Title

Non-Separable Dynamic Nearest-Neighbor Gaussian Process Models for Large Spatio-Temporal Data With An Application to Particulate Matter Analysis

Authors

Datta, Abhirup
Banerjee, Sudipto
Finley, Andrew O
et al.

Publication Date

2015-10-24

Peer reviewed

NON-SEPARABLE DYNAMIC NEAREST-NEIGHBOR GAUSSIAN PROCESS MODELS FOR LARGE SPATIO-TEMPORAL DATA WITH AN APPLICATION TO PARTICULATE MATTER ANALYSIS

By Abhirup Datta and Sudipto Banerjee and Andrew O. Finley and Nicholas A.S. Hamm and Martijn Schaap

University of Minnesota
University of California, Los Angeles
Michigan State University
University of Twente
TNO Built Environment and Geosciences

Particulate matter (PM) is a class of malicious environmental pollutants known to cause detrimental effects on human health. Regulatory efforts aimed at curbing PM levels in different countries require high resolution space-time maps that can identify red-flag regions exceeding statutory concentration limits. Continuous space-time Gaussian Process (GP) models can potentially deliver uncertainty quantified map predictions for PM levels. However, traditional GP based approaches are thwarted by computational challenges posed by large datasets. We construct a novel class of scalable Dynamic Nearest Neighbor Gaussian Process (DNNGP) models that can provide a sparse approximation to any non-separable and possibly non-stationary spatio-temporal GP. The DNNGP can be used as a sparsity-inducing prior for spatio-temporal random effects in any Bayesian hierarchical model to deliver full posterior inference. Storage and memory requirements for a DNNGP model are linear in the size of the dataset thereby delivering massive scalability without sacrificing inferential richness. Extensive numerical studies reveal that the DNNGP provides substantially superior approximations to the underlying process than low rank approximations. Finally, we use the DNNGP to analyze a massive air quality dataset to substantially improve predictions of PM levels across Europe in conjunction with the LOTOS-EUROS chemistry transport models (CTMs).

1. Introduction. Recent years have witnessed considerable growth in statistical modeling of large spatio-temporal datasets; see, for example, the recent books by Gelfand et al. (2010), Cressie and Wikle (2011) and Banerjee, Carlin and Gelfand (2014) and the references therein for a variety of of

Keywords and phrases: Spatio-temporal Models, Scalable Gaussian Process, Nearest Neighbors, Bayesian Inference, Markov Chain Monte Carlo, Environmental Pollutants

methods and applications. An especially important domain of application for such models is environmental public health, where analysts and researchers seek map projections of ambient air pollutants measured at monitoring stations and understand the temporal variation in such maps. When inference is sought at the same scale as the observed data, a popular approach is to model the measurements as a time series of spatial processes. This approach encompasses standard time series models with spatial covariance structures (Pfeifer and Deutsch, 1980a,b; Stoffer, 1986) and dynamic models (Stroud, Muller and Sanso, 2001; Gelfand, Banerjee and Gamerman, 2005) among numerous other alternatives.

On the other hand, when inference is sought at arbitrary scales, possibly finer than the observed data (e.g., interpolation over the entire spatial and temporal domains), one constructs stochastic process models to capture dependence using spatio-temporal covariance functions (see, e.g., Cressie and Huang, 1999; Kyriakidis and Journel, 1999; Gneiting, 2002; Stein, 2005; Allcroft and Glasbey, 2003; Gneiting, Genton and Guttorp, 2007). In modeling ambient air pollution data, it is now customary to meld observed measurements with physical model outputs, where the latter can operate at much finer scales. Inference, therefore, is increasingly being sought at arbitrary resolutions using spatio-temporal process models (see, e.g., Gneiting and Guttorp, 2010) . Henceforth, we focus upon this setting.

While the richness and flexibility of spatio-temporal process models are indisputable, their computational feasibility and implementation pose major challenges for large datasets. Model-based inference usually involves the inverse and determinant of an $n \times n$ spatio-temporal covariance matrix $\mathbf{C}(\boldsymbol{\theta})$, where n is the number of space-time coordinates at which the data have been observed. This typically requires $\sim n^3$ floating point operations (flops) and storage of the order of n^2, which becomes prohibitive if n is large and $\mathbf{C}(\boldsymbol{\theta})$ has no exploitable structure. Approaches for modeling large covariance matrices in purely spatial settings include low rank models (see, e.g., Higdon, 2001; Kammann and Wand, 2003; Stein, 2007, 2008; Banerjee et al., 2008; Cressie and Johannesson, 2008; Crainiceanu, Diggle and Rowlingson, 2008; Rasmussen and Williams, 2005; Finley, Banerjee and McRoberts, 2009), covariance tapering (see, e.g., Furrer, Genton and Nychka, 2006; Kaufman, Scheverish and Nychka, 2008; Du, Zhang and Mandrekar, 2009; Shaby and Ruppert, 2012), approximations using Gaussian Markov Random Fields (GMRF) (see, e.g., Rue and Held, 2005), products of lower dimensional conditional densities (Datta et al., 2015; Vecchia, 1988, 1992; Stein, Chi and Welty, 2004), and composite likelihoods (e.g., Eidsvik et al., 2014). Extensions to spatio-temporal settings include Cressie, Shi and Kang (2010);

Finley, Banerjee and Gelfand (2012); Katzfuss and Cressie (2012) who extend low-rank spatial processes in a dynamic spatio-temporal setup, Fasso, Finazzi and Bevilacqua (2011) who uses tapering in a similar setup, and Xu, Liang and Genton (2014) who opts for a GMRF approach. All these methods use dynamic models defined on fixed temporal lags and do not easily lend themselves to continuous spatio-temporal domains.

Spatio-temporal process models for continuous space-time modeling of large datasets have received relatively scant attention. Bevilacqua et al. (2014) proposed a new class of 'adaptive taper' covariance functions for continuous spatio-temporal setup, while Bai, Song and Raghunathan (2012) used composite likelihoods in similar settings. Both these approaches, like their spatial analogues, have focused upon constructing computationally attractive likelihood approximations and have restricted inference only to parameter estimation. Estimating parameter uncertainty is usually based on asymptotic results which may not hold for irregularly observed datasets. Moreover, prediction at arbitrary locations and time points proceeds by imputing estimates into an interpolator derived from a different process model. This may not reflect accurate estimates of predictive uncertainty and is undesirable.

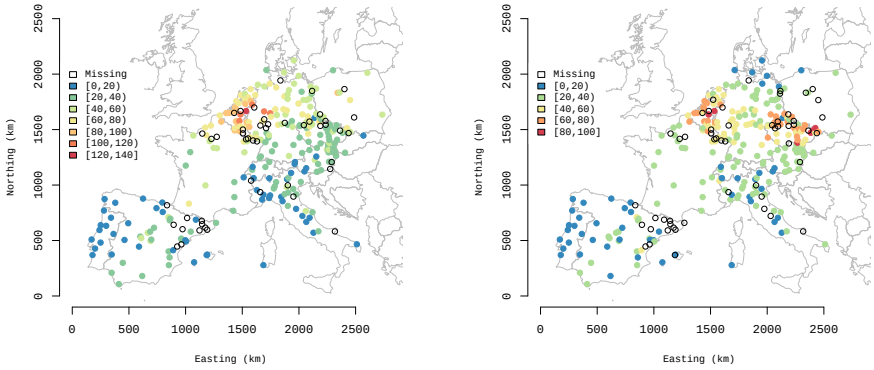
Our current work offers a highly scalable spatio-temporal process for continuous space-time modeling. While several of the aforementioned spatial approaches can, in theory, be extended to spatio-temporal models for large datasets, we opt to expand upon the neighbor-based conditioning set approaches outlined in purely spatial contexts by Vecchia (1988), Stein, Chi and Welty (2004) and Datta et al. (2015). We derive a scalable version of a spatio-temporal process, which we call the Dynamic Nearest-Neighbor Gaussian Process (DNNGP), using information from smaller sets of neighbors over space and time. This approach offers several benefits. The DNNGP is a well-defined spatio-temporal process whose realizations follow Gaussian distributions with sparse precision matrices. Thus, the DNNGP can act as a sparsity-inducing prior for spatio-temporal random effects in any Bayesian hierarchical model and enables full posterior inference, which considerably enhances its applicability. Moreover, it can be used with any spatio-temporal covariance function, thereby accommodating non-separability and non-stationarity. Being a process, importantly, allows the DNNGP to provide inference at arbitrary resolutions and, in particular, enables predictions at new spatial locations and time points in posterior predictive fashion. The DNNGP also delivers substantially superior approximation to the underlying process than, for example, what would be obtained by low rank approximations (see, e.g, Stein, 2014, for problems with low-rank approximations).

Finally, storage and memory requirements for a DNNGP model are linear in the number of observations, so it efficiently scales up to massive datasets without sacrificing richness and flexibility in modeling and inference.

The remainder of the article is organized as follows. In Section 2 we present the details of a massive environmental pollutants dataset and the need for a full Bayesian analysis. Section 3 elucidates a general framework for building scalable spatio-temporal processes and uses it to construct a sparsity-inducing DNNGP over a spatio-temporal domain. Section 4 describes efficient schemes for fixed as well as adaptive neighbor selection, which are used in the DNNGP. Section 5 details a Bayesian hierarchical model with a DNNGP prior and its implementation using Markov Chain Monte Carlo (MCMC) algorithms, including an algorithm for fast updates of the adaptive neighbor sets. Section 6 illustrates the performance of DNNGP using simulated datasets. In Section 7 we present a detailed analysis of our environmental pollutants dataset. We conclude the manuscript in Section 8 with a brief review and pointers to future research.

2. PM_{10} pollution analysis. Exposure to airborne particulate matter (PM) is known to increase human morbidity and mortality (Brunekreef and Holgate, 2002; Loomis et al., 2013; Hoek et al., 2013). In response to these and other health impact studies, regulatory agencies have instigated policies to monitor and regulate PM concentrations. For example, the European Commission's air quality standards limit PM_{10} (PM<10 μm in diameter) concentrations to 50 μg m^{-3} average over 24 hours and 40 μg m^{-3} over a year (European Commission, 2015). Measurements made with standard instruments are considered authoritative; however, these observations are sparse and accurate regional scale maps are needed for monitoring progress with mitigation strategies and for monitoring compliance. Properly quantifying uncertainty in predicted PM concentrations is therefore critical.

Substantial work has been aimed at developing regional scale chemistry transport models (CTM) for use in generating such maps. CTM's, however, have been shown to systematically underestimate observed PM_{10} concentrations, due to lack of information and understanding about emissions and formation pathways (Stern et al., 2008). Empirical regression (Brauer et al., 2011) or geostatistical models (Lloyd and Atkinson, 2004) are an alternative to CTM's for predicting continuous surfaces of PM_{10}. Empirical models may give accurate results, but are restricted to the conditions under which they are developed (Manders, Schaap and Hoogerbrugge, 2009). Assimilating monitoring station observations and CTM output, with appropriate bias adjustments, has been shown to provide improvements over using either data



(a) April 3, 2009 (b) April 5, 2009

Fig 1: Observed PM_{10} μg m^{-3} for two example dates.

source alone (van de Kassteele and Stein, 2006; Denby et al., 2008; Candiani et al., 2013; Hamm et al., 2015). In such setting, the CTM output enters as a model covariate and the measured station observations are the response. In addition to delivering more informed and realistic maps, analyses conducted using the models detailed in Section 5 can provide estimates of spatial and temporal dependence not accounted for by the CTM and hence provide insights useful for improving the transport models.

Our focus was on development and illustration of continuous space-time process models able to deliver map predictions and forecasts for PM_{10} and similar pollutants using sparse monitoring networks and CTM output. We coupled observed PM_{10} measurements across central Europe with corresponding output from the LOTOS-EUROS (Schaap et al., 2008) CTM. Inferential objectives included: *i*) delivering continuous maps of PM_{10} with associated uncertainty; *ii*) producing statistically valid forecast maps given CTM projections, and; *iii*) developing inference about space and time residual structure, i.e., space and time lags, that can help identify processes that are missing in the CTM. The study area and dataset are the same as those used by Hamm et al. (2015) and the reader is referred to that paper for more background information. Note, the current paper works with a 2-year time series, whereas Hamm et al. (2015) focused on daily analysis of a limited number of pollution events.

2.1. *Study area.* The study was mainland European countries with a substantial number of available PM_{10} observations. The countries included

were Portugal, Spain, Italy, France, Switzerland, Belgium, The Netherlands, Germany, Denmark, Austria, Poland, The Czech Republic, Slovakia and Slovenia. All data were projected to the European Terrestrial Reference System 1989 (ETRS) Lambert Azimuthal Equal-Area (LAEA) projection which gives a coordinate reference system for the whole of Europe.

