Bag of DAGs: Flexible Nonstationary Modeling of Spatiotemporal Dependence

Bora Jin\textsuperscript{1,*,} Michele Peruzzi\textsuperscript{1}, and David B. Dunson\textsuperscript{1}

\textsuperscript{1}Department of Statistical Science, Duke University
*email: bora.jin@duke.edu

Abstract

We propose a class of nonstationary processes that characterize varying directional associations in space and time for point-referenced data. Our construction is based on local mixtures of sparse directed acyclic graphs (DAGs). In stochastically choosing DAG edges from a “bag,” we account for uncertainty in directional correlation patterns across a domain. The resulting “bag of DAGs” processes (BAGs) lead to interpretable nonstationarity and scalability for large data due to sparsity of all DAGs in the bag.

We are motivated by spatiotemporal modeling of air pollutants in which a directed edge in a DAG represents a prevailing wind direction causing some associated covariance in the pollutants. We outline Bayesian hierarchical models embedding the resulting nonstationary BAGs and illustrate inferential and performance gains of our methods compared to existing alternatives. We analyze fine particulate matter (PM2.5) in California with high-resolution data from low-cost air quality sensors. The code for all analyses is publicly available at https://github.com/jinbora0720/BAG.

Keywords: Air pollution, Bayesian geostatistics, Directed acyclic graphs, Gaussian process, Particulate matter, Scalability

1 Introduction

In the spread of aqueous or air pollutants, local dynamics of currents or winds influence the strength of spatial and temporal correlations. Figure 1 (left) illustrates an incident in California (CA) in which smoke from wildfires exhibits local patterns of spread in different directions. Realistic models of pollutant spread should allow correlations between spatial locations \( s \) and \( s' \) to peak if one is downwind of the other after a certain temporal lag. In principle, one can construct rich models for any point-referenced data including pollutants via Gaussian processes (GPs). Let \( w(s) \sim GP(0, C(\cdot, \cdot | \theta)) \) denote a zero-mean univariate
GP where \( s \in \mathcal{D} \subseteq \mathbb{R}^d \) is a location in a region of interest \( \mathcal{D} \), and \( d \) is the dimension of coordinates (e.g., \( d = 2 \) if \( s = \) (longitude, latitude)). A covariance function \( C(\cdot, \cdot | \theta) : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R} \) specifies associations across locations, indexed by a parameters \( \theta \). Then, finite realizations of \( w(\cdot) \) at an arbitrary set of locations \( \mathcal{S} = \{s_1, \ldots, s_k\} \) follow a multivariate normal distribution with mean zero and a covariance matrix \( C_{\mathcal{S}} \) whose \((i, j)\) element is \( C(s_i, s_j | \theta) = \text{Cov}(w(s_i), w(s_j)) \).

Customarily GPs assume a stationary covariance function such that \( C(s, s + h) = C(s', s' + h) = C(h) \) for any \( s, s' \in \mathcal{D} \), omitting \( \theta \) for brevity. The stationarity assumption, however, is overly restrictive in our applications of interest because it assigns the same correlation to any pairs of locations with the same lag \( h \) without regard to their respective and relative positions. To overcome the limitation, GPs can rely on nonstationary covariance functions (see Sampson 2010). Moreover, when spatially- and/or temporally-varying covariates are available, one can explicitly model covariance parameters using the covariates (Calder, 2008; Schmidt et al., 2011; Neto et al., 2014; Risser and Calder, 2015). Calder (2008) and Neto et al. (2014) use wind direction data to characterize nonstationarity in air pollutants. However, reliable covariate data (e.g., wind directions) that are recorded at a sufficiently high spatial and/or temporal resolution are seldom available. In this paper,
we aim to construct nonstationary processes that do not require nonstationarity-informing covariates but take into account effects of winds or currents.

Another hurdle in nonstationary GPs is poor scalability to large data. Some computationally efficient nonstationary GP models use the basis graphical lasso (Krock et al., 2021a,b) or fit locally stationary processes for different regions of a domain, allowing for different covariance parameters (Fuentes, 2001; Kim et al., 2005; Lindgren et al., 2011; Kleiber and Nychka, 2012). Several extensions (Datta et al., 2016a; Peruzzi et al., 2020; Katzfuss and Guinness, 2021) of Vecchia (1988) approximation of stationary GPs also induce nonstationary GPs. Vecchia-style approximations achieve scalability by making spatial conditional independence assumptions across locations or knots. Each knot is associated to a node in a sparse directed acyclic graphs (DAG), and the edges pointing to it populate its parent or conditioning set. One can construct a Nearest-Neighbor GP (NNGP; Datta et al. 2016a) by restricting the conditioning set to include a few closest neighbors by Euclidean distance or a Meshed GP (MGP; Peruzzi et al. 2020) by fixing a patterned DAG over domain partitions. Further generalizations are available in Katzfuss and Guinness (2021). DAG-based methods are appealing because they can be implemented with any covariance functions, can lead to standalone processes that facilitate predictions at arbitrary locations, and can be embedded seamlessly within Bayesian hierarchies. Multiple extensions have been proposed for data that are spatiotemporal (Datta et al., 2016b), multivariate (Peruzzi and Dunson, 2020; Zhang and Banerjee, 2021), or non-Gaussian (Peruzzi and Dunson, 2022).

Conditional independence encoded via predetermined DAGs, however, may inhibit inferences on certain types of nonstationarity. In the middle and the right plots of Figure 1, fixed DAGs respectively from nearest neighbors and adjacent partitions disregard the local dynamics in the spread of wildfire smoke and air pollutants. The middle plot illustrates a simple scenario of an NNGP in which four nearest neighbors are connected by pink arrows to a location approximately at (120°W, 37°N). Three of these neighbors, however, are outside the smoke above the location of interest and may be undesirably non-informative despite proximity. Similarly, on the right plot, a Cubic MGP (Q-MGP) model fails to capture locally varying associations due to the fixed and repeated pattern in its DAG. These exemplify potentially incorrect assumptions about spatial and/or temporal dependence previous DAG-based models rest on, which can also result in unreliable estimates of covariance parameters.
and predictions.

In this paper, we propose a novel class of nonstationary processes characterizing varying directional dependence structures in space and time. Conditional independence between partitions of a domain is encoded in a sparse DAG whose directed edges are allowed to vary locally to capture different directional dependencies across the domain. We call this a Bag of directed Acyclic Graphs process (BAG). One may choose observed locations as nodes of DAGs instead of partitions; however, it is unwieldy to infer directions at the level of observed locations unless they lie on a grid. Therefore, we simplify the problem by partitioning the domain into disjoint regions, each of which becomes a node. Importantly, BAGs are valid standalone processes, leading to inferential advantages by incorporating parameter estimation and prediction at arbitrary locations into a coherent framework. BAGs can be employed as a prior in Bayesian latent spatial/spatiotemporal process models and can scale to large data (the number of observations $n > 10^5$). Unlike a recently proposed DAG-based nonstationary model of Gaussian data with replicates (Kidd and Katzfuss, 2021), BAGs do not rely on approximations in posterior sampling and can be easily extended to multivariate non-Gaussian settings following Peruzzi and Dunson (2022). Furthermore, BAGs induce interpretable nonstationarity.

Our construction of BAGs is partly motivated by causal inference in which DAGs have been used as systematic representations of causal relationships (Wright, 1934; Greenland et al., 1999; Shrier and Platt, 2008; Textor et al., 2016). We propose a related idea using a local mixture of directed edges. Directed edges in a DAG are inferred probabilistically and interpreted as directional associations across a domain. Mixture-related methods have been proposed for added flexibility (Griffin and Steel, 2006; Duan et al., 2007) or to generalize underlying stationary GPs to nonstationarity and beyond Gaussian assumptions (Gelfand et al., 2005; Rodríguez et al., 2010). Distinctively, our mixture is local and involves smaller regions of the domain one-at-a-time, instead of the whole domain at once. In the context of air pollution, local mixture components for each region represent different prevailing wind directions which may explain associated local covariance in the pollutants. Through stochastic selection of arrows, BAGs account for uncertainties of directional dependence patterns. Since different prevailing wind directions could be selected for different regions of the domain, BAGs are flexible enough to mimic the potential volatility of wind directions.
The rest of the paper is structured as follows: Section 2 formulates BAGs and Gaussian BAGs (G-BAGs) as a special case, followed by an empirical investigation on nonstationarity of G-BAG induced covariance functions. We demonstrate the performance and inferential benefits of our proposed approach via simulation results (Section 3) and application results to air quality data in CA (Section 4). Section 5 concludes the paper with useful discussions and possible extensions. Supporting Material is separately available including detailed derivations of the properties of BAGs, estimation and prediction steps, computational costs, and more results from simulation studies and data applications.

2 Spatial Process Modeling Using BAGs

Consider a univariate spatial process, \( \{w(s) \in \mathbb{R} \mid s \in \mathcal{D} \subseteq \mathbb{R}^d\} \). The following discussions are identically applied to spatiotemporal processes defined in \( \mathcal{D} \subseteq \mathbb{R}^{d+1} \), so we may proceed with “spatial” processes for brevity. We first divide the domain \( \mathcal{D} \) into \( M \) disjoint partitions, \( \mathcal{D} = \bigcup_{i=1}^{M} \mathcal{D}_i \) and \( \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \) for \( i \neq j \). Let \( \mathcal{S} = \{s_1, \ldots, s_k\} \) denote a fixed “reference set” set of locations. Here, \( \mathcal{S} \) need not coincide or intersect with observations \( \mathcal{T} = \{t_1, \ldots, t_n\} \), but \( \mathcal{T} \) is a practical choice for \( \mathcal{S} \) nonetheless. Figure 2 illustrates examples of domain partitioning using tessellations via rectangles and hexagons. Blue dots represent reference locations in \( \mathcal{S} \), while black dots are other locations. A partition of \( \mathcal{D} \) similarly divides \( \mathcal{S} \) as \( \mathcal{S} = \bigcup_{i=1}^{M} \mathcal{S}_i \) in which \( \mathcal{S}_i = \mathcal{S} \cap \mathcal{D}_i \) are disjoint reference subsets enumerated as \( \mathcal{S}_i = \{s_{i_1}, \ldots, s_{i_{k_i}}\} \) with \( \{i_1, \ldots, i_{k_i}\} \subseteq \{1, \ldots, k\} \) and \( k = \sum_{i=1}^{M} k_i \).

![Figure 2](image-url)

Figure 2: Examples of partitions made of rectangles (left) or hexagons (right). Each dot is a spatial location, and they are omitted for visual simplicity except on the first partition in the left plot. Arrows show a pool of potential directed edges to choose for shaded partitions.
2.1 Fixed DAG

Consider finite realizations of the latent process $w(\cdot)$ over $\mathcal{S}$, $\mathbf{w}_\mathcal{S} = (w_1^T, \ldots, w_M^T)^T$ where $\mathbf{w}_i = (w(s_{i1}), \ldots, w(s_{iN_i}))^T$ are realizations over partition $\mathcal{S}_i$ for $i = 1, \ldots, M$. After choosing an arbitrary order of the $M$ partitions, the joint density $p(\mathbf{w}_\mathcal{S}) = p(w_1, \ldots, w_M)$ can be rewritten as a product of conditional densities:

$$
p(\mathbf{w}_\mathcal{S}) = p(w_1) \prod_{i=2}^M p(w_i | w_1, \ldots, w_{i-1}).
$$

(1)

We can represent (1) as a DAG $G_\mathcal{S} = (A, E)$ with nodes $A = \{a_1, \ldots, a_M\}$ and directed edges $E = \{([a_i] \rightarrow a_i) | i = 1, \ldots, M\}$. The $k_i$ random variables $w_i$ are collectively mapped to a single node $a_i$. As an example, realizations of $w(\cdot)$ at all of the blue dots in Figure 2 are assigned to the node $a_1$. A set of nodes $[a_i]$ is a subset of $A$ whose edges are directed to $a_i$ and is referred to as a parent set. (1) implies $[a_1] = \emptyset$ and $[a_i] = \{a_1, \ldots, a_{i-1}\}$ for $i = 2, \ldots, M$. The directed graph $G_\mathcal{S}$ is acyclic because elements in $E$ do not make cycles.

