The Annals of Applied Statistics
2021, Vol. 15, No. 2, 688–710
https://doi.org/10.1214/20-AOAS1422

SCALABLE PENALIZED SPATIOTEMPORAL LAND-USE REGRESSION FOR GROUND-LEVEL NITROGEN DIOXIDE

BY KYLE P. MESSIER[1] AND MATTHIAS KATZFUSS[2]

[1] *Division of the National Toxicology Program, National Institute of Environmental Health Sciences, kyle.messier@nih.gov*
[2] *Department of Statistics, Texas A&M University, katzfuss@gmail.com*

Nitrogen dioxide (NO_2) is a primary constituent of traffic-related air pollution and has well-established harmful environmental and human-health impacts. Knowledge of the spatiotemporal distribution of NO_2 is critical for exposure and risk assessment. A common approach for assessing air pollution exposure is linear regression involving spatially referenced covariates, known as land-use regression (LUR). We develop a scalable approach for simultaneous variable selection and estimation of LUR models with spatiotemporally correlated errors, by combining a general-Vecchia Gaussian-process approximation with a penalty on the LUR coefficients. In comparison to existing methods using simulated data, our approach resulted in higher model-selection specificity and sensitivity and in better prediction in terms of calibration and sharpness, for a wide range of relevant settings. In our spatiotemporal analysis of daily, US-wide, ground-level NO_2 data, our approach was more accurate, and produced a sparser and more interpretable model. Our daily predictions elucidate spatiotemporal patterns of NO_2 concentrations across the United States, including significant variations between cities and intra-urban variation. Thus, our predictions will be useful for epidemiological and risk-assessment studies seeking daily, national-scale predictions, and they can be used in acute-outcome health-risk assessments.

1. Introduction.

Nitrogen dioxide (NO_2) is a primary constituent of traffic-related air pollution and has well-established harmful environmental and human-health impacts (US Environmental Protection Agency (2016)). For example, exposure to NO_2 is associated with increased all-cause mortality (Hoek et al. (2013)), myocardial infarction (Rosenlund et al. (2006, 2009)), coronary heart disease (Rosenlund et al. (2008)), cardiovascular events (Alexeeff et al. (2018)), asthma (Gauderman et al. (2005)), autism spectrum disorders (Volk et al. (2013)) and impaired neurological development and other neurological disorders (Xu, Ha and Basnet (2016)). Additionally, atmospheric oxides of nitrogen, including NO_2, are precursors to hazardous acid rain (Schindler (1988)), tropospheric ozone (US Environmental Protection Agency (1999)), fine particulate matter ($PM_{2.5}$) (US Environmental Protection Agency (1999)) and can result in negative ecological (Schindler (1988)) and economic impacts (Mauzerall et al. (2005)).

Knowledge of the spatiotemporal distribution of NO_2 is critical for assessing exposure and subsequent risks. A common approach for assessing exposure to outdoor air pollution is linear regression involving spatially referenced covariates, known as land-use regression (LUR). There are many strengths in current implementations of LUR models. First, is the ability to predict a variable of interest in space and time at unmonitored coordinates, including uncertainty quantification. Second, is the use of readily-available, large geospatial datasets such as satellite imagery and census information. Third, is the elucidation and interpretation of coefficients that are possible with linear models, which allows for meaningful policy discussions around factors affecting the distribution of exposure and risk.

Received June 2020; revised November 2020.

Key words and phrases. General Vecchia approximation, spatial statistics, Gaussian process, variable selection, air pollution, Kriging.

Assuming independent and identically distributed (i.i.d.) errors, LUR has been implemented for air-quality-exposure modeling of NO_2 (Briggs et al. (1997), de Hoogh et al. (2018), Hoek et al. (2008), Knibbs et al. (2014), Larkin et al. (2017), Novotny et al. (2011), Ross et al. (2013), Su, Jerrett and Beckerman (2009)) and other air pollutants such as $PM_{2.5}$ (Henderson et al. (2007), Moore et al. (2007), Ross et al. (2013)). Typically, LUR involves model selection or dimension reduction on a large candidate-set of spatially referenced covariates. For example, LUR has been implemented with stepwise model selection for NO_2 (Briggs et al. (1997), de Hoogh et al. (2018), Knibbs et al. (2014), Novotny et al. (2011), Ross et al. (2013), Su, Jerrett and Beckerman (2009)) and partial-least-squares dimension reduction for NO_2 (Young et al. (2016)) and $PM_{2.5}$ (Sampson et al. (2013)). NO_2 LUR models have also employed penalization-based model-selection methods such as the LASSO (Knibbs et al. (2014), Larkin et al. (2017)). Additionally, LUR prediction residuals are often integrated into geostatistical models, such as Kriging (de Hoogh et al. (2018), Wu, Wang and Wu (2013)) and Bayesian maximum entropy (Beckerman et al. (2013), Coulliette et al. (2009), Messier, Akita and Serre (2012), Messier et al. (2014), Reyes and Serre (2014), Messier et al. (2015)), in a two-stage approach with the goal of improving prediction accuracy.

While LURs have undoubtedly been useful for many exposure and risk assessment studies, the assumption of i.i.d. errors is usually violated, because the spatial dependence in the response cannot be captured fully by the covariates, resulting in biased covariate estimates and decreased sensitivity and specificity in the model-selection process. An exception to this case is Holcomb et al. (2018), which implemented backwards model selection in a full Kriging model, but this approach is not feasible for large data sets. Guan et al. (2020) implemented a scalable approach with LUR with spatiotemporal errors, but used principal components instead of model selection to reduce the number of covariates.

In spatial statistics and Gaussian-process modeling, many approaches have been proposed to ensure scalability to large datasets (see, e.g., Heaton et al. (2019), Liu et al. (2020), for recent reviews and comparison) but the focus is often more on prediction based on the (residual) covariance structure, and less on penalized selection from among a large number of spatial or spatiotemporal covariates. Perhaps the most promising approaches for scalable spatial prediction are based on the ordered conditional approximation of Vecchia (1988); here, we use and extend the general Vecchia approximation (Katzfuss and Guinness (2021), Katzfuss et al. (2020a)), which is highly accurate, can guarantee linear complexity with respect to the sample size, and includes many existing Gaussian-process approximations as special cases (e.g., Datta et al. (2016a), Finley et al. (2009), Katzfuss (2017), Katzfuss and Gong (2020), Sang, Jun and Huang (2011), Snelson and Ghahramani (2007), Vecchia (1988)).

We develop an approach for simultaneous variable selection and estimation of LUR models with spatiotemporally correlated errors, extending the general Vecchia approximation to ensure scalability to large datasets. The resulting dependent-error regression problem can be transformed into standard i.i.d.-error regression involving pseudo data, which can be computed rapidly using Vecchia. This approach can be combined with any existing method for fitting penalized regression models with independent errors, such as least-angle regression (Efron et al. (2004)) for LASSO-type L1 penalties (Tibshirani (1996)), and coordinate descent (Breheny and Huang (2011)) for nonconvex (e.g., smoothly clipped absolute deviation) penalties (Fan and Li (2001)). The ordering and conditioning-set selection necessary for the Vecchia approximation is carried out based on appropriately scaled spatiotemporal coordinates. All computations necessary for inference scale linearly in the data size for fixed tuning parameters.

The remainder of this article is organized as follows: Section 2 describes the daily, ground-level NO_2 data and the geographic covariates. Section 3 provides a description of LUR with penalization. Section 4 presents our proposed methodology based on the general Vecchia approximation to Kriging models with SCAD penalty. Section 5 compares approaches in simulation studies. In Section 6, we apply our method to the NO_2 concentrations and discuss the results. Section 7 highlights the main conclusions and discusses areas for future research. A review of the general-Vecchia approximation and an alternative expression of the objective function can found in Appendices A–B. A separate Supplementary Material document contains Sections S1–S3 with additional plots, discussion (Messier and Katzfuss (2021a)) and R code (Messier and Katzfuss (2021b)). The code is based on the `R` package `GPvecchia` (Katzfuss et al. (2020b)) and is also available online at https://github.com/NIEHS/LURK-Vecchia.

2. Ground-level NO_2 data. We consider daily ground-level (i.e., tropospheric) NO_2 concentrations across the conterminous United States, monitored and distributed by the United States Environmental Protection Agency (USEPA) Air Quality System (AQS) (US Environmental Protection Agency (2019)). The date range for our study was July 10, 2018 to May 1, 2019, based on the availability of geographic covariates, primarily the TROPOMI real-time satellite imagery. The final NO_2 dataset contained 76,748 unique spatiotemporal observations distributed across 459 monitoring sites (Figure 1).

2.1. *Geographic covariates.* For our analysis, we calculated 139 spatial and spatiotemporal geographic covariates representing possible NO_2 sources and attenuation factors. A key characteristic of our and the majority of LUR studies is the presence of highly correlated covariates. In particular, many covariates only differ by their spatial resolution. After all of the covariates are calculated, each covariate is standardized to mean 0 and variance 1. The following subsections explain how each potential covariate was calculated.

2.1.1. *TROPOMI.* We utilized data from the TROPOspheric Monitoring Instrument (TROPOMI) to calculate many satellite-based spatiotemporal covariates. TROPOMI is the sensor on-board the Copernicus Sentinel-5 Precursor satellite. The TROPOMI-based covariate calculations are performed in Google Earth Engine, a cloud platform for earth observation data analysis that combines a public data catalog with a large-scale computational facility optimized for parallel processing of geospatial data.

TROPOMI provides output (i.e., Level-2 or L2 products) representing atmospheric air pollution and physical properties with a spatial resolution of approximately 3.5 by 7 km. We calculated daily mean values within 1, 10 and 100 km circular buffers for the following TROPOMI L2 products. Note the 1 km buffer is equivalent to the coincident TROPOMI value

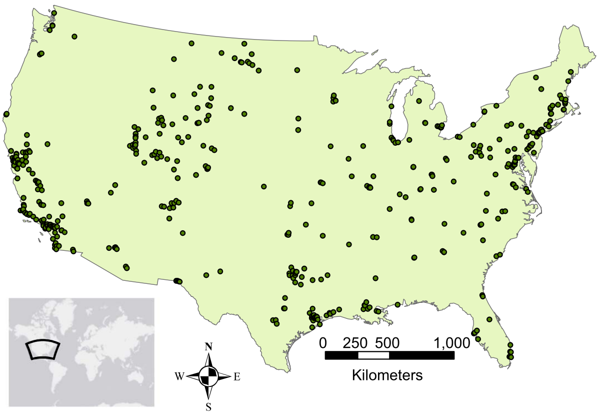

FIG. 1. *Locations of the* 459 *monitoring sites for the USEPA AQS* NO_2 *data over the study domain* (*conterminous United States*).

at the location of the monitor: Total vertical column NO_2 (mol-m^{-2}), tropospheric vertical column NO_2 (mol-m^{-2}), NO_2 slant column density (mol-m^{-2}), tropopause (i.e., boundary between troposphere and stratosphere) pressure (Pa), absorbing aerosol index (AAI; dimensionless), cloud fraction, and the solar azimuth angle (degrees). We used the near real-time TROPOMI product, if available, for the estimation of models. This was the driving factor for the sparsity of the data set within the study range. For prediction, if the real-time data were unavailable, we used the offline data. The near real-time data are available sooner and have small differences with the offline data (Boersma et al. (2007)). Lastly, we used a simple average from 10 nearest-neighbor spatiotemporal coordinates if neither near real-time nor offline were observed at a spatiotemporal coordinate. Potential alternatives for interpolating missing TROPOMI data include longer-time-scale moving-window averages (e.g., monthly) or developing a predictive model (de Hoogh et al. (2019)). The final covariate dataset included 21 TROPOMI-based variables.

2.1.2. *Meteorology.* Spatial and daily time-resolved meteorological covariates were calculated in the Google Earth Engine using the University of Idaho Gridded Surface Meteorological dataset (GRIDMET) (Abatzoglou, Rupp and Mote (2014)). GRIDMET provides daily surface fields at approximately 4 km resolution. We calculated average daily values inside 1, 10 and 100 km buffers for the following variables: precipitation (mm), maximum relative humidity (percent), specific humidity (kg-kg^{-1}), surface downward shortwave radiation (W-m^{-2}), maximum temperature (K) and wind velocity (m-s^{-1}). The 1 km buffer is equivalent to choosing the containing grid cell. The final covariate dataset included 18 meteorology-based variables.