2.2. *Observed measurements.* Air quality observations for the study area were drawn from the Airbase (*Air* quality data*base*)[1]. Daily PM_{10} concentrations were extracted for January 1 2008 through December 30 2009 resulting in a maximum of M=730 observations at each of $N = 308$ monitoring stations. Airbase daily values are averaged over the within-day hourly values when at least 18 hourly measurements are available, otherwise no data are provided. Airbase monitors are classified by type of area (rural, urban, suburban) and by type (background, industrial, traffic or unknown). Only rural background monitors were used in our study. This is common for comparing measured observations to coarse resolution CTM simulations (Denby et al., 2008). Monitoring stations above 800 m altitude were also excluded. These tend to be located in areas of variable topography and the accuracy of the CTM for locations that shift from inside to outside the atmospheric mixing layer is known to be poor. No further quality control was performed on the data. The locations of the 308 stations used in the subsequent analysis are shown in Figure 1 with associated observed and missing PM_{10} for two example dates. Of the 224,840 ($M \times N$) potential observations across 730 day time series and 308 stations, 41,761 observations were missing due to sensor failure or removal, and post-processing removal by Airbase. These missing values were predicted using the proposed models.

2.3. *LOTOS-EUROS CTM data.* LOTOS-EUROS (v1.8) is a 3D CTM that simulates air pollution in the lower troposphere. The simulator geographic projection is longitude-latitude and with a 0.50° longitude $\times$0.25° latitude resolution (approximately 25 km $\times$ 25 km). LOTOS-EUROS simulates the evolution of the components of particulate matter separately. Hence, this CTM incorporates the dispersion, formation and removal of sulfate, nitrate, ammonium, sea salt, dust, primary organic and elemental carbon and non-specified primary material, although, it does not incorporate secondary organic aerosol. Hendriks et al. (2013) provide a detailed description of LOTOS-EUROS.

The hour-by-hour calculations of European air quality in 2008-2009 were driven by the European Centre for Medium Range Weather Forecasting

[1]http://acm.eionet.europa.eu/databases/airbase (accessed 26 September 2014)

(ECMWF). Emissions were taken from the MACC (Monitoring Atmospheric Composition and Climate) emissions database (Pouliot et al., 2012). Boundary conditions were taken from the global MACC service (Flemming et al., 2009). The LOTOS-EUROS hourly model output was averaged to daily mean PM_{10} concentrations. LOTOS-EUROS grid cells that were spatially coincident with the Airbase observations were extracted and used as the covariate in the subsequent model.

CTM grid cell values nearest to station locations were used for subsequent model development. No attempt was made to match the spatial support (resolution) of the CTM simulations and station observations. The support of the CTM is 25 km, but the support of the observations is vague. Rural background observations were deliberately chosen because they are distant from urban areas and pollution sources. They are therefore considered representative of background, ambient pollution conditions and appropriate for matching with moderate resolution CTM-output (Denby et al., 2008; Hamm et al., 2015). This assumption is further backed-up by empirical studies that have shown that, even in urban areas, PM10 concentrations are dominated by rural background values (Eeftens et al., 2012).

3. Scalable Dynamic Nearest-Neighbor Gaussian Processes. Let $\{w(\ell) : \ell \in \mathcal{L}\}$ be a continuous spatio-temporal process (see, e.g., Gneiting and Guttorp, 2010, for details), where $\mathcal{L} = \mathcal{S} \times \mathcal{T}$ with $\mathcal{S} \subset \Re^d$ (usually $d = 2$ or 3), $\mathcal{T} \subset [0, \infty)$ and $\ell = (\mathbf{s}, t)$ is a space-time coordinate with $\mathbf{s} \in \mathcal{S}$ denoting a spatial location and $t \in \mathcal{T}$ is a time point. Such processes are specified with a spatio-temporal *covariance function* $\mathrm{Cov}\{w(\ell_i), w(\ell_j)\} = C(\ell_i, \ell_j \,|\, \boldsymbol{\theta})$. For any finite collection $\mathcal{U} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ in $\mathcal{L}$, let $\mathbf{w}_\mathcal{U} = (w(\ell_1)), w(\ell_2), \ldots, w(\ell_n))'$ be the realizations of the process over $\mathcal{U}$. Also, for two finite sets $\mathcal{U}$ and $\mathcal{V}$ containing n and m points in $\mathcal{L}$, respectively, we define the $n \times m$ matrix $\mathbf{C}_{\mathcal{U},\mathcal{V}}(\boldsymbol{\theta}) = \mathrm{Cov}(\mathbf{w}_\mathcal{U}, \mathbf{w}_\mathcal{V} \,|\, \boldsymbol{\theta})$, where the covariances are evaluated using $C(\cdot, \cdot \,|\, \boldsymbol{\theta})$. A valid spatio-temporal covariance function ensures that $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ is positive definite for any finite set $\mathcal{U}$. In particular, for spatio-temporal Gaussian processes, $\mathbf{w}_\mathcal{U}$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta}))$ and the (i, j)th element of $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ is $C(\ell_i, \ell_j \,|\, \boldsymbol{\theta})$.

As mentioned in the preceding section, storage and computations involving $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ can become impractical when n is large relative to the resources available. For full Bayesian inference on a continuous domain, we seek a scalable (in terms of flops and storage) spatio-temporal Gaussian process that will provide an excellent approximation to a full spatio-temporal process with any specified spatio-temporal covariance function. We outline a general framework that first uses a set of points in $\mathcal{L}$ to construct a com-

putationally efficient approximation for the random field and extends the finite dimensional distribution over this set to a process.

Let $\mathcal{R} = \{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$ be a fixed finite set of r points in $\mathcal{L}$. We refer to $\mathcal{R}$ as a *reference set*. We construct a spatio-temporal process $w(\ell)$ on $\mathcal{L}$ by first specifying $\mathbf{w}_\mathcal{R} = (w(\ell_1^*), w(\ell_2^*), \ldots, w(\ell_r^*))' \sim N(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))$, where $\mathbf{K}(\boldsymbol{\theta})$ is any $r \times r$ positive definite matrix and then defining

$$w(\ell) = \sum_{i=1}^{r} a_{i\ell}(\boldsymbol{\theta}) w(\ell_i^*) + \eta(\ell) \text{ for any } \ell \notin \mathcal{R}\,, \tag{3.1}$$

where $\eta(\ell)$ is a zero-centered Gaussian process independent of $\mathbf{w}_\mathcal{R}$ and such that $\text{Cov}\{\eta(\ell_i), \eta(\ell_j)\} = 0$ for any two distinct points in $\mathcal{L}$.

Observe that $w(\ell)$ in Equation 3.1 is a well defined spatio-temporal Gaussian process on $\mathcal{L}$ for *any* choice of $a_{i\ell}(\boldsymbol{\theta})$'s, as long as $\mathbf{K}(\boldsymbol{\theta})$ is positive definite. Every choice of $\mathbf{K}(\boldsymbol{\theta})$ and the $a_{i\ell}(\boldsymbol{\theta})$'s yield a possible model for continuous spatio-temporal data. For example, $w(\ell)$ is a Gaussian process with covariance function $C(\cdot, \cdot \,|\, \boldsymbol{\theta})$ if we choose $\mathbf{K}(\boldsymbol{\theta}) = \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})$, obtain $a_{i\ell}(\boldsymbol{\theta})$'s from the solution of $\mathbf{C}_{\mathcal{R},\mathcal{R}}\mathbf{a}_\ell(\boldsymbol{\theta}) = \mathbf{C}_{\ell,\mathcal{R}}$, where $\mathbf{a}_\ell(\boldsymbol{\theta})$ is $r \times 1$ with elements $a_{i\ell}(\boldsymbol{\theta})$, and choose $\eta(\ell)$ as an independent Gaussian process with variance $C(\ell, \ell \,|\, \boldsymbol{\theta}) - \mathbf{C}_{\ell,\mathcal{R}}\mathbf{C}_{\mathcal{R},\mathcal{R}}^{-1}\mathbf{C}_{\mathcal{R},\ell}$. Equation 3.1 then represents the 'kriging' equation for a location ℓ based on observations over $\mathcal{R}$ (Cressie and Wikle, 2011). Dimension reduction can be achieved with suitable choices for $\mathbf{K}(\boldsymbol{\theta})$ and $\mathbf{a}_\ell(\boldsymbol{\theta})$. Low-rank spatio-temporal processes emerge when we choose $\mathcal{R}$ to be a smaller set of 'knots' (or 'centers'). Additionally, specifying $\eta(\ell)$ to be a diagonal or sparse residual process yields $w(\ell)$ to be a non-degenerate (or bias-adjusted) low rank Gaussian Process approximating $w(\ell)$ (Banerjee et al., 2008; Finley, Banerjee and McRoberts, 2009; Sang and Huang, 2012).

Because of demonstrably impaired inferential performance of low-rank models in purely spatial contexts (see, e.g., Stein, 2014; Datta et al., 2015) at scales that we purport to work with here, we use the framework in Equation 3.1 to construct a class of sparse spatio-temporal process models. To be specific, let the reference set $\mathcal{R}$ be an enumeration of $r = MN$ points in $\mathcal{L}$, so that each ℓ_i^* in $\mathcal{R}$ corresponds to some $(\mathbf{s}_j, t_k)$ for $j = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, M$. For any $\ell_i^* = (\mathbf{s}_j, t_k)$ in $\mathcal{R}$ we define a *history set* $H(\ell_i^*)$ as the collection of all locations observed at times before t_k and of all points at t_k with spatial locations in $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{j-1}\}$. Thus, $H(\ell_i^*) = \{(\mathbf{s}_p, t_q) \,|\, p = 1, 2, \ldots, N,\ q = 1, 2, \ldots, (k-1)\} \cup \{(\mathbf{s}_p, t_k) \,|\, p = 1, 2, \ldots, (j-1)\}$. For any location ℓ_i^* in $\mathcal{R}$, let $N(\ell_i^*)$ be a subset of the history set $H(\ell_i^*)$. Also, for any location $\ell \notin \mathcal{R}$, let $N(\ell)$ denote any finite subset of $\mathcal{R}$. We refer to the sets $N(\ell)$ as a 'neighbor set' for the location ℓ and describe their construction later. We now turn to our choices for $\mathbf{K}(\boldsymbol{\theta})$ and $\mathbf{a}_\ell(\boldsymbol{\theta})$ in Equation 3.1.

Let $w(\ell) \sim GP(0, C(\cdot,\cdot \,|\, \boldsymbol{\theta}))$. We choose $\mathbf{K}(\boldsymbol{\theta})$ to effectuate a sparse approximation for the joint density of the realizations of $w(\ell)$ over $\mathcal{R}$, i.e., $N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta}))$. Adapting the ideas underlying likelihood approximations in Vecchia (1988) and Datta et al. (2015), we specify $\mathbf{K}(\boldsymbol{\theta})$ to be the $r \times r$ matrix such that

$$\begin{aligned} N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})) &= \prod_{i=1}^{r} p(w(\ell_i^*) \,|\, \mathbf{w}_{H(\ell_i^*)}) \\ &\approx \prod_{i=1}^{r} p(w(\ell_i^*) \,|\, \mathbf{w}_{N(\ell_i^*)}) = N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{K}(\boldsymbol{\theta})) \,, \end{aligned} \tag{3.2}$$

where $H(\ell_1^*)$ is the empty set (hence, so is $N(\ell_1^*)$) and $p(w(\ell_1^*) \,|\, \mathbf{w}_{H(\ell_1^*)}) = p(w(\ell_1^*) \,|\, \mathbf{w}_{N(\ell_1^*)}) = p(w(\ell_1^*))$. The underlying idea behind the approximation in Equation 3.2 is to compress the conditioning sets from $H(\ell_i^*)$ to $N(\ell_i^*)$ so that the resulting approximation is a multivariate normal distribution with a sparse precision matrix $\mathbf{K}(\boldsymbol{\theta})^{-1}$. As a consequence, we have

$$E[w(\ell_i^*) \,|\, \mathbf{w}_{H_{\ell_i^*}}] = \mathrm{E}[w(\ell_i^*) \,|\, \mathbf{w}_{N(\ell_i^*)}] = \mathbf{a}'_{N(\ell_i^*)} \mathbf{w}_{N(\ell_i^*)} \tag{3.3}$$

where $\mathbf{a}_{N(\ell_i^*)} = \mathbf{C}^{-1}_{N(\ell_i^*),N(\ell_i^*)} \mathbf{C}_{N(\ell_i^*),\ell_i^*}$. Additionally, let $f_{\ell_i^*} = \mathrm{Var}(w(\ell_i^*) \,|\, \mathbf{w}_{N(\ell_i^*)}) = C(\ell_i^*, \ell_i^* \,|\, \boldsymbol{\theta}) - \mathbf{C}_{\ell_i^*,N(\ell_i^*)} \mathbf{C}^{-1}_{N(\ell_i^*),N(\ell_i^*)} \mathbf{C}_{N(\ell_i^*),\ell_i^*}$ and $\mathbf{v}_i$ denote the $r \times 1$ vector satisfying $\mathbf{v}_i' \mathbf{w}_{\mathcal{R}} = w(\ell_i^*) - \mathbf{a}'_{N(\ell_i^*)} \mathbf{w}_{N(\ell_i^*)}$ for all values of $\mathbf{w}_R$. Then, we have,

$$\mathbf{K}(\boldsymbol{\theta})^{-1} = \mathbf{V}' diag(f_{\ell_1^*}, f_{\ell_2^*}, \ldots, f_{\ell_r^*})^{-1} \mathbf{V} \text{ where } \mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r) \tag{3.4}$$

If $m(<< r)$ denotes the limiting size of the neighbor sets $N(\ell)$, then $\mathbf{v}_i$'s are sparse with at most $m+1$ non-zero elements and consequently $\mathbf{K}(\boldsymbol{\theta})^{-1}$ has at most $O(rm^2)$ non-zero elements (this is the spatial-temporal analogue of the result in Datta et al., 2015). Hence, the approximation in (3.2) produces a sparsity-inducing proper prior distribution for the spatio-temporal random effects over $\mathcal{R}$ that closely approximates the realizations from a $GP(0, C(\cdot,\cdot \,|\, \boldsymbol{\theta}))$.