Motivated by Vecchia approximation, we can drop some edges in $E$ and build a new DAG $G_\mathcal{S}^*$ that leads to a distinct joint density $p^*$ as a product of conditional densities with reduced conditioning sets

$$
p^*(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^M p(w_i | \mathbf{w}_{[i]}).
$$

(2)

Here, $\mathbf{w}_{[i]}$ is a subset of $\{w_1, \ldots, w_{i-1}\}$ collecting realizations at parent nodes of $a_i$, $\mathbf{w}_{[i]} = \{w_j | a_j \in [a_i] \subset \{a_1, \ldots, a_{i-1}\}\}$. The parent set $[a_i] \subset \{a_1, \ldots, a_{i-1}\}$. Since $G_\mathcal{S}^*$ is a DAG, $p^*(\mathbf{w}_\mathcal{S})$ is a proper joint density (Lauritzen, 1996). Approximating $p$ with $p^*$ coupled with a DAG renders advantages: (a) the approximation is highly accurate in Kullback-Leibler divergence; (b) accuracy can be improved with a careful order of nodes (Guinness, 2018); (c) the DAG can be extended to arbitrary locations to create standalone stochastic processes (Datta et al., 2016a); and (d) DAGs can be designed for faster computations (Peruzzi et al., 2020). In these cases, a pre-specified DAG is introduced only for computational reasons.
2.2 Unknown DAG

Taking a different perspective, we allow the DAG to be inferred from data. We account for uncertainty about dependency patterns across the spatial domain by considering finite mixtures of possible local arrows within each partition. A simplifying assumption is made that each node can have at most one parent from neighboring partitions in the spatial domain. This assumption may be relaxed by letting the number of parents be $>1$, but at the cost of reduced clarity in exposition, computational tractability, and interpretability.

Taking toy examples in Figure 2, directed edges 1, 2, 3, 4 (left) or 1, 2, 3 (right) correspond to competing assumptions on the dependence structure relevant for the shaded partitions given the tessellations using rectangles or hexagons, respectively. Such tessellations of the domain are advantageous because directed edges lead to intuitive interpretations, e.g. the directed edge 1 on the left of Figure 2 is interpreted as the direction from west to east. Moreover, stochastic search of directed edges will proceed analogously for all partitions due to the same neighbor structure and thus the same “bag.” For node $a_i$ (equivalently, reference subset $S_i$), we introduce a latent membership variable $z_i$ that determines the direction along which dependence is allowed to flow and thus the parent node. After enumerating possible directions to each node as $h = 1, \ldots, K$, we let $p(z_i = h) = \pi_{ih}$ be the probability that $a_i$ receives the $h$th directed edge. We denote the resulting parent node as $[a_i|z_i = h]$. We include only half of the possible directions in Figure 2 to ensure acyclicity of inferred directed graphs.

Prior knowledge about the domain may be utilized to choose one direction over the other on each axis. For a spatiotemporal domain, we assume that each node receives an additional directed edge from the partition that covers the same spatial region at the previous time point, and a spatial and the temporal parent are collectively denoted as $[a_i|z_i = h]$.

Using these assumptions, we define a joint density $\tilde{p}$ conditional on a specific configuration of directed edges $z = (z_1, \ldots, z_M)$, modifying equation (2) as follows:

$$
\tilde{p}(w_S | z) = \prod_{i=1}^{M} p(w_i | w_{[i|z_i]})
$$

(3)

with $w_{[i|z_i]} = \{w_j | a_j \in [a_i|z_i] \subset \{a_1, \ldots, a_{i-1}\}\}$. Since $z_1, \ldots, z_M$ are independent a priori,
integrating them out gives
\[ \tilde{p}(w_S) = \sum_{z \in \text{bag of DAGs}} \tilde{p}(w_S | z)p(z) = \sum_z \left\{ \prod_{i=1}^M p(w_i | w_{[i|z_i]}) \pi_i, z_i \right\}. \]

For notational brevity, \( \sum_z \) is hereafter \( \sum_{z \in \text{bag of DAGs}} \) unless otherwise specified. Then proposition 1 holds whose proof is provided in Supporting Material S1.

**Proposition 1.** \( \tilde{p}(w_S) \) is a proper joint density.

### 2.3 BAGs

The above discussion focuses on finite dimensional \( w_S \). We extend our approach to a valid process over all locations in the domain. We label non-reference locations as \( \mathcal{U} = \mathcal{D} \setminus \mathcal{S} \), and we partition \( \mathcal{U} \) into disjoint sets \( \mathcal{U}_1, \ldots, \mathcal{U}_M \) such that \( \bigcup_{i=1}^M \mathcal{U}_i = \mathcal{U} \). We extend the DAG over \( \mathcal{S} \) to a larger DAG \( \tilde{G} \), with each of the \( \mathcal{U}_i \)'s mapped to a node \( b_i \in B \). The construction of \( \tilde{G} \) is completed by assigning directed edges from nodes in \( A \) to those in \( B \), ensuring acyclicity of \( \tilde{G} \). There are several possible ways to place these directed edges. Assuming that \( \mathcal{S}_i \neq \emptyset \) for all \( i \), one can fix the edges as \( a_i \rightarrow b_i \) (\([b_i] = a_i\)), implying that local reference locations become the parent set for the non-reference locations in the same partition. Kolmogorov consistency conditions are then easily verified (see Appendix A in Peruzzi et al. 2020), proving the validity of the resulting process. However, a better alternative lets \([b_i | z_i] = a_i \cup [a_i | z_i]\), which allows modeling at any non-reference locations to depend not only on the local reference locations but also on their parent locations learned from data in our approach. In this way, predictions are made based on inferred wind directions. We prove the Kolmogorov consistency conditions for our proposed stochastic process when there is randomness in choosing the DAG for both \( A \) and \( B \).

First, we assume conditional independence of non-reference locations given the local reference locations and their parents so that
\[ \tilde{p}(w_u | w_S, z) = \prod_{i=1}^M \prod_{u \in \mathcal{U}_i} p(w(u) | w_i, w_{[i|z_i]}). \] (4)

The equations (3) and (4) suffice to describe the joint density of any finite subset of spatial
locations $L \subset D$. With $U_L = L \setminus S$,

$$
\tilde{p}(w_L) = \int \left\{ \sum_z \tilde{p}(w_{U_L} \mid w_S, z) \tilde{p}(w_S \mid z) p(z) \right\} \prod_{s_i \in S \setminus L} d(w(s_i)).
$$

(5)

We confirm that the collection of finite dimensional densities in equation (5) satisfy the Kolmogorov conditions (see Supporting Material S2). This implies that there exists a stochastic process associated with them, which we call a BAG. Therefore, our approach generates a valid spatial process via domain partitioning and local mixtures of DAGs.

When modeling spatial variation, we are interested in understanding the covariance between process realizations at different spatial locations. With BAGs, we find that

$$
\text{Cov} \tilde{p}(w(l_1), w(l_2)) = E_z[\text{Cov} \tilde{p}(w(l_1), w(l_2) \mid z)] + \text{Cov}_z[E \tilde{p}(w(l_1) \mid z), E \tilde{p}(w(l_2) \mid z)]
$$

$$
= \sum_z \text{Cov}_\tilde{p}(w(l_1), w(l_2) \mid z) p(z),
$$

(6)

if $E \tilde{p}(w(l) \mid z) = 0$ for any $l \in D$, where $E \tilde{p}$ and $E_z$ indicate expectation with respect to $\tilde{p}$ and $z$, respectively. This suggests that pairwise covariances between process realizations can be studied starting from the conditional covariances given a DAG configuration $z$. We further study BAGs in the Gaussian case and the induced nonstationary covariance.

### 2.4 Gaussian BAGs

Suppose the spatial process $\{w(s)\}$ is a zero-centered GP, $w(s) \sim GP(0, C(\cdot, \cdot|\theta))$ with a valid covariance function $C$ parametrized by $\theta$. Then equation (3) becomes

$$
\tilde{p}(w_S \mid z) = \prod_{i=1}^M N(w_i; H_{[i]z}, w_{[i]z}, R_{[i]z}),
$$

(7)

where, if $[a_i|z_i] = a_j$, $H_{[i]z} = C_{i,j}C_j^{-1}$ and $R_{[i]z} = C_i - C_{i,j}C_j^{-1}C_{j,i}$. If $[a_i|z_i] = \emptyset$, $H_{[i]z}w_{[i]z} = 0 \in \mathbb{R}^{k_i}$ and $R_{[i]z} = C_i$. The matrix $C_{i,j}$ has dimension $k_i \times k_j$, and its $(p,q)$ element is $C(s_{ip}, s_{jq}|\theta)$. We let $C_i = C_{i,i}$ for $i,j = 1, \ldots, M$. The latent membership variable $z$ determines the spatial DAG which in turn defines $H_{[i]z}$ and $R_{[i]z}$. We also find $\tilde{p}(w_S \mid z) = (2\pi)^{-k/2} \left( \prod_{i=1}^M |R_{[i]z} | \right)^{-1/2} \exp \{ -w_S^T(I_k - H_z)T R_z^{-1}(I_k - H_z)w_S/2 \} = N(0, \tilde{C}_z)$ where
\[
\tilde{C}_z^{-1} = (I_k - H_z)^T R_z^{-1} (I_k - H_z)
\]
with the identity matrix \(I_k\) of size \(k\), a \(k \times k\) sparse block matrix \(H_z\) whose \((i, j)\)th block is \(H_{i|z_i}\) if \([a_i|z_i] = a_j\) and zero otherwise, and a block-diagonal matrix \(R = \text{diag}(R_{1|z_1}, \ldots, R_{M|z_M})\). Given a bag of directed edges, it is always possible to find an ordering of \(\{1, \ldots, M\}\) such that \([a_i|z_i] \subset \{a_1, \ldots, a_{i-1}\}\) for \(i = 1, \ldots, M\), resulting in a lower triangular matrix \(H_z\) with zero diagonals. This renders \(|I_k - H_z| = 1\), and thus \(|\tilde{C}_z| = \prod_{i=1}^{M} |R_{i|z_i}|\). Furthermore, in the spatial only case, the \(i\)th block-row \(H_z[i,:]\) of \(H_z\) has at most one non-zero block for any \(i\) due to the assumption that all nodes in \(\tilde{G}_S(z)\) have at most one parent. As a result, the undirected moral graph found by “marrying the parents” of nodes in \(\tilde{G}_S(z)\) generates no additional edges. This means we can immediately identify the \((i, j)\)th block of \(\tilde{C}_z^{-1}\) as non-zero if and only if there is a directed edge between \(a_i\) and \(a_j\) in \(\tilde{G}_S(z)\) (either \(a_i = [a_j|z_j]\) or \(a_j = [a_i|z_i]\)). Hence, the sparsity structure of lower-triangular blocks (excluding diagonals) of \(\tilde{C}_z^{-1}\) is identical to that of \(I_k - H_z\). For \(i \geq j\), the \((i, j)\)th block of \(\tilde{C}_z^{-1}\) is \(R_{i|z_i}^{-1} + \sum_{j: [a_j|z_j] = a_i} H_{j|z_j}^{-1} H_{j|z_j}^{-1} H_{i|z_i}\) if \(i = j\), \(-R_{i|z_i}^{-1} H_{i|z_i}\) if \(a_j = [a_i|z_i]\), and \(0\) otherwise with \(\tilde{C}_z^{-1}(j, i) = (\tilde{C}_z^{-1}(i, j))^T\) by symmetry. In spatiotemporal cases, the undirected moral graph will include additional edges connecting a spatial parent and a temporal parent for each node in \(\tilde{G}_S(z)\). As a result, \(\tilde{C}_z^{-1}\) will have a non-zero block at \((i, j)\) not only if there is a directed edge between \(a_i\) and \(a_j\), but also if \(a_i\) and \(a_j\) have a common child node (one as the spatial directional parent and the other as the temporal parent).

