Multivariate spatial prediction of air pollutant concentrations with INLA

Wenlong Gong, Brian J Reich and Howard H Chang

1 North Carolina State University, United States of America
2 Emory University, United States of America

E-mail: wenlonggong@gmail.com

Keywords: air pollution, CMAQ, geostatistical model, INLA, linear model of coregionalization

Abstract

Estimates of daily air pollution concentrations with complete spatial and temporal coverage are important for supporting epidemiologic studies and health impact assessments. While numerous approaches have been developed for modeling air pollution, they typically only consider each pollutant separately. We describe a spatial multipollutant data fusion model that combines monitoring measurements and chemical transport model simulations that leverages dependence between pollutants to improve spatial prediction. For the contiguous United States, we created a data product of daily concentration for 12 pollutants (CO, NOx, NO2, SO2, O3, PM10, and PM2.5 species EC, OC, NO3, NH4, SO4) during the period 2005 to 2014. Out-of-sample prediction showed good performance, particularly for daily PM2.5 species EC ($R^2 = 0.64$), OC ($R^2 = 0.75$), NH4 ($R^2 = 0.84$), NO3 ($R^2 = 0.73$), and SO4 ($R^2 = 0.80$). By employing the integrated nested Laplace approximation (INLA) for Bayesian inference, our approach also provides model-based prediction error estimates. The daily data product at 12 km spatial resolution will be publicly available immediately upon publication. To our knowledge this is the first publicly available data product for major PM2.5 species and several gases at this spatial and temporal resolution.

1. Introduction

Ambient air pollution measurements from regulatory monitoring networks are routinely used to support large population-based epidemiologic studies globally (Dai et al 2014, Chen et al 2018, Vicedo-Cabrera et al 2020). However, regulatory monitors are spatially sparse, temporally incomplete, and preferentially located in areas with large populations. There has been considerable effort to estimate air quality by supplementing monitoring measurements with additional data sources, such as numerical model simulations, dispersion model outputs, satellite imagery, portable monitors and low-cost sensors (Özkaynak et al 2013, Chu et al 2016, Yu et al 2018, Morawaska et al 2018). These data products, often with enhanced spatial-temporal coverage and resolution, are increasingly being used to estimate exposures for health effect and health impact analyses (McGuinn et al 2017, Jin et al 2019).

The main objective of this study is to create a publicly available data product of daily pollutant concentrations over the contiguous United States for the period 2005 to 2014. This is accomplished by combining measurements from monitoring networks and outputs from the Community Multiscale Air Quality (CMAQ) model (Binkowski et al 2003, Appel et al 2007, Appel et al 2008). CMAQ is a 3-dimensional chemical transport model (CTM) that simulates pollution concentrations using emission inventory data, meteorology, atmospheric chemistry and physical principles. CMAQ outputs have been used extensively in previous applications to estimate pollutant concentrations in the US (Berrocal et al 2012, Friberg et al 2016, Senthilkumar et al 2019) and China (Liang et al 2017, Lyu et al 2019). One key advantage of using CTM outputs for modeling air quality is its complete spatial-temporal coverage, which has also been leveraged to address missing data in remotely-sensing aerosol optical depth when estimating air quality (Xiao et al 2017, Shaddick et al 2018).
Our approach to utilizing CTM outputs have several unique features. First, CTM simulations are conducted under a multipollutant framework because concentrations of pollutants are often correlated due to shared emission sources and meteorological drivers. Exploiting cross-pollutant dependence may lead to improved prediction performance, as demonstrated by Berrocal et al (2010) in a bi-pollutant model for CMAQ PM$_{2.5}$ and ozone, and by Crooks et al (2014) that imposed a sum-constraint for modeling PM$_{2.5}$ species. Here we take a joint modeling approach that utilizes multipollutant CMAQ outputs as predictors of each other and incorporates spatial cross-pollutant correlation in monitoring measurements. We also expand on previous joint modeling work by considering a larger set of 12 pollutants, including fine and coarse particulate matter (PM$_{2.5}$ and PM$_{10}$), carbon monoxide (CO), nitrogen dioxide (NO$_2$), nitrogen oxides (NOx), sulfur dioxide (SO$_2$), ozone (O$_3$), and 5 major PM$_{2.5}$ species: including nitrate (NO$_3$), ammonium (NH$_4$), sulfate (SO$_4$), elemental carbon (EC) and organic carbon (OC).

A second feature of our approach is the use of Bayesian geostatistical models that allow for complex multipollutant dependence based on spatial locations. Geostatistical models are widely used for modeling air pollution levels (Bergen et al 2013, Hu et al 2013, Keller et al 2015, Berrocal et al 2020), and studies have shown that considering spatial autocorrelation can be particularly beneficial for large study regions. The Bayesian geostatistical approach also provides straight-forward uncertainty quantification for each prediction (e.g., prediction standard error) that can be difficult to obtain from machine learning algorithms. Prediction uncertainty can be incorporated as exposure measurement error in subsequent health impact assessments or health effect analyses (Rushworth et al 2014, Blangiardo et al 2016, Butland et al 2019). Consideration for prediction uncertainty may be particularly important when the modeling domain includes regions with sparse monitors. Finally, to address computational burden typically associated with Bayesian computation, we also describe how to employ integrated nested Laplace approximation (INLA) for model fitting and prediction with large spatial-temporal datasets.