Turning to the vector of coefficients $\mathbf{a}_\ell(\boldsymbol{\theta})$ in Equation 3.1, we extend the idea in (3.3) to any point ℓ outside $\mathcal{R}$ by requiring that $\mathrm{E}[w(\ell) \,|\, \mathbf{w}_{\mathcal{R}}] = \mathrm{E}[w(\ell) \,|\, \mathbf{w}_{N(\ell)}]$. This is achieved by setting $a_{i\ell}(\boldsymbol{\theta}) = 0$ in Equation 3.1 whenever $\ell_i^* \notin N(\ell)$ for any point ℓ outside $\mathcal{R}$. Hence, if $N(\ell)$ contains m points, then at most m of the elements in the $r \times 1$ vector $\mathbf{a}_\ell(\boldsymbol{\theta})$ can be nonzero. These nonzero entries are determined from the above conditional expectation given $N(\ell)$. To be precise, we collect these m possibly nonzero elements into an $m \times 1$ vector $\mathbf{a}_{N(\ell)}$ and solve $\mathbf{C}_{N(\ell),N(\ell)} \mathbf{a}_{N(\ell)} = \mathbf{C}_{N(\ell),\ell}$. Also note

that $\mathbf{a}_\ell(\boldsymbol{\theta})'\mathbf{w}_{\mathcal{R}} = \mathbf{a}'_{N(\ell)}\mathbf{w}_{N(\ell)}$. Finally, to complete the process specifications in Equation 3.1, we specify $\eta(\ell) \overset{ind}{\sim} N(0, f_\ell)$, where $f_\ell = \text{Var}(w(\ell) \,|\, \mathbf{w}_{N(\ell)}) = C(\ell, \ell \,|\, \boldsymbol{\theta}) - \mathbf{C}_{\ell,N(\ell)}\mathbf{C}^{-1}_{N(\ell),N(\ell)}\mathbf{C}_{N(\ell),\ell}$. The covariance function $\tilde{\mathbf{C}}(\cdot,\cdot \,|\, \boldsymbol{\theta})$ of the resulting Gaussian Process is given by:

(3.5)
$$\tilde{\mathbf{C}}(\ell_i, \ell_j \,|\, \boldsymbol{\theta}) = \begin{cases} \mathbf{K}(\boldsymbol{\theta})_{p,q} \text{ if } \ell_i = \ell^*_p \text{ and } \ell_j = \ell^*_q \text{ are both in } \mathcal{R} \\ \mathbf{a}'_{\ell_i}\mathbf{K}(\boldsymbol{\theta})_{*q} \text{ if } \ell_i \notin \mathcal{R} \text{ and } \ell_j = \ell^*_q \in \mathcal{R} \\ \mathbf{a}'_{\ell_i}\mathbf{K}(\boldsymbol{\theta})\mathbf{a}_{\ell_j} + I(\ell_i = \ell_j) f_{\ell_i} \text{ if } \ell_i \text{ and } \ell_j \text{ are outside } \mathcal{R} \end{cases}$$

Owing to the sparsity of $\mathbf{K}(\boldsymbol{\theta})^{-1}$, the likelihood $N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))$ can be evaluated using $O(rm^3)$ flops. Substantial computational savings accrue because m is usually very small (also see later sections). Furthermore as $\eta(\ell)$ yields a diagonal covariance matrix and $\mathbf{a}_\ell(\boldsymbol{\theta})$ has at most m non-zero elements, for any finite set $\mathcal{V}$ outside $\mathcal{R}$, the flop count for computing the density $N(\mathbf{w}_{\mathcal{V}} \,|\, \mathbf{w}_{\mathcal{R}})$ will be linear in the size of $\mathcal{V}$. We have now constructed a scalable Gaussian Process in the spatio-temporal domain using small neighbor sets $N(\ell)$. We refer to this as a *Dynamic Nearest Neighbor Gaussian Process* (DNNGP) in spatio-temporal domains and denote it by $DNNGP(0, \tilde{\mathbf{C}}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$ where $\tilde{\mathbf{C}}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$ denotes the covariance function of this new GP.

4. Constructing Neighbor-Sets.

4.1. *Simple Neighbor Selection.* So far we have not discussed the choice of the neighbor sets apart from a restriction on their size. Spatial correlation functions usually decay with increasing inter-site distance, so the set of nearest neighbors based on the inter-site distances represents locations exhibiting highest correlation with the given location. This has motivated use of nearest neighbors to construct these small neighbor sets (Vecchia, 1988; Datta et al., 2015). On the other hand, spatio-temporal covariances between two points typically depend both on the spatial and temporal lag between the points. To be specific, non-separable isotropic spatio-temporal covariance functions can be written as $C((\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2) \,|\, \boldsymbol{\theta}) = C(h, u \,|\, \boldsymbol{\theta})$ where $h = ||\mathbf{s}_1 - \mathbf{s}_2||$ and $u = |t_1 - t_2|$. This often precludes defining any universal distance function $d : (\mathcal{S} \times \mathcal{T})^2 \to \mathbb{R}^+$ such that $C((\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2) \,|\, \boldsymbol{\theta})$ will be monotonic with respect to $d((\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2))$ for all choices of $\boldsymbol{\theta}$.

In the light of the above discussion, we define 'nearest neighbors' in a spatio-temporal domain using the spatio-temporal covariance functions. To elucidate, for any three points $(\mathbf{s}_1, t_1)$, $(\mathbf{s}_2, t_2)$ and $(\mathbf{s}_3, t_3)$, we say $(\mathbf{s}_2, t_2)$ is nearer to $(\mathbf{s}_1, t_1)$ than $(\mathbf{s}_3, t_3)$ if $C((\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2) \,|\, \boldsymbol{\theta}) > C((\mathbf{s}_1, t_1), (\mathbf{s}_3, t_3) \,|\, \boldsymbol{\theta})$.

Of course, this choice of nearest neighbors depends on the choice of the covariance function C and $\boldsymbol{\theta}$. Since the purpose of the DNNGP is to provide a scalable approximation of the parent GP, we always choose $C(\cdot,\cdot\,|\,\boldsymbol{\theta})$ to be same as the covariance function of the parent GP. However, $\boldsymbol{\theta}$ is unknown in most applications precluding the use of these newly defined neighbor sets to construct DNNGP.

We propose a simple intuitive method to construct neighbor sets. We choose m to be a perfect square. The central idea is to construct a neighbor set of size m using $\sqrt{m}$ spatial nearest neighbors and $\sqrt{m}$ temporal nearest neighbors. For $\ell = (\mathbf{s}_i, t_j)$ in $\mathcal{R}$, the neighbor set $N(\ell) = N(\mathbf{s}_i, t_j)$ needs to be a subset of the history set $H(\mathbf{s}_i, t_j)$. We define $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$, $S_i = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ and $T = \{t_1, t_2, \ldots, t_M\}$. Furthermore, for any finite set of spatial locations V, let $A(\mathbf{s}, V, m)$ denote set of m nearest neighbors of $\mathbf{s}$ in V. For any point $(\mathbf{s}_i, t_j)$ we define

(4.1)
$$N(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{\sqrt{m}-1} \{(\mathbf{s}, t_{j-k}) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S, \sqrt{m})\} \bigcup \{(\mathbf{s}, t_j) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S_i, \sqrt{m})\}$$

The above construction implies that the neighbor set for any point in $\mathcal{R}$ consists of $\sqrt{m}$ nearest spatial neighbors of the preceding $\sqrt{m}$ time points. Extending to any point outside $\mathcal{R}$, $N(\mathbf{s}, t)$ is simply defined as the Cartesian product of $\sqrt{m}$ nearest neighbors of $\mathbf{s}$ in $\mathcal{S}$ with $\sqrt{m}$ nearest neighbors of t in T. In many applications, one desirable property of the spatio-temporal covariance functions is natural monotonicity, i.e. $C(h, u)$ is decreasing in h for fixed u and decreasing in u for fixed h. Many non-separable classes of covariance functions proposed in the literature possess this property (Stein, 2013; Omidi and Mohammadzadeh, 2015). If $C(\cdot,\cdot\,|\,\boldsymbol{\theta})$ possesses natural monotonicity, then $N(\mathbf{s}_i, t_j)$ defined in Equation 4.1 is guaranteed to contain at least $\sqrt{m} - 1$ nearest neighbors of $(\mathbf{s}_i, t_j)$ in $H(\mathbf{s}_i, t_j)$. Thus, the neighbor sets defined above do not depend on any parameter and, for any value of $\boldsymbol{\theta}$, will contain a few nearest neighbors.

4.2. *Adaptive Neighbor Selection.* The neighbor selection scheme defined in Section 4.1 does not depend on the value of the covariance parameters. This is undoubtedly useful for fast implementation of the DNNGP. However, depending on the value of $\boldsymbol{\theta}$, the neighbor sets may often consist of very few nearest neighbors and many points with negligible correlation with the given point. This issue is illustrated in Figure 2. For a given point in a $10 \times 10 \times 20$ grid, we plot the neighbor set as well as the true set of m-nearest neighbors based on the covariance function defined in (5.2) with

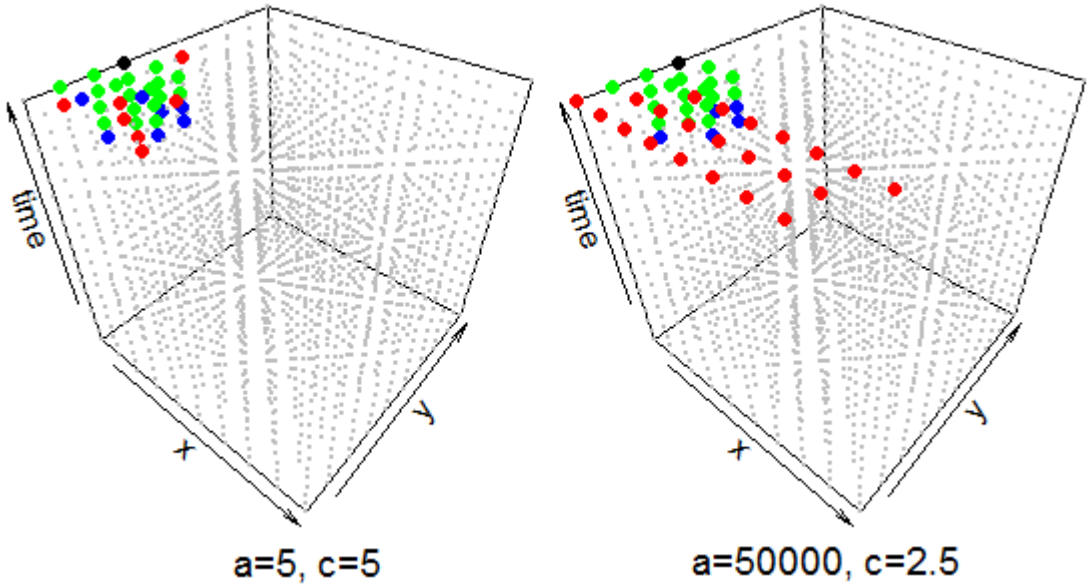


Fig 2: Neighbor sets and true nearest neighbors: Black dot indicates the given point; Green dots indicate true nearest neighbors which are included in the neighbor set; Red dots indicate true nearest neighbors not included in the neighbor set and Blue dots indicate points inside the neighbor set which do not belong to the set of true nearest neighbors

$\sigma^2 = 1$, $\alpha = \delta = 1$, $\kappa = 0$, $\nu = 1/2$. Two different sets of values of a and c are chosen. The figure on the left panel corresponds to a covariance function with significant dependence along both space and time while the one on the right panel corresponds to dependence mostly along space. We see that for different choices of the covariance parameters the neighbor sets contain different proportions of the true nearest neighbors. The problem is exacerbated in extreme cases with variation only along the spatial or temporal direction. In such cases, the neighbor sets defined in (4.1) will contain only about $\sqrt{m}$ nearest neighbors and $m - \sqrt{m}$ uncorrelated points.