Similarly, for non-reference locations, equation (4) becomes

\[
\tilde{p}(\mathbf{w}_U | \mathbf{w}_S, z) = \prod_{i=1}^{M} \prod_{u \in U_i} N(w(u); H_{u|z_i} w_{[u|z_i]}, R_{u|z_i}) = N(\mathbf{w}_U; H_{u|z} \mathbf{w}_S, R_{u|z})
\] (8)

with \(H_{u|z_i} = C_{u,[u|z_i]}C_{[u|z_i]}^{-1}\) and \(R_{u|z_i} = C_u - C_{u,[u|z_i]}C_{[u|z_i]}^{-1} C_{[u|z_i],u}\). If \([b_i|z_i] = a_i \cup [a_i|z_i]\) and \([a_i|z_i] = a_j, [u|z_i]\) becomes \(a_i \cup a_j\) for any \(u \in U_i\), yielding a \(1 \times (k_i + k_j)\) covariance matrix \(C_{u,[u|z_i]}\) whose elements are \(C(u, s | \Theta) \forall s \in S_i \cup S_j\). In vector form, the joint density of \(\mathbf{w}_U\) is multivariate Gaussian with mean \(H_{u|z} \mathbf{w}_S\) and covariance matrix \(R_{u|z}\) in which \(H_{u|z}\) is a \(|U| \times k\) matrix with \(|U|\) being the number of locations in \(U\), and \(R_{u|z}\) is a diagonal matrix. Again, if \([b_i|z_i] = a_i \cup a_j\), the \(i\)th block-row \(H_{u|z}[i,:]\) of \(H_{u|z}\) for all locations in \(U_i\) has non-zero blocks in columns corresponding to \(a_i\) and \(a_j\), and the \(i\)th block matrix of \(R_{u|z}\) is also a diagonal matrix whose elements are \(R_{u|z}\) \(\forall u \in U_i\).

Following equation (5) and the subsequent discussion, we obtain a Gaussian BAG (G-
BAG) such that \( w(s) \sim \text{G-BAG}(0, \tilde{C}(\cdot, \cdot | \theta)) \). The new covariance function \( \tilde{C} \) for G-BAG is given as 

\[
\tilde{C}(l_1, l_2 | \theta) = \sum_z \tilde{C}(l_1, l_2 | z, \theta)p(z) \text{ for any two locations } l_1, l_2 \in D \text{ by equation (6).}
\]

\( \tilde{C}(l_1, l_2 | z, \theta) \) is the induced covariance function of a GP one obtains from fixing a single DAG as \( z \): since equations (7) and (8) are all Gaussian, the density \( \tilde{p}(w_L | z) \) is also Gaussian for any finite subset \( L \subset D \), leading to a GP for any given \( z \). The implied covariance function conditional on \( z \) is then

\[
\tilde{C}(l_1, l_2 | z, \theta) = \begin{cases} 
\tilde{C}_{s_i, s_j}, & \text{if } l_1 = s_i \in S \text{ and } l_2 = s_j \in S \\
H_{l_1|z_i} \tilde{C}_{t_1|z_i, s_j}, & \text{if } l_1 \in U_i \text{ and } l_2 = s_j \in S \\
1(l_1 = l_2) R_{t_1|z_i} H_{l_1|z_i} \tilde{C}_{t_1|z_i, l_2|z_j} H_{l_2|z_j}^T, & \text{if } l_1 \in U_i \text{ and } l_2 \in U_j
\end{cases}
\]

where \( \tilde{C}_{P,Q} \) is a submatrix of \( \tilde{C}_z \) corresponding to locations in the sets \( P \) and \( Q \), and \( 1(\cdot) \) is the indicator function. Here \( \tilde{C}(\cdot, \cdot | z, \theta) \) is nonstationary, as is the marginal covariance function \( \tilde{C}(\cdot, \cdot | \theta) \). Derivation of equation (9) is described in Supporting Material S3.

### 2.5 Bayesian Hierarchical Model with G-BAGs

Consider a general regression model

\[
y(t) = x(t)^T \beta + w(t) + \epsilon(t),
\]

where \( y(t) \) is a response variable, \( x(t) \) is a \( p \)-dimensional vector of point-referenced predictors, \( w(t) \) is a spatial process over domain \( D \), and \( \epsilon(t) \sim N(0, \tau^2) \) is a measurement error. Over the observation set \( T = \{t_1, \ldots, t_n\} \), equation (10) is expressed in vector form as \( y = X \beta + w + \epsilon \) with \( y = (y(t_1), \ldots, y(t_n))^T \), \( w \) and \( \epsilon \) similarly defined, a \( n \times p \) matrix \( X \) having \( x(t_i)^T \) as its \( i \)-th row, and \( \epsilon \sim N(0, \tau^2 I_n) \). We assume \( w(\cdot) \sim \text{G-BAG}(0, \tilde{C}(\cdot, \cdot | \theta)) \) a priori as specified in Section 2.4. We complete our Bayesian hierarchical model by specifying priors for all unknowns. The joint posterior distribution for \( \{w_S, w_U, \{z_i, \pi_i\}_{i=1}^M, \beta, \tau^2, \theta\} \) is proportional to the likelihood times the prior distributions:

\[
N(y; X \beta + w, \tau^2 I_n) N(w_S; 0, \tilde{C}_z) N(w_U; H_{U|z} w_S, R_{U|z}) \prod_{i=1}^M \text{Cat}(z_i; \pi_i) \prod_{i=1}^M \text{Dir}(\pi_i; \alpha) N(\beta; \mu_\beta, V_\beta) IG(\tau^2; a_r, b_r) p(\theta)\]

where \( U = T \setminus S \), and \( \text{Cat}, \text{Dir}, \text{and IG} \) denote Categorical, Dirich-
let, and inverse Gamma distribution, respectively. The vector \( \mathbf{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK}) \) stores the probability for each possible value of \( z_i \), and \( \mathbf{\alpha} = (\alpha, \ldots, \alpha)^T \) is the \( K \times 1 \) Dirichlet hyperparameter vector. The prior \( p(\mathbf{\theta}) \) is left unspecified as it depends on the choice of a base covariance function \( \mathbf{C} \).

Spatiotemporal G-BAGs can use any spatiotemporal covariance function as a base. As a flexible default, we focus on the nonseparable (stationary) spatiotemporal covariance function of Gneiting (2002). For space-time lag \((\mathbf{h}, u) \in \mathbb{R}^{d+1}\), the covariance function is

\[
C(\mathbf{h}, u) = \frac{\sigma^2}{(a|u| + 1)} \exp \left( -\frac{c\|\mathbf{h}\|}{(a|u| + 1)^{\kappa/2}} \right) \tag{11}
\]

where \( \kappa \in [0, 1] \) is a space-time interaction parameter, \( a > 0 \) and \( c > 0 \) are temporal and spatial decays, respectively, and \( |\cdot|, \|\cdot\| \) denote 1, 2 dimensional Euclidean distance. (11) reduces to a separable function when \( \kappa = 0 \). We discuss nonstationarity induced by our approach from this stationary base covariance in Section 2.6. A straightforward Markov chain Monte Carlo (MCMC) sampler to obtain posterior samples is provided in Supporting Material S4. We show that G-BAGs have computational complexity of order \( n \) at each iteration of the MCMC sampler in Supporting Material S5.

### 2.6 Nonstationarity of G-BAGs

Even with a stationary base covariance, G-BAGs induce directional nonstationarity. Here, we consider a simplified setting to elucidate the nonstationarity features of \( \tilde{C}(\cdot, \cdot | \mathbf{\theta}) \) among reference locations. The reference set \( \mathcal{S} \) is a \( 30 \times 30 \times 4 \) grid on \( \mathcal{D} = [0, 1]^3 \) divided into \( M = 36 \) partitions. With a bag of three arrows coming from west (W), northwest (NW), and north (N), we assume that only three DAGs have a positive probability. Each DAG consists of only one of the three directed edges across the whole domain, and their probability is \( p(z_W) = 0.5, p(z_{NW}) = 0.4, \) and \( p(z_N) = 0.1 \) for W, NW, and N, respectively. The stationary covariance function in (11) is adopted with \( a = 0.7, c = 0.8, \kappa = 0, \) and \( \sigma^2 = 1 \). The resulting nonstationary covariances from two partitioning schemes are depicted in Figure 3a. Partition1 has an octagon in the middle with fan-shaped arms, and partition2 consists of axis-parallel rectangles. Each row of Figure 3a is a covariance heat map between the reference point (red point in the middle at time = 0.333) and other locations.
(a) Covariance heat maps between a reference point (red dot around (0.5, 0.5, 0.333)) and the others. G-BAG nonstationary covariances are depicted with two partitioning schemes (first two rows), using the base stationary covariance on the bottom row.

(b) Covariances of a reference point (black dot at (0.5, 0.5, $t^*$)) to (i) locations on its left (blue) and to (ii) those on its right (vermilion) at varying times. The time point $t^*$ of the reference point is given as a black vertical line. A DAG with $W$ arrows is assumed.

Figure 3: Features of nonstationarity induced by G-BAGs compared with stationary base covariance functions.
The induced covariance is orientational. The reference point has higher covariance on west-east and northwest-southeast axes than other axes, while the stationary covariance produces the same value at the same space-time lag regardless of axes. With time components, however, the induced nonstationary covariance becomes directional. The time identifies W over east (E) and NW over southeast (SE) on the corresponding axes, as hinted in partition2 of Figure 3a. On the second row of the figure, covariance from the reference point to \( s_j \) is higher than that to \( s_i \) despite the same space-time lag. This is because the path from the reference point to \( s_j \) aligns with the DAG with NW arrows, whereas \( s_i \) requires SE arrows not specified in any DAGs.

We further examine the directionality in G-BAGs’ nonstationary covariance functions and show that covariances flow along DAGs. The same base covariance function is used with a different temporal decay \( a = 2 \). With \( S \) on a \( 3 \times 3 \times 30 \) grid in \( D = [0, 1]^3 \), we let each grid point be a partition and assume W arrows among them. As drawn with gray arrows on the left plot of Figure 3b, we can imagine steady winds from W over time. Then we compute covariances of a reference point in black at \((0.5, 0.5, t^*)\) to (i) locations on its left and to (ii) those on its right at different time points. As expected, stationarity produces the same covariance values for both groups of locations, peaked at time \( t^* \) and decreasing as time lag increases (see the right plot of Figure 3b). The G-BAG induced nonstationary covariance gives separate curves for the two groups. In the middle plot of Figure 3b, we observe that the reference point at time \( t^* \) has the maximum covariance with locations on its right at time \( t^* + \Delta \), while it achieves its maximum with locations on its left in the past \( (t^* - \Delta) \). Moreover, at each time lag in the future, locations on the right have uniformly higher covariance values with the reference point than locations on the left. These results are intuitive as winds are coming from left to right. The locations on the left affect the reference point first, and in turn, the reference point affects those on the right. Hence, covariance persists longer in the future in directions towards which winds blow.

