Supporting Information

Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: An Application for Attainment Demonstration in North Carolina

Audrey de Nazelle [a,b,c,d], Saravanan Arunachalam[a], Marc L. Serre[a,*]

[a] University of North Carolina, Chapel Hill, NC, USA

[b] Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

[c] Municipal Institute of Medical Research (IMIM-Hospital del Mar), Barcelona, Spain

[d] CIBER Epidemiologia y Salud Pública (CIBERESP), Barcelona, Spain

* corresponding author: marc_serre@unc.edu

Pages: 16
Figures: S1 to S10
Table: S1

This online supporting information provides descriptions of (1) the review of the Bayesian melding approach (2) the parameter estimation of the BME method (3) the ozone data and its mean trend and covariance functions, (3) the calculation of error variance in model performance analysis, (4) the truncated Gaussian distribution used to model the soft data, (5) the results of the BME error variance estimates, (6) the statistical test for significance of reduction in mean square error for the cross-validation analysis, and (7) comparison of BME non attainment determination with EPA classification analysis.

1. Review of the Bayesian melding approach

Our notation for variables will consist in denoting a single random variable Z in capital letter, its realization z in lower case, and vectors or matrices in bold faces, e.g. $\boldsymbol{Z}=[Z_1,Z_2,\ldots]^{\mathrm{T}}$ and $\boldsymbol{z}=[z_1,z_2,\ldots]^{\mathrm{T}}$.

We summarize here the important Bayesian melding approach developed by Fuentes and Raftery [1] in a spatial estimation context. They combine observations and model predictions to estimate the unobserved true ozone concentration $Z(\boldsymbol{s})$ at location $\boldsymbol{s}$.

Their "data" model for an observation $\hat{Z}(\boldsymbol{s})$ at point $\boldsymbol{s}$ is

$$\hat{Z}(\boldsymbol{s}) = Z(\boldsymbol{s}) + \hat{E}(\boldsymbol{s}) \tag{S1}$$

where the observation error $\hat{E}(.) \sim N(0,\sigma^2_{\hat{E}})$ is a white noise process independent from $Z(\boldsymbol{s})$. Similarly the "data" model for a model prediction $\tilde{Z}(B)$ over some region B is

$$\tilde{Z}(B) = \|B\|^{-1}\int_B d\boldsymbol{s}\left[Z(\boldsymbol{s}) + \tilde{E}(\boldsymbol{s})\right] \tag{S2}$$

where the point-prediction error $\tilde{E}(\boldsymbol{s})$ is the sum of a spatial bias with parameters $\boldsymbol{a}$ and a white noise process $N(0,\sigma^2_{\tilde{E}})$ independent from $Z(.)$ and $\hat{E}(.)$.

The "process" model for ozone concentration $Z(\boldsymbol{s})$ is

$$Z(\boldsymbol{s})=\mu(\boldsymbol{s})+Z'(\boldsymbol{s}) \tag{S3}$$

where the mean trend model $\mu(\boldsymbol{s})$ may be a function of parameters $\boldsymbol{\beta}$, and the zero-mean autocorrelated Gaussian process $Z'(\boldsymbol{s})$ has covariance with parameters $\boldsymbol{\theta}$ that may be varying in space.

Let the random variable $Z_k=Z(\boldsymbol{s}_k)$ correspond to ozone at some estimation point $\boldsymbol{s}_k$, and let the column vector of random variables $\boldsymbol{Z}_d = [\hat{\boldsymbol{Z}}_o^T, \tilde{\boldsymbol{Z}}_m^T]^T$ be the collection of random variables corresponding to the observations $\hat{\boldsymbol{Z}}_o = [\hat{Z}_1, \hat{Z}_2, \ldots]^T$ at points $\boldsymbol{s}_o = \{\boldsymbol{s}_{o1}, \boldsymbol{s}_{o2}, \ldots\}$ and model predictions $\tilde{\boldsymbol{Z}}_m = [\tilde{Z}_1, \tilde{Z}_2, \ldots]^T$ at points $\boldsymbol{s}_m = \{\boldsymbol{s}_{m1}, \boldsymbol{s}_{m2}, \ldots\}$. We denote by z_k a realization of Z_k, and by $\boldsymbol{z}_d$ an observed/model-predicted value for $\boldsymbol{Z}_d$. The model equations (S1-3) describe the statistical distribution of $(Z_k,\boldsymbol{Z}_d)$ in terms of the parameters $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2_{\hat{E}}, \sigma^2_{\tilde{E}}, \boldsymbol{a})$. The solution of the spatial estimation problem for some fixed parameters $\boldsymbol{\phi}$ is given by the conditional PDF for $(Z_k|\boldsymbol{Z}_d,\boldsymbol{\phi})$ which we write as