Ideally, if $\boldsymbol{\theta}$ was known, one could have simply used m-nearest neighbors to construct the neighbor sets. In practice, however, as $\boldsymbol{\theta}$ is unknown, from an implementation point of view we encounter a computational roadblock as for every new value of $\boldsymbol{\theta}$ in an iterative optimizer or Markov Chain Monte Carlo run, we need to recalculate the neighbor sets for all the points. For updating neighbor sets of all points in $\mathcal{R}$ and n datapoints outside $\mathcal{R}$,

this step requires $O(r^2 + nr)$ flops at each iteration. The reference set $\mathcal{R}$ is typically chosen to match the scale of the observed dataset to achieve meaningful approximation of the parent GP by DNNGP. Hence, for large datasets this update becomes computationally deterrent. In fact, Vecchia (1988) and Stein, Chi and Welty (2004) admitted that this challenge has inhibited the use of correlation based neighbor sets in a spatial setting. Jones and Zhang (1997) permitted locations within a small prefixed temporal lag of a given location to be eligible for neighbors. However, this assumption will fail to capture any long term temporal dependence present in the datasets.

We now provide an algorithm that efficiently updates the neighbor sets after every update of $\boldsymbol{\theta}$. The central idea is to construct small sets that will contain the set of m-nearest neighbors for all choice of the parameter $\boldsymbol{\theta}$. Recall from Section 4.1 that, $A(\mathbf{s}, V, m)$ is the set of m-nearest neighbors of $\mathbf{s}$ in V. So $\mathbf{s} \in V$ implies that $\mathbf{s} \in A(\mathbf{s}, V, m)$ for all $m \geq 1$. For each $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$, we define the *eligible set*

(4.2)
$$E(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t_{j-k}) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S, [m/k])\} \bigcup \{(\mathbf{s}, t_j) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S_i, m)\}$$

where for any positive number x, $[x]$ denotes the greatest integer not exceeding x. So the eligible set for a point consists of m-nearest neighbors from the time levels j and $j-1$, $[m/2]$ nearest neighbors from time level $j-2$ and so on upto $[m/m] = 1$ nearest neighbor from time level $j-m$. For any point t outside T, let $t[k]$ denote the k^{th} nearest time point of t in T. Then, we define the eligible set for any $(\mathbf{s}, t)$ outside $\mathcal{R}$ as

$$E(\mathbf{s}, t) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t[k]) \,|\, \mathbf{s} \in A(\mathbf{s}, S, [m/k])\} \tag{4.3}$$

The eligible sets do not depend on the covariance parameters $\boldsymbol{\theta}$. We show in Appendix A that for any point $(\mathbf{s}, t)$ in $\mathcal{L}$, the eligible set $E(\mathbf{s}, t)$ defined by Equations 4.2 and 4.3 contains m-nearest neighbors of $(\mathbf{s}, t)$ for all values of $\boldsymbol{\theta}$ as long as the underlying covariance function $C(h, u \,|\, \boldsymbol{\theta})$ possess natural monotonicity. This property is illustrated in Figure 3 where for two different sets of parameter values used in Figure 2, the eligible sets remain same and contains the entire set of m-nearest neighbors for both cases.

This result has substantial consequences because the size of the eligible sets do not exceed $m + \sum_{k=1}^{m} [m/k]$. As m is typically chosen to be around 20, this sum is approximately equal to $4m$. The eligible sets needs to be calculated only once as they are free of any parameter choices. Subsequently, for every new update of $\boldsymbol{\theta}$ in a MCMC sampler or an iterative solver, one

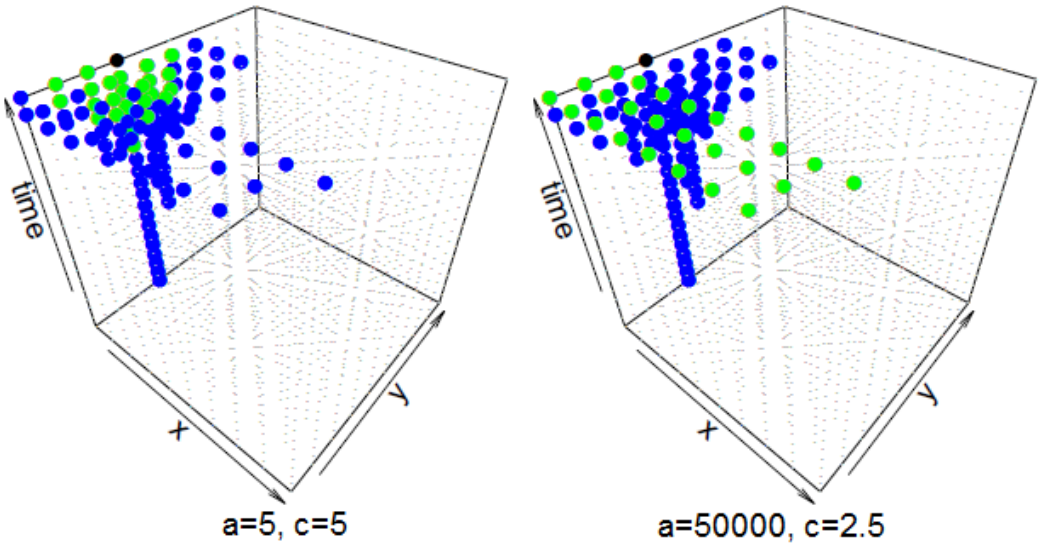


Fig 3: Eligible sets and true nearest neighbors: Black dot indicates the Given point; Green dots indicate true nearest neighbors which are included in the eligible set and Blue dots indicate points inside the eligible set which do not belong to the set of true nearest neighbors.

can search for a new set of m-nearest neighbors only within the eligible sets and use the m-nearest neighbors as the conditioning sets to construct the DNNGP. We summarize the MCMC steps of the dynamic NNGP with adaptive neighbor selection in Algorithm 1.

Algorithm 1 Algorithm for adaptive neighbor selection in dynamic NNGP

1: Compute the eligible sets $E(\mathbf{s}_i, t_j)$ for all $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$ from Eqn. (4.2)
2: At the l^{th} iteration of the MCMC:
   (a) Calculate $C((s,t),(s_i,t_j)\,|\,\boldsymbol{\theta}^{(l)})$ for all (s,t) in $E(\mathbf{s}_i,t_j)$
   (b) Define $N(\mathbf{s}_i,t_j)^{(l)}$ as the set of m locations in $E(\mathbf{s}_i,t_j)$ which minimizes $C((s,t),(s_i,t_j)\,|\,\boldsymbol{\theta}^{(l)})$
   (c) Repeat steps (a) and (b) for all $(\mathbf{s}_i,t_j)$ in $\mathcal{R}$
   (d) Update $\boldsymbol{\theta}^{(l+1)}$ based on the new set of neighbor sets computed in step (c)
3: Repeat Step 2 for N MCMC iterations

As the size of the sets are approximately $4m$, the search is now restricted to eligible sets having $4m$ points instead of history sets having upto r points. So the total computational complexity of the search is now reduced to

$O(4m(n+r))$ from $O(nr+r^2)$. This is at par with the scale of implementing the remainder of the algorithm. With this adaptive neighbor selection scheme we gain the advantage of selecting the set of m-nearest neighbors at every update while retaining the scalability of the DNNGP. Parallel computing resources, if available, can be greatly utilized to further reduce computations as the search for eligible sets for each point can proceed independent of one another.

5. Bayesian DNNGP model. We consider a spatio-temporal dataset observed at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ and at time points $t_1, t_2, \ldots, t_M$. Note that there may not be data for all locations at all time points i.e. we allow missing data. Let $\{\ell_1, \ell_2, \ldots, \ell_n\}$ be an enumeration of $n = MN$ points in $\mathcal{L}$, where each ℓ_i is an ordered pair $(\mathbf{s}_j, t_k)$. Let $y(\ell_i)$ be a univariate response corresponding to ℓ_i and let $\mathbf{x}(\ell_i)$ be a corresponding $p \times 1$ vector of spatio-temporally referenced predictors. A spatio-temporal regression model relates the response and the predictors as

$$y(\ell_i) = \mathbf{x}(\ell_i)'\boldsymbol{\beta} + w(\ell_i) + \epsilon(\ell_i)\ , \quad i = 1, 2, \ldots, MN\ , \tag{5.1}$$

where $\boldsymbol{\beta}$ denotes the coefficient vector for the predictors, $w(\ell_i)$ is the spatio-temporally varying intercept and $\epsilon(\ell_i)$ is the random noise customarily assumed to be independent and identically distributed copies from $N(0, \tau^2)$.

Usually $w(\ell_i)$'s are modeled as realizations of a spatio-temporal GP. To ensure scalability, we will construct a DNNGP from a parent GP with a non-separable spatio-temporal isotropic covariance function $C((\mathbf{s}+\mathbf{h}, t+u), (\mathbf{s}, t) \,|\, \boldsymbol{\theta})$, introduced by Gneiting (2002),

$$\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)(a|u|^{2\alpha}+1)^{\delta+\kappa}} \times \left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\kappa/2}}\right)^{\nu} \times K_{\nu}\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\kappa/2}}\right)\ , \tag{5.2}$$

where $\mathbf{h}$ and u refers to the spatial and temporal lags between $(\mathbf{s}+\mathbf{h}, t+u)$ and $(\mathbf{s}, t)$ and $\boldsymbol{\theta} = \{\sigma^2, \alpha, \kappa, \delta, \nu, a, c\}$. The spatial covariance function at temporal lag zero corresponds to the Whittle-Matern class with marginal variance σ^2, smoothness parameter ν and decay parameter c. The parameters α and a control smoothness and decay, respectively, for the temporal process, while κ captures non-separability between space and time.

A straightforward choice of the reference set $\mathcal{R}$ is the set $\{\ell_1, \ell_2, \ldots, \ell_n\}$. While this set will typically be large, its size does not adversely affect the computations. This choice has been shown to yield excellent approximations to the parent random field (Vecchia, 1988; Stein, Chi and Welty, 2004). Also, while several alternate choices of reference sets (like choosing the points over

a regular grid) are possible, it is unlikely they will provide any additional computational or inferential benefits; this has been demonstrated in purely spatial contexts by Datta et al. (2015). Hence, we choose $\mathcal{R} = \{\ell_1, \ell_2, \ldots, \ell_n\}$, i.e., $\ell_i^* = \ell_i$ for $i = 1, 2, \ldots, n$.

A full hierarchical model with a DNNGP prior on $w(\ell)$ is given by

$$\begin{aligned} & p(\boldsymbol{\theta}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \times N(\mathbf{w}_\mathcal{R} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{R},\mathcal{R}}) \\ & \times \prod_{i=1}^{n} N(\mathbf{y}(\ell_i) \,|\, \mathbf{x}(\ell_i)'\boldsymbol{\beta} + \mathbf{w}(\ell_i), \tau^2) \,, \end{aligned} \tag{5.3}$$

where $p(\boldsymbol{\theta})$ is the prior on $\boldsymbol{\theta}$, and $IG(\tau^2 \,|\, a_\tau, b_\tau)$ denotes the Inverse-Gamma density with shape a_τ and rate b_τ. Below we describe an efficient MCMC algorithm using Gibbs and Metropolis steps only to carry out full inference from the posterior in Equation 5.3.

5.1. *Gibbs' sampler steps.* Let S_o be the points in $\mathcal{R}$ where the $y(\ell_i)$'s is observed and $I(\ell_i)$ denote the binary indicator for presence or absence of data at ℓ_i. Let $\mathbf{y}$ be the $n_o \times 1$ vector formed by stacking the responses observed and $\mathbf{X}$ denotes the corresponding $n_o \times p$ design matrix. The full conditional distribution of $\boldsymbol{\beta}$ is $N(\mathbf{V}_\beta^* \boldsymbol{\mu}_\beta^*, \mathbf{V}_{\boldsymbol{\beta}}^*)$ where $\mathbf{V}_\beta^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}'\mathbf{X}/\tau^2)^{-1}$ and $\boldsymbol{\mu}_\beta^* = (\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'(\mathbf{y} - \mathbf{w}_{S_o})/\tau^2)$. The full conditional distribution of τ^2 follows $IG\left(a_\tau + \frac{n_o}{2}, b_\tau + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o})\right)$.