3 Applications on Simulated Data

We conduct simulation studies to evaluate performance of G-BAGs in prediction and learning DAGs. We consider a scenario in which a G-BAG model is correctly specified, and one in
which a fitted G-BAG model is misspecified relative to the data generating model. In both scenarios, our spatiotemporal domain is \( D = [0,1]^3 \) with locations \( t = (t_1, t_2, t_3) \in D \). For G-BAG models, we assume that a latent process \( w(t) \) follows \( \text{G-BAG}(0, \tilde{C}(\cdot, \cdot | \theta)) \) \textit{a priori} with the base covariance function in equation (11) with \( \theta = (a, c, \kappa, \sigma^2) \).

We compare our G-BAG approach to Q-MGP whose cubic meshes are considered as fixed wind directions on the same domain partitions and SPDE approaches with a stationary (referred to as SPDE-stationary; Lindgren et al. 2011) or a nonstationary (SPDE-nonstationary; Lindgren and Rue 2015) separable spatiotemporal covariance function. The SPDE models are approximated via integrated nested Laplace approximations (INLA). The R-INLA software package (Rue et al., 2009) is used to implement SPDE models, whose basic interface only supports separable spatiotemporal models. Hence, an autoregressive model of order 1 (AR(1)) is assumed for temporal correlations, while nonstationarity is imposed in spatially varying spatial range such that \( \rho(t) \propto \exp(\psi_1 + \psi_2 t_1 + \psi_3 t_2 + \psi_4 (t_1 - 1)(t_2 - 1)) \). The reference sets for G-BAG and MGP models are the observed locations which are 80% of the simulated locations. Twenty percent of data are held-out for validation. The number of threads is set at 10; different methods benefit from multiple threads by different degrees.

### 3.1 Fitted G-BAG is correctly specified

In the first scenario, we let \( y(t) = x(t)\beta + w(t) + \epsilon(t) \) with \( x(t) \sim N(0, 0.1) \) and \( \epsilon(t) \sim N(0, \tau^2) \), on a 40\( \times \)40\( \times \)8 regular grid in \( D \) for a total of 12,800 spatiotemporal locations. Parameters \( \beta \) and \( \tau^2 \) are set at 2 and 0.01, respectively. We partition each of the spatial axes into 6 intervals and the temporal axis into 8 intervals, resulting in \( M = 288 \) partitions. True directions (coming from W, NW, N, and northeast (NE)) vary by space and time as depicted in Figure 4a. Twenty-five synthetic data sets were generated with \( \theta_1 = (5, 0.5, 0.9, 2) \) and another 25 data sets with \( \theta_2 = (10, 0.1, 0.2, 2) \). Full specifications of priors are described in Supporting Material S6.1. G-BAG and Q-MGP results are based on 1,000 MCMC samples obtained from 10,000 samples by discarding first 8,000 samples as burn-in and saving every other sample in the subsequent 2,000 samples.

The true DAG is almost perfectly retrieved through the inferred DAG by posterior probabilities of the four arrows in the bag (Figure 4). Boundary partitions whose true DAG induces no parent often miss the true arrows but properly with high uncertainty. Predic-
Figure 4: Each time frame depicts the whole spatial domain whose partitions are colored by true or inferred directions. In (b), the inferred direction in each partition is the directed edge with the highest posterior probability whose value determines the degree of transparency. The inferred results are based on 25 data sets generated from G-BAG with $\theta_1$. 
tion/estimation performance of different models are summarized in Table 1 & S1 in Supporting Material S6.1. The Root Mean Square Prediction Error (RMSPE), Mean Absolute Prediction Error (MAPE), empirical coverage of 95% posterior predictive credible intervals (CI), and 95% CI width all indicate large predictive gains of G-BAG over the other models. This suggests that when directional associations are believed to shift rapidly over space and time, it is recommended to use G-BAGs that explicitly model directional dependence structures. Its computational time (in minutes), however, is longer than the competitors largely due to difference in coding languages and mixtures that bring heavier computational burden. Although we believe the current running time is acceptable, we expect some improvement in computational time by fully optimizing the code as a future work. The coefficient $\beta$ is estimated correctly by all models. The SPDE models considerably overestimate $\tau^2$, while G-BAG and Q-MGP produce $\tau^2$ estimates close to the truth. The decay parameters $a$ and $c$ are identified by G-BAG using weakly informative priors. The nonstationary SPDE barely improves the stationary SPDE in terms of prediction. This may imply that when directional associations are prevalent, the nonstationarity specified via a location-specific spatial range does not suffice.

Table 1: Simulation results when G-BAG is the true data generating model with $\theta_1$. Mean (standard deviation) over 25 synthetic data sets are provided.

|                  | G-BAG       | Q-MGP       | SPDE - stationary | SPDE - nonstationary |
|------------------|-------------|-------------|-------------------|----------------------|
| $\beta = 2$      | 2.004 (0.019) | 2.013 (0.036) | 2.005 (0.056)     | 2.005 (0.055)        |
| $\tau^2 = 0.01$  | 0.012 (<0.001) | 0.017 (0.002) | 0.178 (0.019)     | 0.177 (0.019)        |
| $\sigma^2 = 2$   | 1.760 (0.240) | 2.523 (0.428) | -                 | -                    |
| $a = 5$          | 7.034 (0.267) | 7.822 (0.301) | -                 | -                    |
| $c = 0.5$        | 0.527 (0.076) | 3.266 (0.569) | -                 | -                    |
| $\kappa = 0.9$   | 0.771 (0.071) | 0.955 (0.080) | -                 | -                    |
| RMSPE            | 0.189 (0.003) | 0.404 (0.028) | 0.479 (0.029)     | 0.480 (0.029)        |
| MAPE             | 0.150 (0.002) | 0.273 (0.015) | 0.321 (0.017)     | 0.322 (0.017)        |
| 95% CI coverage  | 0.951 (0.005) | 0.945 (0.008) | 0.925 (0.008)     | 0.925 (0.009)        |
| 95% CI width     | 0.745 (0.007) | 1.722 (0.082) | 1.795 (0.094)     | 1.792 (0.094)        |
| Run time (min.)  | 26.034 (0.557) | 1.279 (0.023) | 2.639 (0.430)     | 5.800 (1.460)        |
| Language         | R/C++       | C++         | C/C++/Fortran     | C/C++/Fortran        |
3.2 Fitted G-BAG is misspecified

In the misspecified case, we generate data on a regular lattice of size $193 \times 193 \times 59$ assuming a univariate model $y(t) = w(t) + \epsilon(t)$ with $\epsilon(t) \sim N(0, \tau^2)$ and $\tau^2 = 0.1$. We create $w(t)$ with north arrows fixed over $1 \times 193 \times 59$ partitions to mimic a surface under constant wind directions. The following covariance function introduced in Gneiting (2002) is used as a base:

$$C(h, u) = \sigma^2 / \left\{ 2^{\nu - 1} \Gamma(\nu) (a|u| + 1) \right\} \left( -c \|h\| / \{(a|u| + 1)^{\nu/2}\} \right)^\nu K_\nu \left( -c \|h\| / \{(a|u| + 1)^{\nu/2}\} \right)$$

where $\nu$ is the smoothness parameter of space, and $K_\nu$ is the modified Bessel function of the second kind of order $\nu$. Fixing $\nu = 1.5$, we generate 25 synthetic data sets with $\theta_3 = (5, 20, 1, 150)$ and another 25 with $\theta_4 = (10, 20, 1, 150)$. The latter mimics faster wind speed than the former. Figure 5 illustrates simulated examples of the true $w$ with $\theta_3$ and $\theta_4$. Directional dependence from north to south and different speed by $\theta$’s are evident across the spatiotemporal domain.

![Figure 5: Simulated examples of the true w with $\theta_3$ (top row) and $\theta_4$ (bottom row) at different five consecutive time points. Contour lines are drawn at 2.5 (top) and 1.7 (bottom).](image)

We take a subset of data of size $n = 18,750$ by retaining $25 \times 25 \times 30$ grid points. We fit a misspecified G-BAG model with $M = 2 \times 6 \times 30 = 360$ axis-parallel partitions. Based on visual inspection of the data, $K = 4$ directions - NW, N, NE, E - were placed in a bag. Prior choices are explained in Supporting Material S6.2. For G-BAG and Q-MGP, 7,000 MCMC samples were drawn, of which 5,000 samples were discarded as burn-ins, and every second sample was saved in the remaining 2,000 draws.
Prediction results for the 25 data sets with $\theta_3$ are illustrated in Figure S1 in Supporting Material S6.2. In terms of MAPE and RMSPE, Q-MGP (MAPE: 0.358 & RMSPE: 0.450) shows the best performance followed by G-BAG (0.367 & 0.462) and SPDE models (0.376 & 0.475). Both G-BAG (95% CI Coverage: 0.965) and Q-MGP (0.964) have more accurate empirical coverage, while SPDE models (0.895) have lower-than nominal coverage. Slightly better predictive performance of Q-MGP than the G-BAG model can be attributable to (a) the data generating DAG being fixed and (b) two spatial parents in Q-MGP versus one spatial parent in G-BAG. It is encouraging that G-BAG has comparable performance to Q-MGP using less information from parent nodes, by choosing one “best” spatial parent based on inferred directions. We observe similar results with $\theta_4$ which are omitted to save space. Moreover, G-BAG enables inferences on directions along which correlations move across the domain. Figure S2 in Supporting Material S6.2 demonstrates that the true direction (N) is properly found among four directions in the bag by the highest posterior probability in over 80% of the space-time partitions (left). Across time, each spatial partition assigns N an average posterior probability of around 0.7 (right).

4 Air Quality Data Analysis

Among many ambient pollutants, fine particulate matter that is 2.5 microns or less in diameter (PM2.5) is the primary concern due to its abundance and association to adverse health effects. In particular, acute exposures to PM2.5 have been associated with detrimental effects on human health including heart/lung diseases and associated premature deaths (Kloog et al., 2013; Gutiérrez-Avila et al., 2018). However, PM2.5 monitoring sites are in general sparse. For instance, the United States Environmental Protection Agency (EPA) has 165 PM2.5 monitoring sites in CA as of 2020, implying that each site covers 2,570 km$^2$ on average, which is roughly the size of Luxembourg. Considering the association of PM2.5 to various health effects and its influence on daily activities, accurate prediction of PM2.5 in regions without monitoring sites is an imperative task. We focus on CA (Section 4.1) and South Korea (Supporting Material S7) whose residents suffer from acute exposures to high PM2.5 seasonally due to wildfires and long-range movement of air pollutants, respectively.

The spatiotemporal spread of PM2.5 is heavily affected by winds. A large body of
literature has included wind-relevant variables in an attempt to predict PM2.5 (Wu et al., 2006; Calder, 2008; Wang and Ogawa, 2015; Preisler et al., 2015; Kleine Deters et al., 2017; Aguilera et al., 2020). However, there is a fundamental issue in summarizing wind direction at discrete times. The ambiguity of representative values for wind direction is evident in a tendency for different data sources to provide different summaries. For instance, EPA provides average wind direction, while the National Oceanic and Atmospheric Administration (NOAA) provides direction of the fastest wind. Moreover, due to the volatile nature of winds, naive averages of such wind directions over a given time period may rarely be meaningful. Therefore, we use the G-BAG approach on the latent process of log transformed PM2.5 to overcome these difficulties and implicitly incorporate wind effects.

