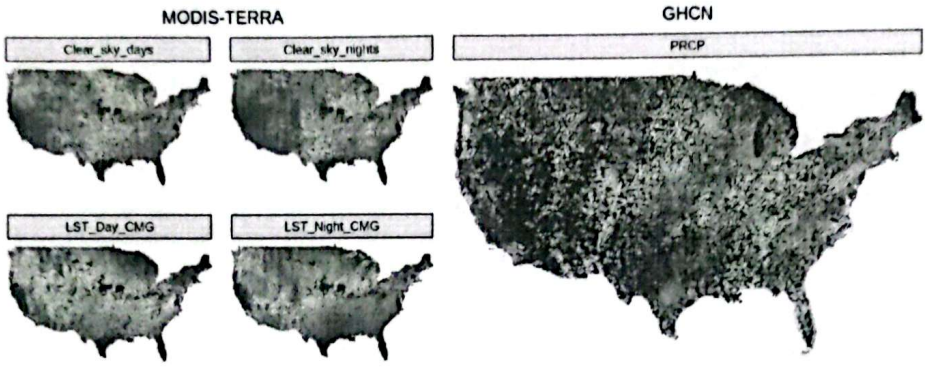

Figure 1: Observed data of Section 4.2. Missing outcomes are in magenta. GHCN data are much more sparsely observed compared to satellite imaging from MODIS.

2021). These methods may be unavailable or perform poorly in geostatistical settings, which focus on small-dimensional inputs, i.e. the spatial coordinates plus time. In these scenarios, low-rank methods oversmooth the spatial surface (Banerjee et al., 2010), Toeplitz-like structures are typically absent, and so-called *separable* covariance functions obtained via tensor products poorly characterize spatial and temporal dependence. To overcome these hurdles, one can use covariance tapering and domain partitioning (Furrer et al., 2006; Kaufman et al., 2008; Sang and Huang, 2012; Stein, 2014; Katzfuss, 2017) or composite likelihood methods and sparse precison matrix approximations (Vecchia, 1988; Rue and Held, 2005; Eidsvik et al., 2014); refer to Sun et al. (2011), Banerjee (2017), Heaton et al. (2019) for reviews of scalable geostatistical methods.

Additional difficulties arise in multivariate (or multi-output) regression settings. Multivariate geostatistical data are commonly misaligned, i.e. observed at non-overlapping spatial locations (Gelfand et al., 2010). Figure 1 shows several variables measured at non-overlapping locations, with one measurement grid considerably sparser than the others. In these settings, replacing a multi-output regression with separate single-output models is a valid option for predicting outcomes at new locations. While single-output models may sometimes perform equally well or even outperform multi-output models, they fail to characterize and estimate cross-dependences across outputs; testing the existence of such dependences may be scientifically more impactful than making predictions. This issue can be solved by modeling the outputs via latent spatial random effects thought of as a realization of an underlying multivariate GP and embedded in a larger hierarchical model.

Unfortunately, GP approximations that do not correspond to a valid stochastic process may inaccurately characterize uncertainty, as the models used for estimation and interpolation may not coincide. Rather than seeking approximations to the full GP, one can develop valid standalone spatial processes by introducing conditional independence across spatial locations as prescribed by a sparse directed acyclic graph (DAG). These models are advantageous because they lead to scalability by construction; in other words, posterior computing algorithms for these methods can be interpreted not only as approximate algorithms for the full GP, but also as exact algorithms for the standalone process.

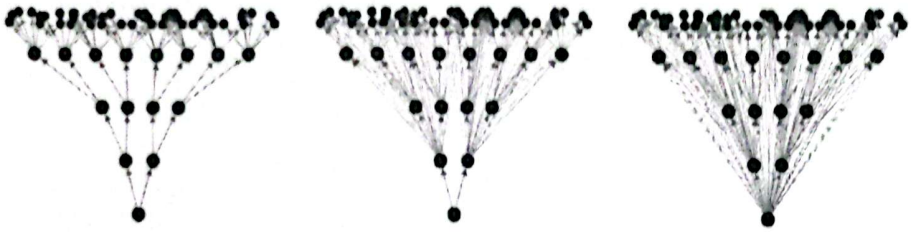

Figure 2: Three SPAMTREES on $M = 4$ levels with depths $\delta = 1$ (left), $\delta = 3$ (center), and $\delta = 4$ (right). Nodes are represented by circles, with branches colored in brown and leaves in green.

than methods based on recursive partitioning and can be built to guarantee similarly-sized conditioning sets at all locations.