2. Air quality data

Daily air pollution concentration measurements are acquired from the US Environmental Protection Agency (EPA) covering the continental US for the period 2005–2014. We include 7 particulate pollutant (PM$_{2.5}$, PM$_{10}$, NO$_2$, NH$_4^+$, EC, OC, and SO$_4^{2-}$), and 5 gases (CO, NOx, NO$_2$, SO$_2$ and O$_3$). Following daily air quality metrics most often used in health studies, particulate pollutants are recorded as daily averages, ozone is recorded as the daily maximum 8 h averages, and other gases are recorded as the maximum 1 h average. The daily AQS dataset (US EPA, 2017) includes a total of 2,644 stations measuring at least one pollutant, as shown in figure 1. Some pollutants are available every day, while others are only measured every 3 or 6 days. The pollutant measurement time and unites are shown in table A1 in supplemental materials. For each pollutant, numerical model outputs from CMAQ version 5.0.2 are acquired from the US EPA. CMAQ simulations were run at a 12 km resolution with $299 \times 459$ grid cells over the contiguous US. Hourly CMAQ concentration simulations were used to calculate daily metrics. For developing the prediction model, each AQS monitor is linked to a CMAQ grid cell. The sample correlation between CMAQ and AQS for the same pollutant ranges from 0.12 for PM$_{10}$ to 0.74 for O$_3$.

Table 1 provides the number of observations for each pollutant and the number of station-days where each pair of pollutants are measured at the same location and same date; additional summary statistics are provided in Senthilkumar et al (2019). The most commonly measured pollutant across days and locations is O$_3$, and the least commonly measured pollutants are the PM$_{2.5}$ species EC and OC. There are strong correlations between several...
pairs of pollutants that motivate using a multivariate model for prediction. For example, more intensely-sampled NO$_3$ and sparsely-sampled EC have sample correlation 0.79 at collocated monitoring sites.

### 3. Multi-pollutant geostatistical models

For each of the $p = 12$ pollutants, let $Y_k(s, t)$ be the AQS measurement of pollutant $k = 1, \ldots, p$ at spatial location $s$ and day $t$. To avoid prediction of negative concentrations, we transform the responses as $\log(1 + Y_k(s, t))$. Further, to put all pollutants on the same scale for multivariate modeling, we standardize across $s$ and $t$ so that each pollutant has mean zero and variance one. Denote $y_k(s, t)$ as the log-transformed and standardized pollution concentration. Similarly, denote $x_k(s, t)$ as the log-transformed and standardized of value of CMAQ for pollutant $k$ for the grid cell containing location $s$ on day $t$.

A hierarchical framework is used to jointly model the spatial variation of multiple air pollutants while accounting for bias and error in the deterministic CMAQ model outputs. For each pollutant interest, we assume

$$y_k(s, t) = \beta_{0k} + \sum_{l=1}^{p} x_l(s, t) \cdot \beta_{lk}(t) + \omega_k(s, t) + \epsilon_k(s, t),$$

where $\beta_{0k}$ is the intercept for pollutant $k$, $\beta_{lk}(t)$ is the association between pollutant $l$ from CMAQ and the pollutant of interest, $\omega_k(s, t)$ is the spatial random effect for pollutant $k$ assumed to be independent over time after accounting for the CMAQ covariates, and the residual error component $\epsilon_k(s, t) \sim N(0, \sigma^2_k)$ are independent over space and time. In the above model, we borrow information across pollutants by imposing between-pollutant dependence in the residual spatial random effects $\omega_k(s, t)$.

The regression coefficients for pollutant $k$ are allowed to vary over time via a random slopes model

$$\beta_k(t) = (\beta_{10k}(t), \beta_{2k}(t), \ldots, \beta_{pk}(t))^T \sim Normal(\mu_k, \Sigma_k),$$

where $\mu_k$ and $\Sigma_k$ are the random effect mean vector and covariance matrix, respectively, and the slopes are independent over time. The multivariate spatial random effects $\omega_k(s, t)$ explains the residual correlation not addressed by CMAQ and is modelled using the Linear Model of Coregionalization (LMC; Gelfand et al 2004). The LMC captures both spatial correlation between nearby locations and cross-correlation between the $p$ pollutants. The LMC is a spatial extension of factor analysis (Everitt and Hothorn, 2011) with the random effect for each pollution modeled as a linear combination of spatially-varying factors. That is,

$$\omega_k(s, t) = \sum_{j=1}^{p} L_{kj} z_j(s, t)$$

where $L_{kj}$ are the factor loadings and $z_j(s, t)$ are the spatial factors (independent over $j$ and $t$). The spatial factors $z_j(s, t)$ are modeled as Gaussian processes with mean zero, variance one, and an isotropic Matérs spatial correlation function (Stein, 2012). The factor loadings determine the cross-covariance with

$$\text{Cov} \{ \omega_k(s, t), \omega_l(s, t) \} = \sum_{j=1}^{p} L_{kj} L_{lj}.$$
4. Estimation procedure

Fitting the model described in section 3 using all of the data pooled over days, sites and pollutants is computationally infeasible. The analysis thus proceeds in two stages: we first estimate the random slopes \( \beta_k(t) \) using a non-spatial method and then analyze the residuals using a multivariate spatial model. Predictions from these two stages are then combined to give the final predictions.