4.1 California, the United States

The year 2020 was the largest wildfire season in CA’s modern history. During fire events, dramatically poor air quality is witnessed due to wildfire emissions in which PM2.5 is the primary pollutant (Liu et al., 2017). These environmental risks have made CA outstanding in its widespread use of low-cost sensors such as PurpleAir. In particular, more than half of the PurpleAir sensors in the United States are concentrated in CA as of February, 2020 (deSouza and Kinney, 2021). Therefore, we analyze daily PM2.5 levels in CA during fire seasons (August 1 to October 22, 2020) using EPA and PurpleAir measurements. Low-cost sensors require appropriate correction for improved comparability to regulatory monitors; e.g., using the Federal Reference Method (FRM) or Federal Equivalent Method (FEM) (Datta et al., 2020). Using recommended practice by EPA (Barkjohn et al., 2021; Evans et al., 2021), PurpleAir data are calibrated relative to FRM/FEM via: $PM_{2.5} = 5.75 + 0.52PA_{CF1} - 0.09RH$ if $PA_{CF1} \leq 343 \mu g/m^3$ and $PM_{2.5} = 2.97 + 0.46PA_{CF1} + 3.93 \times 10^{-4}PA_{CF1}^2$ otherwise. $PM_{2.5}$ is the calibrated value, $PA_{CF1}$ is the PurpleAir measurements from their correction factor labeled as $CF = 1$, and $RH$ is relative humidity in percentage.

The model is $y(t) = x(t)\beta + w(t) + \epsilon(t)$ where $x(t)$ is the Euclidean distance to the nearest fire at spatiotemporal location $t$, and $y(t)$ is the mean centered log(PM2.5). We expect $x(t)$ to capture elevated PM2.5 level due to proximity to wildfires. The covariate $x(t)$ is standardized to have mean 0 and standard deviation 1. A total of 110,473 spatiotemporal locations are observed of which 20% are held out for validation. We additionally predict at
274,564 new locations. The (W, NW, N, NE) edges are chosen in the bag because CA lies within the area of prevailing westerlies. CA is partitioned by \( M = 320 \) rectangles, and each covers approximately 55\(^2\)km\(^2\) (0.5\(^\circ\) resolution). With 8,000 burn-in and 2 thinning, 1,000 posterior samples are analyzed. We compare G-BAG to Q-MGP and SPDE-nonstationary. Detailed prior specifications are available in Supporting Material S8. The total run time for G-BAG, Q-MGP, and SPDE-nonstationary with 10 threads is roughly 22 hours, 32 minutes, and 4 hours, respectively.

Table 2 summarizes the results in comparison to Q-MGP and SPDE-nonstationary. All parameters are similarly estimated by G-BAG and Q-MGP, and the effect of the distance to the nearest fire appears insignificant in that the 95\% CIs of \( \beta \) from both models contain zero. Although the fitted \( \beta \) is significantly negative from SPDE-nonstationary, it is likely misled due to unexplained space-time variations after fitting spatiotemporal random effects (see Figure S4 in Supporting Material S8). Much larger \( \tau^2 \) fitted by the SPDE model than that by G-BAG also corroborates this remaining variability. These suggest that the nonstationarity via location-specific spatial range is insufficient to fully characterize the cause and removal of air pollution affected by wind directions in CA over this time period.

From Table 2, G-BAG yields smaller errors than the other two models and better accuracy of uncertainty quantification in prediction across the whole period compared to Q-MGP. A detailed comparison is illustrated in Figure S5 in Supporting Material S8 by which we confirm that the G-BAG approach improves prediction by a large margin at each time point especially over the fixed-DAG approach. We achieved down to 60\% reduction in prediction errors. These results signify that the unknown DAG approach and construction of nonstationarity via the inferred directions help enhance performance in this application.

The predicted surfaces of log(PM2.5) and observations are presented in Figure 6. We confirm that G-BAG can produce plausible predicted values even in regions lacking data with properly increased prediction uncertainty (Figure S6 in Supporting Material S8). Moreover, we find that W and N are the most prevailing directions causing associations in PM2.5 in CA from August to October, 2020. They are selected in 70\% of the space-time partitions. In particular, the bottom row of Figure 6 shows that the growth and reduction of areas with poor PM2.5 levels (> 35\(\mu g/m^3\)) progress from W and N directions, and the posterior mode of arrows captures the tendency well. Between latitude 36\(^\circ\)N and 40\(^\circ\)N and longitude 121\(^\circ\)W
Table 2: Posterior summaries and prediction performance measures of G-BAG, Q-MGP, and SPDE-nonstationary models on CA PM2.5 data. Posterior mean (95% CI) are provided for parameters.

|                  | G-BAG                  | Q-MGP                  | SPDE-nonstationary |
|------------------|------------------------|------------------------|--------------------|
| $\beta$          | -0.001 (-0.011, 0.007) | 0.006 (-0.003, 0.018)  | -0.038 (-0.054, -0.021) |
| $\tau^2$         | 0.011 (0.010, 0.011)   | 0.011 (0.011, 0.011)   | 0.079 (0.076, 0.081) |
| $\sigma^2$       | 3.730 (3.551, 3.886)   | 4.409 (4.408, 4.410)   | -                  |
| $a$              | 3.124 (3.011, 3.258)   | 1.262 (1.262, 1.263)   | -                  |
| $c$              | 0.009 (0.009, 0.009)   | 0.010 (0.010, 0.010)   | -                  |
| $\kappa$         | 0.010 (0.000, 0.036)   | 0.151 (0.151, 0.151)   | -                  |
| RMSPE            | 0.296                  | 0.435                  | 0.343              |
| MAPE             | 0.154                  | 0.296                  | 0.174              |
| 95% CI coverage  | 0.963                  | 0.906                  | 0.961              |
| 95% CI width     | 1.503                  | 1.388                  | 1.216              |

and 124°W, cleaner air started to flow in from W on September 16 and expanded along NW and N in a few days.

We also examined mid August with a particular interest in the August Complex, the largest wildfire in CA’s history recorded to have burnt 1,032,648 acres according to the department of Forestry and Fire Protection of CA (refer to https://www.fire.ca.gov/media/4jand1hh/top20 acres.pdf). After ignition on August 16 around shared boundaries of Mendocino, Lake, Glenn, and Tehama counties, its effect in PM2.5 started to appear on August 18 (see predicted PM2.5 surfaces by G-BAG and SPDE-nonstationary zoomed in northern CA in Figure 7). Due to westerlies, the left side of Mendocino enjoyed better air quality than the right side where PM2.5 from fires accumulated and exceeded 12 $\mu g/m^3$, the annual standard of EPA. Further elevated PM2.5 over the 24-hour EPA standard (35 $\mu g/m^3$) stayed on the same side until August 20 mainly due to NW directions. On August 22, however, the west Mendocino also had to face poorer ambient conditions because NE directions arose around [39°N, 40°N] at 123°W. Similar progression of PM2.5 is captured by SPDE-nonstationary as well.
Figure 6: Observed (top row) and predicted (middle) log(PM2.5) by G-BAG in CA. On the bottom, the posterior mode of wind directions is overlaid on discretized predicted results for PM2.5 based on EPA standards.

5 Discussion

We have developed a class of nonstationary processes based on flexible selection of DAGs for large point-referenced data. The interpretability of our model lies in its construction of nonstationarity and recovery of true directions among directed edges in the bag. Our model can manage high-resolution data at individual household level (e.g. PurpleAir) which may also be at road (e.g. Google Street View vehicles; Guan et al. 2020) or at pixel (e.g. Satellite imaging) levels. Predictive and inferential gains have been demonstrated compared to state-of-the-art methods. Decisions on domain partitioning should be assisted by computational budget and the desirable resolution at which inferences of correlation directions are to be
made. Although we have focused on univariate applications motivated by PM2.5 analysis, multivariate extensions are immediate with a proper cross-covariance function. In addition, as GPs were used to derive G-BAGs from the more general BAG approach, one may use other processes. For instance, Student-t processes can derive T-BAGs, and such heavier-tail processes increase robustness to outliers.

Our model has broad applicability in several research areas. First, it can be used for inferring environmental risks at individual level which is then linked to various health outcomes. Such imputation is highly valuable because measurements of environmental risks and health outcomes are often misaligned. Second, it can simulate realistic processes of air pollutants under various wind scenarios and provide mechanistic insights of how air quality would behave with abnormal events such as fires. This would contribute to a body of literature assessing impacts of wildfires or smoke plumes on air quality, which currently relies on separate modeling approaches with or without fires (Wu et al., 2006), regressions without spatial effects (Preisler et al., 2015; Larsen et al., 2018) or inverse distance weighting interpolations (Aguilera et al., 2020). Finally, in a multivariate setting, the stochastic selection of DAGs may be applied not only for spatiotemporal dependence of multiple outcomes, but
also for the dependence among the outcomes. The latter can be useful in community ecology applications, for example, where one could use BAGs to scalably infer the relationships among species, choosing from DAGs based on prior information about species’ phylogeny.

Our current implementation of BAGs ensures acyclicity by choosing directed edges from a subset of all possible directions (e.g., either W or E, either N or S, etc.). In addition, we restrict each node in a DAG to have at most one spatial parent. These assumptions help reduce the computational burden caused by the stochastic search of the DAG space. However, they may be too restrictive in some contexts. Therefore, work needs to be done to relax these assumptions in a coherent framework, keeping computational burden low. Another interesting future research is to develop a more realistic model of the time dynamics, rather than fixing a directed edge from a previous time point to the current, for better predictions in longer time lags.

**Acknowledgement**

We are grateful for the financial support from the National Institute of Environmental Health Sciences through grants R01ES027498 and R01ES028804 and from the European Research Council under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 856506). We would like to thank David Buch for helpful discussions.
S1 Proof of Proposition 1

Proof. The joint density in equation (3) is proper because there exists an associated DAG \( \tilde{G}_S(z) \) whose nodes are \( A \) and edges are identified by \( z \). Then

\[
\int \tilde{p}(w_S)dw_S = \int \left\{ \sum_z \tilde{p}(w_S|z)p(z) \right\} dw_S = \sum_z \left\{ \int \tilde{p}(w_S|z)dw_S \right\} p(z) = \sum_z p(z)
\]

\[
= \sum_{z_1=1}^K \cdots \sum_{z_M=1}^K (\pi_{1,z_1} \times \cdots \times \pi_{M,z_M}) = \sum_{z_1=1}^K \pi_{1,z_1} \times \cdots \times \sum_{z_M=1}^K \pi_{M,z_M} = 1
\]

because \( \sum_{z_i=1}^K \pi_{i,z_i} = 1 \) for all \( i = 1, \ldots, M \). Hence, \( \tilde{p}(w_S) \) is a proper joint density. \( \square \)

S2 Kolmogorov Consistency Conditions for BAG

Kolmogorov consistency conditions indicate consistency with (a) permutation of indices and (b) marginalization. First, we show that \( \tilde{p}(w_L) = \tilde{p}(w_{L\pi}) \) for any permutation \( L\pi \) of \( L \). Let \( U_{L\pi} = L\pi \setminus S \) with the fixed reference set \( S \). Given fixed partitions,

\[
\tilde{p}(w_L) = \int \left\{ \sum_z \tilde{p}(w_{U_L} | w_S, z) \tilde{p}(w_S | z)p(z) \right\} \prod_{s_i \in S \setminus L} d(w(s_i))
\]

\[
= \sum_z \left\{ \prod_{i=1}^M \left( \prod_{u \in U_{L_i}} p(w(u) | w_i, w_{[i|z_i]}\pi_{i,z_i})p(w_i | w_{[i|z_i]}\pi_{i,z_i}) \right) \right\} \prod_{s_i \in S \setminus L} d(w(s_i))
\]

\[
= \sum_z \left\{ \prod_{i=1}^M \left( \prod_{u \in U_{L\pi_{s_i}}} p(w(u) | w_i, w_{[i|z_i]}\pi_{i,z_i})p(w_i | w_{[i|z_i]}\pi_{i,z_i}) \right) \right\} \prod_{s_i \in S \setminus L_{\pi}} d(w(s_i))
\]

\[
= \int \left\{ \sum_z \tilde{p}(w_{U_{L\pi}} | w_S, z) \tilde{p}(w_S | z)p(z) \right\} \prod_{s_i \in S \setminus L_{\pi}} d(w(s_i)) = \tilde{p}(w_{L\pi})
\]
because $\mathcal{U}_L$ and $\mathcal{U}_{L_x}$, $S \setminus \mathcal{L}$ and $S \setminus \mathcal{L}_x$ share the same collection of elements, respectively.