The present work adds to the growing literature on spatial processes defined on DAGs by developing a method that targets efficient computations of Bayesian multivariate spatial regression models. SPAMTREES share similarities with MRAs (Katzfuss, 2017); however, while MRAs are defined as a basis function expansion, they can be represented by a treed graph of a SPAMTREE with full "depth" as defined later (the DAG on the right of Figure 2), in univariate settings, and "response" models. All these restrictions are relaxed in this article. In considering spatial proximity to add "leaves" to our treed graph, our methodology also borrows from nearest-neighbor methods (Datta et al., 2016a). However, while we use spatial neighbors to populate the conditioning sets for non-reference locations, the same cannot be said about reference locations for which the treed graph is used instead. Our construction of the SPAMTREE process also borrows from MGPs on tessellated domains (Peruzzi et al., 2020); however, the treed DAG we consider here induces markedly different properties on the resulting spatial process owing to its recursive nature. Finally, a contribution of this article is in developing self-contained sampling algorithms which, based on the graphical model representation of the model, will not require any external libraries.

The article builds SPAMTREES as a standalone process based on a DAG representation in Section 2. A Gaussian base process is considered in Section 3 and the resulting properties outlined, along with sampling algorithms. Simulated data and real-world applications are in Section 4; we conclude with a discussion in Section 5. The Appendix provides more in-depth treatment of several topics and additional algorithms.

2. Spatial Multivariate Trees

Consider a spatial or spatiotemporal domain $\mathcal{D}$. With the temporal dimension, we have $\mathcal{D} \subset \Re^d \times [0, \infty)$, otherwise $\mathcal{D} \subset \Re^d$. A q-variate spatial process is defined as an uncountable set of random variables $\{w(\ell) : \ell \in \mathcal{D}\}$, where $w(\ell)$ is a $q \times 1$ random vector with elements $w_i(\ell)$ for $i = 1, 2, \ldots, q$, paired with a probability law P defining the joint distribution of any finite sample from that set. Let $\{\ell_1, \ell_2, \ldots, \ell_{n_\mathcal{L}}\} = \mathcal{L} \subset \mathcal{D}$ be of size $n_\mathcal{L}$. The $n_\mathcal{L} q \times 1$ random

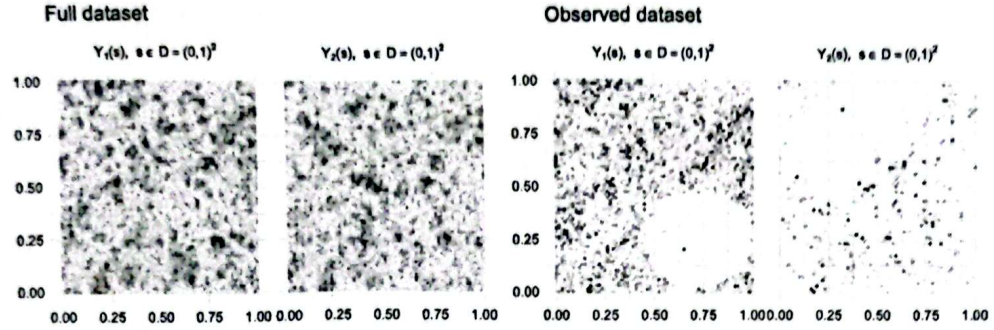

Figure 3: Left half: *Full data set* – a bivariate outcome is generated on 4,900 spatial locations. Right half: *Observed data set* – the training sample is built via independent subsampling of each outcome.