We update the elements of $\mathbf{w}_\mathcal{R}$ sequentially. For any two locations ℓ_1 and ℓ_2 in $\mathcal{L}$, if $\ell_1 \in N(\ell_2)$ and is the j-th member of $N(\ell_2)$, then we define b_{ℓ_2,ℓ_1} as the j-th entry of $\mathbf{a}_{N(\ell_2)}$. Let $U(\ell_1) = \{\ell_2 \in \mathcal{R} \,|\, \ell_1 \in N(\ell_2)\}$ and for every $\ell_2 \in U(\ell_1)$, define, $a_{\ell_2,\ell_1} = w(\ell_2) - \sum_{\ell \in N(\ell_2), \ell \neq \ell_1} w(\ell) b_{\ell_2,\ell}$. Then, for $i = 1, 2, \ldots, n$ the full conditional distribution for $\mathbf{w}(\ell_i)$ is $N\left(v(\ell_i)\mu(\ell_i), v(\ell_i)\right)$, where

$$\begin{aligned} v(\ell_i) &= \left(I(\ell_i)/\tau^2 + 1/f_{\ell_i} + \textstyle\sum_{\ell \in U(\ell_i)} b_{\ell,\ell_i}^2/f_\ell\right)^{-1} \quad \text{and} \\ \mu(\ell_i) &= (y(\ell_i) - \mathbf{x}(\ell_i)'\boldsymbol{\beta})I(\ell_i)/\tau^2 + \mathbf{a}'_{N(\ell_i)}\mathbf{w}_{N(\ell_i)}/f_{\ell_i} + \textstyle\sum_{\ell \in U(\ell_i)} b_{\ell,\ell_i} a_{\ell,\ell_i}/f_\ell \,. \end{aligned} \tag{5.4}$$

If $U(\ell_i)$ is empty for some ℓ_i, then all summations of the form $\sum_{\ell \in U(\ell_i)}$ in Equation 5.4 vanish for that $\mathbf{w}(\ell_i)$.

5.2. *Metropolis step.* We update $\boldsymbol{\theta}$ using a random walk Metropolis step. The full-conditional for $\boldsymbol{\theta}$ is proportional to

$$p(\boldsymbol{\theta})p(\mathbf{w}_\mathcal{R} \,|\, \boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \times \prod_{i=1}^{n} N\left(\mathbf{w}(\ell_i) \,|\, \mathbf{a}'_{N(\ell_i)}\mathbf{w}_{N(\ell_i)}, f_{\ell_i}\right) \,. \tag{5.5}$$

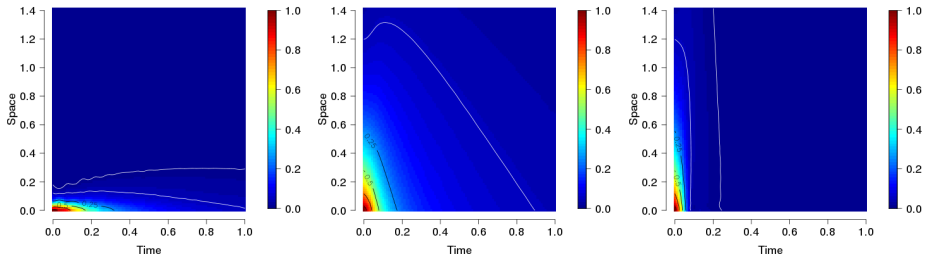
Since none of the above updates involve expensive matrix decompositions, the likelihood can be evaluated very efficiently. The algorithm for updating the parameters for a hierarchical DNNGP model is exactly analogous to the corresponding updates for a purely spatial NNGP model (see Datta et al. (2015)). The only additional computational burden stems from updating the neighbor sets in the adaptive neighbor selection scheme, but even this can be handled efficiently using eligible sets. Hence, the number of floating point operations per update is linear in the number of points in $\mathcal{L}$.

5.3. *Prediction.* Once we have computed the posterior samples of the model parameters and the spatio-temporal random effects over $\mathcal{R}$, we can execute, cheaply and efficiently, full posterior predictive inference at unobserved locations and time points. The Gibbs' sampler in Section 5.1 generates full posterior distributions of the $\mathbf{w}$'s at all locations in $\mathcal{R}$. Let ℓ_i^* denote a point in $\mathcal{R}$ where the response is unobserved i.e. $I(\ell_i^*) = 0$. We already have posterior distributions of $\mathbf{w}(\ell_i^*)$ and the parameters. We can now generate posterior samples of $\mathbf{y}(\ell_i^*)$ from $N(\mathbf{x}(\ell_i^*)'\boldsymbol{\beta} + \mathbf{w}(\ell_i^*), \tau^2)$. Turning to prediction at a location ℓ outside $\mathcal{R}$, we construct $N(\ell)$ from $E(\ell)$ described in Equation 4.3 for every posterior sample of $\boldsymbol{\theta}$. We generate posterior samples of $\mathbf{w}(\ell)$ from $N(\mathbf{a}'_{N(\ell)}\mathbf{w}_{N(\ell)}, f_\ell)$ and, subsequently, draw posterior samples of $y(\ell)$ from $N(\mathbf{x}(\ell)'\boldsymbol{\beta} + \mathbf{w}(\ell), \tau^2)$.

6. Synthetic data analyses. We generated observations over a $n = 15 \times 15 \times 15 = 3375$ grid within a unit cube domain. An additional 500 observations used for out-of-sample prediction validation were also located within the domain. All data were generated using model 5.1 with $\mathbf{x}(\ell)$ comprising an intercept and covariate drawn from $N(0, 1)$. The spatial covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ was constructed using an exponential form of the non-separable spatio-temporal covariance function 5.2. Specifically, the (i, j)-th element of $\mathbf{C}(\boldsymbol{\theta})$ was calculated using

$$\frac{\sigma^2}{(a|u|^2+1)^\kappa}\exp\left(\frac{-c\|h\|}{(a|u|^\alpha+1)^{\kappa/2}}\right), \tag{6.1}$$

where $u = |t_i - t_j|$ and $h = ||\mathbf{s}_i - \mathbf{s}_j||$ are the time and space Euclidean norms, respectively. By specifying different values of the decay and interaction parameters in $\boldsymbol{\theta} = (\sigma^2, \kappa, a, c)$, function 6.1 was used to generated three datasets that exhibited different covariance structures. The first column in Table 1 provides the three specifications for $\boldsymbol{\theta}$ and Figure 4 shows the corresponding space-time correlation surface realizations. As illustrated in Figure 4, the three datasets exhibit: 1) short spatial range and long tem-



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Fig 4: Space-time correlation surface realizations given *true* parameter values in Table 1. Correlation contours are provided, with the two outer white lines corresponding to 0.05 and 0.01.

poral range; 2) long spatial and temporal range, and; 3) long spatial range and short temporal range.

For each dataset, model parameters were estimated using the: *i*) full Gaussian Process (GP); *ii*) DNNGP with simple neighbor set selection (Simple DNNGP) described in Section 4.1; *iii*) DNNGP with adaptive neighbor set selection (Adaptive DNNGP) described in Section 4.2, and; *iv*) bias-corrected Gaussian Predictive Process (GPP) detailed in Banerjee et al. (2008) and Finley, Banerjee and McRoberts (2009). DNNGP models were fit using $m = \{16, 25, 36\}$ and the Gaussian Predictive Process model used a regularly spaced grid of $8 \times 8 \times 8 = 512$ knots within the domain.

For all models, the intercept and slope regression parameters, β_0 and β_1, were assigned *flat* prior distributions. The variance parameters were assumed to follow inverse-Gamma prior distributions width $\sigma^2 \sim IG(2, 1)$ and $\tau^2 \sim IG(2, 0.1)$. The time and space decay parameters received uniform priors that were dataset specific: 1) $a \sim U(1, 100)$, $c \sim U(0, 50)$; 2) $a \sim U(300, 700)$, $c \sim U(0, 10)$, and; 3) $a \sim U(1000, 3000)$, $c \sim U(0, 10)$. The prior for the interaction term matched its theoretical support with $\kappa \sim U(0, 1)$.

Candidate model comparison was based on parameter estimates, fit to the observed data, out-of-sample prediction accuracy, and posterior predictive distribution coverage. Goodness-of-fit was assessed using DIC (Spiegelhalter et al., 2002) and GPD (Gelfand and Ghosh, 1998). Prediction accuracy for the 500 holdout locations was measured using root mean squared prediction error (Yeniay and Goktas, 2002). The percent of holdout locations that fell within the candidate models' posterior predictive distribution's 95% credible

Table 1: Synthetic data analysis parameter estimates and computing time for the candidate models. Parameter posterior summary 50 (2.5, 97.5) percentiles. Bold indicates estimates with 95% credible intervals that do not include the *true* parameter value.

		GP	GPP knots=512	Adaptive DNNGP m=25	Simple DNNGP m=25
			Dataset 1		
β_0	1	0.99 (0.80, 1.12)	1.02 (0.89, 1.16)	0.97 (0.86, 1.11)	0.97 (0.86, 1.11)
β_1	5	4.99 (4.97, 5.01)	4.98 (4.94, 5.02)	4.99 (4.97, 5.01)	4.99 (4.97, 5.01)
a	50	46.46 (38.02, 67.46)	**16.93 (11.91, 29.17)**	48.37 (37.85, 77.93)	53.18 (35.93, 83.78)
c	25	25.69 (22.00, 29.49)	22.73 (13.53, 34.20)	24.77 (20.55, 28.71)	25.16 (21.91, 29.52)
k	0.75	0.83 (0.61, 0.94)	0.78 (0.39, 0.91)	0.80 (0.57, 0.99)	0.75 (0.53, 0.98)
σ^2	1	**1.13 (1.03, 1.24)**	**0.70 (0.56, 0.92)**	**1.13 (1.03, 1.24)**	**1.14 (1.04, 1.25)**
τ^2	0.1	0.09 (0.07, 0.11)	**0.95 (0.89, 1.02)**	0.09 (0.06, 0.11)	0.09 (0.06, 0.11)
pD		2214.57	225.66	2236.81	2258.39
DIC		3700.68	9644.76	3650.55	3567.45
G		104.60	3008.41	100.06	94.50
P		512.29	3436.52	504.66	493.63
D		616.90	6444.93	604.72	588.13
RMSPE		0.84	0.95	0.84	0.84
95% CI cover %		95.6	94.6	95.6	96
			Dataset 2		
β_0	1	0.81 (0.48, 1.26)	0.79 (0.26, 1.16)	1.01 (0.57, 1.27)	0.92 (0.58, 1.26)
β_1	5	4.98 (4.96, 5.00)	4.99 (4.97, 5.02)	4.98 (4.96, 5.00)	4.98 (4.96, 5.00)
a	500	352.82 (301.69, 521.64)	583.59 (391.79, 661.36)	480.28 (331.63, 662.12)	410.84 (317.29, 602.21)
c	2.5	2.52 (1.93, 3.13)	**1.67 (1.03, 2.31)**	2.91 (2.49, 3.37)	2.59 (1.91, 3.37)
k	0.5	0.56 (0.44, 0.67)	0.39 (0.26, 0.53)	0.46 (0.36, 0.62)	0.53 (0.42, 0.67)
σ^2	1	1.01 (0.85, 1.31)	1.14 (0.83, 1.77)	0.94 (0.81, 1.10)	1.03 (0.83, 1.32)
τ^2	0.1	0.11 (0.09, 0.13)	**0.44 (0.41, 0.47)**	0.10 (0.08, 0.12)	0.10 (0.08, 0.12)
pD		1913.96	312.76	1999.98	2014.78
DIC		3988.36	7091.84	3866.38	3871.09
G		157.52	1336.94	139.28	137.33
P		576.02	1609.99	550.28	550.36
D		733.53	2946.94	689.56	687.68
RMSPE		0.53	0.71	0.53	0.53
95% CI cover %		96.4	93	94	93.8
			Dataset 3		
β_0	1	0.94 (0.66, 1.14)	**0.55 (0.32, 0.84)**	0.93 (0.74, 1.17)	0.87 (0.68, 1.12)
β_1	5	4.98 (4.96, 5.00)	4.98 (4.95, 5.02)	4.98 (4.96, 5.00)	4.98 (4.96, 5.00)
a	2000	1214.02 (1008.23, 2141.16)	1590.77 (1151.78, 2118.63)	1635.46 (1046.76, 2922.59)	1495.94 (1019.16, 2751.17)
c	2.5	2.38 (1.79, 2.95)	**1.36 (0.73, 2.16)**	2.25 (1.62, 2.81)	2.27 (1.60, 2.97)
k	0.95	0.91 (0.72, 0.98)	**0.68 (0.40, 0.90)**	0.71 (0.46, 0.98)	0.81 (0.58, 0.97)
σ^2	1	1.03 (0.86, 1.35)	0.91 (0.67, 1.83)	1.09 (0.89, 1.44)	1.07 (0.87, 1.42)
τ^2	0.1	0.11 (0.09, 0.13)	**0.68 (0.62, 0.74)**	0.11 (0.09, 0.14)	0.11 (0.09, 0.14)
pD		1990.41	210.11	1982.28	1994.77
DIC		4210.71	8463.33	4214.68	4223.88
G		155.87	2137.55	157.24	155.61
P		610.01	2424.66	611.89	612.98
D		765.89	4562.21	769.13	768.59
RMSPE		0.78	0.92	0.77	0.77
95% CI cover %		92.8	91.4	95.6	95.6
CPU (min)		7646.96	856.54	496.12	430.88

interval (CI) was also computed. Candidate model inference was based on 15,000 MCMC samples comprising post burn-in samples from three chains of 25,000 iterations (i.e., 5,000 samples from each chain).