Second, we show that $\tilde{p}(w_L) = \int \tilde{p}(w_{L \cup \{l_0\}}) dw(l_0)$ for a new location $l_0 \in \mathcal{D} \setminus \mathcal{L}$. Take $\mathcal{L}_1 = \mathcal{L} \cup \{l_0\}$. We consider two cases: when $l_0 \in S$ and when $l_0 \notin S$. If $l_0 \in S$, then $\mathcal{U}_L = \mathcal{L} \setminus S = \mathcal{L}_1 \setminus S = \mathcal{U}_{L_1}$ and $\mathcal{S} = (S \setminus \mathcal{L}_1) \cup \{l_0\}$, resulting in

$$
\int \tilde{p}(w_{L \cup \{l_0\}}) dw(l_0) = \int \left[ \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \right\} \prod_{s_i \in S \setminus \mathcal{L}_1} d(w(s_i)) \right] dw(l_0)
$$

$$
= \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \right\} \prod_{s_i \in S \setminus \mathcal{L}_1} d(w(s_i)) dw(l_0)
$$

If $l_0 \notin S$, then $\mathcal{U}_L \cup \{l_0\} = \mathcal{U}_{L_1}$ and $\mathcal{S} = \mathcal{S} \setminus \mathcal{L}_1$, yielding

$$
\int \tilde{p}(w_{L \cup \{l_0\}}) dw(l_0) = \int \left[ \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \right\} \prod_{s_i \in S \setminus \mathcal{L}_1} d(w(s_i)) \right] dw(l_0)
$$

$$
= \int \left[ \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \right\} \prod_{s_i \in S \setminus \mathcal{L}_1} d(w(s_i)) \right] dw(l_0)
$$

$$
= \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z)dw(l_0) \right\} \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \prod_{s_i \in S \setminus \mathcal{L}_1} d(w(s_i))
$$

$$
= \int \left\{ \sum_z \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z)p(z) \right\} \prod_{s_i \in S \setminus \mathcal{L}} d(w(s_i)) = \tilde{p}(w_L)
$$

where $\tilde{p}(w_{L_1} \mid w_S, z) = \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_S \mid z) = \tilde{p}(w_{L_1} \mid w_S, z) \tilde{p}(w_{L_1} \mid w_S, z)$ due to the conditional independence assumption at non-reference locations given the reference set. Note that a finite sum is interchangeable with an integral.

### S3 Conditional Nonstationary Covariance Function

The conditional nonstationary covariance function in equation (9) is computed as follows. $\text{Cov}_{\tilde{p}}$ and $E_{\tilde{p}}$ implicitly depend on $\theta$. When $l_1 = s_i \in S$ and $l_2 = s_j \in S$, the result is trivial.
When $l_1$ is not in $S$ but $l_2$ is, i.e. $l_1 \in \mathcal{U}_i$ and $l_2 = s_j \in S$,

$$
\text{Cov}_{\tilde{p}}(w(l_1), w(s_j)|z) \\
= E_{\tilde{p}}[\text{Cov}_{\tilde{p}}(w(l_1), w(s_j)|w_S, z)|z] + \text{Cov}_{\tilde{p}}[E_{\tilde{p}}(w(l_1)|w_S, z), E_{\tilde{p}}(w(s_j)|w_S, z)|z] \\
= E_{\tilde{p}}[0|z] + \text{Cov}_{\tilde{p}}[H_{t_1|z_i}w_{[t_1|z_i]}, w(s_j)|z] = H_{t_1|z_i} \tilde{C}_{[t_1|z_i], s_j}.
$$

When neither $l_1$ nor $l_2$ is in $S$, i.e. $l_1 \in \mathcal{U}_i$ and $l_2 \in \mathcal{U}_j$,

$$
\text{Cov}_{\tilde{p}}(w(l_1), w(l_2)|z) \\
= E_{\tilde{p}}[\text{Cov}_{\tilde{p}}(w(l_1), w(l_2)|w_S, z)|z] + \text{Cov}_{\tilde{p}}[E_{\tilde{p}}(w(l_1)|w_S, z), E_{\tilde{p}}(w(l_2)|w_S, z)|z] \\
= 1(1 = l_2)R_{t_1|z_i} + \text{Cov}_{\tilde{p}}[H_{t_1|z_i}w_{[t_1|z_i]}, H_{t_2|z_j}w_{[t_2|z_j]}|z] \\
= 1(1 = l_2)R_{t_1|z_i} + H_{t_1|z_i} \tilde{C}_{[t_1|z_i], [t_2|z_j]}H_{t_2|z_j}^T.
$$

**S4 Estimation and Prediction of G-BAGs**

The straightforward Markov chain Monte Carlo (MCMC) sampler to obtain posterior samples with a general $C$ is provided below:

(a) Update $\beta$ from

$$(\beta | \cdot) \sim N((V^{-1}_\beta \cdot X^TX/\tau^2)^{-1}(V^{-1}_\beta \cdot \mu_\beta + X^T(y - w)/\tau^2), (V^{-1}_\beta + X^TX/\tau^2)^{-1}).$$

(b) Update $\tau^2$ from

$$(\tau^2 | \cdot) \sim IG\left(a_r + \frac{n}{2}, b_r + \frac{1}{2}(y - X\beta - w)^T(y - X\beta - w)\right).$$

(c) Update $z_i$ from

$$p(z_i = h | w_S, w_{\mathcal{U}}) = \frac{\pi_{ih}N(w_i; H_{i|h}w_{[i|h]}, R_{i|h}) \prod_{u \in \mathcal{U}_i}N(w(u); H_{u|h}w_{[u|h]}, R_{u|h})}{\sum_{l = 1}^K \pi_{il}N(w_i; H_{i|l}w_{[i|l]}, R_{i|l}) \prod_{u \in \mathcal{U}_i}N(w(u); H_{u|l}w_{[u|l]}, R_{u|l})}$$

for $h = 1, \ldots, K$ and $i = 1, \ldots, M$. 

3
(d) Update $\pi_i$ for $i=1,\ldots,M$ from
\[
(\pi_i \mid z_i) \sim \text{Dir}(\alpha + 1(z_i = 1), \ldots, \alpha + 1(z_i = K)).
\]

(e) Update $w_i$ for $i=1,\ldots,M$ from $(w_i \mid --) \sim N(\Sigma_i \eta_i, \Sigma_i)$ where
\[
\Sigma_i^{-1} = \sum_{u:a_i \in [u]} H_{u,i}^T R_u^{-1} H_{u,i} + \sum_{j:[a_j|z_j]=a_i} H_{j|z_j}^T R_{j|z_j}^{-1} H_{j|z_j} + R_{i|z_i}^{-1} + I_{k_i}/\tau^2,
\]
\[
\eta_i = \sum_{u:a_i \in [u]} H_{u,i}^T R_u^{-1} e_u + \sum_{j:[a_j|z_j]=a_i} H_{j|z_j}^T R_{j|z_j}^{-1} w_j + R_{i|z_i}^{-1} H_{i|z_i} w_{i|z_i} + (y_i - X^*_i \beta)/\tau^2.
\]
Let $S_i^* = T \cap S_i$ and $|S_i^*| = n_i \leq k_i = |S_i|$. The response $y_i^*$ is a $k_i \times 1$ vector whose $j$th element is $y(s_{ij})$ if $s_{ij} \in S_i^*$ or 0 otherwise, leaving $n_i$ non-zero elements. Similarly, $X_i$ is a $k_i \times p$ matrix with zeros at rows corresponding to locations outside $S_i^*$ and $I_{k_i} = \text{diag}(1(s_{ij} \in S_i^*), \ldots, 1(s_{ik_i} \in S_i^*))$. Lastly, if $u \in U_j$, $R_u$ is $R_{u|z_j}$, and $H_{u,i}$ is a submatrix of $H_{u|z_j}$ by choosing columns that correspond to $w_i$ given that $w_i \subseteq w_{[u|z_j]}$. With $w_{-i} = w_{[u|z_j]\setminus i}$ and $H_{u,-i}$ being a submatrix of $H_{u|z_j}$ for $w_{-i}$, $e_u = w(u) - H_{u,-i} w_{-i}$.

(f) Update $w(u)$ for $u \in U_i$ for $i=1,\ldots,M$ from
\[
(w(u) \mid x(u), y(u), \tau^2, z) \sim N(\sigma_u^2 \mu_u, \sigma_u^2) \text{ where}
\]
\[
\sigma_u^2 = (1/R_{u|z_i} + 1/\tau^2)^{-1}, \quad \mu_u = H_{u|z_i} w_{[u|z_i]} / R_{u|z_i} + (y(u) - x(u)^T \beta) / \tau^2.
\]

(g) Finally, $\theta$ is updated from a Metropolis-Hastings step with target density
\[
p(\theta)N(w_S; 0, \tilde{C}_z)N(w_{\tilde{d}; H_{\tilde{d}|z} w_S, H_{\tilde{d}|z}}).
\]

(i) In the covariance function in equation (11), $\theta$ is $(a, c, \kappa, \sigma^2)$.

(ii) We use the robust adaptive Metropolis algorithm proposed by Vihola (2012) for $g(a, c, \kappa)$ with target acceptance rate of 0.234. A link function $g$ maps $(a, c, \kappa)$ to $(-\infty, \infty)$ range. Uniform priors are used.

(iii) A Gibbs step to update $\sigma^2$ can be easily derived with a prior $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. 

4
Posterior predictive samples at an arbitrary location \( l \in \mathcal{D} \) are obtained as follows: at each MCMC iteration, sample \( y(l) | y \sim N(x(l)^T \beta + w(l), \tau^2) \) if \( l \in \mathcal{S} \cup \mathcal{T} \). If \( l \notin \mathcal{S} \cup \mathcal{T} \), then first sample \( w(l) \) from \( N(H_l w(l); R_l) \) conditioned on \( w_S \) and \( z \) if desired, and then sample \( y(l) | y \sim N(x(l)^T \beta + w(l), \tau^2) \).

S5 Computational Cost of G-BAGs

In this section, we show that G-BAGs have computational complexity of order \( n \) at each iteration of the MCMC sampler in Section S4. For explanatory purposes, we assume \( |\mathcal{S}_i| = |\mathcal{U}_i| = m \) for \( i = 1, \ldots, M \). The number of locations in \( \mathcal{S} \cup \mathcal{U} \) are at maximum \( n + k \), in which case \( m = (n + k)/(2M) \). Reference nodes have at most two parent nodes (in spatiotemporal cases), while non-reference nodes can have, say, \( J \) parent nodes or less as there is more freedom in placing directed edges between \( A \) and \( B \). As in previous sections, we assume all locations in \( \mathcal{U}_i \) share the same parent set. The number of directions available in a bag is \( K \).