Estimation of the random slopes begins by performing a least squares regression separately for each day and each pollutant. For pollutant \( k \) on day \( t \), let \( \hat{\beta}_k(t) = \{ \hat{\beta}_{k1}(t), \hat{\beta}_{k2}(t), \ldots, \hat{\beta}_{kp}(t) \} \) be the least squares estimate of \( \beta_k(t) \) and the associated \( \text{covariance matrix} \) (of the sampling distribution) be \( V_k(t) \). Using these estimates we then fit the Bayesian random effect model

\[
\hat{\beta}_k(t) | \beta_k(t) \sim MVN(\beta_k(t), V_k(t)) \quad \text{and} \quad \beta_k(t) \sim MVN(\mu_k, \Sigma_k)
\]

with uninformative prior distributions \( \mu_k \sim \text{Normal}(0, 100I_{p+1}) \) and \( \Sigma_k \sim \text{InvWishart}(p + 2, I_{p+1}) \). This stage is fit separately by pollutant using JAGS (Plummer 2003) with 10,000 Markov Chain Monte Carlo (MCMC) iterations after a burn-in of 1,000 iterations. Let \( \hat{\beta}_k(t) \) denote the mean from the posterior samples. Since these regression coefficients are estimated without accounting for correlation between observations, they are suboptimal estimates of covariate effects and their posterior variances likely underestimate their true uncertainty. However, since our main goal is spatial prediction and not inference on the effects of the CMAQ predictors, this non-spatial first step is a reasonable computational compromise.

In the second stage, we fit a multivariate spatial model in equation (1) separately by day with \( \beta_k(t) \) fixed at its first-stage estimate, \( \hat{\beta}_k(t) \). Even analyzing data from a single day is difficult because of the large number of pollutants and spatial locations. Therefore, rather than using MCMC, we use the Integrated Nested Laplace Approximation (INLA; Rue et al. 2009) with the Stochastic Partial Differential equation (SPDE; Lindgren et al. 2011) approximation, which aims to approximate the marginal distributions of the parameters by exploiting Gaussian Markov Random Field (GMRF) property with finite element methods and Laplace approximation for multi-dimensional integration (Blangiardo and Cameletti 2015). When defining the SPDE in INLA, we used a penalized complexity prior (PC prior; Fuglstad et al. 2019) for the spatial parameters. The PC prior is weakly informative and penalizes complexity by shrinking the Matérn spatial range parameter toward infinity, and the marginal standard deviation and error precision toward zero. The smoothness parameter is fixed at 1. The prior distributions for the coefficients \( \lambda_j \)’s are N(0, 0.1). The model is fit using the R-INLA package (Lindgren and Rue, 2015).

As mentioned in section 3, the temporal trend is captured by the random slope model in the first stage. The final prediction will use its first-stage posterior estimate, \( \hat{\beta}_k(t) \) with CMAQ covariates and the multivariate spatial prediction results. The spatial prediction is performed in INLA at the mesh locations (constructed by Delaunay triangles) and then projected on a set of dense regular prediction grids or nodes covering the US domain. As it is quite computationally expensive for dense mesh over a large domain, we employ a moderate number of mesh nodes (varying depending on the number of observed locations for all pollutants in a day) to achieve computation efficiency. Figure 2 shows the comparison of observed AQS locations for PM2.5 species in

Figure 2. Mesh locations for INLA and AQS observed locations for a particular day is in red for EC, OC, NH4, NO3, SO4.
2014 denoted in red and the mesh locations to fit in the multivariate INLA model. The predictions on mesh locations within the US are then interpolated to the 12 km resolution of CMAQ.

Exploratory analysis suggests that many of the pollutants are uncorrelated after accounting for the CMAQ outputs in the mean structure. Therefore, as a final computational simplification, we fit the multivariate spatial model separately to correlated groups of pollutants. Since not all monitors are collocated, to determine which pollutants are strongly correlated, we first calculate the cross-covariogram (CCG; Nychka et al. 2017) for each pair of pollutants using the residuals after removing the first-stage CMAQ mean estimates. The strength of association between each pair of pollutants is measured using the CCG estimate (averaged over days) at distance 0.5 degrees (about 50 kilometers). The number of pollutants to be included in a multivariate model is determined by cross-validation; the remaining pollutants are analyzed using a univariate spatial model in INLA. In the cross-validation study, data with all 12 pollutants measured on the same day in the first 9 years are used to calculate and rank the averaged pairwise CCG. We use the final year of data (all days included) to select models. The training set consists 90% of the data which is randomly separated by day across sites and pollutants and we use it to fit the multivariate models and the remaining 10% treated as the testing set estimates root mean square error (RMSE) and R-square between predicted and the true values. The comparison is summarized in section 5.2. And figure 3 shows a flowchart of the modeling and estimation procedures with the final selected model.

