HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2018 April 13.

Published in final edited form as:

Ann Appl Stat. 2016 September ; 10(3): 1286–1316. doi:10.1214/16-AOAS931.

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

ABHIRUP DATTA[*], **SUDIPTO BANERJEE**[1,†], **ANDREW O. FINLEY**[‡,2], **NICHOLAS A. S. HAMM**[§], and **MARTIJN SCHAAP**[¶]

[*]Johns Hopkins University

[†]University of California, Los Angeles

[‡]Michigan State University

[§]University of Twente

[¶]TNO

Abstract

Particulate matter (PM) is a class of malicious environmental pollutants known to be detrimental to human health. Regulatory efforts aimed at curbing PM levels in different countries often require high resolution space–time maps that can identify red-flag regions exceeding statutory concentration limits. Continuous spatio-temporal Gaussian Process (GP) models can deliver maps depicting predicted PM levels and quantify predictive uncertainty. However, GP-based approaches are usually 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 spatio-temporal GP (e.g., with nonseparable covariance structures). The DNNGP we develop here 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

[1]Supported in part by NSF Grant DMS-15-13654 and by CDC/NIOSH R01OH010093.

[2]Supported in part by NSF Grants DMS-1513481, EF-1137309, EF-1241874, and EF-1253225, as well as NASA Carbon Monitoring System grants.

A. Datta, Department OF Biostatistics, Johns hopkins University, Baltimore, Maryland 55455, USA, abhidatta@jhu.edu

S. Banerjee, Department OF Biostatistics, University OF California, Los Angeles, Los Angeles, California 90095, USA, sudipto@ucla.edu

A. O. Finley, Departments of Forestry and Geography, Michigan State University, East Lansing, Michigan 48824, USA, finleya@msu.edu

N. A. S. Hamm, University OF Twente, Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, 7500 AE, The Netherlands, n.hamm@utwente.nl

M. Schaap, TNO, Department OF Climate, Air and Sustainability, Utrecht, 3508 TA, The Netherlands, martijn.schaap@tno.nl

SUPPLEMENTARY MATERIAL

Supplement to "Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis" (DOI: 10.1214/16-AOAS931SUPP;.pdf). File containing supplementary materials including a formal construction of eligible sets, additional simulation experiments and possible extension of DNNGP to model nonstationary covariances.

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).

keywords and phrases

Nonseparable spatio-temporal models; scalable Gaussian process; nearest neighbors; Bayesian inference; Markov chain Monte Carlo; environmental pollutants

1. Introduction

Recent years have witnessed considerable growth in statistical modeling of large spatio-temporal datasets; see, for example, the recent books by Cressie and Wikle (2011), Gelfand et al. (2010) and Banerjee, Carlin and Gelfand (2014) and the references therein for a variety of methods and applications. An especially important domain of application for such models is environmental public health, where analysts and researchers seek map projections for 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, one 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, 1980b), Stoffer (1986)] and dynamic models [Gelfand, Banerjee and Gamerman (2005), Stroud, Müller and Sansó (2001)], 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., Allcroft and Glasbey (2003), Cressie and Huang (1999), Gneiting (2002), Gneiting, Genton and Guttorp (2007), Kyriakidis and Journel (1999), Stein (2005)]. 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. When $\mathbf{C}(\boldsymbol{\theta})$ has no exploitable structure, matrix computations typically require $\sim n^3$ floating point operations (flops) and storage in the order of n^2 which becomes prohibitive if n is large. Approaches for modeling large covariance matrices in purely spatial settings include low-rank models [see, e.g., Banerjee et al. (2008), Crainiceanu, Diggle and Rowlingson (2008), Cressie and Johannesson (2008), Finley,

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

Banerjee and McRoberts (2009), Higdon (2001), Kammann and Wand (2003), Katzfuss (2016), Rasmussen and Williams (2005), Stein (2007, 2008)], covariance tapering [see, e.g., Bevilacqua et al. (2015), Du, Zhang and Mandrekar (2009), Furrer, Genton and Nychka (2006), Kaufman, Schervish and Nychka (2008), 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. (2016a), Stein, Chi and Welty (2004), Vecchia (1988, 1992)] 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) and Katzfuss and Cressie (2012) who extend low-rank spatial processes to dynamic spatio-temporal settings, while Xu, Liang and Genton (2014) opt for a GMRF approach. All these methods use dynamic models defined on fixed temporal lags and do not lend themselves easily to continuous spatio-temporal domains.

Spatio-temporal process models for continuous space–time modeling of large datasets have received relatively scant attention. Bai, Song and Raghunathan (2012) and Bevilacqua et al. (2012) used composite likelihoods for parameter estimation in a continuous space–time setup. Both these approaches, like their spatial analogues, have focused upon constructing computationally attractive likelihood approximations and have restricted inference only to parameter estimation. Uncertainty estimates are usually based on asymptotic results which are usually inappropriate 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 remains expensive for large n and may not reflect predictive uncertainty accurately.

Our current work offers a highly scalable spatio-temporal process for continuous space–time modeling. We expand upon the neighbor-based conditioning set approaches outlined in purely spatial contexts by Stein, Chi and Welty (2004), Vecchia (1988) and Datta et al. (2016a). 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, considerably enhancing its applicability. Moreover, it can be used with any spatio-temporal covariance function, thereby accommodating nonseparability. 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 a substantially superior approximation to the underlying process than, for example, 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. 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), Hoek et al. (2013), Loomis et al. (2013)]. In response to these and other health impact studies, regulatory agencies have introduced 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 an average of 50 μg m^{-3} over 24 hours and of 40 μg m^{-3} over a year [European Commission (2015)]. Measurements made with standard instruments are considered authoritative, but these observations are sparse and maps at finer scales are needed for monitoring progress with mitigation strategies and for monitoring compliance. Hence, accurately quantifying uncertainty in predicted PM concentrations is 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 source alone [Candiani et al. (2013), Denby et al. (2008), Hamm et al. (2015), van de Kassteele and Stein (2006)]. In such settings, 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.

We focus on the development and illustration of continuous space–time process models capable of delivering predictive maps 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 on space and time residual structure, that is, space and time lags, that can help identify lurking processes 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 that 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 domain comprises 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*).[3] 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's geographic projection is longitude–latitude at a resolution of 0.50° longitude × 0.25° latitude (approximately 25 km × 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 nonspecified 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 (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 indicating that PM_{10} concentrations are dominated by rural background values even in urban areas [Eeftens et al. (2012)].

3. Scalable dynamic nearest neighbor Gaussian processes

Let $\{w(\ell) : \ell \in \mathcal{L}\}$ be a zero-centered 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) is the spatial region, $\mathcal{T} \subset [0,\infty)$ is the time domain and $\ell = (\mathbf{s}, t)$ is a space–time coordinate with spatial location $\mathbf{s} \in \mathcal{S}$ and time point $t \in \mathcal{T}$. 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 $C(\cdot, \cdot|\boldsymbol{\theta})$. When $\mathcal{U}$ or $\mathcal{V}$ contains a single point, $\mathbf{C}_{\mathcal{U},\mathcal{V}}$ is a row or column vector. 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 $C(\ell_i, \ell_j|\boldsymbol{\theta})$.