2.1.3. *Vegetative indices.* Spatial covariates of vegetative indices were calculated in the Google Earth Engine using the MODIS/Terra Vegetative Indices 16-Day L3 Global 500 m SIN Grid (NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team (2019)). We calculated spatial averages of the normalized difference vegetative index (NDVI) and the enhanced vegative index (EVI) in 1, 10 and 100 km circular buffers. The final covariate dataset included 6 vegetative index based variables.

2.1.4. *Population*, *traffic and roads.* Population density (people-km^{-2}) was calculated in the Google Earth Engine from the Gridded Population of World Version 4 (Center for International Earth Science Information Network—CIESIN—Columbia University (2018)). Average population (2015-equivalent) density was calculated in 1, 10 and 100 km circular buffers.

A surrogate for traffic was calculated using the University of Oxford Malaria Atlas Project global travel friction dataset (Weiss et al. (2018)). Average travel friction (min-m^{-1}), or travel time, was calculated in 1, 10 and 100 km circular buffers.

Road length variables were calculated in ArcMap 10.6.1 and MATLAB R2018a using the ESRI major roads shapefile (Esri, TomTom North America, Inc.). The road length in 1, 10 and 100 km circular buffers was calculated for the following road classifications (FRC code in ESRI shapefile): all roads classes, highway (0), major roads (1, 2) and secondary roads (3, 4, 5). The final covariate dataset included 18 population, traffic, or road variables.

2.1.5. *Land cover.* Spatial land-cover attributes were calculated in the Google Earth Engine from the National Land Cover Database (Homer et al. (2015)). The percent of each land cover class (e.g., water, low developed, deciduous trees, etc.) were calculated in 1, 10 and 100 km circular buffers.

Average elevation was calculated in 1, 10 and 100 km circular buffers using the Japan Aerospace Exploration Agency Advanced Land Observing Satellite global digital surface model with a horizontal resolution of approximately 30 meters (Tadono et al. (2014)). The final covariate dataset included 48 land-cover or elevation variables.

2.1.6. *National emissions inventory.* Data on point source emissions was downloaded from the USEPA National Emissions Inventory (US Environmental Protection Agency (2017)) for the year 2017. Following Messier, Akita and Serre (2012), we calculated NO_2 point source emissions as the sum of isotropic, exponentially decaying contributions from the point sources. The initial value was the total 2017 NO_2 emissions from the emissions inventory and decay ranges were a series of ranges from short to long distance decay ranges: 1 to 10 km by 1 km increments; 20 to 100 km by 10 km increments; and 200 to 1000 km by 100 km increments, resulting in 28 NEI-based covariates.

3. Land-use regression with penalization.

Let $\mathbf{z} = (z_1, \dots, z_n)'$ denote the response vector, where $z_i = z(\mathbf{s}_i, t_i)$ is the log-transformed NO_2 measured on day t_i at spatial location $\mathbf{s}_i$. (We denote the response by $\mathbf{z}$ here for consistency with papers on general Vecchia.) We have the values $\mathbf{x}_i = \mathbf{x}(\mathbf{s}_i, t_i) = (x_1(\mathbf{s}_i, t_i), \dots, x_p(\mathbf{s}_i, t_i))'$ of the $p = 139$ covariates (described in Section 2.1) at the same (space, time)-coordinate pairs $\mathcal{S} = \{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\}$. Spatial-only covariates are repeated in time as needed. We assume a linear relationship between the response and covariates,

$$z_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \eta_i + \delta_i = y_i + \delta_i, \quad i = 1, \dots, n, \tag{1}$$

where $\epsilon_i = \eta_i + \delta_i$ is the regression error consisting of a spatiotemporally dependent component $\eta_i = \eta(\mathbf{s}_i, t_i)$ and an independent measurement-noise component $\delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$, $i = 1, \dots, n$, and $y_i = z_i - \delta_i = \mathbf{x}_i'\boldsymbol{\beta} + \eta_i$ is the (noise-free) true log-NO_2. We assume that $\eta(\cdot) \sim \mathcal{GP}(0, C_{\boldsymbol{\theta}})$ follows a Gaussian process with covariance function $C_{\boldsymbol{\theta}}$. Throughout, we assume a nonseparable spatiotemporal exponential covariance function, $C_{\boldsymbol{\theta}}((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = \sigma^2 \exp(-d_{\boldsymbol{\theta}}((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)))$, where

$$d_{\boldsymbol{\theta}}\big((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)\big) = \sqrt{\frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2^2}{\gamma_s^2} + \frac{(t_i - t_j)^2}{\gamma_t^2}}, \tag{2}$$

and $\boldsymbol{\theta} = (\sigma, \gamma_s, \gamma_t, \tau)$ contains the unknown parameters in the model. We also considered and dismissed a Matérn covariance with estimated smoothness parameter, as this resulted in a smoothness parameter near that of the exponential model (i.e., 0.5), nearly identical negative log-likelihood values and increased model run time due to evaluation of Bessel functions and larger parameter space. Note that standard LUR models (e.g., Briggs et al. (1997), de Hoogh et al. (2018), Hoek et al. (2008)) do not include the dependent component $\eta(\cdot)$ and assume the error terms $\epsilon_1, \dots, \epsilon_n$ to be i.i.d.

Stacking the quantities in (1), we obtain the regression model

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathbf{C}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}_n$, with $\mathbf{C}_{\boldsymbol{\theta}} = (C_{\boldsymbol{\theta}}((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)))_{i,j=1,\dots,n}$. Equivalently, we can write this in terms of a multivariate Gaussian density for the response,

$$f(\mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{N}_n(\mathbf{z}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{3}$$

The goal is to estimate the p-vector $\boldsymbol{\beta}$ and determine its nonzero elements, which also requires estimation of the covariance parameters $\boldsymbol{\theta}$. Further, given parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, we would like to predict the process $y(\cdot)$ at unobserved coordinates.

A standard approach for parameter estimation is to maximize the likelihood in (3) with respect to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. However, we have a large number $p = 139$ of (correlated) covariates, which makes the least-squares or maximum likelihood estimates of $\boldsymbol{\beta}$ unstable. To alleviate this issue, and to be able to select certain variables and set the coefficients corresponding to the other variables to zero, we instead consider optimizing an objective function consisting of the negative loglikelihood plus a penalization term $\mathrm{p}(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$:

$$(4)\quad Q(\boldsymbol{\beta},\boldsymbol{\theta}) = -2\log f(\mathbf{z};\boldsymbol{\beta},\boldsymbol{\theta}) + \lambda\,\mathrm{p}(\boldsymbol{\beta}) = (\mathbf{z}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}) + \log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \lambda\,\mathrm{p}(\boldsymbol{\beta}),$$

where λ is a shrinkage or tuning parameter, and we have omitted an additive constant in the last equation. In our numerical examples and application, we will use the popular nonconvex, smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)):

$$(5)\qquad \mathrm{p}(\beta) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \le \lambda, \\ \dfrac{2a\lambda|\beta| - \beta^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \le a\lambda, \\ \dfrac{\lambda^2(a+1)}{2} & \text{otherwise,} \end{cases}$$

where $a = 3.7$, a popular choice that performs comparably to values based on generalized cross-validation (Fan and Li (2001)). We use the SCAD penalty for its oracle property, but other penalties can be easily be swapped in our framework. Li and Sudjianto (2005) demonstrate that a SCAD-penalized likelihood as in (4) and (5) reduces the variance in the estimates of $\boldsymbol{\theta}$; however, their discussion did not address model selection or large sample sizes.

4. Our methodology.

4.1. *A general Vecchia approximation of the objective function.* Evaluation or optimization of the objective function Q in (4) requires decomposition of the $n \times n$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ for many different values of $\boldsymbol{\theta}$, each of which takes $\mathcal{O}(n^3)$ time. This is computationally infeasible for the large $n = 76{,}748$ in our application.

Hence, we will extend the sparse general Vecchia (SGV) approximation (Katzfuss and Guinness (2021)), which we briefly review here, with more details given in Appendix A. SGV applies the approximation of Vecchia (1988) to the vector $\mathbf{u} = (y_1, z_1, \dots, y_n, z_n)'$, which interweaves the latent true-process realizations $y_1, \dots, y_n$ and the observed noisy data $z_1, \dots, z_n$. This approximation essentially replaces the conditioning sets in the exact factorization $f(\mathbf{u}) = \prod_{j=1}^{2n} f(u_j|u_1, \dots, u_{j-1})$ by small subsets, resulting in the approximation

$$(6)\qquad \widehat{f}(\mathbf{u}) = \prod_{i=1}^{2n} p(u_i|\mathbf{u}_{g(i)}) = \mathcal{N}_{2n}\big(\mathbf{u}|(\mathbf{X}\otimes\mathbf{1}_2)\boldsymbol{\beta}, (\mathbf{U}\mathbf{U}')^{-1}\big),$$

where each $g(i) \subset (1, \dots, i-1)$ is a conditioning index set of size $|g(i)| \le m$, $\otimes$ is the Kronecker product, $\mathbf{1}_2$ is a vector consisting of two 1s, and $\mathbf{U} = \mathbf{U}_{\boldsymbol{\theta}}$ is a sparse upper triangular matrix whose nonzero entries can be computed easily based on $C_{\boldsymbol{\theta}}$ and τ^2. Recent results (Schäfer, Katzfuss and Owhadi (2020)) indicate that the approximation error can be bounded with the conditioning-set size m increasing logarithmically in n in some settings; in practice, $m \approx 30$ is often sufficient for accurate approximations. We further define $\mathbf{A}$ and $\mathbf{B}$ as the submatrices of $\mathbf{U}$ consisting of the odd- and even-numbered rows of $\mathbf{U}$, corresponding to $\mathbf{y}$ and $\mathbf{z}$, respectively. Then $\mathbf{W} = \mathbf{A}\mathbf{A}'$ is the implied posterior precision matrix of $\mathbf{y}$ given $\mathbf{z}$, and we define $\mathbf{V}$ as the Cholesky factor based on reverse row-column ordering of $\mathbf{W}$.

Our approximation $\widehat{f}(\mathbf{u})$ is an extension of the SGV approach for spatial processes described in Katzfuss and Guinness (2021); to approximate the spatiotemporal covariance function $C_{\boldsymbol{\theta}}$, we modify the ordering and conditioning scheme here to be carried out based on the scaled spatiotemporal distance (2), which depends on unknown parameters and must be updated along with the parameters. Again, more details are given in Appendix A.

The SGV approximation of the density of $\mathbf{u} = (y_1, z_1, \dots, y_n, z_n)'$ in (6) implies an approximation of the distribution for the response:

$$\widehat{f}(\mathbf{z}) = \int \widehat{f}(\mathbf{u})\, d\mathbf{y},$$

which is also multivariate normal. This concludes our review of Katzfuss and Guinness (2021). Plugging the approximation $\widehat{f}(\mathbf{z})$ into (4) results in the Vecchia objective function

$$\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta}) = -2 \log \widehat{f}_{\boldsymbol{\beta},\boldsymbol{\theta}}(\mathbf{z}) + \lambda\, \mathrm{p}(\boldsymbol{\beta}), \tag{7}$$

where we have now made explicit the dependence of the distribution of $\mathbf{z}$ on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We will use and optimize the Vecchia objective function $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta})$ in the remainder of the manuscript.

4.2. *Inference.* The most straightforward way to optimize the objective function is to optimize iteratively with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, while holding the respective other parameter vector fixed. In practice, it is usually sufficient to do this just a small number of times, after which there is little change in the parameter values.