5. Results

5.1. Associations between CMAQ Outputs and AQS Measurements
Table 2 presents mean regression coefficient \( \hat{\mu}_k \) in equation (2) as a measure of the impact of CMAQ output for different pollutants on predicting a specific pollutant concentration. For most pollutants, their corresponding CMAQ output is the strongest predictor. For example, for O3, the CMAQ O3 output has average coefficient 0.75 while the other 11 CMAQ pollutants have average coefficients less than 0.29. However, many of the cross-pollutant relationships are statistically significant with the 95% interval of \( \hat{\mu}_k \) excluding zero. To support this conclusion, we also considered the simple linear regression model with only the corresponding CMAQ pollutant and found the multivariate model with all CMAQ pollutant predictors gave better fit (table A2 in supplemental materials).
| Outcome | CO         | EC         | NH₄        | NO₂        | NO₃        | NOₓ        | O₃         | OC         | PM₁₀       | PM₂.₅      | SO₂        | SO₄        |
|---------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| CO      | 0.23(0.00) | 0.17(0.01) | −0.11(0.01)| 0.14(0.01)| 0.02(0.00)| −0.06(0.01)| 0.09(0.00)| 0.03(0.01)| 0.03(0.00)| −0.01(0.01)| 0.01(0.00)| −0.09(0.01)|
| EC      | 0.09(0.01) | 0.43(0.01)| 0.00(0.02)| 0.20(0.02)| 0.06(0.01)| −0.20(0.02)| 0.13(0.01)| 0.05(0.01)| 0.20(0.01)| −0.23(0.02)| 0.07(0.01)| 0.06(0.01)|
| NH₄     | −0.03(0.01)| 0.02(0.01)| 0.43(0.02)| 0.09(0.01)| 0.14(0.01)| −0.04(0.02)| 0.07(0.01)| −0.20(0.01)| 0.10(0.01)| 0.14(0.02)| −0.01(0.00)| 0.08(0.02)|
| NO₂     | 0.06(0.00)| 0.03(0.00)| −0.03(0.00)| 0.51(0.00)| −0.02(0.00)| 0.09(0.01)| 0.11(0.00)| 0.00(0.00)| −0.01(0.00)| 0.17(0.01)| 0.00(0.00)| −0.11(0.00)|
| NO₃     | −0.09(0.01)| 0.00(0.01)| 0.05(0.02)| 0.02(0.01)| 0.56(0.01)| 0.15(0.01)| −0.02(0.01)| −0.14(0.01)| 0.23(0.01)| −0.04(0.02)| −0.08(0.00)| −0.08(0.01)|
| NOₓ     | 0.09(0.00)| 0.01(0.00)| −0.06(0.00)| 0.25(0.00)| 0.05(0.00)| 0.27(0.01)| 0.09(0.00)| 0.01(0.00)| 0.00(0.00)| 0.13(0.01)| 0.05(0.00)| −0.13(0.00)|
| O₃      | 0.01(0.00)| 0.19(0.00)| 0.11(0.01)| −0.02(0.00)| −0.10(0.00)| −0.08(0.01)| 0.75(0.01)| −0.10(0.00)| −0.08(0.00)| 0.29(0.01)| 0.00(0.00)| −0.19(0.01)|
| OC      | 0.02(0.01)| 0.2(0.01)| 0.17(0.02)| 0.26(0.02)| −0.08(0.01)| −0.33(0.02)| 0.19(0.01)| 0.42(0.02)| 0.20(0.02)| −0.13(0.03)| −0.05(0.01)| 0.01(0.02)|
| PM₁₀    | −0.12(0.00)| 0.29(0.01)| −0.10(0.01)| 0.07(0.01)| −0.03(0.01)| 0.00(0.01)| 0.31(0.00)| 0.00(0.01)| 0.13(0.01)| 0.11(0.01)| −0.11(0.00)| 0.10(0.01)|
| PM₂.₅   | −0.04(0.00)| 0.08(0.01)| 0.18(0.01)| 0.11(0.01)| −0.02(0.01)| −0.14(0.01)| 0.12(0.00)| −0.04(0.01)| 0.05(0.01)| 0.37(0.01)| −0.06(0.00)| 0.14(0.01)|
| SO₂     | −0.04(0.00)| −0.07(0.00)| −0.11(0.01)| 0.14(0.00)| 0.04(0.00)| −0.16(0.01)| 0.02(0.00)| −0.01(0.00)| −0.21(0.00)| 0.26(0.01)| 0.50(0.00)| 0.09(0.00)|
| SO₄     | 0.05(0.01)| 0.05(0.01)| 0.15(0.01)| 0.15(0.01)| 0.05(0.01)| −0.25(0.01)| 0.08(0.01)| −0.10(0.01)| 0.00(0.01)| −0.02(0.02)| 0.03(0.00)| 0.64(0.01)|

Table 2. Posterior mean (posterior standard deviation) of the average regression coefficient $\mu$ for each CMAQ predictor (column) and each response (row). All variables are standardized so the coefficients are unitless and the largest CMAQ output coefficients for each pollutant are marked in bold.
5.2. Model selection
Table 3 shows the prediction RMSE and R^2 for the multivariate models with 4 pollutants. Model selection is based on the lower prediction error and higher R^2 than other models; these three pollutants show clear spatial patterns and are selected as our final model for multivariate INLA fitting. Results for multivariate models with 7 or more pollutants are not shown in the table due to poor performance with over-parameterization.