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 covariance function. We outline a general framework that first uses a set of points in $\mathcal{L}$ to construct a computationally efficient approximation for the random field and extends the finite-dimensional distribution over this set to a process. To ease the notation, we will suppress the explicit dependence of matrices and vectors on $\boldsymbol{\theta}$ whenever the context is clear.

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

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

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 (3.1) is a well-defined spatio-temporal Gaussian process on $\mathcal{L}$ for *any* choice of $a_i(\ell)$'s, as long as $\mathbf{K}(\boldsymbol{\theta})$ is positive definite. For example, $w(\ell)$ is a Gaussian process with covariance function $C(\cdot, \cdot|\boldsymbol{\theta})$ if we set $\mathbf{K}(\boldsymbol{\theta}) = \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})$, $\mathbf{a}(\ell) = \mathbf{C}^{-1}_{\mathcal{R},\mathcal{R}}\mathbf{C}_{\mathcal{R},\ell}$, where $\mathbf{a}(\ell)$ is $r \times 1$ with elements $a_i(\ell)$, and $\eta(\ell) \overset{\text{ind}}{\sim} N(0, C(\ell, \ell \mid \boldsymbol{\theta}) - \mathbf{C}_{\ell,\mathcal{R}}\mathbf{C}^{-1}_{\mathcal{R},\mathcal{R}}\mathbf{C}_{\mathcal{R},\ell})$. Now (3.1) 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 $\mathbf{a}(\ell)$. 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 nondegenerate (or bias-adjusted) low-rank Gaussian Process [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 at scales similar to ours [see, e.g., Datta et al. (2016a), Stein (2014)], we use the framework in (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 time t_k with spatial locations in $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{j-1}\}$. Thus, $H(\ell_i^*) = \{(\mathbf{s}_p, t_q) \mid p = 1, 2, \ldots, N, q = 1, 2, \ldots, (k-1)\} \cup \{(\mathbf{s}_p, t_k) \mid p = 1, 2, \ldots, (j-1)\}$. 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 $\mathbf{a}(\ell)$ in (3.1). Let $w(\ell) \sim \text{GP}(0, C(\cdot, \cdot|\boldsymbol{\theta}))$. We choose $\mathbf{K}(\boldsymbol{\theta})$ to effectuate a sparse approximation for the joint density of the realizations of $w(\ell)$ over $\mathcal{R}$, that is, $N(\mathbf{w}_{\mathcal{R}}|\mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta}))$. Adapting the ideas underlying likelihood approximations in Vecchia (1988) and Datta et al. (2016a), we specify $\mathbf{K}(\boldsymbol{\theta})$ to be the $r \times r$ matrix such that

$$N\left(\mathbf{w}_{\mathcal{R}} \mid \mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})\right) = \prod_{i=1}^{r} p\left(w(\ell_i^*) \mid \mathbf{w}_{H(\ell_i^*)}\right) \quad (3.2)$$
$$\approx \prod_{i=1}^{r} p\left(w(\ell_i^*) \mid \mathbf{w}_{N(\ell_i^*)}\right) = N\left(\mathbf{w}_{\mathcal{R}} \mid \mathbf{0}, \mathbf{K}(\boldsymbol{\theta})\right).$$

Here, $H(\ell_1^*)$ is the empty set [hence, so is $N(\ell_1^*)$] and $p(w(\ell_1^*) \mid \mathbf{w}_{H(\ell_1^*)}) = p(w(\ell_1^*) \mid \mathbf{w}_{N(\ell_1^*)}) = p(w(\ell_1^*))$. 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 $\mathbf{K}^{-1}$. This implies

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

where $\mathbf{a}_{N(\ell_i^*)} = \mathbf{C}^{-1}_{N(\ell_i^*), N(\ell_i^*)} \mathbf{C}_{N(\ell_i^*), \ell_i^*}$. Also, $\mathbf{K}$ is determined by $\mathbf{C}_{\mathcal{R},\mathcal{R}}$ because $\mathbf{K}^{-1} = \mathbf{V}'\mathbf{F}^{-1}\mathbf{V}$, where $\mathbf{F}$ is a diagonal matrix with diagonal entries $f_{\ell_i^*} = \mathrm{Var}(w(\ell_i^*) \mid \mathbf{w}_{N(\ell_i^*)}) = C(\ell_i^*, \ell_i^* \mid \boldsymbol{\theta}) - \mathbf{C}_{\ell_i^*, N(\ell_i^*)} \mathbf{C}^{-1}_{N(\ell_i^*), N(\ell_i^*)} \mathbf{C}_{N(\ell_i^*), \ell_i^*}$ and $\mathbf{V}$ is the $r \times r$ matrix with entries $v_{i,j}$ such that $v_{i,i} = 1$ and $v_{i,j} = 0$ whenever $\ell_i^* \notin N(\ell_j^*)$. The remaining entries in column j of $\mathbf{V}$ are specified by setting the subvector $\mathbf{V}_{c(\ell_j^*), j} = -\mathbf{a}_{N(\ell_j^*)}$, where $c(\ell_j^*) = \{i \mid \ell_i^* \in N(\ell_j^*)\}$. If $m (\ll r)$ denotes the limiting size of the neighbor sets $N(\ell)$, then the columns of $\mathbf{V}$ are sparse with at most $m+1$ nonzero elements. Consequently, $\mathbf{K}^{-1}$ has at most $O(rm^2)$ nonzero elements [this is the spatial-temporal analogue of the result in Datta et al. (2016a)]. 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$(0, C(\cdot, \cdot \mid \boldsymbol{\theta}))$.

Turning to the vector of coefficients $\mathbf{a}(\ell)$ in (3.1), we extend the idea in (3.3) to any point $\ell \notin \mathcal{R}$ by requiring that $\mathrm{E}[w(\ell) \mid \mathbf{w}_{\mathcal{R}}] = \mathrm{E}[w(\ell) \mid \mathbf{w}_{N(\ell)}]$. This is achieved by setting $a_i(\ell) = 0$ in (3.1) whenever $\ell_i^* \notin N(\ell)$ for any point $\ell \notin \mathcal{R}$. Hence, if $N(\ell)$ contains m points, then at most m of the elements in the $r \times 1$ vector $\mathbf{a}(\ell)$ can be nonzero. These nonzero entries are determined from the above conditional expectation given $N(\ell)$. To be precise, if $\mathbf{a}_{N(\ell)}$ is the $m \times 1$ vector of these m entries, then we solve $\mathbf{C}_{N(\ell),N(\ell)}\mathbf{a}_{N(\ell)} = \mathbf{C}_{N(\ell),\ell}$ for $\mathbf{a}_{N(\ell)}$. Also note that $\mathbf{a}'(\ell)\mathbf{w}_{\mathcal{R}} = \mathbf{a}'_{N(\ell)}\mathbf{w}_{N(\ell)}$. Finally, to complete the process specifications in (3.1), we specify $\eta(\ell) \overset{\text{ind}}{\sim} N(0, f_\ell)$, where $f_\ell = \mathrm{Var}(w(\ell) \mid \mathbf{w}_{N(\ell)}) = C(\ell, \ell \mid \boldsymbol{\theta}) - \mathbf{C}_{\ell, N(\ell)}\mathbf{C}^{-1}_{N(\ell), N(\ell)}\mathbf{C}_{N(\ell), \ell}$. The covariance function $\tilde{C}(\cdot, \cdot \mid \boldsymbol{\theta})$ of the resulting Gaussian Process is given by