As we prove in Appendix B, $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (7) can be written as

$$\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \|\tilde{\mathbf{z}}_{\boldsymbol{\theta}} - \tilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta}\|_2^2 + \lambda\, \mathrm{p}(\boldsymbol{\beta}) - 2\sum_i \log\bigl((\mathbf{U}_{\boldsymbol{\theta}})_{ii}\bigr) + 2\sum_i \log\bigl((\mathbf{V}_{\boldsymbol{\theta}})_{ii}\bigr), \tag{8}$$

where $\tilde{\mathbf{z}}_{\boldsymbol{\theta}} = \mathbf{B}'\mathbf{z} + \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{z}$, $\tilde{\mathbf{X}}_{\boldsymbol{\theta}} = \mathbf{B}'\mathbf{X} + \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{X}$, and $\mathbf{V}_{\boldsymbol{\theta}}$ depend on $\boldsymbol{\theta}$ through the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{W}$ computed from $\mathbf{U} = \mathbf{U}_{\boldsymbol{\theta}}$.

4.2.1. *Estimation of the regression coefficients.* Optimizing $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (8) with respect to $\boldsymbol{\beta}$ for fixed $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is equivalent to solving a standard penalized regression problem with i.i.d. errors, except based on the pseudo-data $\tilde{\mathbf{z}}_{\hat{\boldsymbol{\theta}}}$ and $\tilde{\mathbf{X}}_{\hat{\boldsymbol{\theta}}}$.

To develop some intuition, consider briefly the case $m = n - 1$, in which case the approximation $\widehat{f}(\mathbf{u})$ in (6) becomes exact. Then the pseudo-data $\tilde{\mathbf{z}}_{\hat{\boldsymbol{\theta}}}$ are obtained by first creating an augmented data vector of length $2n$ consisting of $\mathbf{z}$ and $\mathbb{E}(\mathbf{y}|\mathbf{z})$, and then transforming this vector to a vector of i.i.d. normal variables based on the joint distribution or covariance matrix of $\mathbf{z}$ and $\mathbf{y}$. Interestingly, the resulting inference on $\boldsymbol{\beta}$ is unchanged relative to simply transforming the data $\mathbf{z}$ alone, as is often done for general linear models. In the case of small m, our sparse general Vecchia approximation allows us to carry out this inference on $\boldsymbol{\beta}$ more accurately based on the first approach.

For example, in the case of the SCAD (Fan and Li (2001)) or L1 penalties (Tibshirani (1996)), solution paths of optimal $\boldsymbol{\beta}$ values for each value of λ can be computed rapidly using coordinate descent (Breheny and Huang (2011)) or least angle regression (Efron et al. (2004)), respectively. We select the optimal λ and the corresponding $\hat{\boldsymbol{\beta}}$ based on the lowest cross-validated mean square error, which can be performed in many software packages such as *ncvreg* (Breheny and Huang (2011)) or *glmnet* (Friedman, Hastie and Tibshirani (2010)). Breheny and Huang (2011) demonstrated for high-dimensional problems that the SCAD penalty estimated with coordinate descent, in combination with cross-validation, leads to the global minimum solution as it likely resides in the locally convex region of λ.

We considered a simple example to illustrate how quickly the Vecchia solution can converge to the exact solution in the estimation of trend parameters, $\boldsymbol{\beta}$. We simulated $p = 5$ correlated covariates with correlations ranging from 0.45 to 0.82, set the true $\boldsymbol{\beta} = \mathbf{0}$, and simulated data with spatially dependent error at $n = 500$ locations. We then computed the exact generalized least-squares (GLS) estimates and the Vecchia GLS estimates implied by the pseudo-data in (8). As shown in Figure 2, the Vecchia solution quickly approached the exact solution going from $m = 0$ (i.e., assuming independent errors) to $m = 10$.

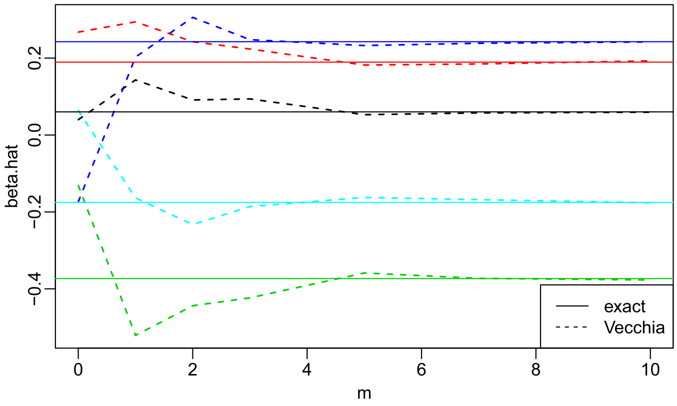

FIG. 2. *Demonstration of the quick convergence of general Vecchia estimates of $\boldsymbol{\beta}$ with increasing m toward the exact* (*i.e., without approximation*) *generalized-least squares* (*GLS*) *estimate.*

4.2.2. *Estimation of the covariance parameters.* Defining $\boldsymbol{\epsilon}_{\boldsymbol{\beta}} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$, (8) can be rearranged to yield

$$\hat{Q}(\boldsymbol{\beta},\boldsymbol{\theta}) = \|\mathbf{B}'\boldsymbol{\epsilon}_{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{V}^{-1}\mathbf{A}\mathbf{B}'\boldsymbol{\epsilon}_{\boldsymbol{\beta}}\|_2^2 - 2\sum_i \log \mathbf{U}_{ii} + 2\sum_i \log \mathbf{V}_{ii} + \lambda\,\mathrm{p}(\boldsymbol{\beta}), \tag{9}$$

where $\mathbf{V}$, $\mathbf{A}$ and $\mathbf{B}$ implicitly depend on $\boldsymbol{\theta}$ through $\mathbf{U} = \mathbf{U}_{\boldsymbol{\theta}}$. Note that this expression is an extension of the Vecchia log-likelihood in Katzfuss and Guinness (2021); we replaced their zero-mean data $\mathbf{z}$ with our residuals $\boldsymbol{\epsilon}_{\boldsymbol{\beta}}$, and we have added the penalization term $\lambda\,\mathrm{p}(\boldsymbol{\beta})$. This alternative expression of the objective function is important, as it avoids having to compute the pseudo-data for every evaluation as in (8).

For a given $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, (9) can be evaluated cheaply for any given parameter value $\boldsymbol{\theta}$, and hence we can optimize the objective function with respect to $\boldsymbol{\theta}$ using standard numerical optimization algorithms (e.g., Nelder–Mead). As in most Gaussian-process models, there is no guarantee that this procedure will find the global optimum, but we have not observed any negative consequences. Similarly, we have not observed any numerical issues due to nonidentifiability between the variance and range parameters, which theoretically holds under in-fill asymptotics (Tang, Zhang and Banerjee (2019), Zhang (2004)) but not under the increasing-domain asymptotics that may be more appropriate for our real-data application with small effective ranges relative to the domain size.

We monitor convergence of the overall algorithm by considering the minimum value of (9) achieved at each iteration.

4.2.3. *Prediction.* Often, interest is in prediction of the noise-free process $\mathbf{y}_P$ at a set of n_P spatiotemporal coordinates $\mathcal{S}_P$, which is equivalent to obtaining the conditional distribution of $\mathbf{y}_P$ given the data $\mathbf{z}$. To do so, we extend the response-first full-conditioning (RF-full) approach of Katzfuss et al. (2020a), which essentially consists of a general Vecchia approximation $\widehat{f}(\tilde{\mathbf{u}})$, similar to (6), but now applied to the vector $\tilde{\mathbf{u}} = (\mathbf{z}', \mathbf{y}'_{\mathrm{all}})'$, where $\mathbf{y}_{\mathrm{all}} = (\mathbf{y}', \mathbf{y}'_P)'$. We have

$$\widehat{f}(\mathbf{y}_{\mathrm{all}}|\mathbf{z}) = \frac{\widehat{f}(\mathbf{z}, \mathbf{y}_{\mathrm{all}})}{\int \widehat{f}(\mathbf{z}, \mathbf{y}_{\mathrm{all}})\, d\mathbf{y}_{\mathrm{all}}} =: \mathcal{N}_n(\boldsymbol{\mu}_{\mathrm{all}}, \mathbf{W}_{\mathrm{all}}^{-1}).$$

Any quantities of interest can be extracted from this *joint* distribution, after computing $\mathbf{V}_{\mathrm{all}}$ as the Cholesky factor based on reverse row-column ordering of $\mathbf{W}_{\mathrm{all}}$. For example, the prediction mean, also referred to as the kriging predictor, is obtained by subsetting the vector $\boldsymbol{\mu}_{\text{all}} = \mathbf{X}_{\text{all}}\hat{\boldsymbol{\beta}} - (\mathbf{V}'_{\text{all}})^{-1}\mathbf{V}_{\text{all}}^{-1}\mathbf{A}_{\text{all}}\mathbf{B}'_{\text{all}}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{\mu}', \boldsymbol{\mu}'_P)'$, where $\mathbf{X}_{\text{all}} = (\mathbf{X}', \mathbf{X}'_P)'$, while the prediction or kriging variances are given by a subset of the diagonal elements of $\mathbf{W}_{\text{all}}^{-1}$, which can be obtained using selected inversion based on the Takahashi recursions for $\mathbf{V}_{\text{all}}$. Note that we ignore uncertainty in $\hat{\boldsymbol{\beta}}$ in the predictions, but we conducted experiments that showed this uncertainty is often small relative to the uncertainty in $\eta(\cdot)$. If prediction of $\mathbf{z}_P$ is desired, we simply need to add τ^2 to the prediction variances. Our approximation is an extension of the spatial RF-full prediction in Katzfuss et al. (2020a), in that the nonzero mean has to be added and subtracted in the kriging predictor, and we carry out the ordering and conditioning in the scaled spatiotemporal domain. Details are given in Appendix A.

4.2.4. *Complexity.* Our proposed inference procedure is summarized in Algorithm 1. If each conditioning index vector in our Vecchia approximations is at most of size m, SGV and RF-full ensure that $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{V}_{\text{all}}$ are all highly sparse with at most m nonzero off-diagonal entries per column. As a consequence, for fixed m and p, our entire inference procedure requires linear time in the number of observed and prediction coordinates.

More precisely, assuming that $m, p, n_P \leq n$, evaluation of the likelihood and prediction for each parameter value requires $\mathcal{O}(nm^3)$ time, coordinate descent for SCAD requires $\mathcal{O}(np)$ time per iteration, one triangular solve involving $\mathbf{V}$ requires $\mathcal{O}(nm)$ time, and hence computing the pseudo-data $\tilde{\mathbf{z}}_{\hat{\boldsymbol{\theta}}}$ and $\tilde{\mathbf{X}}_{\hat{\boldsymbol{\theta}}}$ requires $\mathcal{O}(nmp)$ time. If we require L iterations going back and forth between estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, L_θ iterations to estimate $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\beta}}$, and L_β iterations in the coordinate descent (including selecting tuning parameters using cross-validation) to estimate $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\theta}}$, the overall cost of our algorithm is $\mathcal{O}(nL(L_\theta m^3 + L_\beta p))$. We utilize a tolerance $\texttt{tol} = 10^{-6}$ for the stopping criterion; a less stringent tolerance may be used, which will likely result in a smaller number of iterations L and a less accurate approximation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Algorithm 1: LURK-Vecchia: Land-use regression Kriging with Vecchia approx

```
Input: $\mathbf{z}$, $\mathcal{S}$, $\mathbf{X}$, $\mathcal{S}_P$, $\mathbf{X}_P$, $C_{\boldsymbol{\theta}}$, tol
Result: Parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$; prediction $\widehat{f}(\mathbf{y}_{\text{all}}|\mathbf{z})$
Initialize $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, prev.objective $= \infty$, and converged $=$ FALSE
OC: Maxmin ordering and nearest-neighbor conditioning for $\mathcal{S}$ based on $d_{\hat{\boldsymbol{\theta}}}$ in (2)
while converged = FALSE do
    Compute $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \hat{Q}(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta})$ based on (9) using OC
    Update OC based on $d_{\hat{\boldsymbol{\theta}}}$ in (2)
    Compute pseudo-data $\tilde{\mathbf{z}}_{\hat{\boldsymbol{\theta}}}$ and $\tilde{\mathbf{X}}_{\hat{\boldsymbol{\theta}}}$ as in (8) using OC
    Estimate $\boldsymbol{\beta}$ using standard (i.i.d.) SCAD based on $\tilde{\mathbf{z}}_{\hat{\boldsymbol{\theta}}}$ and $\tilde{\mathbf{X}}_{\hat{\boldsymbol{\theta}}}$ (see Section 4.2.1)
    new.objective $= \hat{Q}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$
    if new.objective $>$ (prev.objective $\times$ (1-tol)) then
        converged $=$ TRUE
    else
        prev.objective $=$ new.objective
    end if
end while
OC$_P$: Ordering and conditioning for $\mathcal{S}$ and $\mathcal{S}_P$ based on $d_{\hat{\boldsymbol{\theta}}}$ (see Appendix A)
Prediction: Compute relevant summaries of $\widehat{f}(\mathbf{y}_{\text{all}}|\mathbf{z})$ using OC$_P$ (see Section 4.2.3)
```

5. Simulation study.

5.1. *Simulation scenarios.* We sampled 2000 spatiotemporal coordinates from the possible combinations of 276 unique days and 50 unique spatial locations randomly distributed across the United States. The unique days correspond to the set of dates with complete geographic covariate datasets. The 2000 coordinates are randomly divided in half for a training set of size $n = 1000$, and a test set of size $n_P = 1000$ that was never used in any model development.