$$f(z_k|\boldsymbol{z}_d,\boldsymbol{\phi}). \tag{S4}$$

If we assume that the random variables $(Z_k,\boldsymbol{Z}_d|\boldsymbol{\phi})$ are jointly normally distributed, then it follows that $(Z_k|\boldsymbol{Z}_d,\boldsymbol{\phi})$ is normal with mean

$$\mu_{k|d}(\boldsymbol{\phi}) = \mu(\boldsymbol{s}_k) + \boldsymbol{C}_{kd}\boldsymbol{C}_{dd}^{-1}(\boldsymbol{z}_d - \boldsymbol{\mu}_d) \tag{S5}$$

and variance

$$\sigma_{k|d}^2(\boldsymbol{\phi}) = \sigma_k^2 - \boldsymbol{C}_{kd}\boldsymbol{C}_{dd}^{-1}\boldsymbol{C}_{kd}^{T} \tag{S6}$$

where $\boldsymbol{\mu}_d$=E[$\boldsymbol{Z}_d$] is the expected value of $\boldsymbol{Z}_d$, σ^2_k is the variance of Z_k, C_{kd}=cov(Z_k,$\boldsymbol{Z}_d$) is a row vector of covariances between Z_k and each element in $\boldsymbol{Z}_d$, and C_{dd}=cov($\boldsymbol{Z}_d$,$\boldsymbol{Z}_d$) is a square matrix of covariances between each pairs of elements in $\boldsymbol{Z}_d$. As shown in Fuentes and Raftery[1], Eqs. (S1-3) can easily be used to express $\boldsymbol{\mu}_d$, σ^2_k, C_{kd} and C_{dd} as a function of the parameters $\boldsymbol{\phi}$.

Eqs. (S4-6) basically correspond to the solution of the classical kriging method when $\boldsymbol{\phi}$ is known. Various strategies can be used to remove the dependency of Eqs. (S4-6) on $\boldsymbol{\phi}$. The Bayesian approach taken by Fuentes and Raftery[1] consists in removing the conditionalization on $\boldsymbol{\phi}$ in Eq. (S4) by taking the marginal PDF of $f(z_k,\boldsymbol{\phi}\,|\,\boldsymbol{z}_d) = f(z_k\,|\,\boldsymbol{z}_d,\boldsymbol{\phi})f(\boldsymbol{\phi}\,|\,\boldsymbol{z}_d)$ with respect to z_k, which leads to

$$f(z_k\,|\,\boldsymbol{z}_d) = \int d\boldsymbol{\phi}\, f(z_k\,|\,\boldsymbol{z}_d,\boldsymbol{\phi})\, f(\boldsymbol{\phi}\,|\,\boldsymbol{z}_d) \tag{S7}$$

Fuentes and Raftery[1] calculate this integral using a Gibbs sampling approach, which involves defining a prior PDF for $\boldsymbol{\phi}$, and taking the likelihood $f(\boldsymbol{z}_d\,|\,\boldsymbol{\phi})$ as a multivariate normal PDF with mean and covariance derived from the model Eqs. (S1-3). The expected value and variance of $(Z_k|\boldsymbol{Z}_d)$ can then at once be written as

$$\mu_{k|d} = \int d\boldsymbol{\phi}\ \mu_{k|d}(\boldsymbol{\phi})\ f(\boldsymbol{\phi}\,|\,\boldsymbol{z}_d) \tag{S8}$$

and

$$\sigma^2_{k|d} = \int d\boldsymbol{\phi}\ \sigma^2_{k|d}(\boldsymbol{\phi})\ f(\boldsymbol{\phi}\,|\,\boldsymbol{z}_d) \tag{S9}$$

These equations show that, basically, the Bayesian approach corresponds to taking the average of the kriging mean $\mu_{k|d}(\boldsymbol{\phi})$ and variance $\sigma^2_{k|d}(\boldsymbol{\phi})$ weighted by $f(\boldsymbol{\phi}|\boldsymbol{z}_d)$, which expresses how likely each values of the parameters $\boldsymbol{\phi}$ are given the observed data $\boldsymbol{z}_d$.

A powerful feature of this Bayesian approach is that the kriging solution can be viewed as an approximate limiting case of the more general Bayesian case, in that the kriging case consists in using an empirical or a likelihood method to obtain an estimate $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$, and then treating $\hat{\boldsymbol{\phi}}$ as the known value for $\boldsymbol{\phi}$ by "plugging" it in Eqs. (S5-6), i.e. $\mu_{k|d} \approx \mu_{k|d}(\hat{\boldsymbol{\phi}})$ and $\sigma^2_{k|d} \approx \sigma^2_{k|d}(\hat{\boldsymbol{\phi}})$. As pointed by Fuentes and Raftery[1], this approximation is not robust for the Bayesian melding problem they consider, probably because in their problem both the observation and model-prediction error variances $\sigma^2_{\hat{E}}$ and $\sigma^2_{\tilde{E}}$ are unknown parameters. However, that is not the case for our problem where we take $\sigma^2_{\hat{E}}$=0, which essentially means that $\hat{Z}(\boldsymbol{s}) = Z(\boldsymbol{s})$, or stated in other words, that in a regulatory context, the observations must be used as a proxy for the true concentration of ozone. We conducted a simulation study that