$$\tilde{C}(\ell_i, \ell_j \mid \boldsymbol{\theta}) = \begin{cases} K_{p,q}, & \text{if } \ell_i = \ell_p^* \text{ and } \ell_j = \ell_q^* \text{ are both in } \mathcal{R}, \\ \mathbf{a}'(\ell_i)\mathbf{K}_{*q}, & \text{if } \ell_i \notin \mathcal{R} \text{ and } \ell_j = \ell_q^* \in \mathcal{R}, \\ \mathbf{a}'(\ell_i)\mathbf{K}\mathbf{a}(\ell_j) + I(\ell_i = \ell_j) f_{\ell_i}, & \text{if } \ell_i \notin \mathcal{R} \text{ and } \ell_j \notin \mathcal{R}, \end{cases} \quad (3.4)$$

where $K_{p,q}$ is element (p, q) and $\mathbf{K}_{*q}$ is column q in $\mathbf{K}$.

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

4. Constructing neighbor sets

4.1. Simple neighbor selection

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 [Datta et al. (2016a), Vecchia (1988)]. On the other hand, spatio-temporal covariances between two points typically depend on the spatial as well as the temporal lag between the points. To be specific, nonseparable 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 $d: (\mathscr{S} \times \mathscr{T})^2 \to \mathbb{R}^+$ 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 light of the above discussion, we define "nearest neighbors" in a spatio-temporal domain using the spatio-temporal covariance function itself as a proxy for distance. To elucidate, for any three points $(\mathbf{s}_1, t_1)$, $(\mathbf{s}_2, t_2)$ and $(\mathbf{s}_3, t_3)$, we say that $(\mathbf{s}_1, t_1)$ is nearer to $(\mathbf{s}_2, t_2)$ than to $(\mathbf{s}_3, t_3)$ if $C((\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2)|\boldsymbol{\theta}) > C((\mathbf{s}_1, t_1), (\mathbf{s}_3, t_3)|\boldsymbol{\theta})$. Subsequently, this definition of "distance" is used to find m nearest neighbors for any location.

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 the same as the covariance function of the parent GP. However, for every location $(\mathbf{s}_i, t_j)$, its neighbor set, denoted by $N_\theta(\mathbf{s}_i, t_j)$, still depends on $\boldsymbol{\theta}$. This is illustrated in Figures 2(a) and 2(b) which show how neighbor sets can differ drastically based on the choice of $\boldsymbol{\theta}$.

In most applications, $\boldsymbol{\theta}$ is unknown, precluding the use of these newly defined neighbor sets $N_\theta(\mathbf{s}_i, t_j)$ to construct the DNNGP. We propose a simple intuitive method to construct neighbor sets. We choose m to be a perfect square and construct a simple neighbor set of size m using $\sqrt{m}$ spatial nearest neighbors and $\sqrt{m}$ temporal nearest neighbors. Figure 2(c) illustrates the simple neighbor set of size $m = 9$ for the red point. In order to formally define the simple neighbor sets, we denote $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$, $S_i = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ and $T = \{t_1, t_2, \ldots, t_M\}$. Furthermore, for any finite set of spatial locations $V \subseteq S$, let $A(\mathbf{s}, V, m)$ denote the

set of m nearest spatial neighbors in V for the location $\mathbf{s}$. For any point $(\mathbf{s}_i, t_j) \in \mathcal{R}$, we define the simple neighbor sets

$$N(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{\sqrt{m}-1} \left\{(\mathbf{s}, t_{j-k}) \mid \mathbf{s} \in A(\mathbf{s}_i, S, \sqrt{m})\right\} \cup \left\{(\mathbf{s}, t_j) \mid \mathbf{s} \in A(\mathbf{s}_i, S_i, \sqrt{m})\right\}. \quad (4.1)$$

The above construction implies that the neighbor set for any point in $\mathcal{R}$ consists of $\sqrt{m}$ spatial nearest neighbors of the preceding $\sqrt{m}$ time points. For arbitrary $(\mathbf{s}, t) \notin \mathcal{R}$, $N(\mathbf{s}, t)$ is simply defined as the Cartesian product of $\sqrt{m}$ nearest neighbors for $\mathbf{s}$ in $\mathcal{S}$ with $\sqrt{m}$ nearest neighbors of t in $\mathcal{T}$.

In many applications, one desirable property of the spatio-temporal covariance functions is *natural monotonicity*, that is, $C(h, u)$ is decreasing in h for fixed u and decreasing in u for fixed h. All Matèrn-based space–time separable covariances and many nonseparable classes of covariance functions possess this property [Omidi and Mohammadzadeh (2015), Stein (2013)]. 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 simple neighbor selection scheme described in Section 4.1 does not depend on $\boldsymbol{\theta}$ and is undoubtedly useful for fast implementation of the DNNGP. However, for some values of $\boldsymbol{\theta}$, the neighbor sets may often consist of very few nearest neighbors. This issue is illustrated in Figure 2 where the simple neighbor set (blue points) contained 7 out of 9 true nearest neighbors (green points) for $\theta = 1$, but only 3 out of 9 true nearest neighbors for $\theta = 2$. We see that for different choices of the covariance parameters the simple 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 evaluated the pairwise correlations between any point $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$ and all points in its history set $H(\mathbf{s}_i, t_j)$ to obtain $N_\theta(\mathbf{s}_i, t_j)$—the set of m true nearest neighbors. In practice, however, we encounter a computational roadblock because $\boldsymbol{\theta}$ is unknown and for every new value of $\boldsymbol{\theta}$ in an iterative optimizer or Markov chain Monte Carlo sampler, we need to redo the search for the neighbor sets within the history sets. As the history sets are typically large, this is computationally challenging. For example, in Figure 2, the history set for the red point is composed of all points below the red horizontal line, so evaluating the pairwise correlations required for updating neighbor sets of all points in $\mathcal{R}$ and n datapoints outside $\mathcal{R}$ will use $O(r^2 + nr)$ flops at each iteration. The reference set $\mathcal{R}$ is typically chosen to match the scale of the observed dataset to achieve a reasonable approximation of the parent GP by DNNGP. Hence, for large datasets this updating becomes a deterrent. In fact, Vecchia (1988) and Stein, Chi and Welty (2004) admit

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 underlying idea is to restrict the search for the neighbor sets to carefully constructed small subsets of the history sets. These small *eligible sets* $E(\mathbf{s}_i, t_j)$ are constructed in such a manner that, despite being much smaller than the history sets, they are guaranteed to contain the true nearest neighbor sets $N_\theta(\mathbf{s}_i, t_j)$ for all choices of the parameter $\boldsymbol{\theta}$. So, for each $\boldsymbol{\theta}$ we can evaluate the pairwise correlations between $(\mathbf{s}_i, t_j)$ and only the points in $E(\mathbf{s}_i, t_j)$ and still find the true set of m-nearest neighbors.