For the spatiotemporal regression errors ϵ_i in (1), we specified a baseline (minimum; maximum) scenario of the model, with spatial range parameter $\gamma_s = 1000$ (200; 3000) km, temporal range $\theta_t = 30$ (7; 365) days, total variance (i.e., sill) $\sigma^2_{\text{total}} = \sigma^2 + \tau^2 = s^2_{\text{trend}}$ ($0.5 \times s^2_{\text{trend}}$; $5 \times s^2_{\text{trend}}$), and nugget-to-sill ratio $\tau^2/\sigma^2_{\text{total}} = 0.25$ (0.01; 0.99), where s^2_{trend} is the sample variance of the entries of the regression term $\mathbf{X}\boldsymbol{\beta}$, evaluated at the true value of $\boldsymbol{\beta}$ (see below). The nugget-to-sill ratio is the ratio of the noise to total variance. We considered a large number of simulation scenarios in which we varied, in turn, each of these variables, while holding the other variables fixed at their baseline levels. (Results for additional scenarios in which the variables varied jointly, including a spatial range of 30 km, are shown in Section S1.)

For the simulated spatiotemporal coordinates, we created spatiotemporal covariate matrices $\mathbf{X}$ and $\mathbf{X}_P$ based on the methods described for the NO_2 data in Section 2, in order to obtain a realistic simulation setting. To provide a unique set of covariates from the NO_2 analysis, we included and removed variables as follows. Ozone TROPOMI satellite data (air mass factor (AMF), total column and slant) were included with 1, 10, and 100 km buffers. Randomly generated point sources with isotropic exponentially decaying contribution with decay ranges of 1, 10 and 100 km (Messier, Akita and Serre (2012)) and randomly generated spatiotemporal random fields were included in lieu of the NEI and road covariates from the NO_2 dataset. The final candidate set for the simulation included 123 potential covariates. The true trend coefficients $\boldsymbol{\beta}$ were assumed to have 8 nonzero coefficients: NO_2 Slant 1 km, cloud fraction 10 km, ozone AMF 100 km, precipitation 10 km, NDVI 100 km, developed high intensity 1 km, point sources with 100 km decay range and a smoothly varying spatiotemporal random field; the corresponding true coefficient values were 5, 5, 3, -3, -5, 10, 3, 5, respectively. The true covariates exhibited low to moderate correlation ($|\rho| \leq 0.41$) with the other true covariates, and low to extremely high correlation ($|\rho| \leq 0.99$) with the extraneous covariates.

5.2. *Approaches under comparison.* We compared our proposed method to several popular land-use regression approaches:

LUR-i.i.d.: An i.i.d. land-use regression model, which can be viewed as a special case of (1) with $\eta(\cdot) \equiv 0$. Point predictions are then simply given by $\mathbf{X}_P\hat{\boldsymbol{\beta}}$.

LURK-Local: Land-use regression Kriging with a local neighborhood, which is the current state-of-the-art approach in land-use regression (de Hoogh et al. (2018)). LURK-local consists of the following steps:

1. Estimate $\hat{\boldsymbol{\beta}}$ as in LUR-i.i.d., and compute residuals $\boldsymbol{\epsilon}_{\hat{\boldsymbol{\beta}}}$.
2. Estimate $\boldsymbol{\theta}$ as an average of estimates based on $k = 10$ samples of size $l = (m\frac{n}{k})^{1/3}$ from $\boldsymbol{\epsilon}_{\hat{\boldsymbol{\beta}}}$
3. Carry out local kriging at each prediction coordinate using the m nearest (in terms of (2)) space-time neighbors among $\boldsymbol{\epsilon}_{\hat{\boldsymbol{\beta}}}$.

LURK-Vecchia: Our proposed methodology, summarized in Algorithm 1.

LURK-Full: The full Kriging and SCAD penalized method based on (3). This is equivalent to the proposed LURK-Vecchia approach with $m = n - 1$.

Local-Kriging: Does not use geographic covariates. Similar to LURK-local, we estimate $\boldsymbol{\theta}$ as an average of estimates based on $k = 10$ samples of size $l = (m\frac{n}{k})^{1/3}$ from $\mathbf{z}$, and then make predictions using the m nearest spatiotemporal observations.

LURK-Full can be considered the most accurate approach, but it is computationally infeasible for large n (in the tens of thousands or more). For the other spatiotemporal approaches, we ensure similar computational complexity by using the same $m = 25$.

5.3. *Prediction scores.* For the prediction at unobserved coordinates, we considered three proper scoring rules (e.g., Gneiting and Katzfuss (2014)) that all compare the true simulated test data $\mathbf{y}_P^\star$ to the predictive distribution $\widehat{f}(\mathbf{y}_P|\mathbf{z}) = \mathcal{N}(\mathbf{y}_P|\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ (see Section 4.2.3) as approximated by each method. The mean squared error (MSE) is given by $(1/n_P)\sum_{i=1}^{n_P}(\mathbf{y}_{P,i}^\star - \boldsymbol{\mu}_{P,i})^2$, the log-score is given by $-(1/n_P)\sum_{i=1}^{n_P}\mathcal{N}(\mathbf{y}_{P,i}^\star|\boldsymbol{\mu}_{P,i}, \boldsymbol{\Sigma}_{P,ii})$ and the continuous ranked probability score is given by $(1/n_P)\sum_{i=1}^{n_P}\int(F_i(x) - 1\{\mathbf{y}_{P,i}^\star \leq x\})^2\,dx$, where F_i is the cumulative distribution function of $\mathcal{N}(\boldsymbol{\mu}_{P,i}, \boldsymbol{\Sigma}_{P,ii})$. Each score is averaged over 20 simulations.

5.4. *Simulation results.*

5.4.1. *Out-of-sample prediction.* Figure 3 shows ridgeline density plots of the prediction scores in Section 5.3 for the methods in Section 5.2 (except LUR-i.i.d., which was not competitive) for the different simulation scenarios described in Section 5.1. The vertically oriented densities are generated as trimmed (inner 98%) density functions using the geom_density_ridges function in the ggplot2 and ggridges packages of R. The results for our

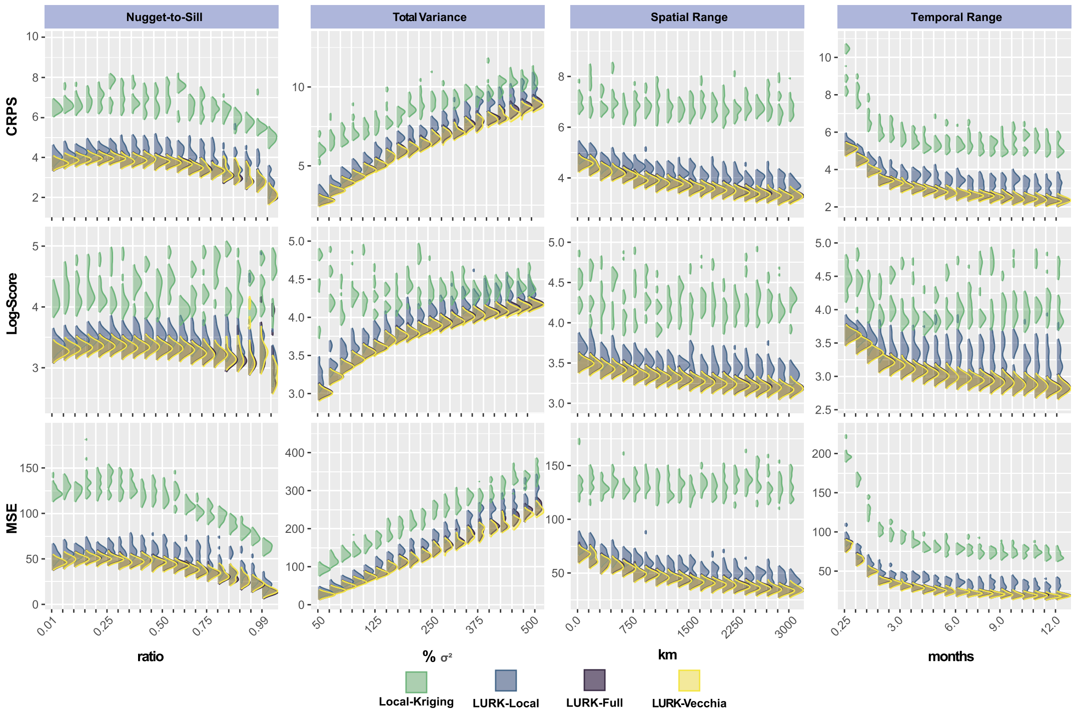

FIG. 3. *Ridgeline density plots of prediction scores (the lower the better) for different simulation scenarios (Section 5.1), in which the spatiotemporal parameters are singly varied with others held constant at a baseline level. The LURK-Vecchia results were very close to, and hence largely cover the results for LURK-Full. Note that each scenario has its own y-axis scale.*

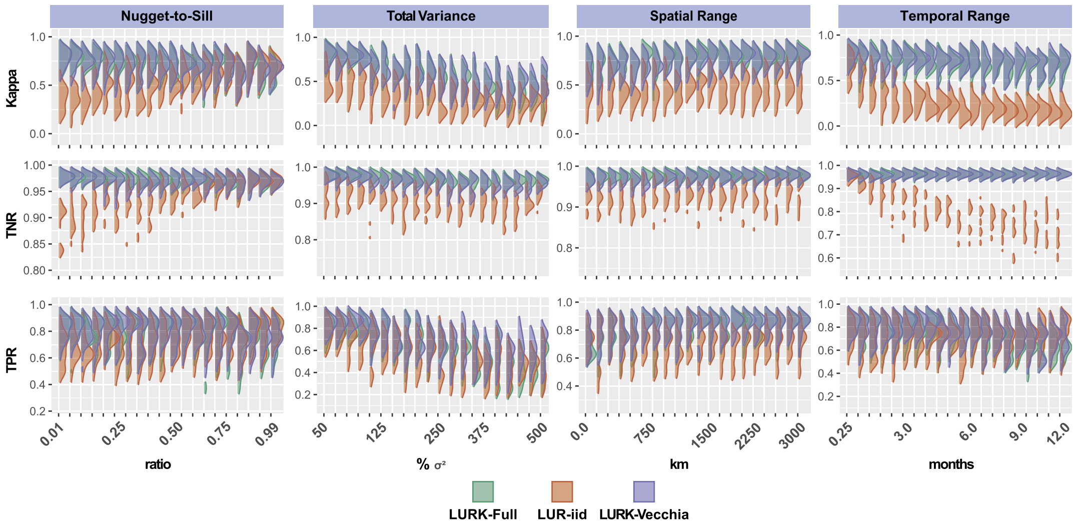

FIG. 4. *Ridgeline density plot comparison of model-selection statistics for simulation scenarios* (*Section* 5.1), *in which the spatiotemporal parameters are singly varied with others held constant at a baseline level.*

LURK-Vecchia approach were nearly identical to those for LURK-Full. Across all scenarios, the average LURK-Vecchia scores were consistently between 5 and 60% better than those for LURK-Local. Local-Kriging was much worse. (Plots showing percent differences are shown in Section S1.)