5.3. Multi-pollutant cross-correlations
Relationships between pollutants not explained by CMAQ can be captured in the residual correlation matrix, \( \Omega = \text{Cor}(\omega_1(s), \ldots, \omega_p(s)) \), derived from the covariance in equation (4) (where \( p = 5 \) with the chosen model in section 4). The cross-correlation is allowed to vary by day, and table 4 shows the mean (standard deviation) across days. The strongest residual cross-correlations are between EC/OC, NH₄/NO₃ and NH₄/SO₄, which is consistent with the ranking of CCG in section 4.

5.4. Spatial predictions and uncertainty quantification
As described in section 4.1, predictions are made at mesh locations and then interpolated to CMAQ grid which is of 12 km resolution for all pollutants. The time series plots in figure A1 show that all pollutants have a seasonal periodical trend. For example, EC and NO₃ tend to be higher in winter and lower in summer, while others behave the opposite. Besides the seasonality, most of the pollutant concentrations have decreased over the years, except that OC remains at a flat level. Figure 4 gives season-specific maps (median concentrations) and corresponding standard error for EC, OC and NO₃ in 2014. These three pollutants show clear spatial patterns with higher levels in the central valley of California and Midwest and Eastern regions of the US. As a primary pollutant, EC shows less spatial smoothness compared to OC and NO₃. The high level of NO₃ in the Midwest during winter has been attributed to high emission of ammonia associated with animal agriculture (Pitchford et al 2009). The prediction standard error is plotted for comparison across seasons and their variation is mainly driven by the number of AQS observations associated with the model. For example, NO₃ is measured slightly more sparsely in winter and spring, which leads to an elevated uncertainty for the prediction. The regions with

| Pollutants | UV | MV-6 | MV-5 | MV-4 |
|------------|----|------|------|------|
|            | RMSE | R^2 | RMSE | R^2 | RMSE | R^2 | RMSE | R^2 |
| EC         | 0.623 | 0.555 | 0.537 | 0.643 | 0.549 | 0.639 | 0.532 | 0.645 |
| OC         | 0.632 | 0.603 | 0.450 | 0.711 | 0.416 | 0.746 | 0.329 | 0.741 |
| NH₄        | 0.431 | 0.673 | 0.366 | 0.791 | 0.302 | 0.834 | 0.346 | 0.794 |
| NO₃        | 0.388 | 0.619 | 0.356 | 0.678 | 0.330 | 0.735 | 0.416 | 0.730 |
| SO₂        | 0.420 | 0.738 | 0.421 | 0.748 | 0.367 | 0.798 |         |      |
| CO         | 0.619 | 0.429 | 0.648 | 0.397 |         |      |         |      |
| PM₁₀       | 0.629 | 0.495 |         |      |         |      |         |      |
| PM₂.₅      | 0.493 | 0.670 |         |      |         |      |         |      |
| NO₂        | 0.594 | 0.566 |         |      |         |      |         |      |
| NOₓ        | 0.621 | 0.456 |         |      |         |      |         |      |
| SO₄        | 0.792 | 0.268 |         |      |         |      |         |      |
| O₃         | 0.373 | 0.773 |         |      |         |      |         |      |

Table 4. Estimated residual spatial correlation between pollutants (i.e., \( \text{cor}(\omega_i(s, t), \omega_j(s, t)) \)) for the multivariate spatial model with 5 pollutants averaged over days in 2014.
fewer monitors in the west and mountainous areas also have higher standard prediction error. Similar season concentration maps for NH₄ and SO₄ are given in supplementary materials figure A2.

Figure 5 compares seasonal predictions of NO₃ and SO₄ from the multi-pollutant model with raw CMAQ outputs, showing that predictions from the multivariate model are slightly higher overall. while EC and NH₄ are very similar with CMAQ, NO₃ predictions in the Fall from multivariate models show less variation compared to CMAQ; Spring and Summer predictions are slightly higher overall. The comparisons for other pollutants are included in appendix figure A3.