Figure 3(a) and 3(b) illustrates how to determine which points belong to $E(\mathbf{s}_i, t_j)$. Let h and u denote the spatial and temporal lags with the black point and the red point in Figure 3(a). All other points in the black rectangle have spatial lag $\leq h$ and temporal lag $\leq u$ with the red point. So if the covariance function $C(h, u|\boldsymbol{\theta})$ possess natural monotonicity, the black point has the lowest correlation with the red point among all the points in the black rectangle. For the black point to be in the set of m nearest neighbors $N_\theta(\mathbf{s}_i, t_j)$ for any $\boldsymbol{\theta}$, all other points in the black rectangle should also be included. Since this is not possible as the black rectangle contains 12 points and $m = 9$, the black point becomes ineligible. By a similar logic, the blue rectangle in Figure 3(b) contains $8(< m)$ points and is included in $E(\mathbf{s}_i, t_j)$. Proceeding like this, we can easily determine the entire eligible set [Figure 3(c)] without any knowledge of the parameter $\boldsymbol{\theta}$.

A formal construction of eligible sets is provided in Section S1 of the supplemental article Datta et al. (2016b). Proposition S1 proves that eligible sets are guaranteed to contain the neighbor sets for all choices of $\boldsymbol{\theta}$. This result has substantial consequences because the size of the eligible sets is approximately equal to $4m$. The eligible sets need to be calculated only once before the MCMC 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 can search for a new set of m-nearest neighbors $N_\theta(\mathbf{s}_i, t_j)$ only within the eligible sets and use $N_\theta(\mathbf{s}_i, t_j)$ as the conditioning sets to construct the DNNGP. We summarize the MCMC steps of the DNNGP with adaptive neighbor selection in Algorithm 1.

As the size of the sets are approximately $4m$, for every $(\mathbf{s}_i, t_j)$ we need to evaluate only $4m$ pairwise correlations. 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 [Algorithm 1: Step (c)] can proceed independently of one another.

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 equation (S1).

2: At the l^{th} iteration of the MCMC:

- **1** Calculate $C((\mathbf{s}, t), (\mathbf{s}_i, t_j)|\boldsymbol{\theta}^{(l)})$ for all $(\mathbf{s}, t)$ in $E(\mathbf{s}_i, t_j)$.
- **2** Define $N_\theta(\mathbf{s}_i, t_j)^{(l)}$ as the set of m locations in $E(\mathbf{s}_i, t_j)$ which maximizes $C((\mathbf{s}, t), (\mathbf{s}_i, t_j)|\boldsymbol{\theta}^{(l)})$.
- **3** Repeat steps (a) and (b) for all $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$.
- **4** Update $\boldsymbol{\theta}^{(l+1)}$ based on the new set of neighbor sets computed in step (c) using the Metropolis step specified in (5.5).

3: Repeat Step 2 for N MCMC iterations.

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, that is, 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

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

where $\boldsymbol{\beta}$ denotes the coefficient vector for the predictors, $w(\ell_i)$ is the spatio-temporally varying intercept and $\varepsilon(\ell_i)$ 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 nonseparable spatio-temporal isotropic covariance function $C((\mathbf{s}+\mathbf{h}, t+u), (\mathbf{s}, t)|\boldsymbol{\theta})$, introduced by Gneiting (2002),

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

where $\mathbf{h}$ and u refer 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 nonseparability 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 [Stein, Chi and Welty (2004), Vecchia (1988)]. 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. (2016a). Hence, we choose $\mathcal{R} = \{\ell_1, \ell_2, \ldots, \ell_n\}$, that is, $\ell_i^* = \ell_i$ for $i = 1, 2, \ldots, n$.

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

$$p(\boldsymbol{\theta}) \times \text{IG}(\tau^2 \mid a_\tau, b_\tau) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \times N(\mathbf{w}_{\mathcal{R}} \mid \mathbf{0}, \widetilde{\mathbf{C}}_{\mathcal{R},\mathcal{R}}) \quad (5.3)$$
$$\times \prod_{i=1}^{n} N(\mathbf{y}(\ell_i) \mid \mathbf{x}(\ell_i)'\boldsymbol{\beta} + \mathbf{w}(\ell_i), \tau^2),$$

where $p(\boldsymbol{\theta})$ is the prior on $\boldsymbol{\theta}$, and $\text{IG}(\tau^2 | a_\tau, b_\tau)$ 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 are observed and let $I(\ell_i)$ 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 $\mathbf{X}$ be the corresponding $n_o \times p$ design matrix. The full conditional distribution of $\boldsymbol{\beta}$ is $N(\mathbf{V}_\beta^* \boldsymbol{\mu}_\beta^*, \mathbf{V}_\beta^*)$, 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 $\text{IG}(a_\tau + \frac{n_o}{2}, b_\tau + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o}))$.

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 jth member of $N(\ell_2)$, then we define b_{ℓ_2,ℓ_1} as the jth 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 $a_{\ell_2,\ell_1} = w(\ell_2) - \Sigma_{\ell \in N(\ell_2), \ell \neq \ell_1} w(\ell) b_{\ell_2,\ell}$ Then, for $i = 1, 2, \ldots, n$, the full conditional distribution for $\mathbf{w}(\ell_i)$ is $N(v(\ell_i)\mu(\ell_i), v(\ell_i))$, where

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

If $U(\ell_i)$ is empty for some ℓ_i, then all instances of $\Sigma_{\ell \in U(\ell_i)}$ in (5.4) disappear 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

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

Since none of the above updates involve expensive matrix decompositions, the likelihood can be evaluated very efficiently. The algorithm for updating the parameters of a hierarchical DNNGP model is analogous to the corresponding updates for a purely spatial NNGP model [see Datta et al. (2016a)]. 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 (Algorithm 1). 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, that is, $I(\ell_i^*) = 0$. 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 (S2) 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

In this section we compare the DNNGP, the full-rank GP and the low-rank Gaussian Predictive Process [Banerjee et al. (2008)]. Additional simulation experiments comparing the predictive performance of DNNGP with Local Approximation GP [Gramacy and Apley (2015)] are provided in Section S2 of the supplemental article Datta et al. (2016b). 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 nonseparable spatio-temporal covariance function (5.2), viz,

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

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 $\boldsymbol{\theta} = (\sigma^2, \kappa, a, c)$, we 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 the following: (1) short spatial range and long temporal 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 following: (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 8×8×8 = 512 knots within the domain.