5.4.2. *Model selection.* In terms of model selection, we considered the true negative rate (TNR), true positive rate (TPR), and Cohen's Kappa (Banerjee et al. (1999)), $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is the observed agreement of coefficient selections and p_e is the expected agreement based on random chance. Of the methods in Section 5.2, we omitted LURK-Local (because its model selection is identical to LUR-i.i.d.) and Local-Kriging (because it does not perform model selection). Figure 4 shows the model-selection statistics for the scenarios in Section 5.1. Similar to the prediction scores, the distributions of the LURK-Vecchia and LURK-Full were very similar, indicating that the LURK-Vecchia approach approximated the full model well in terms of model selection. Compared to LUR-i.i.d., LURK-Vecchia had 5 to 80% higher average TPR. The difference in terms of TNR and Kappa was even greater, with LUR-i.i.d. selecting a large number of erroneous nonzero coefficients (Figure S3). The results were only comparable for scenarios with negligible spatiotemporal dependence (i.e., high nugget-to-sill ratio or small ranges). An additional plot showing the percent differences in Kappa is shown in Section S1.

6. Ground-level NO$_2$ analysis. We now return to the daily, US-wide, ground-level NO$_2$ data described in Section 6. The minimum, maximum, mean (standard deviation), and median (interquartile range) observed concentrations were 0.004, 62.9, 8.5 (7.4), and 6.4 (9.0) parts-per-billion (ppb), respectively. Because NO$_2$ is positive and right-skewed, NO$_2$ was natural-log-transformed prior to the analyses.

6.1. *Comparison using cross-validation.* We compared the predictive accuracy using 10-fold cross-validation for the methods described in Section 5.2; LURK-Full was omitted because it is intractable for the large sample size. We used the same predictive scores as detailed in Section 5.3, except that we considered $\widehat{f}(\mathbf{z}_P|\mathbf{z}) = \mathcal{N}(\mathbf{z}_P|\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P + \tau^2\mathbf{I}_{n_P})$, because the error-free $\mathbf{y}_P$ was unknown.

TABLE 1
Predictive scores for NO_2 10*-fold cross-validation*

Method	MSE (ppb^2)	CRPS	Log-score
Local-Kriging	0.30	0.30	0.83
LUR-i.i.d.	0.42	0.38	3.90
LURK-Local	0.22	0.25	0.68
LURK-Vecchia	0.20	0.24	0.61

The results are shown in Table 1. LURK-Vecchia outperformed all other methods in terms of all three scores, resulting in a roughly ten percent decrease in MSE and log-score compared to the next best approach, LURK-Local. Further, LURK-Vecchia resulted in a 20% or greater decrease in all of the scores compared to Local-Kriging and LUR-i.i.d.

In terms of model selection, the mean (standard deviation) number of nonzero coefficients was 71 (1.1) and 24 (1.7) for LUR-i.i.d. and LURK-Vecchia, respectively. The LUR-i.i.d. models were severely affected by multicollinearity; the median (mean) variance inflation factor (VIF) for all 10 cross-validation models was 2.3 (6.7) and 5.7 (18.6) for LURK-Vecchia and LUR-i.i.d., respectively. LURK-Local uses the same model-selection procedure as LUR-i.i.d. Other i.i.d. model-selection approaches are likely to result in similarly large numbers of covariates (e.g., Kerckhoffs et al. (2019)).

Thus, by appropriately accounting for spatiotemporal dependence, LURK-Vecchia resulted in more accurate, sparser, and more interpretable models than LUR-i.i.d. and LURK-Local.

6.2. *Prediction maps.* Having shown using cross-validation that our proposed LURK-Vecchia approach can outperform the competing methods, we fitted LURK-Vecchia to the entire dataset. The covariance parameters $\boldsymbol{\theta} = (\sigma^2, \gamma_s, \gamma_t, \tau^2)$ were estimated as (2.2 ppb^2, 1.4 km, 0.63 yr, 0.15 ppb^2), and the trend coefficients are given in Table 1 and discussed in Section 6.3. The entire estimation algorithm required $L = 4$ iterations and took approximately 86 minutes on a machine with 16GB RAM and an Intel(R) i7-8665U processor (4 cores, 1.90 GHz).

Figure 5 shows the prediction geometric mean, $\exp(\boldsymbol{\mu}_P)$, for two distinct days, for the entire US domain and for a more detailed 5-county area surrounding Houston, Texas. (Corresponding prediction uncertainties are shown in Figure S4.) For the US domain, predictions were produced on a 200 by 100 grid (10–20 km resolution) across the conterminous United States. For evaluating fine-scale prediction patterns, a 1–2 km grid was produced in the Houston, TX, five-county area. Distinct spatiotemporal patterns emerged. Cities and developed areas, such as roadways, showed elevated NO_2 concentrations, as expected for a traffic-related pollutant. However, there was temporal variability in the spatial patterns around the cities and roads. Comparing the the upper, midwest cities, such as Chicago and Cleveland (blue box, Figure 5), predicted NO_2 was lower on July 11, 2018, than on February 11, 2019. In contrast, in the Houston area subfigure, predicted NO_2 was higher on July 11, 2018, than on February 11, 2019. Visually inspecting the predictors in the final model can reveal the primary drivers of the spatiotemporal variability, which is easy in linear models with interpretable covariates. Complex machine learning models and dimension-reduction techniques do not allow for such intuitive visual comparisons. Figure S5 shows the predictions and select covariates in the Houston area on July 11, 2018, and February 11, 2019. Visual inspection and correlation of covariate-only predictions with the final predictions show that the TROPOMI NO_2 data are driving the patterns observed on July 11, 2018. Contrarily, on February 11, 2019,

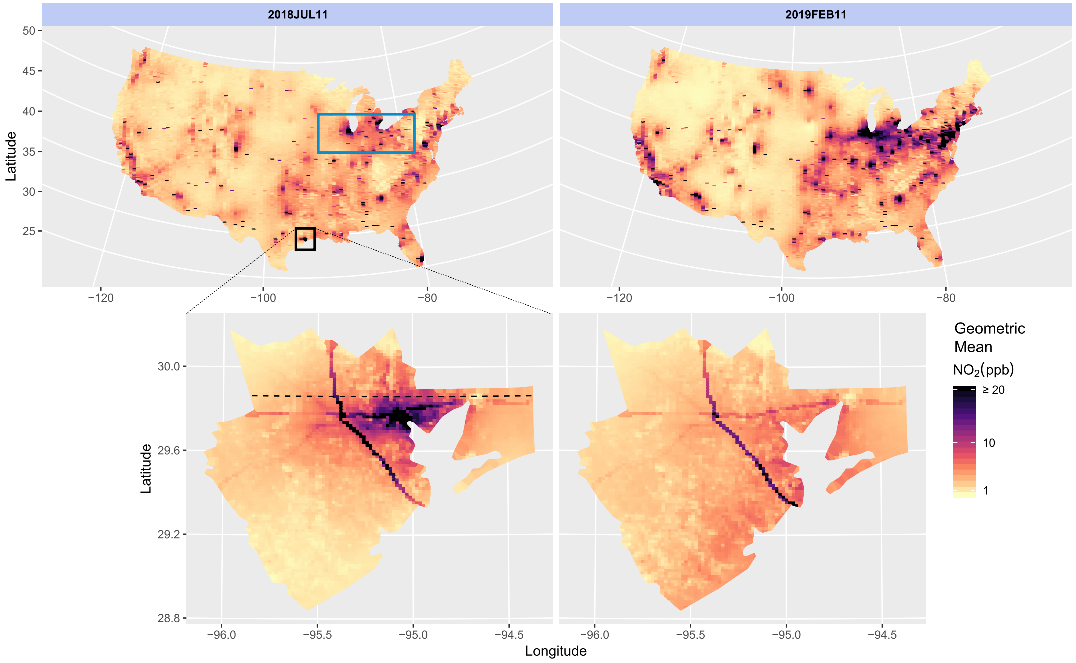

FIG. 5. *LURK-Vecchia prediction geometric mean (exponentiated log-scale predictions) across the conterminous United States (top row) and a five-county subset around Houston, Texas, for July* 11, 2018, *and February* 11, 2019. *The horizontal transect in the Texas subfigure is used in Figure* 6 *to show spatiotemporal patterns in more detail.*

other factors such as the specific humidity display similar general spatial patterns and have high correlation with the final predictions.

Figure 6 shows spatiotemporal predictions in more detail for a transect through the Houston panel of Figure 5. Moving along the transect, we see a general spatial pattern with modulations in time. For instance, the highway consistently had the highest observed concentrations, but the magnitude of the maximum fluctuated daily, driven by time-varying covariates such as the TROPOMI and meteorological variables (see Section 6.3). July 12, 2018, had consistently higher concentrations than other days across most of the transect locations, including the highway. We also showed prediction uncertainties in terms of geometric standard deviations (SDs), $\exp(\mathrm{diag}(\boldsymbol{\Sigma}_P)^{1/2})$. The SD varied over longer time periods than the mean, as evidenced by minor differences within the 2018 and 2019 ranges, but considerable differences between them.

Our predictions will be useful for epidemiological and risk-assessment studies seeking daily, national-scale predictions. For example, Mills et al. (2015) provide meta-analysis results for the impacts of 24-hour NO_2 exposure on all-age-group, all-cause mortality, cardiovascular mortality, respiratory mortality, cardiovascular hospital admissions and respiratory hospital admissions. Our daily NO_2 exposure predictions, combined with population information and with the Mills et al. (2015) relative-risk estimates, may be used to develop NO_2 acute-health impact assessments, such as an attributable-fraction of mortality. Please contact the authors to request predictions at the desired spatiotemporal coordinates.

6.3. *Interpretation of selected covariates.* Table 2 shows the 25 variables selected by our LURK-Vecchia procedure, along with their estimated coefficients. We now discuss interpretations and context for each selected variable, grouped by variable category (see Section 6.3):

- *TROPOMI.* Similar to many LUR studies (Larkin et al. (2017), Novotny et al. (2011), Young et al. (2016), de Hoogh et al. (2018)), we found satellite observations of NO_2 se-

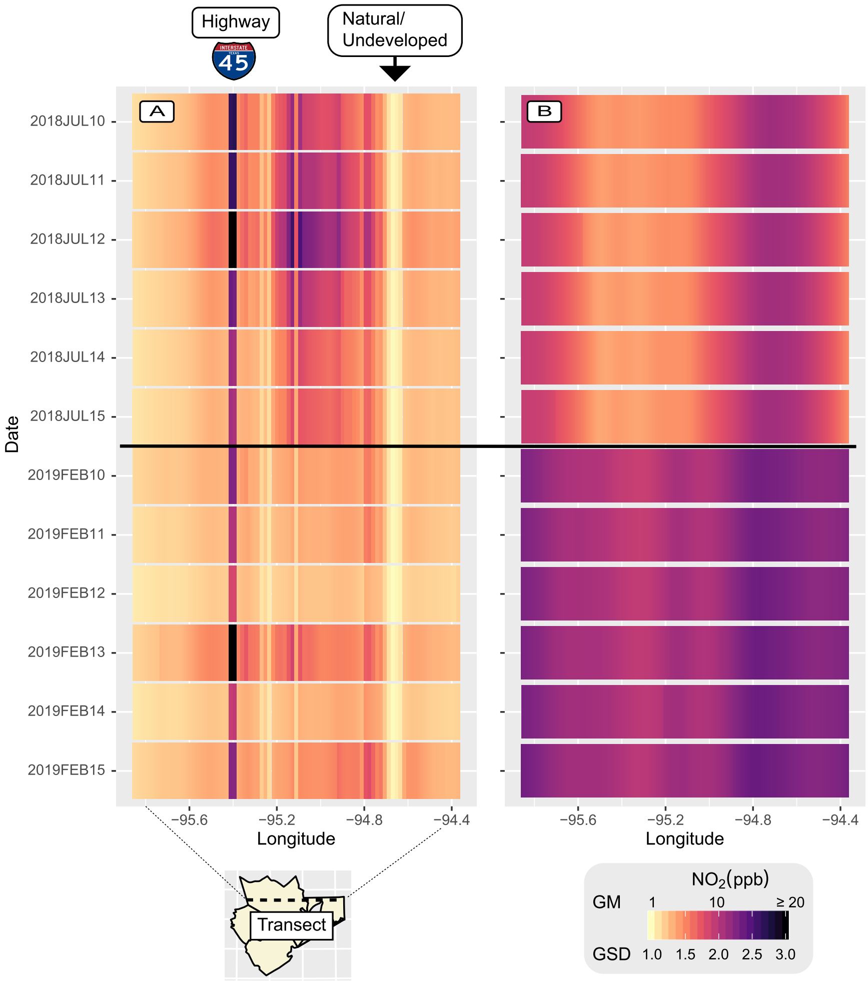

FIG. 6. *For LURK-Vecchia predictions, geometric mean (A) and geometric standard deviation (B) for two sets of dates (y-axis) and along a transect (x-axis) shown in the Houston panel of Figure 5.*

lected to the LURK-Vecchia model. The two covariates for NO_2 Slant combine for a net positive effect on ground-level NO_2 while providing moderation between local and regional scale effects from the 1 and 100 km circular buffer hyperparameters, respectively. Similarly, the tropospheric NO_2 variables have two variables at different spatial scales that contribute a net positive effect to ground-level NO_2.