First, updates of \( z \) involve (a) \( N(w_i; H_{i|l} w_{i|l}, R_{i|l}) \); and (b) \( \prod_{u \in \mathcal{U}_i} N(w(u); H_{u|l} w_{u|l}, R_{u|l}) \) for \( h = 1, \ldots, K \). At a fixed \( h \), (a) requires \( C_{i|l}^{-1} \) for \( H_{i|l} \) and \( R_{i|l} \), and \( R_{i|l}^{-1} \) for density. Since \( C_{i|l} \) is of size \( 2m \times 2m \) and \( R_{i|l} \) is of size \( m \times m \), their inversion leads to complexity \( O(9m^3) \). Repeating for each \( i \) and \( h \), the overall complexity regarding (a) becomes \( O(9MKm^3) \). Similarly, (b) requires \( C_{u|l}^{-1} \) for \( H_{u|l} \) and \( R_{u|l} \), and \( R_{u|l}^{-1} \) for density. The \( Jm \times Jm \) matrix \( C_{i|l} \) is common for all \( u \in \mathcal{U}_i \), and \( R_{u|l} \) is a \( 1 \times 1 \) matrix. Thus, the total floating point operations (flops) pertaining to (b) is \( (Jm)^3 + m \times 1^3)MK \) considering repetitions over \( i \) and \( h \). Second, with \( H \)’s and \( R^{-1} \)’s stored from the first step, posterior updates of \( w_S \) use \( Mm^3 \) flops in computing \( \Sigma_i \in \mathbb{R}^{m \times m} \) for all \( i \)’s, while updates of \( w_{td} \) use \( Mm^3 \) flops in computing \( \sigma^2_u \in \mathbb{R} \\ for \( Mm \) \( u \)’s.

Adding all these steps, each iteration has approximate computational complexity of \( O(9MKm^3 + (J^3m^3 + m)MK + Mm^3 + Mm) \). Note \( (9K + J^3K + 1 + (K + 1)/m^2)Mm^3 \leq (9K + J^3K + 2)M((n + k)/(2M))^3 \). It is reasonable to choose the number of partitions \( M \) proportionally to the sample size \( n \). Therefore, with \( M \sim n \) and even \( k \approx n \), \( O((9K + J^3K + 2)n) \approx O(n) \), given that the fixed values of \( K \) and \( J \) are relatively small.

During the Gibbs iteration, \( H_{i|z}, R_{i|z}^{-1}, H_{u|z}, \) and \( R_{u|z}^{-1} \) need to be stored for \( i = 1, \ldots, M \),
\[ z_i = 1, \ldots, K, \text{ and } m \text{'s in } U_i. \text{ These matrices are of size } m \times 2m, m \times m, 1 \times Jm, \text{ and } 1 \times 1, \text{ respectively, causing storage cost } O(2m^2MK + m^2MK + Jm^2MK + mMK). \text{ With } M \propto n \text{ and } k \approx n, O(((3 + J)K + K/m)m^2M) \leq O(((3 + J)K + 1)n^2/M) \approx O(n). \]

### S6 Applications on Simulated Data

#### S6.1 Fitted G-BAG is correctly specified

To fit a G-BAG model, a vague normal prior \( N(0, 100) \) is specified for \( \beta \). IG priors are specified for nugget (\( \tau^2 \)) and spatial variability (\( \sigma^2 \)) such that \( \tau^2 \sim IG(2, 0.1) \) and \( \sigma^2 \sim IG(2, 1) \). The covariance parameters have weakly informative uniform priors: \( a \sim Unif(4, 8) \) and \( c \sim Unif(0.158, 0.789) \) under \( \theta_1 \) and \( a \sim Unif(7.330, 14.667) \) and \( c \sim Unif(0.075, 0.373) \) under \( \theta_2 \). In both cases, \( \kappa \sim Unif(0, 1) \). The lower and upper bounds for \( a \) and \( c \) correspond to various correlation values ranging from 0.1 to 0.9 at varying spatial/temporal distances. For instance, \( a \sim Unif(7.330, 14.667) \) corresponds to a temporal range between roughly 0.5 and 1 at which the correlation drops to 0.1. The prior distribution for probabilities to choose one of the four directions is given as \( \pi_i \sim Dir(\alpha) \) with \( \alpha = 1/4 \) for \( i = 1, \ldots, 288 \). In QMGP, its default priors (\( \sigma^2, \tau^2 \sim IG(2, 1), \beta \sim N(0, 100) \)) are used except for the covariance decay parameters which have the common uniform prior whose lower bound is the minimum of lower bounds for \( a \) and \( c \) in the G-BAG model, and an upper bound is the maximum of upper bounds for \( a \) and \( c \). Penalized Complexity (PC) priors \( \text{Simpson et al., 2017} \) are used for SPDE model parameters. For AR(1), the autocorrelation parameter is assumed to be larger than 0.05 with prior probability 0.99, and the precision of white noise is larger than 1 with prior probability 0.01. Under stationarity, the spatial standard deviation is larger than 1.5 with prior probability 0.01, and the range parameter is smaller than 1.8 or 9 when the truth is \( \theta_1 \) or \( \theta_2 \), respectively. In a nonstationary case, \( \psi \)'s are modeled with i.i.d. \( N(0, 3) \) priors.

#### S6.2 Fitted G-BAG is misspecified

We generate 25 syntetic data sets with \( \theta_3 = (5, 20, 1, 150) \) and another 25 with \( \theta_4 = (10, 20, 1, 150) \). In the G-BAG model, prior distributions for other parameters are identi-
Table S1: Simulation results when G-BAG is the true data generating model with $\theta_2$. Mean (standard deviation) over 25 synthetic data sets are provided.

|                  | G-BAG  | Q-MGP  | SPDE-stationary | SPDE-nonstationary |
|------------------|--------|--------|-----------------|--------------------|
| $\beta = 2$      | 2.003  | 2.013  | 2.019           | 2.019              |
| $\tau^2 = 0.01$ | 0.011  | 0.015  | 0.150           | 0.148              |
| $\sigma^2 = 2$  | 1.635  | 2.181  | -               | -                  |
| $a = 10$         | 12.449 | 14.177 | -               | -                  |
| $c = 0.1$        | 0.117  | 3.097  | -               | -                  |
| $\kappa = 0.2$  | 0.620  | 0.934  | -               | -                  |
| RMSPE            | 0.129  | 0.356  | 0.428           | 0.427              |
| MAPE             | 0.103  | 0.229  | 0.274           | 0.273              |
| 95% CI coverage | 0.950  | 0.944  | 0.927           | 0.926              |
| 95% CI width     | 0.507  | 1.568  | 1.648           | 1.638              |
| Run time (min.) | 25.899 | 1.291  | 3.358           | 4.806              |

...cal as before except $a \sim Unif(0.1, 9)$ or $a \sim Unif(0.1, 15)$ under $\theta_3$ or $\theta_4$, respectively, while $\kappa \sim Unif(0, 1)$, $c \sim Unif(0.363, 21.183)$ in both cases. The prior for decay parameters in Q-MGP changes accordingly. The PC priors are modified for SPDE-stationary so that the range parameter is assumed to be smaller than 0.1 with prior probability 0.01, while the spatial standard deviation is higher than 1 with prior probability 0.01.

Figure S1: Simulation results when the fitted G-BAG is misspecified, and the true covariance parameters are $\theta_3$. Box plots over 25 synthetic data sets are presented.
South Korea offers locational benefits to study flow of air quality because it is embedded in a region where air quality exhibits rapid changes in time and space and strongly impacted by transboundary pollution, dust, or smoke from China (Crawford et al., 2021). We analyze hourly PM2.5 levels in Korea between 8pm on February 4, 2020 and 7am on February 5, 2020. This time of a day is chosen to minimize traffic effects that appear to significantly influence local air quality. AirKorea data provided by the Korean Ministry Of Environment (MOE) are used, which have 409 monitoring stations at 12 time points resulting in 4,908 observed spatiotemporal locations. For validation, 30% of the locations are left out. Predictions are made at an additional 72,060 locations. We partition Korea by $M = 17 \times 23 = 391$ rectangles, each covering approximately $26 \times 27 km^2$. The model $y(t) = w(t) + \epsilon(t)$, $\epsilon(t) \sim N(0, \tau^2)$ is assumed with the mean centered log(PM2.5) as $y(t)$. We assume $\tau^2 \sim IG(2, 0.1)$, $\sigma^2 \sim IG(2, 1)$ $a \sim Unif(0.052, 38)$, $c \sim Unif(0.001, 0.081)$, and $\kappa \sim Unif(0, 1)$ as prior distributions. The following four directions (SW, W, NW, N) are placed in a bag because these have been identified as the prevailing winds over Korea (Heo et al., 2010), also aligning with the directions of long-range movement from China. Analysis results are reported based on a posterior sample of size 1,000 obtained from 10,000 iterations after 8,000 burn-ins and

Figure S2: Histogram of wind directions chosen by the highest posterior probability in each partition (left). Average posterior probabilities of wind directions in the bag across 30 time points (right). The right plot illustrates the spatial domain $[0, 1]^2$ with 12 partitions, within each of which arrows have different length and thickness proportionally to their posterior probabilities. This result is averaged over 25 synthetic data with $\theta_3$. 

S7 Air Quality Analysis in South Korea
saving every second sample in the subsequent 2,000 samples. The total run time with 10 threads is reported in minutes in Table S2.

In predicted surfaces of Figure S3, it is visible that the PM2.5 level is gradually cleared over night from west and northwest directions, which is well captured by the posterior mean direction of winds. This exemplifies that the cause and removal of air pollution is strongly affected by wind directions. Among four directions in the bag, W and N tend to be selected with higher posterior probabilities (less uncertainty) over Korea. Table S2 shows that G-BAG outperforms Q-MGP in terms of lowering prediction errors and quantifying prediction uncertainty. This suggests that allowing for flexible selection of DAGs can improve prediction over a fixed DAG when directional association is evident or believed to exist in space and time.

Table S2: Posterior summaries and prediction performance measures of G-BAG and Q-MGP models on South Korea PM2.5 data. Posterior mean (95% CI) are provided for parameters.

|          | G-BAG         | Q-MGP         |
|----------|---------------|---------------|
| $\tau^2$ | 0.024 (0.020, 0.027) | 0.028 (0.025, 0.031) |
| $\sigma^2$ | 0.390 (0.337, 0.460) | 2.358 (1.842, 2.847) |
| $\alpha$ | 1.403 (1.152, 1.674) | 1.146 (0.967, 1.340) |
| $c$      | 0.058 (0.047, 0.072) | 0.039 (0.033, 0.045) |
| $\kappa$ | 0.142 (0.011, 0.411) | 0.061 (0.002, 0.199) |
| RMSPE    | 0.403         | 0.407         |
| MAPE     | 0.284         | 0.288         |
| 95% CI coverage | **0.954** | 0.943         |
| 95% CI width  | 1.764         | **1.662**    |
| Run time (min.) | 41.315       | **7.072**    |
Figure S3: Observed (left column) and predicted (middle) log(PM2.5) by G-BAG in South Korea. On the right, the posterior mean direction of winds is overlaid on discretized predicted results for PM2.5 based on its standards by Korean MOE.
S8 Air Quality Analysis in CA

Same priors are used as in Section S7 except $a, c \sim Unif(0, 1000)$ and $\beta \sim N(0, 100)$. The G-BAG and Q-MGP models have the same priors for all parameters. The location-specific spatial range is assumed in the SPDE model as $\rho(t) \propto \exp(\psi_1 + \psi_2 t'_1 + \psi_3 t'_2)$ where $t'_1$ and $t'_2$ are scaled eastings and northings in $[0, 1]$, respectively. PC priors are given identically as in Section 3.1 except independent $N(0, 1)$ priors for $\psi_1$, $\psi_2$, and $\psi_3$.