For all models, the intercept β_0 and slope regression parameters, β_1, were assigned *flat* prior distributions. The variance parameters were assumed to follow inverse-Gamma prior distributions with $\sigma^2 \sim \text{IG}(2, 1)$ and $\tau^2 \sim \text{IG}(2, 0.1)$. 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 posterior predictive loss [Gelfand and Ghosh (1998)]. The DIC is reported along with an estimate of model complexity, p_D, while the posterior predictive loss is computed as D = G + P, where G is a goodness-of-fit measure and P measures the number of model parameters. Predictive accuracy for the 500 holdout locations was measured using root mean squared prediction error [Yeniay and Göktaş (2002)]. The percent of holdout locations that fell within the candidate models' posterior predictive distribution's 95% credible interval (CI) was also computed. Inference was based on 15,000 MCMC samples comprising post burn-in samples from three chains of 25,000 iterations (i.e., 5000 samples from each chain).

Table 1 presents parameter estimation and model assessment metrics. With the exception of τ^2 for Dataset 1, the full GP model recovered the parameter values used to generate the datasets, that is, 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. Hence, 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. 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, computing 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 advantages 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 and $x(\ell_i)$ 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 nonseparable spatio-temporal covariance function (6.1). Exploratory analysis—consisting of semivariogram plots and autocorrelation function plots for simple ordinary least square model residuals—helped guide choice of prior and hyper-parameters for the variance and decay parameters. Specifically, $\sigma^2 \sim \text{IG}(2, 1)$, $\tau^2 \sim \text{IG}(2, 0.1)$, $a \sim U(0.1, 5)$ and $c \sim U(0.01, 0.5)$, with κ fixed at 0.5.

Candidate models included the following: (i) LOTOS-EUROS CTM, (ii) simple linear regression model with no spatio-temporal effects, that is, $w(\ell) = 0$, 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 underestimates PM_{10} [Hamm et al. (2015), Stern et al. (2008)]. 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 suggests the Simple model estimates a moderately longer spatial and temporal range, that is, ~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, that is, 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 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.

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, but rather RMSPE consistently decreased with the 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 stations. 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 nonspatial regression, and (iii) appropriate widening of CIs for missing station observations.

Table 3 provides out-of-sample prediction validation metrics for the nonspace–time and 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. We also added results for the simple linear regression (SLR) model in the first column of Table 3. The simple linear regression model does not consider spatio-temporal effects, nor does it consider a time-varying intercept [unlike the day-specific results presented in Hamm et al. (2015)], which may explain the poor predictive performance—it is more meaningful to compare the DNNGP model prediction metrics to the day-specific metrics presented in Hamm et al. (2015).

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 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 the 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 were implemented in C++ and leveraged OpenMP [Dagum and Menon (1998)] and Intel Math Kernel Library's (MKL) threaded BLAS and LAPACK routines for the 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 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, for example, 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. While 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 lifetime 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 are associated with particular (stagnant) meteorological conditions or occasional large emissions from, for example, 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. We have developed DNNGP assuming an isotropic nonstationary spatio-temporal covariance structure. However, it can also be potentially extended to certain classes of nonstationary space–time covariances [see Section S3 of the supplemental article Datta et al. (2016b)]. 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. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Associate Editor and anonymous reviewers for their suggestions which considerably improved the manuscript. The authors also acknowledge Arjo Segers (TNO) for his support with the LOTOS-EUROS CTM.

References

Allcroft DJ, Glasbey CA. A latent Gaussian Markov random-field model for spatio-temporal rainfall disaggregation. J Roy Statist Soc Ser C. 2003; 52:487–498.

Bai Y, Song PXK, Raghunathan TE. Bayesian dynamic modeling for large space–time datasets using Gaussian predictive processes. J Roy Statist Soc Ser B. 2012; 74:799–824.

Banerjee, S., Carlin, BP., Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. 2. Chapman & Hall; Boca Raton, FL: 2014.

Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B Stat Methodol. 2008; 70:825–848.

Bevilacqua M, Gaetan C, Mateu J, Porcu E. Estimating space and space–time covariance functions for large data sets: A weighted composite likelihood approach. J Amer Statist Assoc. 2012; 107:268–280.

Bevilacqua M, Fass ÒA, Gaetan C, Porcu E, Velandia D. Covariance tapering for multivariate Gaussian random fields estimation. Stat Methods Appl. 2015; 25:21–37.

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, Loschau G. A case of extreme particulate matter concentrations over central Europe caused by dust emitted over the southern Ukraine. Atmos Chem Phys. 2008; 8:997–1016.

Brauer M, Amann M, Burnett RT, Cohen A, Dentener F, Ezzati M, Henderson SB, Krzyzanowski M, Martin RV, Van Dingenen R, Van Donkelaar A, Thurston GD. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. Environ Sci Technol. 2011; 46:652–660.

Brunekreef B, Holgate ST. Air pollution and health. Lancet. 2002; 360:1233–1242. [PubMed: 12401268]

Candiani G, Carnevale C, Finzi G, Pisoni E, Volta M. A comparison of reanalysis techniques: Applying optimal interpolation and ensemble Kalman filtering to improve air quality monitoring at mesoscale. Sci Total Environ. 2013; 458–460:7–14.

Crainiceanu CM, Diggle PJ, Rowlingson B. Bivariate binomial spatial modeling of *Loa loa* prevalence in tropical Africa. J Amer Statist Assoc. 2008; 103:21–37.

Cressie N, Huang HC. Classes of nonseparable, spatio-temporal stationary covariance functions. J Amer Statist Assoc. 1999; 94:1330–1340.

Cressie N, Johannesson G. Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B Stat Methodol. 2008; 70:209–226.

Cressie N, Shi T, Kang EL. Fixed rank filtering for spatio-temporal data. J Comput Graph Statist. 2010; 19:724–745.

Cressie, N., Wikle, CK. Statistics for Spatio-Temporal Data. Wiley; Hoboken, NJ: 2011.

Dagum L, Menon R. OpenMP: An industry standard API for shared-memory programming. IEEE Comput Sci Eng. 1998; 5:46–55.

Datta A, Banerjee S, Finley AO, Gelfand AE. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J Amer Statist Assoc. 2016a; 111:800–812.

Datta A, Banerjee S, Finley AO, Hamm NS, Schaap M. Supplement to “Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. 2016b; doi: 10.1214/16-AOAS931SUPP

Denby B, Schaap M, Segers A, Builtjes P, Horalek J. Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. Atmos Environ. 2008; 42:7122–7134.

Denby B, Sundvor I, Cassiani M, de Smet P, de Leeuw F, Horalek J. Spatial mapping of ozone and SO2 trends in Europe. Sci Total Environ. 2010; 408:4795–4806. [PubMed: 20619880]

Du J, Zhang H, Mandrekar VS. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. Ann Statist. 2009; 37:3330–3361.

Eeftens M, Tsai MY, 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, Gri-vas 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, Hoek G. 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. Atmos Environ. 2012; 62:303–317.

Eidsvik J, Shaby BA, Reich BJ, Wheeler M, Niemi J. Estimation and prediction in spatial models with block composite likelihoods. J Comput Graph Statist. 2014; 23:295–315.