Dataset analysis results are given in Table 1. With the exception of τ^2 for Dataset 1, the full GP model recovered the parameter values used to generate the datasets, i.e., the 95% CIs cover the *true* parameter values. For the DNNGP models, there was negligible difference among parameter estimates for the 15, 25, and 36 neighbor sets and therefore we report only the $m = 25$ cases. There was very little difference between the estimates produced by the Adaptive and Simple DNNGP models, and, like the full GP model, they captured the *true* mean and process parameters, with the exception of τ^2 for Dataset 1. Given the extremes in the space and time decay in Datasets 1 and 3, we anticipated the Simple DNNGP model—with at most 5 neighbors in any given time point—would not be able to estimate the covariance parameters. Extensive analysis of simulated data, some of which is reported in Table 1, suggested the Simple DNNGP model performed as well as the Adaptive DNNGP and full GP models. Model goodness-of-fit and out-of-sample prediction validation metrics in Table 1 also show the full GP and DNNGP models provided comparable results. In contrast the GPP model did not capture many of the process parameters and provided worse fit and prediction than the GP and DNNGP models. The quality of the GPP results would improve with additional knots, however, computational time would also increase. The last row in Table 1 provides the CPU time required for each candidate model to generate 25,000 MCMC samples for the $n = 3375$ observations. Even with the substantial dimension reduction, the GPP model required about twice the CPU time as the DNNGP models. Compared to the full GP model, the DNNGP models provided substantial computational advantage while delivering comparable results.

7. Analysis of Airbase and LOTOS-EUROS CTM data. We consider the model in Equation 5.3, where $y(\ell_i)$ is a square-root transformed measurement of PM_{10} at space-time coordinate ℓ_i, $x(\ell_i)$ is the coinciding square-root transformed output from the LOTOS-EUROS CTM. Given the large dimension of the dataset, $n = N \times M = 308 \times 730 = 224{,}840$, the spatio-temporal random effects were modeled as a DNNGP prior derived from a zero-centered GP with the non-separable spatio-temporal covariance function (6.1). Exploratory analysis—consisting of semivariogram and autocorrelation function plots of non-spatial model residuals—helped guide choice of prior and hyper-parameters for the variance and decay parameters. Specifically, $\sigma^2 \sim IG(2, 1)$, $\tau^2 \sim IG(2, 0.1)$, $a \sim U(0.1, 5)$, and $c \sim U(0.01, 0.5)$,

with κ fixed at 0.5.

Candidate models included the: *i*) LOTOS-EUROS CTM; *ii*) simple linear regression model with no spatio-temporal effects, i.e., $w(\ell) = 0$, and; *iii*) Adaptive and Simple DNNGP with $m = \{16, 25, 36\}$. Following Section 6, candidate model goodness-of-fit to the observed data was assessed using DIC and GPD, whereas predictive performance was assessed using RMSPE and 95% posterior predictive CI coverage rate for out-of-sample prediction. The holdout set comprised blocks of five days per station—five days of continuous observations were withheld at random from each station's 730 day time series.

Additionally, prediction using the Adaptive and Simple DNNGP models for a 25% holdout set selected from April 1-14, 2009 was compared with results from Hamm et al. (2015) who considered time invariant spatial regression models for the same two-week period and comparable prediction validation approach.

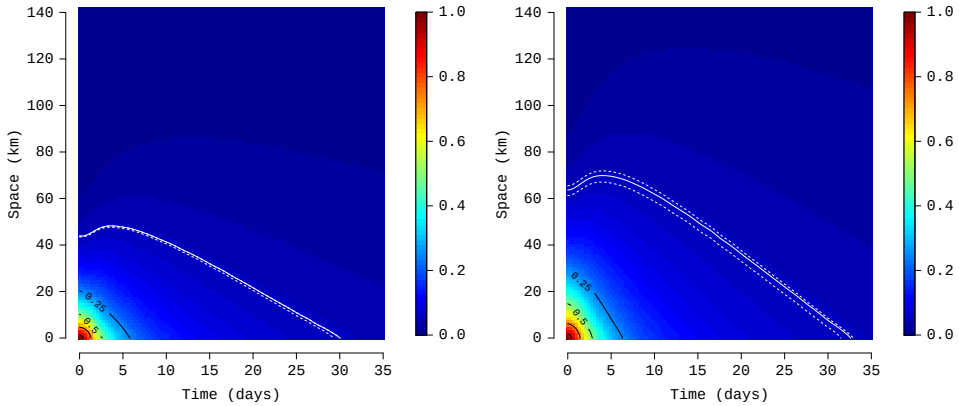
A subset of analysis results are given in Table 2. Parameter estimates for the model intercept and regression slope coefficient associated with the CTM output are consistent across the candidate models. For an accurate CTM it would be expected that $\beta_0 \approx 0$ and $\beta_1 \approx 1$. The finding that $\beta_0 > 0$ and $0 < \beta_1 < 1$ corroborate previous findings that showed the CTM consistently under estimates PM_{10} (Stern et al., 2008; Hamm et al., 2015). The spatial and temporal decay parameters differed between the Adaptive and Simple DNNGP models. Figure 5 provides correlation surfaces generated using posterior median values of a and c from the $m = 36$ Adaptive and Simple DNNGP models (using values given in Table 2). The 0.05 correlation contour on these surfaces suggest the Simple model estimates a moderately longer spatial and temporal range, i.e., ∼60 km and ∼33 days, versus ∼45 km and ∼30 days for the Adaptive model. Within a given DNNGP neighbor selection algorithm there is only marginal difference between the covariance parameters estimates when comparing m of 25 and 36. Neighbor sets of less than 25 provided consistently larger temporal decay parameter estimates, i.e., shorter temporal correlation estimates, although even with such few neighbors the models seemed to produce consistent estimates of the spatial decay.

The spatial range of 45 to 60 km is an order of magnitude less than that observed by Hamm et al. (2015), who estimated median spatial ranges of 500 to 1500 km. This is attributed to the inclusion of temporal correlation in the model, which itself accounts for a large amount of the residual spatial structure. The temporal range is physically reasonable considering the lifetime of PM_{10} is in the order of days and its variability is driven by alternating

Table 2: PM_{10} analysis parameter posterior 50 (2.5, 97.5) percentiles, model fit and prediction metrics, and run time for 25,000 MCMC samples.

		Adaptive			Simple
Parameter	Non space-time	m=16	m=25	m=36	m=36
β_0	1.66 (1.64, 1.68)	2.56 (2.53, 2.59)	2.62 (2.59, 2.65)	2.61 (2.58, 2.64)	2.64 (2.61, 2.68)
β_1	0.76 (0.75, 0.76)	0.47 (0.46, 0.47)	0.45 (0.44, 0.46)	0.45 (0.44, 0.46)	0.44 (0.43, 0.45)
a	–	0.57 (0.57, 0.57)	0.44 (0.44, 0.44)	0.46 (0.46, 0.46)	0.37 (0.37, 0.39)
c	–	0.08 (0.08, 0.08)	0.07 (0.07, 0.07)	0.07 (0.07, 0.07)	0.05 (0.05, 0.05)
σ^2	–	1.49 (1.48, 1.51)	1.64 (1.62, 1.66)	1.56 (1.54, 1.58)	2.06 (2.01, 2.11)
τ^2	1.48 (1.47, 1.48)	0.12 (0.12, 0.12)	0.14 (0.14, 0.14)	0.14 (0.14, 0.14)	0.15 (0.15, 0.16)
P_D	2.75	110266.2	122466.2	111190.6	103038.3
DIC	586,135.8	279077.3	265720.6	277383.9	286,922.9
G	432811.9	11538.63	8707.79	11249.11	13521.63
P	268036.7	40994.19	36711.28	40532.25	43728.23
D	700848.6	52532.82	45419.07	51781.37	57249.86
RMSPE	12.75	8.28	8.24	8.2	8.11
95% CI cover %	93.4	93.33	93.06	93.15	92.86
CPU (min)	–	6182.89	15681.8	27660.5	25819

synoptic meteorological conditions, with certain conditions usually lasting for several days to weeks.



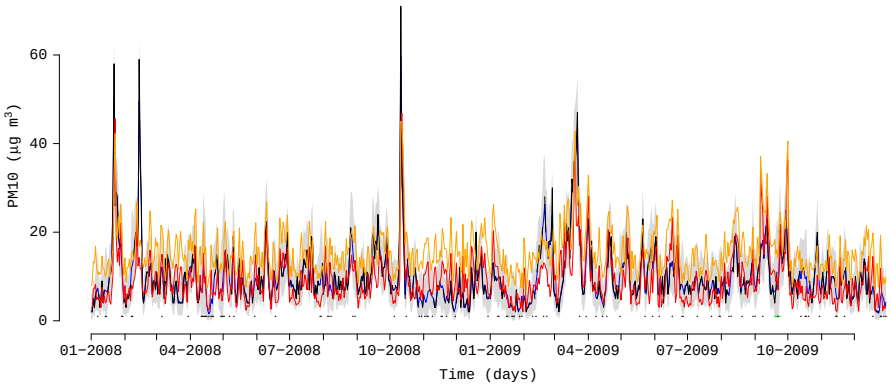
(a) Adaptive m=36 (b) Simple m=36

Fig 5: Space-time correlation posterior distribution median surfaces. Median (white lines) and associated 95% credible intervals (dotted white lines) for correlation contours of 0.05.

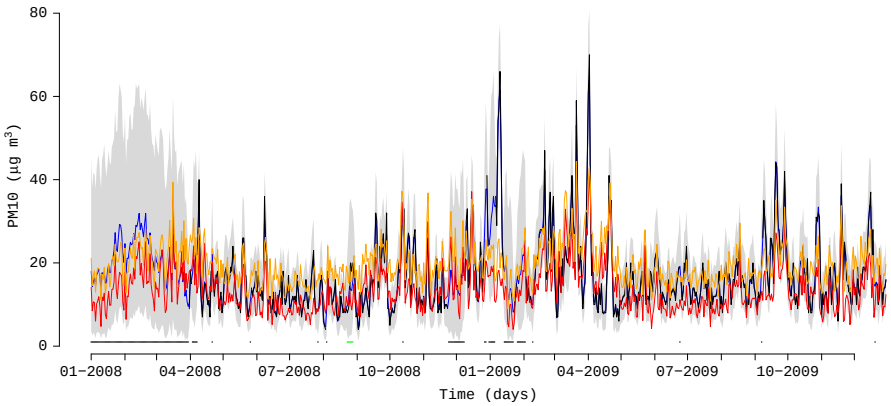
Across all candidate models the Adaptive with m=25 provided the lowest values of DIC and D suggesting improved fit to the observed data. This improved fit did not correspond to increased out-of-sample prediction accuracy. Rather, RMSPE consistently decreased with increasing number of neighbors within the Adaptive and Simple model sets. The smallest RMSPE was achieved using the simple neighbor selection with m=36. All models achieved reasonable coverage rates.

Figure 6 illustrates the observed and candidate model fitted/predicted PM_{10} for three station. These figures are representative of other stations and show: *i*) the downward bias in CTM output; *ii*) improved fit and prediction with the addition of spatio-temporal random effects over non-spatial regression, and; *iii*) appropriate widening of CIs for missing station observations.

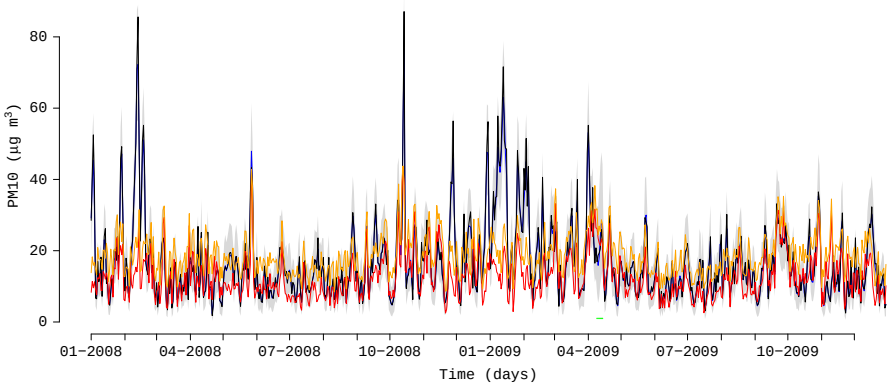
Table 3 provides out-of-sample prediction validation metrics for DNNGP Adaptive and Simple models that can be compared with April 1-14, 2009 holdout validation metrics presented in Hamm et al. (2015, Table 1). Compared to the time invariant (day specific) space-varying intercept (SVI) and space-varying coefficients (SVC) models considered in Hamm et al. (2015), the DNNGP models' RMSPE and bias are lower (more accurate, less biased) while the R^2 values are comparable.