The TROPOMI variables for tropopause pressure at 1 and 100 km were selected with small negative coefficients, indicating areas of reduced ground-level NO_2. These variables, which have a relatively small impact on the final prediction concentrations, may represent a minor changes in the mixing volume for pollutant molecules.

The dynamic relationship between NO_2 and aerosols is complex and not completely understood. For instance, Grundström et al. (2015) observe weak to moderate correlations between total NO_x ($NO + NO_2$) and particle number concentrations (PNC) depending on meteorological conditions such as wind velocity. Apte et al. (2019) found PNC to have a consistent diurnal pattern of midday new particle formation that is poorly approximated with NO_x. Without a priori expectations of the AAI coefficient, we observe a 2.6 percent decrease in NO_2 concentrations for every one SD increase in AAI. We find a positive coefficient for cloud fraction, which is likely due to the protective effect of clouds on incoming solar radiation.

TABLE 2

Selected covariates for NO_2 *data: rhmax = maximum relative humidity, vs = wind velocity, sph = specific humidity, tmax = maximum temperature. For every* 1*-standard-deviation increase in a covariate, we expect an estimated* $(e^{\hat{\beta}} - 1) \times 100\%$ *increase in* NO_2

Variable category	Variable name	Res. (km)	$\hat{\beta}$	$(e^{\hat{\beta}} - 1) \times 100$
Intercept	–	–	0.002	–
TROPOMI (Section 2.1.1)	NO_2 Slant	1	0.086	8.9
	NO_2 Slant	100	−0.002	−0.21
	NO_2 Tropospheric	10	0.036	3.6
	NO_2 Tropospheric	100	0.044	4.5
	Tropo Pressure	1	−0.003	−0.31
	Tropo Pressure	100	−0.0004	−0.04
	AAI	100	−0.025	−2.6
	cloud fraction	1	0.057	5.9
Meteorology (Section 2.1.2)	rhmax	1	0.12	13.1
	rhmax	100	−0.090	−8.6
	sph	100	−0.27	−23.6
	tmax	100	0.44	55.5
	vs	100	−0.21	−19.2
Vegetation (Section 2.1.3)	NDVI	1	−0.052	−5.1
Population & Roads (Section 2.1.4)	Travel Friction	10	−0.038	−3.7
	Total Road	1	0.18	20.3
Land cover (Section 2.1.5)	Water	1	0.0078	0.79
	Mixed Forest	1	−0.21	−18.3
	Mixed Forest	10	0.017	1.7
	Shrub	1	−0.035	−3.5
	Herbaceous	1	−0.081	−7.8
	Dev Open	10	0.0031	0.31
	Dev Low	10	0.056	5.7
	Elevation	1	−0.28	−24.6
Emissions (Section 2.1.6)	NEI	1	0.0092	0.92

- *Meteorology*. Two relative humidity and one specific humidity variable contribute a net negative effect on NO_2 concentrations. Similar to AAI, we expect that water vapor and aerosols to impede solar radiation and breakdown of NO_2 to NO.

  We observe a 55.5% increase in NO_2 concentrations for every 1 SD increase in the maximum daily temperature. Hot days are associated with increased solar radiation and O_3 formation. The significant increase in NO_2 concentrations is likely capturing O_3 mediated conver3sion of NO ($NO \xrightarrow{O_3} NO_2$) (Seinfeld and Pandis (2016)).

  We observe a 19.2% decrease in NO_2 concentration with a 1 SD increase in wind velocity, which is expected as this increases transport of NO_2 and its precursors from the given location.
- *Vegetation*. For every 1 SD increase in NDVI, we observe a 5.1% decrease in NO_2 concentrations. NDVI represents vegetative greenness, thus this is consistent with the lack of NO_2 or NO_x sources.
- *Land cover*. Open water has a small, positive impact on NO_2, which is likely due to the concentration of cities and sources near water sources and coastlines or as a proxy variable for ports. Mixed forest (the net sum of short and medium ranges), shrub-land, and herbaceous wetlands have negative contributions to NO_2 predictions, which is expected due to

the lack of sources. Developed open and low have positive coefficients, while developed low is larger as it represents an increased anthropogenic presence.

For every 1 SD increase in elevation in 1 km buffer, there is a 24.6% decrease in NO_2 concentrations, which is consistent with other LUR models of NO_2 (de Hoogh et al. (2018)) and can be due to a combination of atmospheric mixing, fewer sources, decreased average temperature and increased wind velocity at higher elevations.

- *Population and Roads*. We find travel friction within a 10 km buffer and total road length within a 1 km buffer to result in a 3.7 decrease and 20.3 increase of NO_2 concentrations for every 1 SD increase, respectively. Travel friction is the average travel time, and total road length is a good approximation of vehicle sources, and so they are expected to decrease and increase traffic-related pollutants, respectively.
- *Emissions*. The NEI variable with a 1 km decay range was selected with a small, positive coefficient. Clearly, we expect a covariate representing source emissions of the dependent variable to be contribute positively.

7. Conclusions. We analyzed daily ground-level NO_2 concentrations across the United States, using a novel penalized land-use regression approach with spatiotemporally correlated errors that is also scalable to large datasets via a sparse general Vecchia approximation. Our methodological advances can be used in future human health exposure and risk assessment to improve model selection and prediction characteristics. Key results from the NO_2 analysis include: the development of daily NO_2 concentration predictions that can be used for epidemiological analyses of acute health effects such as asthma and increased hospitalizations; the potential to develop annual average concentrations that propagate uncertainty from daily predictions as opposed to those based on direct annual averages; the elucidation of spatiotemporal patterns of NO_2 concentrations across the United States, including significant variations between cities and intra-urban variation; and the resolving of a parsimonious group of geographic covariates describing the spatiotemporal distribution of daily NO_2 concentrations, including satellite imagery, meteorological data, land cover, population distributions, road networks and point source emissions.

Our methods also offer a scalable way to analyze other large spatiotemporal datasets in environmental and human health risk assessment. For example, in the air-quality research community, mobile monitoring of air pollutants is leading to high-resolution datasets with millions of observations, including campaigns in Zurich, Switzerland (Li et al. (2012)), Boston, MA (Padró-Martínez et al. (2012)), Oakland, CA (Apte et al. (2017), Guan et al. (2020)), Houston, TX (Miller et al. (2020)) and the Netherlands (Kerckhoffs et al. (2019)).

Our methods could also be extended to non-Gaussian data (Zilber and Katzfuss (2021)) or online spatiotemporal filtering (Jurek and Katzfuss (2018)) using extensions or variations of the general-Vecchia framework.

As mentioned in Section 3, the SCAD penalty used for model selection in Line 7 of Algorithm 1 could be replaced by other penalties, such as LASSO (Tibshirani (1996)), elastic net (Zou and Hastie (2005)), or relaxed LASSO (Hastie, Tibshirani and Tibshirani (2017)), which may result in improvements in prediction accuracy or model selection. While accurate uncertainty quantification and significance assessment is difficult in the context of penalized regression, a potential extension of our approach would be to combine it with existing methods proposed for this purpose (e.g., Chatterjee and Lahiri (2011), Meinshausen, Meier and Bühlmann (2009), Xie et al. (2019)). This would likely come at an increased computational cost, but it would also allow for the inclusion of covariate uncertainty in predictions. Lastly, another possible avenue is to adjust for spatial confounding as proposed in Hughes and Haran (2013).

APPENDIX A: REVIEW OF GENERAL VECCHIA

We now provide some further details of the general Vecchia approximation (Katzfuss and Guinness (2021), Katzfuss et al. (2020a)) that we extended and briefly reviewed in Section 4. Because model (1) implies conditional independence in (6) between z_i and all other variables in $\mathbf{u}$ given y_i, we assume that z_i always conditions on only y_i. Hence, we can write the approximation (6) as

$$\widehat{f}(\mathbf{u}) = \prod_{i=1}^{n} p(z_i|y_i)p(y_i|\mathbf{y}_{q_y(i)}, \mathbf{z}_{q_z(i)}),$$

where $q(i) = q_y(i) \cup q_z(i)$ with $q(i) \subset (1, \dots, i-1)$ is the conditioning index vector of size $|q(i)| \leq m$, and we assume $q_y(i) \cap q_z(i) = \varnothing$. The ordering of the variables and the choice of conditioning sets can have a strong effect on the approximation accuracy and computational speed.

The ordering of the spatiotemporal coordinates $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$ implies an ordering of the variables in $\mathbf{u}$. We assume here that the coordinates are ordered and numbered according to a maximum-minimum distance ordering (Guinness (2018)), which sequentially picks each coordinate in the ordering to maximize the minimum distance to previous coordinate in the ordering. The conditioning index vectors $q(i)$ are chosen here as the indices of the nearest m coordinates previous to i in this ordering. To determine the ordering and the conditioning sets, we use the scaled spatiotemporal distance (2) as our measure of distance (cf. Datta et al. (2016b)). However, this measure of distance depends on the unknown parameters $\boldsymbol{\theta}$ (specifically, on γ_s and γ_t), and so we update the ordering and conditioning at each iteration (in Line 5) of Algorithm 1 based on the current estimate of $\boldsymbol{\theta}$.

Different strategies for splitting $q(i)$ into $q_y(i)$ and $q_z(i)$ can also result in vastly different approximation accuracies. In general, conditioning on y_j is often more accurate but also potentially more computationally expensive than conditioning on z_j. Katzfuss and Guinness (2021) proposed a fast and accurate sparse general Vecchia (SGV) approach that chooses $q_y(i) \subset q(i)$ such that $j < k$ can only both be in $q_y(i)$ if $j \in q_y(k)$, with the remaining conditioning indices in $q(i)$ assigned to $q_t(i) = q(i) \setminus q_y(i)$. Specifically, for $i = 1, \dots, n$, SGV finds $p(i) = \arg\max_{j \in q(i)} |q_y(j) \cap q(i)|$ and $k_i = \arg\min_{\ell \in p(i)} \|\mathbf{s}_i - \mathbf{s}_\ell\|$, and then sets $q_y(i) = (k_i) \cup (q_y(k_i) \cap q(i))$. We use SGV in all our numerical examples.