European Commission. European Union Air Quality Standards. 2015. Available at http://ec.europa.eu/environment/air/quality/standards.htm

Finley AO, Banerjee S, Gelfand AE. Bayesian dynamic modeling for large space–time datasets using Gaussian predictive processes. J Geogr Syst. 2012; 14:29–47.

Finley AO, Banerjee S, McRoberts RE. Hierarchical spatial models for predicting tree species assemblages across large domains. Ann Appl Stat. 2009; 3:1052–1079. [PubMed: 20352037]

Flemming J, Inness A, Flentje H, Huijnen V, Moinat P, Schultz MG, Stein O. Coupling global chemistry transport models to ECMWF's integrated forecast system. Geosci Model Dev. 2009; 2:253–265.

Furrer R, Genton MG, Nychka D. Covariance tapering for interpolation of large spatial datasets. J Comput Graph Statist. 2006; 15:502–523.

Gelfand AE, Banerjee S, Gamerman D. Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics. 2005; 16:465–479.

Gelfand AE, Ghosh SK. Model choice: A minimum posterior predictive loss approach. Biometrika. 1998; 85:1–11.

Gelfand, AE.Diggle, PJ.Fuentes, M., Guttorp, P., editors. Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press; Boca Raton, FL: 2010.

Gneiting T. Nonseparable, stationary covariance functions for space–time data. J Amer Statist Assoc. 2002; 97:590–600.

Gneiting, T., Genton, MG., Guttorp, P. Geostatistical space–time models, stationarity, separability and full symmetry. In: Finken-staedt, B.Held, L., Isham, V., editors. Statistics of SpatioTemporal Systems. Chapman & Hall; London: 2007. p. 151-175.

Gneiting, T., Guttorp, P. Continuous parameter spatio-temporal processes. In: Gelfand, AE.Diggle, P.Fuentes, M., Guttorp, P., editors. Handbook of Spatial Statistics. Handb Mod Stat Methods. CRC Press; Boca Raton, FL: 2010. p. 427-436.

Gräler, B., Gerharz, L., Pebesma, E. Spatio-temporal analysis and interpolation of PM10 measurements in Europe. European Topic Centre on Air Pollution and Climate Change Mitigation; Bilthoven, The Netherlands: 2011. ETC/ACM Technical Paper 2011/10

Gramacy RB, Apley DW. Local Gaussian process approximation for large computer experiments. J Comput Graph Statist. 2015; 24:561–578.

Hamm NAS, Finley AO, Schaap M, Stein A. A spatially varying coefficient model for mapping PM10 air quality at the European scale. Atmos Environ. 2015; 102:393–405.

Hendriks C, Kranenburg R, Kuenen J, Van Gijlswijk R, Kruit RW, Segers A, van der Gon HD, Schaap M. The origin of ambient particulate matter concentrations in the Netherlands. Atmospheric Environment. 2013; 69:289–303.

Higdon, D. Technical report. Institute of Statistics and Decision Sciences, Duke Univ; Durham, NC: 2001. Space and space time modeling using process convolutions.

Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD. Long-term air pollution exposure and cardio-respiratory mortality: A review. Environ Health. 2013; 12:43. [PubMed: 23714370]

Intel. Math Kernel Library. 2015. Available at http://developer.intel.com/software/products/mkl/

Jones, RH., Zhang, Y. Models for continuous stationary space–time processes. In: Gregoire, TG.Brillinger, DR.Diggle, PJ.Russek-Cohen, E.Warren, WG., Wolfinger, RD., editors. Modelling Longitudinal and Spatially Correlated Data. Springer; New York: 1997. p. 289-298.

Kammann EE, WMP. Geoadditive models. J Roy Statist Soc Ser C. 2003; 52:1–18.

Katzfuss M. A multi-resolution approximation for massive spatial datasets. J Amer Statist Assoc. 2016 Available at arXiv:1507.04789.

Katzfuss M, Cressie N. Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics. 2012; 23:94–107.

Kaufman CG, Schervish MJ, Nychka DW. Covariance tapering for likelihood-based estimation in large spatial data sets. J Amer Statist Assoc. 2008; 103:1545–1555.

Kyriakidis PC, Journel AG. Geostatistical space–time models: A review. Math Geol. 1999; 31:651–684.

Lloyd CD, Atkinson PM. Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data. International Journal of Applied Earth Observation and Geoinformation. 2004; 5:293–305.

Loomis D, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif S. The carcinogenicity of outdoor air pollution. Lancet Oncol. 2013; 14:1262–1263. [PubMed: 25035875]

Manders AMM, Schaap M, Hoogerbrugge R. Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands. Atmos Environ. 2009; 43:4050–4059.

Mues A, Kuenen J, Hendriks C, Manders A, Segers A, Scholz Y, Hueglin C, Builtjes P, Schaap M. Sensitivity of air pollution simulations with LOTOS-EUROS to the temporal distribution of anthropogenic emissions. Atmos Chem Phys. 2014; 14:939–955.

Omidi M, Mohammadzadeh M. A new method to build spatio-temporal covariance functions: Analysis of ozone data. Statist Papers. 2015:1–15.

Pfeifer PE, Deutsch SJ. Independence and sphericity tests for the residuals of space–time ARMA models. Comm Statist Simulation Comput. 1980a; 9:533–549.

Pfeifer PE, Deutsch SJ. Stationarity and invertibility regions for low order STARMA models. Comm Statist Simulation Comput. 1980b; 9:551–562.

Pouliot G, Pierce T, van der Gon HD, Schaap M, Moran M, Nopmongcol U. Comparing emission inventories and model-ready emission datasets between Europe and North America for the AQMEII project. Atmos Environ. 2012; 53:4–14.

R'Honi Y, Clarisse L, Clerbaux C, Hurtmans D, Duflot V, Turquety S, Ngadi Y, Coheur PF. Exceptional emissions of NH3 and HCOOH in the 2010 Russian wildfires. Atmos Chem Phys. 2013; 13:4171–4181.

Rasmussen, CE., Williams, CKI. Gaussian Processes for Machine Learning. 1. MIT Press; Cambridge, MA: 2005.

Rue, H., Held, L. Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability. Chapman & Hall; Boca Raton, FL: 2005. p. 104

Sang H, Huang JZ. A full scale approximation of covariance functions for large spatial data sets. J R Stat Soc Ser B Stat Methodol. 2012; 74:111–132.

Schaap M, Timmermans RMA, Roemer M, Boersen GAC, Builtjes P, Sauter F, Velders G, Beck J. The LOTOS-EUROS model: Description, validation and latest developments. Int J Environ Pollut. 2008; 32:270–290.

Shaby B, Ruppert D. Tapered covariance: Bayesian estimation and asymptotics. J Comput Graph Statist. 2012; 21:433–452.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol. 2002; 64:583–639.

Stein ML. Space–time covariance functions. J Amer Statist Assoc. 2005; 100:310–321.

Stein ML. Spatial variation of total column ozone on a global scale. Ann Appl Stat. 2007; 1:191–210.

Stein ML. A modeling approach for large spatial datasets. J Korean Statist Soc. 2008; 37:3–10.