(a) Station 14



(b) Station 47



(c) Station 151

Fig 6: Fitted and observed PM_{10} for several example stations. Lines correspond to PM_{10} observed (black), CTM output (red), non space-time, regression (orange), and $m = 36$ Adaptive DNNGP (blue) with associated 95% CI band (gray). Prediction assessment holdout and actual missing observations are indicated with green and black points respectively.

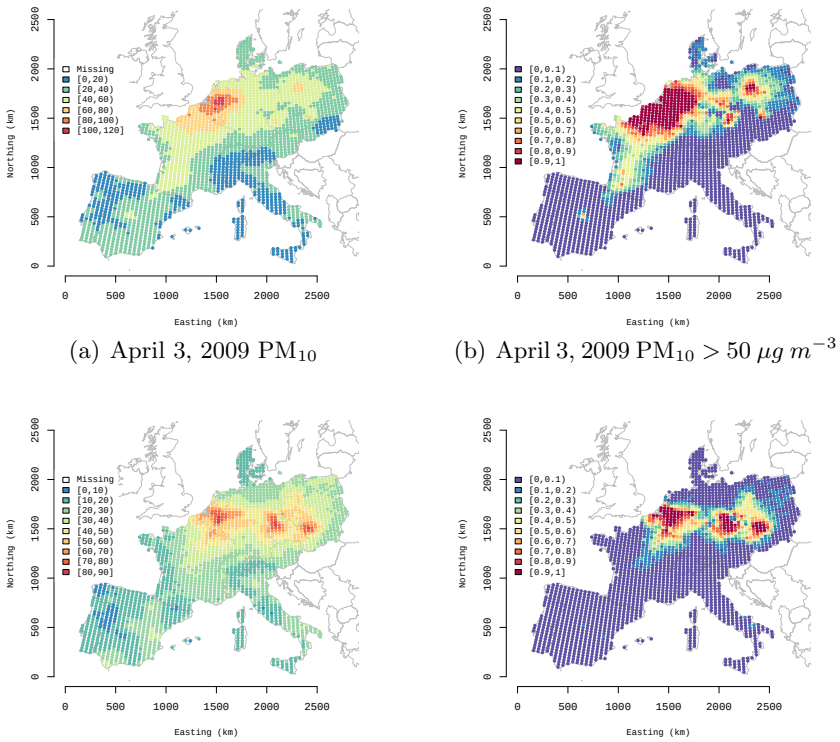
TABLE 3
April 1-14, 2009 25% holdout set prediction summary for comparison with time invariant spatial regression models presented in (Hamm et al., 2015, Table 1).

	Adaptive		Simple	
	m=25	m=36	m=25	m=36
RMSPE	4.97	5.05	5.06	5.04
Bias	0.20	0.20	0.23	0.22
R^2	0.69	0.68	0.68	0.68

In addition to these prediction metrics, maps of posterior predictive summaries at CTM output locations are key inputs to pollution monitoring and mitigation programs. For example, Figure 7 provides maps of the posterior predictive prediction median and the probability of exceeding the 50 μg m^{-3} regulatory threshold for two example dates. These dates were also examined in Hamm et al. (2015, Figure 8) and the resulting maps are directly comparable. The DNNGP, Figure 7, and SVC maps in Hamm et al. (2015) show broadly similar patterns, although there are some differences. For example the high pollution over western France and northern Spain on April 3, 2009 is captured more clearly by Hamm et al. (2015). The SVI and SVC models in Hamm et al. (2015) did not account for temporal correlation over days—clearly not an accurate assumption. In contrast the DNNGP models smooth over days, which can provide improved predictive performance although the details of highly dynamic events may be less well captured than by the daily specific models used in Hamm et al. (2015).

The last row in Table 2 provides the CPU time for delivering 25,000 MCMC iterations. As detailed in Section 4.2 particular components of the algorithm are easily distributed across multiple CPUs. In particular, partitioning the update of $w(\ell_i)$'s across multiple CPUs yields substantial computational gains. The DNNGP samplers where implemented in C++ and leveraged OpenMP (Dagum and Menon, 1998) and Intel Math Kernel Library's (MKL) threaded BLAS and LAPACK routines for matrix (Intel, 2015). Running on a single CPU the Adaptive m=25 model would require approximately 260 hours. However, when distributed across a 10-core Xeon CPU the total run time was approximately 24 hours.

8. Conclusion. We have addressed the problem of modeling large spatio-temporal datasets, specifically for settings where full inference (with proper accounting for uncertainty) is required at arbitrary resolutions. We presented a new class of dynamic nearest-neighbor Gaussian Process (DNNGP) models over a continuous space-time domain. The DNNGP is a legitimate Gaussian process whose realizations over finite sets enjoy sparse precision



(a) April 3, 2009 PM_{10}

(b) April 3, 2009 $PM_{10} > 50\ \mu g\ m^{-3}$

(c) April 5, 2009 PM_{10}

(d) April 5, 2009 $PM_{10} > 50\ \mu g\ m^{-3}$

Fig 7: Predicted PM_{10} and probability of exceeding 50 $\mu g\ m^{-3}$ for two example dates.

matrices, thereby accruing massive computational savings in terms of storage and flops. The DNNGP depends upon the conditional independence of the random effects given its neighbors. We used the strength of a correlation function to construct a parametric distance metric in a spatio-temporal domain. Using monotonicity of covariance functions we showed that it is possible to update neighbor sets using a scalable search algorithm and outlined the steps of a Gibbs' sampler that avoids expensive matrix decompositions and is linear in the number of measurements in terms of storage and flops.

Analyses combining European CTM outputs and observed data has, to date, focused mainly on spatial analysis per day Denby et al. (2008, 2010); Hamm et al. (2015), few studies implement full space-time geostatistical models, e.g., Gräler, Gerharz and Pebesma (2011), and none consider such a long time series. The work presented in this paper focuses on DNNGP development to facilitate novel analyses of spatially-indexed time-series data such as PM_{10} concentrations. Here, in addition to improved predictive performance, inference on model covariance parameters provided insight into space-time structures not captured by the LOTOS-EUROS CTM. Whilst previous analyses of individual days had shown strong residual spatial structure, analysis of this long time-series with explicit time correlation parameters reveals the residual temporal structure dominates. The temporal range is physically reasonable considering the life-time of PM_{10} is in the order of days and its variability is driven by alternating synoptic meteorological conditions, with certain conditions usually lasting for several days to weeks.

Reproducing the observed variability with a CTM remains challenging, especially for episodic conditions which associated with particular (stagnant) meteorological conditions or occasional large emissions from, e.g., large wild fires (R'Honi et al., 2013) or dust events (Birmili et al., 2008). A particular issue to be resolved is the lack of detail in the anthropogenic emission variability. This variability is prescribed using static emission profiles for the month of the year, day of the week, and hour of the day. Further detailing through inclusion of meteorological effects may improve the modeling (Mues et al., 2014) and remove the monthly signature found in this analysis.

The type of analysis that is performed depends on the study objective. Analysis of individual days is important for the study of individual air pollution events and the associated performance of the CTM Hamm et al. (2015). The analysis presented in this paper affords a different perspective by identifying long-term space-time structures that offer insight into the performance of the CTM. The DNNGP also yields more accurate predictions than previous studies of these same data.

Apart from massive scalability, the DNNGP retains the versatility of

process-based modeling and can be used as a sparsity-inducing proper prior in any Bayesian hierarchical model designed to deliver full inference at arbitrary spatio-temporal resolutions for massive spatio-temporal datasets. Even more generally, the DNNGP can be used for any spatio-temporal random effect in the second stage of specification in hierarchical models for non-Gaussian responses. The continuous spatio-temporal framework for the DNNGP model ensures that we can interpolate and extrapolate at arbitrary spatial or temporal resolutions. Full posterior distributions for the underlying spatio-temporal process are available at any arbitrary location and time point. Thus, DNNGP can potentially be deployed for statistical downscaling of spatio-temporal datasets obtained at coarser resolutions (e.g. climate downscaling). We also plan to migrate our lower level C++ code into the `spBayes` R package for wider and friendlier accessibility to DNNGP models.

APPENDIX A: ELIGIBLE SETS

For any point ℓ and any set $\mathcal{U}$ in $\mathcal{L}$ and an isotropic covariance function $C(\cdot,\cdot\,|\,\boldsymbol{\theta})$, let $NN(\ell,\mathcal{U},\boldsymbol{\theta},m)$ denote the set of m-nearest neighbors of ℓ in $\mathcal{U}$ based on $C(\cdot,\cdot\,|\,\boldsymbol{\theta})$. Then we have the following result:

PROPOSITION. *If $C(h,u\,|\,\boldsymbol{\theta})$ posses natural monotonicity defined in Section 4.1 for every value of $\boldsymbol{\theta}$, then*

(a) For every ℓ^* in $\mathcal{R}$, the eligible set $E(\ell^*)$ defined in Equation 4.2 contains $NN(\ell^*,H(\ell^*),m,\boldsymbol{\theta})$ for every value of $\boldsymbol{\theta}$

(b) For every ℓ outside $\mathcal{R}$, the eligible set $E(\ell)$ defined in Equation 4.3 contains $NN(\ell,\mathcal{R},m,\boldsymbol{\theta})$ for every value of $\boldsymbol{\theta}$

PROOF. We only prove part (a). Part (b) can be proved similarly. Without loss of generality, we write $\ell^* = (\mathbf{s}_i,t_j)$ for some $1 \le i \le N$ and $1 \le j \le M$. We assume that $(\mathbf{s}_u,t_{j-k}) \in NN(\ell_i^*,H(\ell_i^*),m,\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$, $u \le N$ and $k \ge 1$. Also let $\mathbf{s}_i[l]$ denote the l^{th} nearest neighbor of $\mathbf{s}_i$ among $\{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_N\}$. So, $\mathbf{s}_u = \mathbf{s}_i[l]$ for some $l \ge 1$. Therefore, by natural monotonicity of C, we have $C((\mathbf{s}_i,t_j),(s_i[a],t_{j-k})\,|\,\boldsymbol{\theta}) \ge C((\mathbf{s}_i,t_j),(s_i[l],t_{j-k})\,|\,\boldsymbol{\theta})$ for all $1 \le a \le l$. One more application of natural monotonicity implies that $C((\mathbf{s}_i,t_j),(s_i[a],t_{j-b})\,|\,\boldsymbol{\theta}) > C((\mathbf{s}_i,t_j),(s_i[a],t_{j-k})\,|\,\boldsymbol{\theta})$ for all $1 \le b \le k$. As $(\mathbf{s}_u,t_{j-k}) \in NN(\ell,\mathcal{U},\boldsymbol{\theta},m)$, we have $(s_i[a],t_{j-b}) \in NN(\ell,\mathcal{U},\boldsymbol{\theta},m)$ for all $a \le l$ and $b \le k$. Therefore, $lk \le m$ i.e. $l \le [m/k]$. □

REFERENCES

ALLCROFT, D. J. and GLASBEY, C. A. (2003). A Latent Gaussian Markov Random Field Model for Spatio-temporal Rainfall Disaggregation. *Journal of the Royal Statistical society, Series C* **52** 487-498.

Bai, Y., Song, P. X. K. and Raghunathan, T. E. (2012). Bayesian Dynamic Modeling for Large Space-time Datasets using Gaussian Predictive Processes. *Journal of the Royal Statistical society, Series B* **74** 799-824.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, second ed. Chapman & Hall/CRC, Boca Raton, FL.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian Predictive Process Models for Large Spatial Datasets. *Journal of the Royal Statistical society, Series B* **70** 825-848.

Bevilacqua, M., Genton, M., Porcu, E. and Zastavnyi, V. (2014). Adaptive Tapering for Space-time Covariance Functions. *Submitted*.

Birmili, W., Schepanski, K., Ansmann, A., Spindler, G., Tegen, I., Wehner, B., Nowak, A., Reimer, E., Mattis, I., Muller, K., Bruggemann, E., Gnauk, T., Herrmann, H., Wiedensohler, A., Althausen, D., Schladitz, A., Tuch, T. and Loschau, G. (2008). A Case of Extreme Particulate Matter Concentrations over Central Europe Caused by Dust Emitted over the Southern Ukraine. *Atmospheric Chemistry and Physics* **8** 997-1016.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Van Dingenen, R., van Donkelaar, A. and Thurston, G. D. (2011). Exposure Assessment for Estimation of the Global Burden of Disease Attributable to Outdoor Air Pollution. *Environmental Science and Technology* **46** 652-660.

Brunekreef, B. and Holgate, S. T. (2002). Air Pollution and Health. *The Lancet* **360** 1233-1242.

Candiani, G., Carnevale, C., Finzi, G., Pisoni, E. and Volta, M. (2013). A Comparison of Reanalysis Techniques: Applying Optimal Interpolation and Ensemble Kalman Filtering to Improve Air Quality Monitoring at Mesoscale. *Science of the Total Environment* **458-460** 7-14.