The restriction of conditioning only on previous variables in the ordering, $q_y(i) \subset (1, \dots, i-1)$, ensures that the implied joint distribution is multivariate normal as indicated in (6). To compute the sparse upper-triangular matrix $\mathbf{U}$, let $g(i)$ denote the vector of indices of the elements in $\mathbf{u}$ on which u_i conditions (e.g., if $u_i = z_k$ then $g(i) = (i-1)$). Also define $K(y_i, y_j) = K(z_i, y_j) = C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ and $K(z_i, z_j) = C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) + \mathbb{1}_{i=j}\tau_i^2$. Then, the (j, i)th element of $\mathbf{U}$ is

$$\mathbf{U}_{ji} = \begin{cases} r_i^{-1/2} & i = j, \\ -b_i^{(j)} r_i^{-1/2} & j \in g(i), \\ 0 & \text{otherwise}, \end{cases}$$

where $\mathbf{b}_i' = K(u_i, \mathbf{u}_{g(i)})K(\mathbf{u}_{g(i)}, \mathbf{u}_{g(i)})^{-1}$, $r_i = K(u_i, u_i) - \mathbf{b}_i' K(\mathbf{u}_{g(i)}, u_i)$, and $b_i^{(j)}$ denotes the ℓth element of $\mathbf{b}_i$ if j is the ℓth element in $g(i)$ (i.e., $b_i^{(j)}$ is the element of $\mathbf{b}_i$ corresponding to u_j).

For prediction of $\mathbf{y}_P$, we employ a spatiotemporal extension of the response-first ordering full-conditioning (RF-full) approach of Katzfuss et al. (2020a), which applies a general Vecchia approximation of the form (6) to $\tilde{\mathbf{u}} = (\mathbf{z}', \mathbf{y}_{\text{all}}')'$, where $\mathbf{y}_{\text{all}} = (\mathbf{y}', \mathbf{y}_P')' =: (y_1, \dots, y_{n_{\text{all}}})'$.

This results in the approximation

$$\widehat{f}(\mathbf{z}, \mathbf{y}_{\text{all}}) = \prod_{i=1}^{n} f(z_i) \times \prod_{i=1}^{n_{\text{all}}} f(y_i | \mathbf{y}_{q_y(i)}, \mathbf{z}_{q_z(i)}),$$

where $\mathbf{y}_{q_y(i)}$ and $\mathbf{z}_{q_z(i)}$ are chosen as the m variables closest in scaled distance (2) to y_i, among those that are previously ordered in $\tilde{\mathbf{u}}$, where we condition on y_j instead of z_j whenever possible. Specifically, we set $q(i)$ to consist of the indices corresponding to the m nearest spatiotemporal coordinates, including i for $i \leq n$, and not including i for $i > n$. Then, for any $j \in q(i)$, we let y_i condition on y_j if it is ordered previously in $\mathbf{u}$, and condition on z_j otherwise. More precisely, we set $q_y(i) = \{j \in q(i) : j < i\}$ and $q_z(i) = \{j \in q(i) : j \geq i\}$.

APPENDIX B: ALTERNATIVE EXPRESSION OF THE OBJECTIVE FUNCTION

We now show that the objective function $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be written as in (8) as

$$\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \|\tilde{\mathbf{z}}_{\boldsymbol{\theta}} - \tilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta}\|_2^2 + \lambda\, \mathrm{p}(\boldsymbol{\beta}) - 2\sum_i \log\bigl((\mathbf{U}_{\boldsymbol{\theta}})_{ii}\bigr) + 2\sum_i \log\bigl((\mathbf{V}_{\boldsymbol{\theta}})_{ii}\bigr).$$

As in Katzfuss and Guinness ((2021), proof of Proposition 2), note that, for any value of $\mathbf{y}$, $\widehat{f}(\mathbf{z}) = \widehat{f}(\mathbf{u})/\widehat{f}(\mathbf{y}|\mathbf{z})$, where $\widehat{f}(\mathbf{u})$ is given in (6), and $\widehat{f}(\mathbf{y}|\mathbf{z}) = \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{W}^{-1})$ with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} - \mathbf{W}^{-1}\mathbf{A}\mathbf{B}'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$. Thus, setting $\mathbf{y} = \boldsymbol{\mu}$, and denoting by $\mathbf{u}_{\boldsymbol{\mu}} = (\mu_1, z_1, \ldots, \mu_n, z_n)$ the resulting vector $\mathbf{u}$, we obtain

$$\begin{aligned}
-2\log \widehat{f}(\mathbf{z}) &= -2\log \mathcal{N}_{2n}\bigl(\mathbf{u}_{\boldsymbol{\mu}} | (\mathbf{X} \otimes \mathbf{1}_2)\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}\bigr) + 2\log \mathcal{N}_n\bigl(\boldsymbol{\mu}|\boldsymbol{\mu}, \mathbf{W}^{-1}\bigr) \\
&= \bigl(\mathbf{u}_{\boldsymbol{\mu}} - (\mathbf{X} \otimes \mathbf{1}_2)\boldsymbol{\beta}\bigr)'\mathbf{U}\mathbf{U}'\bigl(\mathbf{u}_{\boldsymbol{\mu}} - (\mathbf{X} \otimes \mathbf{1}_2)\boldsymbol{\beta}\bigr) - \log\bigl|\mathbf{U}\mathbf{U}'\bigr| + \log|\mathbf{W}| \\
&= \|\mathbf{d}\|_2^2 - 2\sum_i \log \mathbf{U}_{ii} + 2\sum_i \log \mathbf{V}_{ii},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{d} = \mathbf{U}'\mathbf{u}_{\boldsymbol{\mu}} - \mathbf{U}'(\mathbf{X} \otimes \mathbf{1}_2)\boldsymbol{\beta} &= \mathbf{B}'\mathbf{z} + \mathbf{A}'\boldsymbol{\mu} - \bigl(\mathbf{B}'\mathbf{X} + \mathbf{A}'\mathbf{X}\bigr)\boldsymbol{\beta} \\
&= \mathbf{B}'\mathbf{z} - \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{z} - \bigl(\mathbf{B}'\mathbf{X} + \mathbf{A}'\mathbf{X} - \mathbf{A}'\mathbf{X} + \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{X}\bigr)\boldsymbol{\beta} \\
&= \tilde{\mathbf{z}} - \tilde{\mathbf{X}}\boldsymbol{\beta},
\end{aligned}$$

where $\tilde{\mathbf{z}} = \mathbf{B}'\mathbf{z} + \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{z}$ and $\tilde{\mathbf{X}} = \mathbf{B}'\mathbf{X} + \mathbf{A}'\mathbf{W}^{-1}\mathbf{A}\mathbf{B}'\mathbf{X}$.

Acknowledgments. The authors would like to thank Shahzad Gani, Jianhua Huang, Irina Gaynanova, Anirban Bhattacharya and Joe Guinness for helpful comments and suggestions.

Simulations were run on computing resources at the Oregon State University Center for Genome Research and Biocomputing and the NIEHS/NTP Office of Data Science computing cluster.

Funding. Messier's research was partially conducted while at Oregon State University, Department of Environmental and Molecular Toxicology, and supported by NIEHS K99 ES029523. Messier is currently supported by NIH institutes NIEHS/NTP and NIMHD as an intramural investigator. Katzfuss' research was partially supported by National Science Foundation (NSF) Grants DMS-1654083 and DMS-1953005.

SUPPLEMENTARY MATERIAL

Supplementary material to Scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide (DOI: 10.1214/20-AOAS1422SUPPA; .pdf). We provide the results for (1) Simulations with jointly varying covariance parameters; (2) A comparison of the number of nonzero coefficients for a simulation scenario; (3) Prediction uncertainty for the NO_2 application and (4) a comparison of the covariate-only predictions with the full model for select covariates.

Supplementary material: R code (DOI: 10.1214/20-AOAS1422SUPPB; .zip). R code to implement to proposed method and reproduce results.

REFERENCES

ABATZOGLOU, J. T., RUPP, D. E. and MOTE, P. W. (2014). Seasonal climate variability and change in the Pacific Northwest of the United States. *J. Climate* **27** 2125–2142.

ALEXEEFF, S. E., ROY, A., SHAN, J., LIU, X., MESSIER, K., APTE, J. S., PORTIER, C., SIDNEY, S. and VAN DEN EEDEN, S. K. (2018). High-resolution mapping of traffic related air pollution with Google Street View cars and incidence of cardiovascular events within neighborhoods in Oakland, CA. *Environ. Health* **17** 1–13. https://doi.org/10.1186/s12940-018-0382-1

APTE, J. S., MESSIER, K. P., GANI, S., BRAUER, M., KIRCHSTETTER, T. W., LUNDEN, M. M., MARSHALL, J. D., PORTIER, C. J., VERMEULEN, R. C. et al. (2017). High-resolution air pollution mapping with Google street view cars: Exploiting big data. *Environ. Sci. Technol.* **51** 6999–7008.

APTE, J., GANI, S., CHAMBLISS, S., MESSIER, K., LUNDEN, M. et al. (2019). Potential underestimation of ultrafine particle exposure when using proxy pollutants: Lessons from long-term measurements at fixed sites and mobile monitoring. *Environ. Epidemiol.* **3** 13–14.

BANERJEE, M., CAPOZZOLI, M., MCSWEENEY, L. and SINHA, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.* **27** 3–23. MR1703616 https://doi.org/10.2307/3315487

BECKERMAN, B. S., JERRETT, M., SERRE, M. L., MARTIN, R. V., LEE, S., DONKELAAR, A. V., ROSS, Z., SU, J. and BURNETT, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of $PM_{2.5}$ in the contiguous United States. *Environ. Sci. Technol.* **47** 7233–7241. https://doi.org/10.1021/es400039u.A

BOERSMA, K. F., ESKES, H. J., VEEFKIND, J. P., BRINKSMA, E. J., VAN DER A, R. J., SNEEP, M., VAN DEN OORD, G. H. J., LEVELT, P. F., STAMMES, P. et al. (2007). Near-real time retrieval of tropospheric NO_2 from OMI. *Atmos. Chem. Phys.* **7** 2103–2118.

BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396 https://doi.org/10.1214/10-AOAS388

BRIGGS, D. J., COLLINS, S., ELLIOTT, P., FISCHER, P., KINGHAM, S., LEBRET, E., PRYL, K., VAN REEUWIJK, H., SMALLBONE, K. et al. (1997). Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **11** 699–718. https://doi.org/10.1080/136588197242158

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK—CIESIN—COLUMBIA UNIVERSITY (2018). Gridded population of the world, version 4 (GPWv4): Population density, revision 11 [data set]. https://doi.org/10.7927/H49C6VHW

CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974 https://doi.org/10.1198/jasa.2011.tm10159

COULLIETTE, A. D., MONEY, E. S., SERRE, M. L. and NOBLE, R. T. (2009). Space/time analysis of fecal pollution and rainfall in an eastern North Carolina estuary. *Environ. Sci. Technol.* **43** 3728–35.

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. MR3553225 https://doi.org/10.1214/16-AOAS931

DE HOOGH, K., CHEN, J., GULLIVER, J., HOFFMANN, B., HERTEL, O., KETZEL, M., BAUWELINCK, M., VAN DONKELAAR, A., HVIDTFELDT, U. A. et al. (2018). Spatial $PM_{2.5}$, NO_2, O_3, and BC models for Western Europe—evaluation of spatiotemporal stability. *Environ. Int.* **120** 81–92. https://doi.org/10.1016/j.envint.2018.07.036

DE HOOGH, K., SAUCY, A., SHTEIN, A., SCHWARTZ, J., WEST, E. A., STRASSMANN, A., PUHAN, M., ROOSLI, M., STAFOGGIA, M. et al. (2019). Predicting fine-scale daily NO2 for 2005–2016 incorporating OMI satellite data across Switzerland. *Environ. Sci. Technol.* **53** 10279–10287.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166 https://doi.org/10.1214/009053604000000067

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FINLEY, A. O., SANG, H., BANERJEE, S. and GELFAND, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53** 2873–2884. MR2667597 https://doi.org/10.1016/j.csda.2008.09.008

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GAUDERMAN, W. J., AVOL, E., LURMANN, F., KUENZLI, N., GILLILAND, F., PETERS, J. and MCCONNELL, R. (2005). Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology* 737–743.

GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1** 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

GRUNDSTRÖM, M., HAK, C., CHEN, D., HALLQUIST, M. and PLEIJEL, H. (2015). Variation and co-variation of PM_{10}, particle number concentration, NO_x and NO_2 in the urban air-relationships with wind speed, vertical temperature gradient and weather type. *Atmos. Environ.* **120** 317–327.