Stein ML. On a class of space–time intrinsic random functions. Bernoulli. 2013; 19:387–408.

Stein ML. Limitations on low rank approximations for covariance matrices of spatial data. Spat Stat. 2014; 8:1–19.

Stein ML, Chi Z, Welty LJ. Approximating likelihoods for large spatial data sets. J R Stat Soc Ser B Stat Methodol. 2004; 66:275–296.

Stern R, Builtjes P, Schaap M, Timmermans R, Vautard R, Hodzic A, Memmesheimer M, Feldmann H, Renner E, Wolke R, Kerschbaumer A. A model inter-comparison study focussing on episodes with elevated PM10 concentrations. Atmos Environ. 2008; 42:4567–4588.

Stoffer DS. Estimation and identification of space–time ARMAX models in the presence of missing data. J Amer Statist Assoc. 1986; 81:762–772.

Stroud JR, Müller P, Sansó B. Dynamic models for spatio-temporal data. J R Stat Soc Ser B Stat Methodol. 2001; 63:673–689.

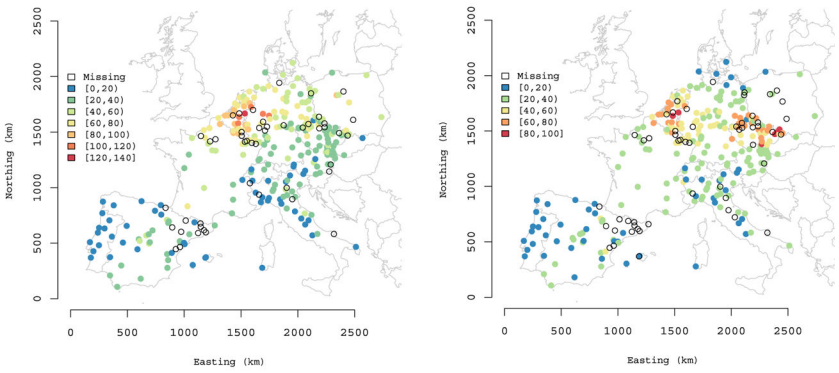
van de Kassteele J, Stein A. A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. Environmetrics. 2006; 17:309–322.

Vecchia AV. Estimation and model identification for continuous spatial processes. J Roy Statist Soc Ser B. 1988; 50:297–312.

Vecchia AV. A new method of prediction for spatial regression models with correlated errors. J Roy Statist Soc Ser B. 1992; 54:813–830.

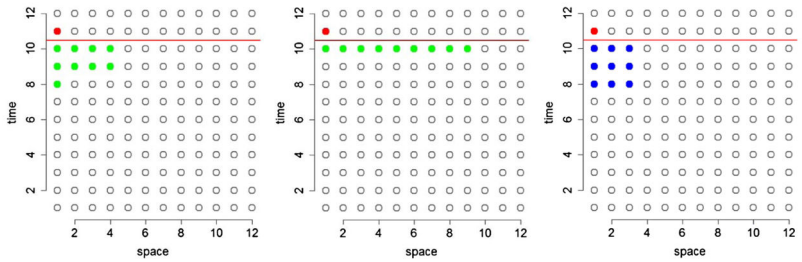
Xu G, Liang F, Genton MG. A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. Statist Sinica. 2015; 25:61–79.

Yeniay Ö, Göktaş A. A comparison of partial least squares regression with other prediction methods. Hacet J Math Stat. 2002; 31:99–111.



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

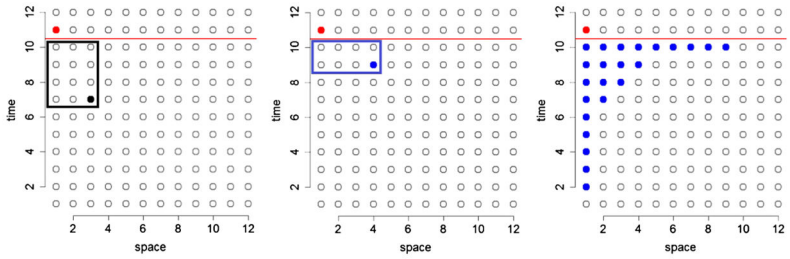
Fig. 1.
Observed PM_{10} μg m^{-3} for two example dates.



(a) $N_{\theta=1}(s_i, t_j)$ (b) $N_{\theta=2}(s_i, t_j)$ (c) Simple neighbor sets

Fig. 2.

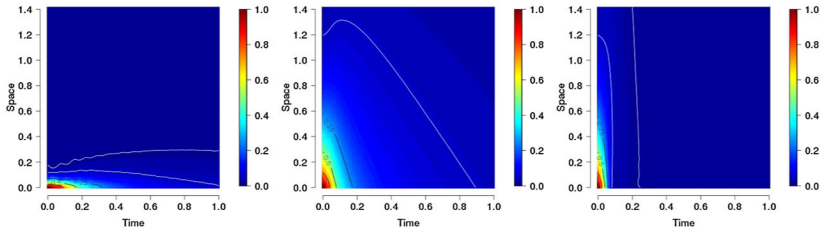
True and simple neighbor sets for a 12×12 spatio-temporal dataset with one-dimensional spatial domain and covariance function $C((s_1, t_1), (s_2, t_2)|\theta) = \exp(-|s_1 - s_2|^2 - \theta|t_1 - t_2|^2)$. All points below the red horizontal line constitute the history set for the red point (s_i, t_j). Green points denote $N_\theta(s_i, t_j)$—the sets of $m(= 9)$ true nearest neighbors with $\theta = 1$ [figure (a)] and $\theta = 2$ [figure (b)]. The blue points in figure (c) denote the simple neighbor set.



(a) Ineligible point (b) Eligible point (c) Full eligible set

Fig. 3.

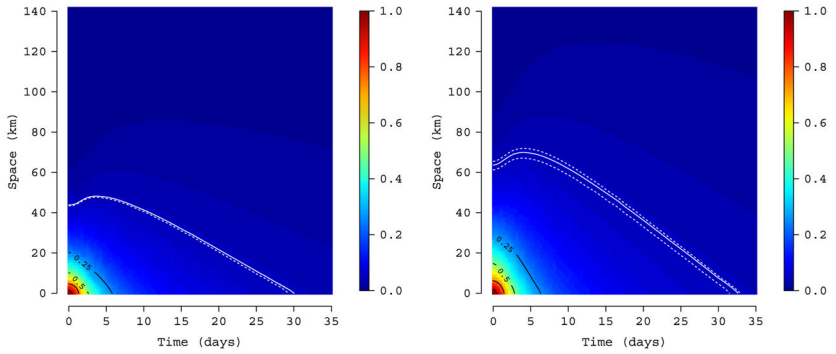
Construction of eligible sets for finding nearest neighbor sets of size $m = 9$: In figure (a) the black point is ineligible because the black rectangle contains more than $m = 9$ points. In figure (b) the blue point will belong to $E(\mathbf{s}_i, t_j)$ as the blue rectangle contains less than $m = 9$ points. Figure (c) shows the final eligible set obtained by repeating this algorithm for all points in the history set (below the red line).