4. Applications

We consider Gaussian SPAMTREES for the multivariate regression model (15). Consider the spatial locations $\ell, \ell' \in \mathcal{D}$ and the locations of variables i and j in the latent domain of variables $\xi_i, \xi_j \in \Xi$, then denote $h = \|\ell - \ell'\|$, $\Delta = \delta_{ij} = \|\xi_i - \xi_j\|$, and

$$C(h, \Delta) = \frac{\exp\left\{-\phi\|h\| / \exp\left\{\frac{1}{2}\beta \log(1+\alpha\Delta)\right\}\right\}}{\exp\left\{\beta \log(1+\alpha\Delta)\right\}}.$$

For $j = 1, \ldots, q$ we also introduce $C_j(h) = \exp\{-\phi_j\|h\|\}$. A non-separable cross-covariance function for a multivariate process can be defined as

$$\text{Cov}(w(\ell, \xi_i), w(\ell', \xi_j)) = C_{ij}(h) = \begin{cases} \sigma_{i1}^2 C(h, \delta_{ij}) + \sigma_{i2}^2 C_i(h) & \text{if } i = j \\ \sigma_{i1}\sigma_{j1} C(h, \delta_{ij}) & \text{if } i \neq j, \end{cases} \tag{18}$$

which is derived from eq. (7) of Apanasovich and Genton (2010); locations of variables in the latent domain are unknown, therefore $\theta = \{\sigma_{i1}, \sigma_{i2}, \phi_i\}_{i=1,\ldots,q} \cup \{\delta_{ij}\}_{i=1,\ldots,q}^{j<i} \cup \{\alpha, \beta, \phi\}$ for a total of $3q + q(q-1)/2 + 3$ unknown parameters.

4.1 Synthetic Data

In this section we focus on bivariate outcomes ($q = 2$). We simulate data from model (15), setting $\beta = 0$, $Z = I_q$ and take the measurement locations on a regular grid of size 70×70 for a total of 4,900 spatial locations. We simulate the bivariate spatial field by sampling from the full GP using (18) as cross-covariance function; the nuggets for the two outcomes are set to $\tau_1^2 = 0.01$ and $\tau_2^2 = 0.1$. For $j = 1, 2$ we fix $\sigma_{j2} = 1, \alpha = 1, \beta = 1$ and independently sample $\sigma_{j1} \sim U(-3, 3), \phi_j \sim U(0.1, 3), \phi \sim U(0.1, 30), \delta_{12} \sim Exp(1)$, generating a total of 500 bivariate data sets. This setup leads to empirical spatial correlations between the two outcomes smaller than 0.25, between 0.25 and 0.75, and larger than 0.75 in absolute value in 107, 330, and 63 of the 500 data sets, respectively. We introduce misalignment and make the outcomes imbalanced by replacing the first outcome with missing values at $\approx 50\%$ of the spatial locations chosen uniformly at random, and then repeating this procedure for the

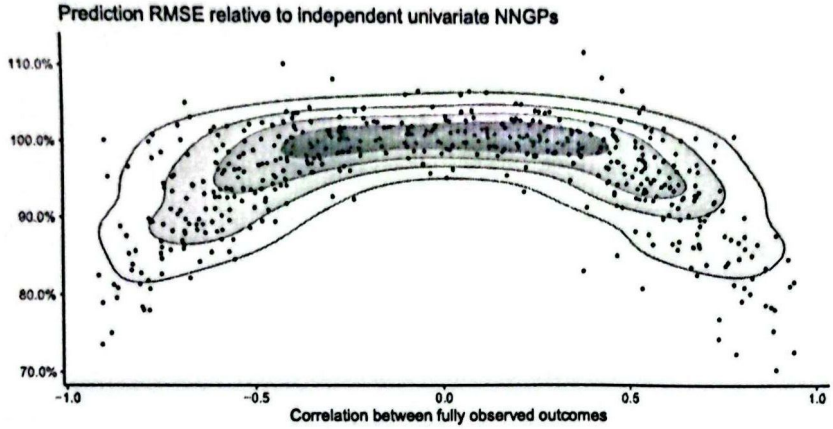

Figure 4: Predictive RMSE of the best-performing SPAMTREE of Table 1 relative to independent univariate NNGP models of the two outcomes, for different empirical correlations between the two outcomes in the full data. Lower values indicate smaller errors of SPAMTREES in predictions.