6. Discussion

Our prediction accuracy compares favorably to previous efforts to map air pollution. Senthilkumar et al (2019) also conducted national data fusions of CMAQ outputs and AQS monitor measurements but using one pollutant at a time. While it is difficult to compare prediction performance metrics due to different
cross-validation designs, there is evidence that our multipollutant approach results in improved performance, especially for pollutants that have sparse monitoring networks when we are able to leverage cross-pollutant spatial dependence. For example, in Senthilkumar et al. (2019), the $R^2$ values for out-of-sample predictions (10% left-out samples) over 2005 to 2014 are 0.55 for EC, 0.59 for OC, 0.64 for NO$_3$, 0.68 for NH$_4$ and 0.68 for SO$_4$. Our corresponding $R^2$ values for 2014 are consistently higher: 0.64 for EC, 0.75 for OC, 0.73 for NO$_3$, 0.83 for NH$_4$ and 0.80 for SO$_4$. Other pollutants also show improvement when CMAQ outputs for multiple pollutants are used as predictors. For example, our $R^2$ values versus Senthilkumar et al. (2019) for NO$_2$ are 0.56 versus 0.46, for NOx are 0.62 versus 0.46, and for PM$_{10}$ are 0.49 versus 0.42. Our results are also comparable to a previous multivariate spatial CMAQ data fusion model for PM$_{2.5}$ components that reported $R^2$ of 0.58 for NO$_x$ and 0.76 for SO$_4$ (Rundel et al. 2015).

There are several previous efforts to estimate national PM$_{2.5}$ and O$_3$ concentrations and our univariate model approach under-performs. For example, Berrocal et al. (2020) compared various geostatistical and machine learning methods to estimate national daily PM$_{2.5}$ levels in 2011. The best performing model (universal kriging with CMAQ output) has an out-of-sample $R^2$ value of 0.76. Other studies have shown that this may be further improved by incorporating other predictions such as land use, meteorology, and satellite-derived aerosol optical depth (AOD). For example, Hu et al. (2017) reported an out-of-sample $R^2$ of 0.80 for PM$_{2.5}$ prediction from a random forest model for the US; Di et al. (2019) reported an $R^2$ of 0.86 for PM$_{2.5}$ at 1 km spatial resolution using an ensemble of machine learning algorithms; more recently, Kianian et al. (2021) incorporated gap-filled satellite AOD and reported an $R^2$ of 0.80 for PM$_{2.5}$. For national ozone predictions, Senthilkumar et al. (2019) reported an $R^2$ of 0.83 and achieved similar prediction performance ($R^2$ of 0.76) in their US model. Our lower $R^2$ values for PM$_{2.5}$ ($R^2 = 0.67$) and O$_3$ ($R^2 = 0.77$) are likely due to the lack of extensive additional predictors, especially at fine-scale spatial resolutions. The use of multiple CMAQ predictors might also result in model overfit. Moreover, previous studies have demonstrated the usefulness of borrowing spatial information across nearby CMAQ grid cells (Berrocal et al. 2012, Reich et al. 2014) and allowing the regression coefficients to be spatially-varying (Berrocal et al. 2012, Chang et al. 2014) that are not considered here.

One of the main contributions of this work is the creation of a US national daily pollution product for PM$_{2.5}$ components. The growing studies on PM$_{2.5}$ components have typically been restricted to specific regions, e.g., northeastern US (Di et al. 2016) and California (Geng et al. 2020), as well as for long-term averages.

Figure 5. Comparison of multivariate model predictions with CMAQ outputs.
Similar to previous studies, among the PM$_{2.5}$ components examined, we found that prediction is more challenging for EC, likely due to its high spatial heterogeneity and limited monitoring stations. One important distinction of our modeling approach is the ability to provide uncertainty estimates for each prediction.

In summary, we describe a multi-pollutant modeling framework and computational approaches to estimate daily ambient pollution concentrations. We utilize numerical model simulations of air quality, observed monitoring measurements, and correlations between pollutants to create a daily data product at 12 km spatial resolutions for the contiguous US. Model evaluations have demonstrated the benefits of considering multiple pollutants simultaneously to improve predictive performance. Pollutant concentration estimates and their associated predictions uncertainties have been made accessible at zenodo.org. To our knowledge this is the first publicly available data product for major PM$_{2.5}$ species and several gases (e.g., CO, NO$_x$) at this spatial and temporal resolution.

**Acknowledgments**

The authors would like to thank Dr James Mulholland and Niru Senthilkumar for providing the datasets. This work was supported by the National Institutes of Health (ES027892 and ES032344). The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**Data availability statement**

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.4784700.

**Appendix**

![Figure A1. Time series plots of the pollutant concentrations averaged over locations.](image-url)
Table A1. Pollutant measurement time range and units.

| Pollutant | Average time | Unit |
|-----------|--------------|------|
| PM10      | 24 h         | μg m$^{-3}$ |
| PM2.5     | 24 h         | μg m$^{-3}$ |
| EC        | 24 h         | μg m$^{-3}$ |
| OC        | 24 h         | μg m$^{-3}$ |
| SO4       | 24 h         | μg m$^{-3}$ |
| NO3       | 24 h         | μg m$^{-3}$ |
| NH4       | 24 h         | μg m$^{-3}$ |
| NOx       | 1 h max      | ppb  |
| NO2       | 1 h max      | ppb  |
| CO        | 1 h max      | ppm  |
| O3        | 8 h max      | ppb  |
| SO2       | 1 h max      | ppb  |
**Figure A2.** Seasonal average prediction (top row) and standard error (bottom) for each pollutant in 2014.