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

Fig. 4.

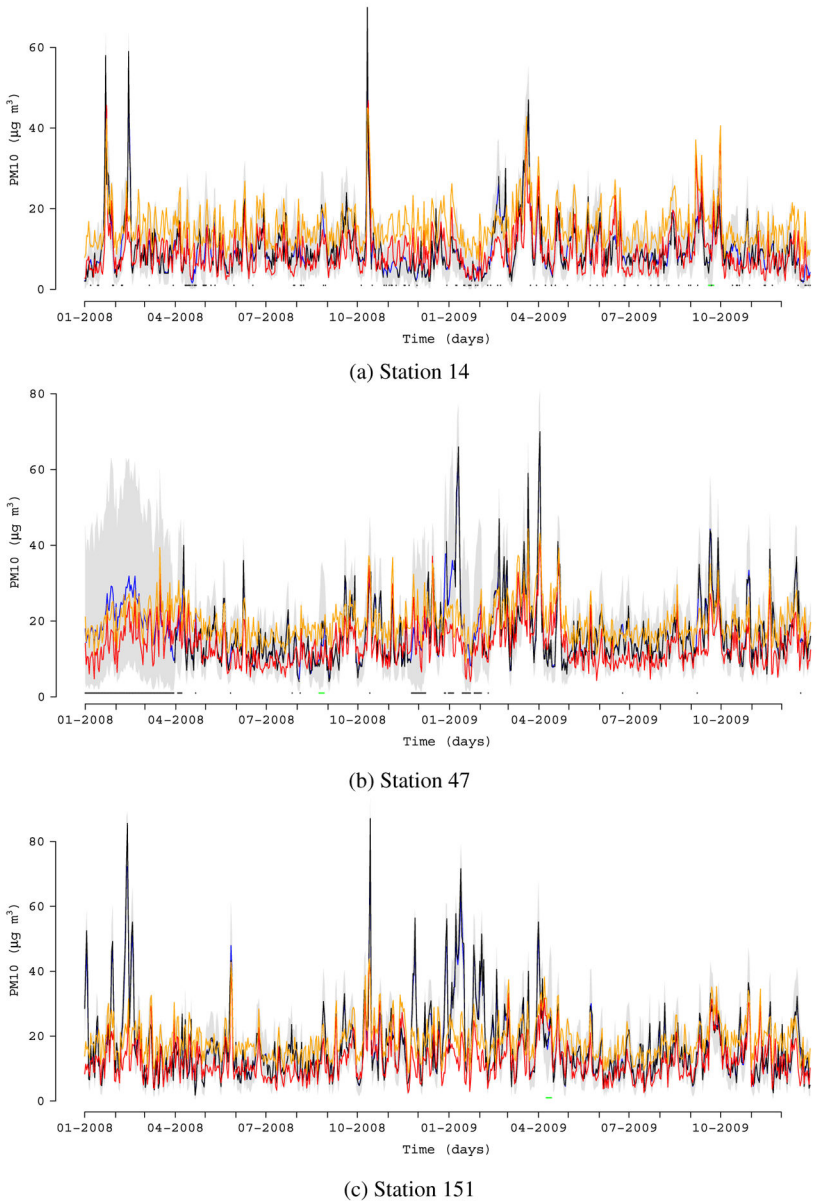
Space–time correlation surface realizations given true parameter values in Table 1. Correlation contours are provided, with the two outer white lines corresponding to 0.05 and 0.01.



(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.



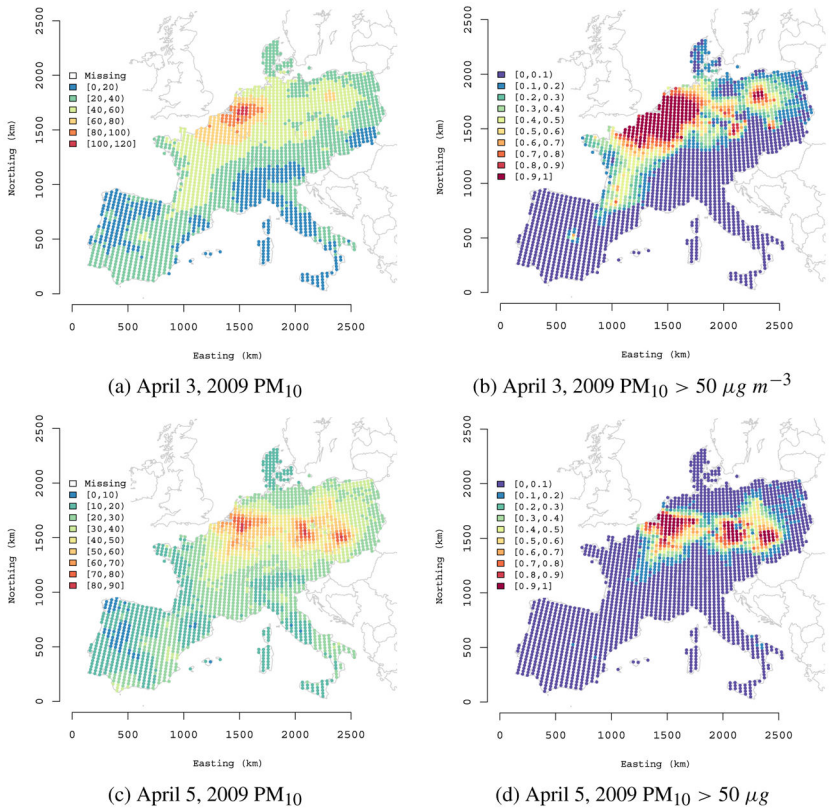
(a) Station 14

(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), nonspace–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.



(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 $PM_{10} > 50\ \mu g$

Fig. 7.

Predicted PM_{10} and probability of exceeding 50 μg m^{-3} for two example dates.

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 = G + P		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

			GP	GPP knots = 512	Adaptive DNNGP $m = 25$	Simple DNNGP $m = 25$
P			576.02	1609.99	550.28	550.36
D = G + P			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 = G + P			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

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			
Parameter	**Nonspace–time**	***m* = 16**	***m* = 25**	***m* = 36**	**Simple *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	110,266.2	122,466.2	111,190.6	103,038.3
DIC	586,135.8	279,077.3	265,720.6	277,383.9	286,922.9
G	432,811.9	11,538.63	8707.79	11,249.11	13,521.63
P	268,036.7	40,994.19	36,711.28	40,532.25	43,728.23
D	700,848.6	52,532.82	45,419.07	51,781.37	57,249.86
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	15,681.8	27,660.5	25,819

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	
	SLR	***m* = 25**	***m* = 36**	***m* = 25**	***m* = 36**
RMSPE	8.48	4.97	5.05	5.06	5.04
Bias	0.71	0.20	0.20	0.23	0.22
R_2	0.14	0.69	0.68	0.68	0.68

Ann Appl Stat. Author manuscript; available in PMC 2018 April 13.