GUAN, Y., JOHNSON, M. C., KATZFUSS, M., MANNSHARDT, E., MESSIER, K. P., REICH, B. J. and SONG, J. J. (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. *J. Amer. Statist. Assoc.* **115** 1111–1124. MR4143453 https://doi.org/10.1080/01621459.2019.1665526

GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics* **60** 415–429. MR3878098 https://doi.org/10.1080/00401706.2018.1437476

HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint. Available at arXiv:1707.08692.

HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. MR3996451 https://doi.org/10.1007/s13253-018-00348-w

HENDERSON, S. B., BECKERMAN, B., JERRETT, M. and BRAUER, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environ. Sci. Technol.* **41** 2422–2428. https://doi.org/10.1021/es0606780

HOEK, G., BEELEN, R., DE HOOGH, K., VIENNEAU, D., GULLIVER, J., FISCHER, P. and BRIGGS, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42** 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057

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. *Environ. Health* **12** 43.

HOLCOMB, D. A., MESSIER, K. P., SERRE, M. L., ROWNY, J. G. and STEWART, J. R. (2018). Geostatistical prediction of microbial water quality throughout a stream network using meteorology, land cover, and spatiotemporal autocorrelation. *Environ. Sci. Technol.* **52**. https://doi.org/10.1021/acs.est.8b01178

HOMER, C., DEWITZ, J., YANG, L., JIN, S., DANIELSON, P., XIAN, G., COULSTON, J., HEROLD, N., WICKHAM, J. et al. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **81** 345–354.

HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275 https://doi.org/10.1111/j.1467-9868.2012.01041.x

JUREK, M. and KATZFUSS, M. (2018). Multi-resolution filters for massive spatio-temporal data. Available at arXiv:1810.04200.

KATZFUSS, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc.* **112** 201–214. MR3646566 https://doi.org/10.1080/01621459.2015.1123632

KATZFUSS, M. and GONG, W. (2020). A class of multi-resolution approximations for large spatial datasets. *Statist. Sinica* **30** 2203–2226. https://doi.org/10.1007/s13253-020-00401-7

KATZFUSS, M. and GUINNESS, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statist. Sci.* **36** 124–141. MR4194207 https://doi.org/10.1214/19-STS755

KATZFUSS, M., GUINNESS, J., GONG, W. and ZILBER, D. (2020a). Vecchia approximations of Gaussian-process predictions. *J. Agric. Biol. Environ. Stat.* **25** 383–414. MR4139037 https://doi.org/10.1007/s13253-020-00401-7

KATZFUSS, M., JUREK, M., ZILBER, D., GONG, W., GUINNESS, J., ZHANG, J. and SCHAEFER, F. (2020b). GPvecchia: Fast Gaussian-process inference using Vecchia approximations. R package version 0.1.3.

KERCKHOFFS, J., HOEK, G., PORTENGEN, L., BRUNEKREEF, B. and VERMEULEN, R. C. (2019). Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environ. Sci. Technol.* **53** 1413–1421.

KNIBBS, L. D., HEWSON, M. G., BECHLE, M. J., MARSHALL, J. D. and BARNETT, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environ. Res.* **135** 204–211. https://doi.org/10.1016/j.envres.2014.09.011

LARKIN, A., GEDDES, J. A., MARTIN, R. V., XIAO, Q., LIU, Y., MARSHALL, J. D., BRAUER, M. and HYSTAD, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environ. Sci. Technol.* **51** 6957–6964.

LI, R. and SUDJIANTO, A. (2005). Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics* **47** 111–120. MR2188073 https://doi.org/10.1198/004017004000000671

LI, J. J., FALTINGS, B., SAUKH, O., HASENFRATZ, D. and BEUTEL, J. (2012). Sensing the air we breathe—the OpenSense Zurich dataset. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

LIU, H., ONG, Y.-S., SHEN, X. and CAI, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 4405–4423. MR4169962

MAUZERALL, D. L., SULTAN, B., KIM, N. and BRADFORD, D. F. (2005). NOx emissions from large point sources: Variability in ozone production, resulting health damages and economic costs. *Atmos. Environ.* **39** 2851–2866.

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 https://doi.org/10.1198/jasa.2009.tm08647

MESSIER, K. P., AKITA, Y. and SERRE, M. L. (2012). Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **46** 2772–80. https://doi.org/10.1021/es203152a

MESSIER, K. P. and KATZFUSS, M. (2021a). Supplement to scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide. *Ann. Appl. Stat.* https://doi.org/10.1214/20-AOAS1422SUPPA

MESSIER, K. P. and KATZFUSS, M. (2021b). R-code for scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide. *Ann. Appl. Stat.* https://doi.org/10.1214/20-AOAS1422SUPPB

MESSIER, K. P., KANE, E., BOLICH, R. and SERRE, M. L. (2014). Nitrate variability in groundwater of North Carolina using monitoring and private well data models. *Environ. Sci. Technol.* **48**. https://doi.org/10.1021/es502725f

MESSIER, K. P., CAMPBELL, T., BRADLEY, P. J. and SERRE, M. L. (2015). Estimation of groundwater Radon in North Carolina using land use regression and Bayesian maximum entropy. *Environ. Sci. Technol.* **49** 9817–9825. https://doi.org/10.1021/acs.est.5b01503

MILLER, D. J., ACTKINSON, B., PADILLA, L., GRIFFIN, R. J., MOORE, K., LEWIS, P. G. T., GARDNER-FROLICK, R., CRAFT, E., PORTIER, C. J. et al. (2020). Characterizing elevated urban air pollutant spatial patterns with mobile monitoring in Houston, Texas. *Environ. Sci. Technol.*.

MILLS, I. C., ATKINSON, R. W., KANG, S., WALTON, H. and ANDERSON, H. R. (2015). Quantitative systematic review of the associations between short-term exposure to nitrogen dioxide and mortality and hospital admissions. *BMJ Open* **5** e006946. https://doi.org/10.1136/bmjopen-2014-006946

MOORE, D. K., JERRETT, M., MACK, W. J. and KÜNZLI, N. (2007). A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *J. Environ. Monit.* **9** 246–252. https://doi.org/10.1039/b615795e

NASA/METI/AIST/JAPAN SPACESYSTEMS and U.S./JAPAN ASTER SCIENCE TEAM (2019). ASTER global digital elevation model V003 [data set]. https://doi.org/10.5067/ASTER/ASTGTM.003

NOVOTNY, E. V., BECHLE, M. J., MILLET, D. B. and MARSHALL, J. D. (2011). National satellite-based land-use regression: NO_2 in the United States. *Environ. Sci. Technol.* **45** 4407–4414.

PADRÓ-MARTÍNEZ, L. T., PATTON, A. P., TRULL, J. B., ZAMORE, W., BRUGGE, D. and DURANT, J. L. (2012). Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmos. Environ.* **61** 253–264.

REYES, J. M. and SERRE, M. L. (2014). An LUR/BME framework to estimate $PM_{2.5}$ explained by on road, mobile and stationary sources. *Environ. Sci. Technol.* **48** 1736–44. https://doi.org/10.1021/es4040528

ROSENLUND, M., BERGLIND, N., PERSHAGEN, G., HALLQVIST, J., JONSON, T. and BELLANDER, T. (2006). Long-term exposure to urban air pollution and myocardial infarction. *Epidemiology* 383–390.

ROSENLUND, M., PICCIOTTO, S., FORASTIERE, F., STAFOGGIA, M. and PERUCCI, C. A. (2008). Traffic-related air pollution in relation to incidence and prognosis of coronary heart disease. *Epidemiology* 121–128.

ROSENLUND, M., BELLANDER, T., NORDQUIST, T. and ALFREDSSON, L. (2009). Traffic-generated air pollution and myocardial infarction. *Epidemiology* 265–271.

ROSS, Z., ITO, K., JOHNSON, S., YEE, M., PEZESHKI, G., CLOUGHERTY, J. E., SAVITZ, D. and MATTE, T. (2013). Spatial and temporal estimation of air pollutants in New York City: Exposure assignment for use in a birth outcomes study. *Environ. Health* **12** 51.

SAMPSON, P. D., RICHARDS, M., SZPIRO, A. A., BERGEN, S., SHEPPARD, L., LARSON, T. V. and KAUFMAN, J. D. (2013). A regionalized national universal kriging model using partial least squares regression for estimating annual $PM_{2.5}$ concentrations in epidemiology. *Atmos. Environ.* **75** 383–392. https://doi.org/10.1016/j.atmosenv.2013.04.015

SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. MR2907125 https://doi.org/10.1214/11-AOAS478

SCHÄFER, F., KATZFUSS, M. and OWHADI, H. (2020). Sparse Cholesky factorization by Kullback–Leibler minimization. Available at arXiv:2004.14455.

SCHINDLER, D. W. (1988). Effects of acid rain on freshwater ecosystems. *Science* **239** 149–157.

SEINFELD, J. H. and PANDIS, S. N. (2016). *Atmospheric Chemistry and Physics*: *From Air Pollution to Climate Change*. Wiley, New York.

SNELSON, E. and GHAHRAMANI, Z. (2007). Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics* 11 (*AISTATS*).

SU, J. G., JERRETT, M. and BECKERMAN, B. (2009). A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Sci. Total Environ.* **407** 3890–3898. https://doi.org/10.1016/j.scitotenv.2009.01.061

TADONO, T., ISHIDA, H., ODA, F., NAITO, S., MINAKAWA, K. and IWAMOTO, H. (2014). Precise global DEM generation by ALOS PRISM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2** 71.

TANG, W., ZHANG, L. and BANERJEE, S. (2019). On identifiability and consistency of the nugget in Gaussian spatial process models. Available at arXiv:1908.05726.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

US ENVIRONMENTAL PROTECTION AGENCY (1999). U.S. EPA technical bulletin: Nitrogen oxides (NO_x), why and how they are controlled.

US ENVIRONMENTAL PROTECTION AGENCY (2016). Integrated science assessment for oxides of nitrogen (final report). Technical report, EPA/600/R-15/068, US Environmental Protection Agency, National Center for Environmental Assessment Research, Research Triangle Park, NC.

US ENVIRONMENTAL PROTECTION AGENCY (2017). 2017 national emissions inventory.

US ENVIRONMENTAL PROTECTION AGENCY (2019). Air quality system pre-generated data files. Available at https://www.epa.gov/outdoor-air-quality-data/download-daily-data.

VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. MR0964183

VOLK, H. E., LURMANN, F., PENFOLD, B., HERTZ-PICCIOTTO, I. and MCCONNELL, R. (2013). Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatr.* **70** 71–77. https://doi.org/10.1001/jamapsychiatry.2013.266

WEISS, D. J., NELSON, A., GIBSON, H. S., TEMPERLEY, W., PEEDELL, S., LIEBER, A., HANCHER, M., POYART, E., BELCHIOR, S. et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553** 333.

WU, H., WANG, C. and WU, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14** 232–243. https://doi.org/10.1093/biostatistics/kxs033

XIE, Y., XU, L., DENG, X., HONG, Y., KOLIVRAS, K. and GAINES, D. N. (2019). Spatial variable selection and an application to Virginia Lyme disease emergence. *J. Amer. Statist. Assoc.* **114** 1466–1480. MR4047274 https://doi.org/10.1080/01621459.2018.1564670

XU, X., HA, S. U. and BASNET, R. (2016). A review of epidemiological research on adverse neurological effects of exposure to ambient air pollution. *Front Public Health* **4** 157. https://doi.org/10.3389/fpubh.2016.00157

YOUNG, M. T., BECHLE, M. J., SAMPSON, P. D., SZPIRO, A. A., MARSHALL, J. D., SHEPPARD, L. and KAUFMAN, J. D. (2016). Satellite-based NO_2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environ. Sci. Technol.* **50** 3686–3694. https://doi.org/10.1021/acs.est.5b05099

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

ZILBER, D. and KATZFUSS, M. (2021). Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. *Comput. Statist. Data Anal.* **153** 107081. MR4146817 https://doi.org/10.1016/j.csda.2020.107081

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x