**Table A2.** Mean adjusted R-square for simple linear regression model with only corresponding CMAQ pollutant and the multivariate linear model with all CMAQ pollutants.

| AQS pollutants | Multivariate linear model | Simple linear model |
|----------------|----------------------------|---------------------|
| CO             | 0.25                       | 0.16                |
| EC             | 0.42                       | 0.29                |
| NH4            | 0.62                       | 0.50                |
| NO2            | 0.55                       | 0.50                |
| NO3            | 0.56                       | 0.43                |
| NOx            | 0.51                       | 0.44                |
| O3             | 0.53                       | 0.39                |
| OC             | 0.48                       | 0.29                |
| PM10           | 0.31                       | 0.15                |
| PM25           | 0.56                       | 0.38                |
| SO2            | 0.29                       | 0.24                |
| SO4            | 0.67                       | 0.35                |
ORCID iDs

Wenlong Gong @ https://orcid.org/0000-0003-2374-3248

References

Appel K W, Gilliland A B, Sarwar G and Gilliam R C 2007 Evaluation of the community multiscale air quality (CMAQ) model version 4.5: sensitivities impacting model performance: I. Ozone Atmos. Environ. 41 9603–15
Appel K W, Bhave P V, Gilliland A B, Sarwar G and Roselle S J 2008 Evaluation of the community multiscale air quality (CMAQ) model version 4.5: sensitivities impacting model performance: II. Particulate matter Atmos. Environ. 42 6057–66
Bergen S, Sheppard L, Sampson P D, Kim S Y, Richards M, Vedal S, Kaufman J D and Szpiro A A 2013 A national prediction model for PM2.5 component exposures and measurement error–corrected health effect inference Environ. Health Perspect. 121 1017–PM25
Berrocal V J, Gelfand A E and Holland D M 2010 A bivariate space-time downscaler under space and time misalignment. Annals of Applied Statistics 4 1942–1975
Berrocal V J, Gelfand A E and Holland D M 2012 Space-time data fusion under error in computer model output: an application to modeling air quality Biometrics 68 837–48
Berrocal V J, Guan Y, Muyskens A, Wang H, Reich B J, Mulholland J A and Chang H H 2020 A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration Atmos. Environ. 222 117130
Binkowski F S and Roselle S J 2003 Models–3 community multiscale air quality (CMAQ) model aerosol component 1. Model description J. Geophys. Res. Space Phys. 108 U6
Blangiardo M and Cameletti M 2015 Spatial and Spatio-Temporal Bayesian Models With R-INLA (New York, NY: Wiley)
Blangiardo M, Finazzi F and Cameletti M 2016 Two-stage bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions *Spatial and Spatio-Temporal Epidemiology* **18** 1–2

Butland B K, Samoli E, Atkinson R W, Barratt B R and Katsouyanni K 2019 Measurement error in a multi-level analysis of air pollution and health: a simulation study *Environmental Health* **18** 1–0

Chang Howard H, Hu Xuexi and Liu Yang 2014 Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downsampling *Journal of Exposure Science & Environmental Epidemiology* **24** 398–404

Chen R et al 2018 Associations between ambient nitrogen dioxide and daily cause-specific mortality: evidence from 272 Chinese cities *Epidemiology* **29** 482–9

Chu Y et al 2016 A review on predicting ground PM2.5 concentration using satellite aerosol optical depth *Atmosphere* **7** 129

Crooks J L and Özkaynak H 2014 Simultaneous statistical bias correction of multiple PM2.5 species from a regional photochemical grid model *Atmos. Environ.* **95** 126–141

Dai L, Zanobetti A, Koutrakis P and Schwartz J D 2014 Associations of fine particulate matter species with mortality in the United States: a multicity time-series analysis *Environ. Health Perspect.* **122** 837–42

Di Qian, Koutrakis, Petros and Schwartz Joel 2016 A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression *Atmospheric Environment* **131** 390–99

Di Qian et al 2019 An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution *Environ. Inter.**130** 104999

Diao M et al 2019 Methods, application, and availabilities of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models *J. Air Waste Manage. Assoc.* **69** 1391–PM414

Everitt B and Hothorn T 2011 An Introduction to Applied Multivariate Analysis With R (New York, NY: Springer)

Friborg M D, Zhai X D, Holmes H, Chang H H, Strickland M, Sarnat S E, Tolbert P E, Russell A G and Mulholland J A 2016 Method for fusing observational data and chemical transport model simulations to estimate spatiotemporally-resolved ambient air pollution *Environ. Sci. Technol.* **50** 7069–705

Fuglstad G A, Simpson D, Lindgren F and Rue H 2019 Constructing priors that penalize the complexity of gaussian random fields and Gaussian Markov random fields *J. Am. Stat. Assoc.* **114** 445–52

Gelfand Alan E., Schmidt Alexandra M., Banerjee Sudipto and Sirmans C. F. 2004 Nonstationary multivariate process modeling through a differential equation approach *Journal of the Royal Statistical Society: Series B* **66** 187–208

Geng G, Meng X, He K and Liu Y 2020 Random forest models for PM2.5 speciation concentrations using MISR fractional AODs *Environ. Res. Lett.* **15** 034056