European Commission (2015). European Union Air Quality Standards. *http://ec.europa.eu/environment/air/quality/standards.htm*.

Crainiceanu, C. M., Diggle, P. J. and Rowlingson, B. (2008). Bivariate Binomial Spatial Modeling of Loa Loa Prevalence in Tropical Africa. *Journal of the American Statistical Association* **103** 21-37.

Cressie, N. and Huang, H. C. (1999). Classes of Nonseparable, Spatio-temporal Stationary Covariance Functions. *Journal of the American Statistical Association* **94** 1330-1340.

Cressie, N. and Johannesson, G. (2008). Fixed Rank Kriging for Very Large Data Sets. *Journal of the Royal Statistical society, Series B* **70** 209-226.

Cressie, N., Shi, T. and Kang, E. L. (2010). Fixed Rank Filtering for Spatio-temporal Data. *Journal of Computational and Graphical Statistics* **19** 724-745.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-temporal Data. Wiley series in probability and statistics*. Hoboken, N.J. Wiley.

Dagum, L. and Menon, R. (1998). OpenMP: An Industry Standard API for Shared-memory Programming. *Computational Science & Engineering, IEEE* **5** 46–55.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2015). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association* In Press.

Denby, B., Schaap, M., Segers, A., Builtjes, P. and Horalek, J. (2008). Comparison of Two Data Assimilation Methods for Assessing PM10 Exceedances on the European Scale. *Atmospheric Environment* **42** 7122-7134.

Denby, B., Sundvor, I., Cassiani, M., de Smet, P., de Leeuw, F. and Horalek, J.

(2010). Spatial Mapping of Ozone and SO2 Trends in Europe. *Science Of The Total Environment* **408** 4795-4806.

Du, J., Zhang, H. and Mandrekar, V. S. (2009). Fixed-domain Asymptotic Properties of Tapered Maximum Likelihood Estimators. *Annals of Statistics* **37** 3330-3361.

Eeftens, M., Tsai, M. Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., De Nazelle, A., de Vocht, F., Declercq, C., Dedele, A., Eriksen, K., Galassi, C., Grazuleviciene, R., Grivas, G., Heinrich, J., Hoffmann, B., Iakovides, M., Ineichen, A., Katsouyanni, K., Korek, M., Kramer, U., Kuhlbusch, T., Lanki, T., Madsen, C., Meliefste, K., Molter, A., Mosler, G., Nieuwenhuijsen, M., Oldenwening, M., Pennanen, A., Probst-Hensch, N., Quass, U., Raaschou-Nielsen, O., Ranzi, A., Stephanou, E., Sugiri, D., Udvardy, O., Vaskoevi, E., Weinmayr, G., Brunekreef, B. and Hoek, G. (2012). Spatial Variation of PM2.5, PM10, PM2.5 Absorbance and PMcoarse Concentrations Between and Within 20 European Study Areas and the Relationship with NO2 - Results of the ESCAPE Project. *Atmospheric Environment* **62** 303-317.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (2014). Estimation and Prediction in Spatial Models with Block Composite Likelihoods. *Journal of Computational and Graphical Statistics* **23** 295-315.

Fasso, A., Finazzi, F. and Bevilacqua, M. (2011). Tapering Spatio Temporal Models. In *Spatial Data Methods for Environmental and Ecological Processes 2nd Edition,* (ed Cafarelli).

Finley, A. O., Banerjee, S. and McRoberts, R. E. (2009). Hierarchical Spatial Models for Predicting Tree Species Assemblages across Large Domains. *Ann. Appl. Stat.* **3** 1052–1079.

Finley, A. O., Banerjee, S. and Gelfand, A. E. (2012). Bayesian Dynamic Modeling for Large Space-time Datasets using Gaussian Predictive Processes. *Journal of Geographical Systems* **14** 29-47.

Flemming, J., Inness, A., Flentje, H., Huijnen, V., Moinat, P., Schultz, M. G. and Stein, O. (2009). Coupling Global Chemistry Transport Models to ECMWF's Integrated Forecast System. *Geoscientific Model Development* **2** 253-265.

Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics* **15** 503-523.

Gelfand, A. E., Banerjee, S. and Gamerman, D. (2005). Spatial Process Modelling for Univariate and Multivariate Dynamic Spatial Data. *Environmetrics* **16** 465-479.

Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **85** 1–11.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Spacetime Data. *Journal of the American Statistical Association* **97** 590–600.

Gneiting, T., Genton, M. G. and Guttorp, P. (2007). Geostatistical Space-time Models, Stationarity, Separability and Full Symmetry. In *Statistics of SpatioTemporal Systems* 151-175. Chapman and Hall CRC Press (eds Finkenstaedt, B. and Held, L. and Isham, V.).

Gneiting, T. and Guttorp, P. (2010). Continuous-parameter Spatio-temporal Processes. *Handbook of Spatial Statistics* 427-436. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 427-436.

Gräler, B., Gerharz, L. and Pebesma, E. (2011). Spatio-temporal Analysis and In-

terpolation of PM10 Measurements in Europe. *ETC/ACM Technical Paper* **10**.

Hamm, N. A. A., Finley, A. O., Schaap, M. and Stein, A. (2015). A Spatially Varying Coefficient Model for Mapping PM10 Air Quality at the European scale. *Atmospheric Environment* **102** 393–405.

Hendriks, C., Kranenburg, R., Kuenen, J., van Gijlswijk, R., Kruit, R. W., Segers, A., van der Gon, H. D. and Schaap, M. (2013). The Origin of Ambient Particulate Matter Concentrations in the Netherlands. *Atmospheric Environment* **69** 289-303.

Higdon, D. (2001). Space and Space Time Modeling using Process Convolutions. *Technical Report, Institute of Statistics and Decision Sciences, Duke University, Durham.*

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B. and Kaufman, J. D. (2013). Long-term Air Pollution Exposure and Cardio- respiratory Mortality: A Review. *Environmental Health* **12** 43.

Intel (2015). Math Kernel Library. http://developer.intel.com/software/products/mkl/.

Jones, R. H. and Zhang, Y. (1997). Models for Continuous Stationary Space-time Processes. In *Modelling Longitudinal and Spatially Correlated Data* 289–298. New York: Springer (eds T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger).

Kammann, E. E. and Wand, M. P. (2003). Geoadditive Models. *Applied Statistics* **52** 1-18.

Katzfuss, M. and Cressie, N. (2012). Bayesian Hierarchical Spatio-temporal Smoothing for Very Large Datasets. *Environmetrics* **23** 94–107.

Kaufman, C. G., Scheverish, M. J. and Nychka, D. W. (2008). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. *Journal of the American Statistical Association* **103** 1545-1555.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical Space-time Models: A Review. *Mathematical Geology* **31** 651-684.

Lloyd, C. D. and Atkinson, P. M. (2004). Increased Accuracy of Geostatistical Prediction of Nitrogen Dioxide in the United Kingdom with Secondary Data. *International Journal of Applied Earth Observation and Geoinformation* **5** 293-305.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H. and Straif, S. (2013). The Carcinogenicity of Outdoor Air Pollution. *The Lancet Oncology* **14** 1262-1263.

Manders, A. M. M., Schaap, M. and Hoogerbrugge, R. (2009). Testing the Capability of the Chemistry Transport Model LOTOS-EUROS to Forecast PM10 Levels in the Netherlands. *Atmospheric Environment* **43** 4050-4059.

Mues, A., Kuenen, J., Hendriks, C., Manders, A., Segers, A., Scholz, Y., Hueglin, C., Builtjes, P. and Schaap, M. (2014). Sensitivity of air pollution simulations with LOTOS-EUROS to the temporal distribution of anthropogenic emissions. *Atmospheric Chemistry and Physics* **14** 939-955.

Omidi, M. and Mohammadzadeh, M. (2015). A New Method to Build Spatio-temporal Covariance Functions: Analysis of Ozone Data. *Statistical Papers* 1-15.

Pfeifer, P. E. and Deutsch, S. J. (1980a). Independence and Sphericity Tests for the Residuals of Spacetime ARMA Models. *Communications in Statistics - Simulation and Computation* **9** 533-549.

Pfeifer, P. E. and Deutsch, S. J. (1980b). Stationarity and Invertibility Regions for Low Order STARMA Models. *Communications in Statistics - Simulation and Computation* **9** 551-562.

Pouliot, G., Pierce, T., van der Gon, H. D., Schaap, M., Moran, M. and Nop-

MONGCOL, U. (2012). Comparing Emission Inventories and Model-ready Emission Datasets between Europe and North America for the AQMEII Project. *Atmospheric Environment* **53** 4-14.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2005). *Gaussian Processes for Machine Learning*, first ed. The MIT Press, Cambridge, MA.

R'HONI, Y., CLARISSE, L., CLERBAUX, C., HURTMANS, D., DUFLOT, V., TURQUETY, S., NGADI, Y. and COHEUR, P. F. (2013). Exceptional Emissions of NH3 and HCOOH in the 2010 Russian Wildfires. *Atmospheric Chemistry and Physics* **13** 4171-4181.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields : Theory and Applications. Monographs on statistics and applied probability.* Chapman & Hall/CRC, Boca Raton, FL.

SANG, H. and HUANG, J. Z. (2012). A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets. *Journal of the Royal Statistical society, Series B* **74** 111-132.

SCHAAP, M., TIMMERMANS, R. M. A., ROEMER, M., BOERSEN, G. A. C., BUILTJES, P., SAUTER, F., VELDERS, G. and BECK, J. (2008). The LOTOS-EUROS Model: Description, Validation and Latest Developments. *International Journal of Environment and Pollution* **32** 270-290.

SHABY, B. A. and RUPPERT, D. (2012). Tapered Covariance: Bayesian Estimation and Asymptotics. *Journal of Computational and Graphical Statistics* **21** 433–452.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society B* **64** 583–639.

STEIN, M. L. (2005). Spacetime Covariance functions. *Journal of the American Statistical Association* **100** 310-321.

STEIN, M. L. (2007). Spatial Variation of Total Column Ozone on a Global Scale. *Annals of Applied Statistics* **1** 191-210.

STEIN, M. L. (2008). A Modeling Approach for Large Spatial Datasets. *Journal of the Korean Statistical Society* **37** 3-10.

STEIN, M. L. (2013). On a Class of Spacetime Intrinsic Random Functions. *Bernoulli* **19** 387–408.

STEIN, M. L. (2014). Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data. *Spatial Statistics* **8** 1–19.

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating Likelihoods for Large Spatial Data Sets. *Journal of the Royal Statistical society, Series B* **66** 275-296.

STERN, R., BUILTJES, P., SCHAAP, M., TIMMERMANS, R., VAUTARD, R., HODZIC, A., MEMMESHEIMER, M., FELDMANN, H., RENNER, E., WOLKE, R. and KERSCHBAUMER, A. (2008). A Model Inter-comparison Study Focussing on Episodes with Elevated PM10 Concentrations. *Atmospheric Environment* **42** 4567-4588.

STOFFER, D. S. (1986). Estimation and Identification of Spacetime ARMAX Models in the Presence of Missing Data. *Journal of the American Statistical Association* **81** 762-772.

STROUD, J. R., MULLER, P. and SANSO, B. (2001). Dynamic Models for Spatiotemporal Data. *Journal of the Royal Statistical society, Series B* **63** 673-689.

VAN DE KASSTEELE, J. and STEIN, A. (2006). A Model for External Drift Kriging with Uncertain Covariates applied to Air Quality Measurements and Dispersion Model Output. *Environmetrics* **17** 309-322.

VECCHIA, A. V. (1988). Estimation and Model Identification for Continuous Spatial Processes. *Journal of the Royal Statistical society, Series B* **50** 297-312.

VECCHIA, A. V. (1992). A New Method of Prediction for Spatial Regression Models with Correlated Errors. *Journal of the Royal Statistical society, Series B* **54** 813-830.

Xu, G., Liang, F. and Genton, M. G. (2014). A Bayesian Spatio-Temporal Geostatistical Model with an Auxiliary Lattice for Large Datasets. *Statistica Sinica* In press.

Yeniay, O. and Goktas, A. (2002). A Comparison of Partial Least Squares Regression with Other Prediction Methods. *Hacettepe Journal of Mathematics and Statistics* **31** 99-111.

University of Minnesota
Division of Biostatistics
Minneapolis, MN 55455
USA
E-mail: datta013@umn.edu

University of California, Los Angeles
Department of Biostatistics
Los Angeles, CA 90095
USA
E-mail: sudipto@ucla.edu

Michigan State University
Departments of Forestry and Geography
East Lansing, MI 48824
USA
E-mail: finleya@msu.edu

University of Twente
Faculty of Geo-Information Science and Earth Observation (ITC)
Enschede, 7500 AE
Netherlands
E-mail: n.hamm@utwente.nl

TNO Built Environment and Geosciences
Utrecht, 3508 TA
Netherlands
E-mail: martijn.schaap@tno.nl