thoughtful and critical comments on a previous version of the manuscript. HM (DMS–1638521) and RM (DMS–1811802) are supported by the National Science Foundation. BJR is supported by the National Institutes of Health (R01ES031651 and R01ES027892) and King Abdullah University of Science and Technology (3800.2).

References

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019), "The Limits of Distribution-Free Conditional Predictive Inference," arXiv preprint arXiv:1903.04684. [2,4]

Cella, L. and Martin, R. (2022), "Validity, Consonant Plausibility Measures, and Conformal Prediction," *International Journal of Approximate Reasoning*, 141, 110–130. [3]

Chernozukov, V., Wüthrich, K., and Zhu, Y. (2021), "Distributional Conformal Prediction," *Proceedings of the National Academy of Sciences*, 118, e2107794118. [2,4]

Cook, B., Nelson, R., Middleton, E., Morton, D., McCorkel, J., Masek, J., Ranson, K., Ly, V., Montesano, P., et al. (2013), "Nasa Goddard's Lidar, Hyperspectral and Thermal (g-liht) Airborne Imager," *Remote Sensing*, 5, 4045–4066. [7]

Cressie, N. (1992), "Statistics for Spatial Data," *Terra Nova*, 4, 613–617. [1,4,6]

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association*, 111, 800–812. [7]

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics," *Journal of the Royal Statistical Society*, Series C, 47, 299–350. [10]

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007), "Generalized Spatial Dirichlet Process Models," *Biometrika*, 94, 809–825. [1]

Finley, A., Datta, A., and Banerjee, S. (2017), "spNNGP: Spatial Regression Models for Large Datasets Using Nearest Neighbor Gaussian Processes," R package version 0.1.0, 1. [7]

Franchi, G., Yao, A., and Kolb, A. (2018), "Supervised Deep Kriging for Single-Image Super-Resolution," in *German Conference on Pattern Recognition*, pages 638–649. Springer. [1,3]

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2015), "Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields?" *Spatial Statistics*, 14, 505–531. [1]

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035. [1]

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [6]

Gramacy, R. B. and Apley, D. W. (2015), "Local Gaussian Process Approximation for Large Computer Experiments," *Journal of Computational and Graphical Statistics*, 24, 561–578. [2,6]

Guan, L. (2019), "Conformal Prediction with Localization," arXiv preprint arXiv:1908.08558. [1]

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2019), "A Case Study Competition among Methods for Analyzing Large Spatial Data," *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425. [1]

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B. (2018), "Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables," *PeerJ*, 6, e5518. [1]

Henley, S. (2012), *Nonparametric Geostatistics*, Springer Science & Business Media. [3]

Kim, J., Kwon Lee, J., and Mu Lee, K. (2016a), "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654. [1]

Kim, J., Kwon Lee, J., and Mu Lee, K. (2016b), "Deeply-Recursive Convolutional Network for Image Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645. [1]

Krupskii, P. and Genton, M. G. (2019), "A Copula Model for Non-Gaussian Multivariate Spatial Data," *Journal of Multivariate Analysis*, 169, 264–277. [10]

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111. [1,4,6]

Lei, J. and Wasserman, L. (2014), "Distribution-Free Prediction Bands for Non-Parametric Regression," *Journal of the Royal Statistical Society*, Series B, 76, 71–96. [1,4,10]

Li, Y., Sun, Y., and Reich, B. J. (2020), "Deepkriging: Spatially Dependent Deep Neural Networks for Spatial Prediction," arXiv preprint arXiv:2007.11972. [1]

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017), "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144. [1]

Reich, B. J. and Fuentes, M. (2007), "A Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields," *The Annals of Applied Statistics*, 1, 249–264. [1]

Reich, B. J. and Shaby, B. A. (2012), "A Hierarchical Max-Stable Spatial Model for Extreme Precipitation," *The Annals of Applied Statistics*, 6, 1430. [10]

Risser, M. D. (2016), "Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate–Driven Approaches," arXiv preprint arXiv:1610.02447. [1]

Rodriguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian Models through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6. [1]

Romano, Y., Patterson, E., and Candes, E. (2019), "Conformalized Quantile Regression," in *Advances in Neural Information Processing Systems*, pages 3538–3548. [1]

Shafer, G. and Vovk, V. (2008), "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, 9, 371–421. [1,2,3]

Stein, M. L. (1990), "A Comparison of Generalized Cross Validation and Modified Maximum Likelihood for Estimating the Parameters of a Stochastic Process," *The Annals of Statistics*, 18, 1139–1157. [4]

Stein, M. L. (2002), "The Screening Effect in Kriging," *The Annals of Statistics*, 30, 298–323. [5]

Tai, Y., Yang, J., and Liu, X. (2017), "Image Super-Resolution Via Deep Recursive Residual Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3147–3155. [1]

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019), "Conformal Prediction Under Covariate Shift," in Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 2530–2540. Curran Associates, Inc. [1,4]

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, Springer Science & Business Media. [1,2,6]

Wang, H., Guan, Y., and Reich, B. (2019), "Nearest-Neighbor Neural Networks for Geostatistics," in *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 196–205. IEEE. [1]

Xu, C. and Xie, Y. (2020), "Conformal Prediction for Dynamic Time-Series," arXiv preprint arXiv:2010.09107. [10]

Zaffran, M., Dieuleveut, A., Féron, O., Goude, Y., and Josse, J. (2022), "Adaptive Conformal Predictions for Time Series," arXiv preprint arXiv:2202.07282. [10]