![Figure S4](image)

Figure S4: Comparison of residuals by SPDE-nonstationary (middle row) and by G-BAG (bottom row) on randomly chosen dates. Residuals are computed as $y - \hat{w}$ where $\hat{w}$ indicates fitted latent spatiotemporal random effects by two models. The observed log(PM2.5) in CA (top row) is given for reference.

Comparing residuals of SPDE-nonstationary and those of G-BAG, subtracting fitted $w$ from $y$, the former exhibit apparent spatial patterns unlike the latter (Figure S4). These patterns are rarely explained by $x(t)\beta$ either because the fitted $\beta$ by SPDE-nonstationary (-0.038) is small compared to the residuals ranging from -6 to 6. We observe that the spatiotemporal random effects from SPDE-nonstationary tend to underestimate PM2.5 when
the true PM2.5 is high and overestimate when it is low.

Figure S5: Comparison of prediction performance measures by G-BAG, Q-MGP, and SPDE-nonstationary at each time point for CA data.

Figure S6: The 95% posterior predictive CI widths on November 13, 2020. The surfaces bear much resemblance at any time point.
References

Aguilera, R., Gershunov, A., Ilango, S. D., Guzman-Morales, J., and Benmarhnia, T. (2020). Santa Ana Winds of Southern California impact PM2.5 with and without smoke from wildfires. *GeoHealth*, 4(1):e2019GH000225. DOI: [http://doi.org/10.1029/2019GH000225](http://doi.org/10.1029/2019GH000225).

Barkjohn, K. K., Gantt, B., and Clements, A. L. (2021). Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques*, 14(6):4617–4637. DOI: [http://doi.org/10.5194/amt-14-4617-2021](http://doi.org/10.5194/amt-14-4617-2021).

Calder, C. A. (2008). A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics*, 19(1):39–48. DOI: [http://doi.org/10.1002/env.852](http://doi.org/10.1002/env.852).

Crawford, J. H., Ahn, J.-Y., Al-Saadi, J., Chang, L., Emmons, L. K., Kim, J., Lee, G., Park, J.-H., Park, R. J., Woo, J. H., Song, C.-K., Hong, J.-H., Hong, Y.-D., Lefer, B. L., Lee, M., Lee, T., Kim, S., Min, K.-E., Yum, S. S., Shin, H. J., Kim, Y.-W., Choi, J.-S., Park, J.-S., Szykman, J. J., Long, R. W., Jordan, C. E., Simpson, I. J., Fried, A., Dibb, J. E., Cho, S., and Kim, Y. P. (2021). The Korea–United States Air Quality (KORUS-AQ) field study. *Elementa: Science of the Anthropocene*, 9(1). DOI: [http://doi.org/10.1525/elementa.2020.00163](http://doi.org/10.1525/elementa.2020.00163).

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812. DOI: [http://doi.org/10.1080/01621459.2015.1044091](http://doi.org/10.1080/01621459.2015.1044091).

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016b). Non-separable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, 10(3):1286–1316. DOI: [http://doi.org/10.1214/16-AOAS931](http://doi.org/10.1214/16-AOAS931).
Datta, A., Saha, A., Zamora, M. L., Buehler, C., Hao, L., Xiong, F., Gentner, D. R., and Koehler, K. (2020). Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmospheric Environment, 242:117761. DOI: http://doi.org/10.1016/j.atmosenv.2020.117761.

deSouza, P. and Kinney, P. L. (2021). On the distribution of low-cost PM2.5 sensors in the US: demographic and air quality associations. Journal of Exposure Science & Environmental Epidemiology, 31(3):514–524. DOI: http://doi.org/10.1038/s41370-021-00328-2.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. Biometrika, 94(4):809–825. DOI: http://doi.org/10.1093/biomet/asm071.

Evans, R., Larkin, S., Barkjohn, K. J., Clements, A., and Holder, A. (2021). AirNow fire and smoke map: extension of the US-wide correction for Purple PM2.5 sensors. Presentation slides, United States Environmental Protection Agency. URL: https://www.epa.gov/sites/default/files/2021-05/documents/toolsresourceswebinar_purpleairsmoke_210519b.pdf.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. Environmetrics, 12(5):469–483. DOI: http://doi.org/10.1002/env.473.

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(471):1021–1035. DOI: http://doi.org/10.1198/0162145050000002078.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. Journal of the American Statistical Association, 97(458):590–600. DOI: http://doi.org/10.1198/016214502760047113.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. Epidemiology, 10(1):37–48. URL: https://www.jstor.org/stable/3702180.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. Journal of the American Statistical Association, 101(473):179–194. DOI: http://doi.org/10.1198/016214505000000727.
Guan, Y., Johnson, M., Katzfuss, M., Mannshardt, E., Messier, K. P., Reich, B. J., and Song, J. J. (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. *Journal of the American Statistical Association*, 115(531):1111–1124. DOI: http://doi.org/10.1080/01621459.2019.1665526.

Guinness, J. (2018). Permutation and grouping methods for sharpening Gaussian Process approximations. *Technometrics*, 60(4):415–429. DOI: http://doi.org/10.1080/00401706.2018.1437476.

Gutiérrez-Avila, I., Rojas-Bracho, L., Riojas-Rodríguez, H., Kloog, I., Just, A. C., and Rothenberg, S. J. (2018). Cardiovascular and cerebrovascular mortality associated with acute exposure to PM2.5 in Mexico City. *Stroke*, 49(7):1734–1736. DOI: http://doi.org/10.1161/STROKEAHA.118.021034.

Heo, C.-U., Lee, Y., and Lee, E.-J. (2010). Spatial distribution characteristics of wind map over Korea using meteorological resources. *Atmosphere*, 20(2):63–71. URL: https://www.koreascience.or.kr/article/JAKO201028839764873.page.

Katzfuss, M. and Guinness, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, 36(1):124–141. DOI: http://doi.org/10.1214/19-STS755.

Kidd, B. and Katzfuss, M. (2021). Bayesian nonstationary and nonparametric covariance estimation for large spatial data. *Bayesian Analysis*, -1(-1):1–22. DOI: http://doi.org/10.1214/21-BA1273.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668. DOI: http://doi.org/10.1198/016214504000002014.

Kleiber, W. and Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, 112:76–91. DOI: http://doi.org/10.1016/j.jmva.2012.05.011.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., and Rybarczyk, Y. (2017). Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters.
Kloog, I., Ridgway, B., Koutrakis, P., Coull, B. A., and Schwartz, J. D. (2013). Long- and short-term exposure to PM2.5 and mortality. *Epidemiology*, 24(4):555–561. DOI: [http://doi.org/10.1097/EDE.0b013e318294beaa](http://doi.org/10.1097/EDE.0b013e318294beaa).

Krock, M., Kleiber, W., and Becker, S. (2021a). Nonstationary modeling with sparsity for spatial data via the basis graphical lasso. *Journal of Computational and Graphical Statistics*, 30(2):375–389. DOI: [http://doi.org/10.1080/10618600.2020.1811103](http://doi.org/10.1080/10618600.2020.1811103).

Krock, M., Kleiber, W., Hammerling, D., and Becker, S. (2021b). Modeling massive highly-multivariate nonstationary spatial data with the basis graphical lasso. *arXiv:2101.02404 [cs, stat]*. URL: [http://arxiv.org/abs/2101.02404](http://arxiv.org/abs/2101.02404).

Larsen, A. E., Reich, B. J., Ruminski, M., and Rappold, A. G. (2018). Impacts of fire smoke plumes on regional air quality, 2006-2013. *Journal of Exposure Science & Environmental Epidemiology*, 28(4):319–327. DOI: [http://doi.org/10.1038/s41370-017-0013-x](http://doi.org/10.1038/s41370-017-0013-x).

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63:1–25. DOI: [http://doi.org/10.18637/jss.v063.i19](http://doi.org/10.18637/jss.v063.i19).

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498. DOI: [http://doi.org/10.1111/j.1467-9868.2011.00777.x](http://doi.org/10.1111/j.1467-9868.2011.00777.x).

Liu, X., Huey, L. G., Yokelson, R. J., Selimovic, V., Simpson, I. J., Müller, M., Jimenez, J. L., Campuzano-Jost, P., Beyersdorf, A. J., Blake, D. R., Butterfield, Z., Choi, Y., Crounse, J. D., Day, D. A., Diskin, G. S., Dubey, M. K., Fortner, E., Hanisco, T. F., Hu, W., King, L. E., Kleinman, L., Meinardi, S., Mikoviny, T., Onasch, T. B., Palm, B. B., Peischl, J., Pollack, I. B., Ryerson, T. B., Sachse, G. W., Sedlacek, A. J., Shilling, J. E., Springston, S.,
St. Clair, J. M., Tanner, D. J., Teng, A. P., Wennberg, P. O., Wisthaler, A., and Wolfe, G. M. (2017). Airborne measurements of western U.S. wildfire emissions: comparison with prescribed burning and air quality implications. *Journal of Geophysical Research: Atmospheres*, 122(11):6108–6129. DOI: http://doi.org/10.1002/2016JD026315.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 63(1):103–122. URL: https://www.jstor.org/stable/24771834.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020). Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *Journal of the American Statistical Association*, pages 1–14. DOI: http://doi.org/10.1080/01621459.2020.1833889.

Peruzzi, M. and Dunson, D. B. (2020). Spatial multivariate trees for big data Bayesian regression. *arXiv:2012.00943 [stat]*. URL: http://arxiv.org/abs/2012.00943.

Peruzzi, M. and Dunson, D. B. (2022). Spatial meshing for general Bayesian multivariate models. *arXiv:2201.10080 [stat]*. URL: http://arxiv.org/abs/2201.10080.

Preisler, H. K., Schweizer, D., Cisneros, R., Procter, T., Ruminski, M., and Tarnay, L. (2015). A statistical model for determining impact of wildland fires on Particulate Matter (PM2.5) in Central California aided by satellite imagery of smoke. *Environmental Pollution*, 205:340–349. DOI: http://doi.org/10.1016/j.envpol.2015.06.018.

Risser, M. D. and Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26(4):284–297. DOI: http://doi.org/10.1002/env.2336.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association*, 105(490):647–659. DOI: http://doi.org/10.1198/jasa.2010.tm08241.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal
Sampson, P. D. (2010). Constructions for nonstationary spatial processes. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of Spatial Statistics*. CRC Press, Boca Raton. DOI: http://doi.org/10.1201/9781420072884.

Schmidt, A. M., Guttorp, P., and O’Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500. DOI: http://doi.org/10.1002/env.1101.

Shrier, I. and Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, 8:70. DOI: http://doi.org/10.1186/1471-2288-8-70.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28. DOI: http://doi.org/10.1214/16-STS576.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liškiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *International Journal of Epidemiology*, 45(6):1887–1894. DOI: http://doi.org/10.1093/ije/dyw341.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312. DOI: http://doi.org/10.1111/j.2517-6161.1988.tb01729.x.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008. DOI: http://doi.org/10.1007/s11222-011-9269-5.

Wang, J. and Ogawa, S. (2015). Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan. *International Journal of Environmental Research and Public Health*, 12(8):9089–9101. DOI: http://doi.org/10.3390/ijerph120809089.
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215. DOI: [http://doi.org/10.1214/aoms/1177732676](http://doi.org/10.1214/aoms/1177732676).

Wu, J., M Winer, A., and J Delfino, R. (2006). Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. *Atmospheric Environment*, 40(18):3333–3348. DOI: [http://doi.org/10.1016/j.atmosenv.2006.01.056](http://doi.org/10.1016/j.atmosenv.2006.01.056).

Zhang, L. and Banerjee, S. (2021). Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data. *Biometrics*, pages 1–14. DOI: [http://doi.org/10.1111/biom.13452](http://doi.org/10.1111/biom.13452).