Hu X, Belle J H, Meng X, Wildani A, Waller L A, Strickland M J and Liu Y 2017 Estimating PM2.5 concentrations in the contiguous United States using the random forest approach *Environmental Science & Technology* **51** 6936–PM414

Hu X, Waller L A, Al-Hamdan M Z, Crosson W L, Estes M G Jr, Estes S M, Quattrochi D A, Sarnat J A and Liu Y 2013 Estimating ground-level PM2.5 concentrations in the southeastern US using geographically weighted regression *Environ. Res.* **121**

Jin X et al 2019 Comparison of multiple PM2.5 exposure products for estimating health benefits of emission controls over New York State, USA. *Environ. Res. Lett.* **14** 040823

Keller J P, Olives C, Kim SY, Sheppard L, Sampson P D, Sipio A A, Oron A P, Lindström J, Vedal S and Kaufman J D 2015 A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution *Environ. Health Perspect.* **123** 301–9

Kianian B, Liu Y and Chang H H 2021 Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for PM2.5 prediction *Remote Sensing* **13** 126

Liang F, Gao M, Xiao Q, Carmichael G R, Pan X and Liu Y 2017 Evaluation of a data fusion approach to estimate daily PM2.5 levels in North China *Environ. Res.* **158** 54–60

Lindgren F and Rue H 2015 Bayesian spatial modelling with R-INLA *Journal of Statistical Software* **63** 1–25

Lindgren F, Rue H and Lindström J 2011 An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach *Journal of the Royal Statistical Society: Series B* **73** 423–98

Lyu B et al 2019 Fusion method combining ground-level observations with chemical transport model predictions using an ensemble deep learning framework: application in China to estimate spatiotemporally-resolved PM2.5 exposure fields in 2014–2017 *Environmental Science & Technology* **53** 7306–19

McGuinn L A et al 2017 Fine particulate matter and cardiovascular disease: comparison of assessment methods for long-term exposure *Environ. Res.* **159** 16–23

Morawska L et al 2018 Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? *Environ. Int.* **116** 286–99

Nychka D, Furrer R, Paige J and Sain S 2017 Fields: tools for spatial data *R package version* **9** 8–6

Özkaynak H, Baxter L K and Burke J 2013 Evaluation and application of alternative air pollution exposure metrics in air pollution epidemiology studies *Journal Of Exposure Science & Environmental Epidemiology* **23** 565

Pitchford M L, Poirot R L, Schichtel B A and Malm W C 2009 Characterization of the winter midwestern particulate nitrate bulge *J. Air Waste Manage. Assoc.* **59** 1061–9

Plummer M 2003 JAGS: a program for analysis of bayesian graphical models using gibbs sampling *Proc. of the 3rd Int. Workshop on Distributed Statistical Computing (DSC) (Vienna, Austria)* **2003**

Reich B J, Chang H H and Foley K M 2014 A spectral method for spatial downsampling *Biometrics* **70** 932–42

Rue H, Martino S and Chopin N 2009 Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations *Journal of the Royal Statistical Society: Series B* **71** 319–92

Rundd C W, Schlep E M, Gelfand A E and Holland D M 2015 A data fusion approach for spatial analysis of speciated PM2.5 across time *Environmetrics* **26** 515–32

Rushworth A, Lee D and Mitchell R 2014 A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London *Spatial and Spatio-Temporal Epidemiology* **10** 29–38

Senthilkumar N, Gilfether M, Metcalfe F, Russell A G, Mulholland J A and Chang H H 2019 Application of a fusion method for gas and particle air pollutants between observational data and chemical transport model simulations over the contiguous United States for 2005–2014 *International Journal of Environmental Research and Public Health* **16** 3314

Shaddick G et al 2018 Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment *Environmental Science & Technology* **52** 9069–78

Stein ML 2012 Interpolation of Spatial Data: Some Theory for Kriging (New York, NY: Springer)
US EPA 2014 Office of Research and Development. CMAQv5.0.2. Zenodo. 30 May. Available online: (http://10.5281/zenodo.1079898) (accessed on 3 August 2017)

US EPA 2017 Air Quality System Database. Available online: (http://epa.gov/ttn/airs/aqsdatamaty) (accessed on 3 August 2017)

Van Donkelaar A, Martin R V, Li C and Burnett R T 2019 Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors Environmental Science & Technology 53 2595–611

Vicedo-Cabrera A M et al 2020 Short term association between ozone and mortality: global two stage time series study in 406 locations in 20 countries Brit. Med. J. 368

Xiao Q, Wang Y, Chang H H, Meng X, Geng G, Lyapustin A and Liu Y 2017 Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China Remote Sens. Environ. 199 437-446

Wu X, Braun D, Schwartz J, Kioumourtzoglou M A and Dominici F 2020 Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly Science Advances 6 eaba5692

Yu H, Russell A, Mulholland J, Odman T, Hu Y, Chang H H and Kumar N 2018 Cross-comparison and evaluation of air pollution field estimation methods Atmos. Environ. 179 49–60