C.2.1. Finally, we show in Figure 4 that the relative gains of SPAMTREES compared to independent univariate NNGP model of the outcomes are increasing with the magnitude of the correlations between the two outcomes, which are only available due to the simulated nature of the data sets.

	All outcomes at ℓ	Cherry pick same outcome	Root bias	RMSE($\boldsymbol{y}$)	MAE($\boldsymbol{y}$)	COVG($\boldsymbol{y}$)	RMSE($\boldsymbol{\theta}$)
SPAMTREES $\delta = M$	No	No	No	1.078	0.795	0.955	4.168
	No	No	Yes	**1.065**	**0.786**	0.955	4.138
	No	Yes	No	1.083	0.799	0.954	4.168
	No	Yes	Yes	1.085	0.799	0.954	4.138
	Yes	Yes	No	1.081	0.797	0.954	**4.080**
	Yes	Yes	Yes	1.087	0.801	0.954	4.188
SPAMTREES $\delta = 1$	Yes	Yes	No	1.198	0.880	0.956	4.221
Q-MGP	Yes	–	–	1.125	0.819	**0.951**	4.389
IND-PART	Yes	–	–	1.624	1.229	0.948	8.064
LOWRANK	Yes	–	–	1.552	1.173	0.952	5.647
SPDE-INLA	Yes	–	–	1.152	0.862	0.913	
SPAMTREES Univariate	–	–	–	1.147	0.846	0.953	
NNGP Univariate	–	–	–	1.129	0.832	0.952	
BART	–	–	–	1.375	1.036	0.488	

Table 1: Prediction and estimation performance on multivariate synthetic data. The four columns on the right refer to root mean square error (RMSE) and mean absolute error (MAE) in out-of-sample predictions, average coverage of empirical 95% prediction intervals, and RMSE in the estimation of $\boldsymbol{\theta}$.

4.2 Climate Data: MODIS-TERRA and GHCN

Climate data are collected from multiple sources in large quantities; when originating from satellites and remote sensing, they are typically collected at high spatial and relatively

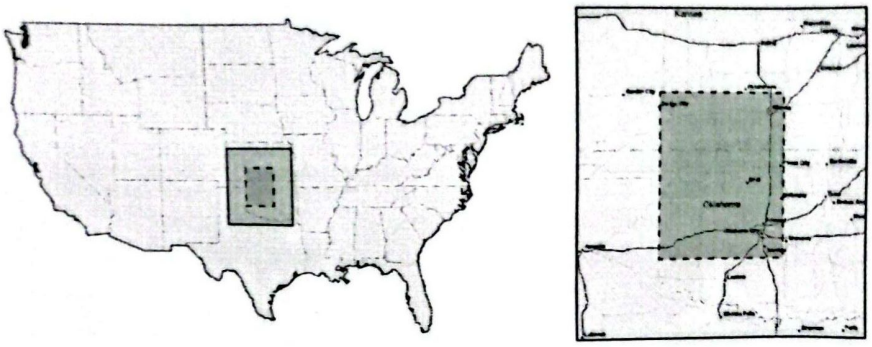

Figure 5: Prediction area

MODIS/GHCN variables		*Multivariate*		*Univariate*	
		SPAMTREE	Q-MGP	SPAMTREE	NNGP
Clear_sky_days	RMSE	1.611	1.928	1.466	1.825
	COVG	0.980	0.866	0.984	0.986
Clear_sky_nights	RMSE	1.621	1.766	2.002	2.216
	COVG	0.989	0.943	0.992	0.992
LST_Day_CMG	RMSE	1.255	1.699	1.645	1.666
	COVG	1.000	1.000	1.000	1.000
LST_Night_CMG	RMSE	1.076	1.402	0.795	1.352
	COVG	0.999	0.999	1.000	1.000
PRCP	RMSE	0.517	0.632	0.490	0.497
	COVG	0.972	1.000	0.969	0.958

Table 2: Prediction results over the 3×3 degree area shown in Figure 5

chains to 15,000 for a total compute time of less than 7 hours for both models. We provide additional details about the models we implemented at Appendix C.

Table 2 reports predictive performance of all models, and Figure 6 maps the predictions at all locations from SPAMTREES and the corresponding posterior uncertainties. Multivariate models appear advantageous in predicting some, but not all outcomes in this real world illustration; nevertheless, SPAMTREES outperformed a Q-MGP model using the same cross-covariance function. Univariate models perform well and remain valid for predictions, but cannot estimate multivariate relationships. We report posterior summaries of θ in Appendix C.2.2. Opposite signs of σ_{i1} and σ_{j1} for pairs of variables $i, j \in \{1, \ldots, q\}$ imply a negative relationship; however, the degree of spatial decay of these correlations is different for each pair as prescribed by the latent distances in the domain of variables δ_{ij}. Figure 7 depicts the resulting cross-covariance function for three pairs of variables.

5. Discussion

In this article, we introduced SPAMTREES for Bayesian spatial multivariate regression modeling and provided algorithms for scalable estimation and prediction. SPAMTREES add

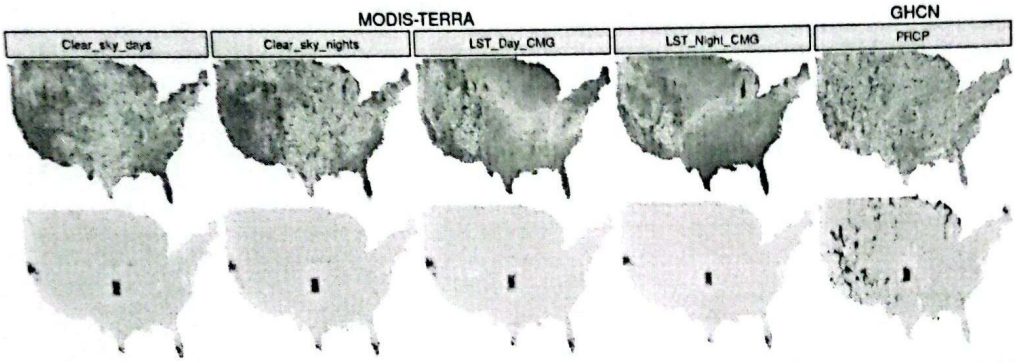

Figure 6: Predicted values of the outcomes at all locations (top row) and associated 95% uncertainty (bottom row), with darker spots corresponding to wider credible intervals.

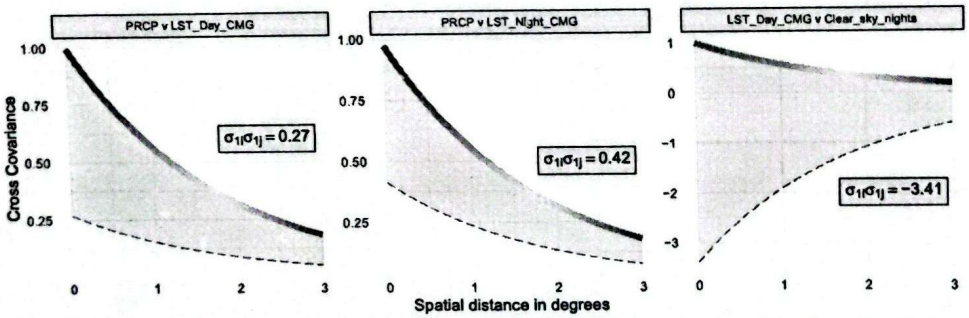

Figure 7: Given the latent dimensions δ_{ij}, the color-coded lines represent $C(\boldsymbol{h}, \delta_{ij})$ whereas $\boldsymbol{C_{ij}(h)} = \sigma_{1i}\sigma_{1j}C(\boldsymbol{h}, \delta_{ij})$ is shown as a dashed grey line.