revealed that in this case the kriging approximation $\mu_{k|d} \approx \mu_{k|d}(\hat{\boldsymbol{\phi}})$ and $\sigma^2_{k|d} \approx \sigma^2_{k|d}(\hat{\boldsymbol{\phi}})$ is reasonable when taking $\hat{\boldsymbol{\phi}}$ as the Maximum a-posterior (MAP) estimate of $\boldsymbol{\phi}$, i.e. by selecting the $\boldsymbol{\phi}$ that maximizes $f(\boldsymbol{\phi}|z_d)$, and acceptable for other empirical estimates $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$.

The Bayesian melding approach described above can surely be used for our problem by simply setting $\sigma^2_{\hat{E}}$=0, considering the reduced vector of parameters $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2_{\tilde{E}}, \boldsymbol{a})$ (notice that $\sigma^2_{\hat{E}}$ was removed from $\boldsymbol{\phi}$), and then either obtaining the full Bayesian estimator $\mu_{k|d}$ and $\sigma^2_{k|d}$ (Eqs. S8-9), or its kriging approximation $\mu_{k|d}(\hat{\boldsymbol{\phi}})$ and $\sigma^2_{k|d}(\hat{\boldsymbol{\phi}})$ where $\hat{\boldsymbol{\phi}}$ might be the MAP estimator, a Maximum Likelihood estimator, or even an empirical estimator. However, this approach relies on some strong assumptions which, we believe, may lead to two important limitations.

The first limitation pertains to the assumption that the complex stochastic relationship between the process $Z(\boldsymbol{s})$ and its model prediction $\tilde{Z}(B)$ can be adequately represented by the linearized form expressed in Eq. (S2) with a Gaussian prediction error. The second limitation is the assumption that the covariance matrices $\boldsymbol{C}_{kd}$ and $\boldsymbol{C}_{dd}$ defined in Eqs. (S5-6) capture the full knowledge of the linearized relationship (S2).

These limitations of Bayesian melding, and more generally the similar limitations of Bayesian hierarchical modeling, are the motivation for proposing the BME approach presented in this paper. The BME approach provides a sound mathematical framework for non-linear, non-Gaussian estimation that is not limited to the linear form of Eq. (S2) and may better capture the complex non-Gaussian stochastic relationship between $Z(\boldsymbol{s})$ and $\tilde{Z}(B)$.

2. Parameter estimation for the BME method

Let the S/TRF $Z(\boldsymbol{p})$ represent ozone at space/time point $\boldsymbol{p}=(\boldsymbol{s},t)$, where $\boldsymbol{s}$ is the spatial location and t is the time. The BME method requires calculation of the G–KB based PDF $f_G(\mathbf{z}) = \int d\boldsymbol{\beta} \int d\boldsymbol{\theta} f(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\mathbf{z}}_o) f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \hat{\mathbf{z}}_o)$ (Eq. 4), where $\mathbf{z} = (z_k, \hat{\mathbf{z}}_o, \mathbf{z}_m)$ is a realization of the vector of random variables $(Z_k, \mathbf{Z}_o, \mathbf{Z}_m)$ representing the ozone S/TRF $Z(\boldsymbol{p})$ at the estimation point $\boldsymbol{p}_k$, the observation points $\boldsymbol{p}_o$, and model prediction grid points $\boldsymbol{p}_m$, respectively. The conditioning on $\hat{\mathbf{z}}_o$ was noted in the integrand to reflect that the ozone concentration is known at the observation points $\boldsymbol{p}_o$. $\boldsymbol{\beta}$ are the parameters of the ozone mean trend model $\mu(\boldsymbol{p};\boldsymbol{\beta})$=E[$Z(\boldsymbol{p})$], while $\boldsymbol{\theta}$ are the parameters of the ozone covariance model $c_Z(\boldsymbol{p},\boldsymbol{p}';\boldsymbol{\theta})$=cov($Z(\boldsymbol{p})$,$Z(\boldsymbol{p}')$). $f(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\mathbf{z}}_o)$ is the multivariate Gaussian PDF for $\mathbf{z}$ with mean and variance given by $\mu(\boldsymbol{p};\boldsymbol{\beta})$ and $c_Z(\boldsymbol{p},\boldsymbol{p}';\boldsymbol{\theta})$, respectively. The PDF $f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \hat{\mathbf{z}}_o)$ is proportional to the product of the likelihood $f(\hat{\mathbf{z}}_o \mid \boldsymbol{\beta}, \boldsymbol{\theta})$ and some prior PDF $f(\boldsymbol{\beta}, \boldsymbol{\theta})$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

An approximation for Eq. 4 is provided by $f_G(\mathbf{z}) \approx f(\mathbf{z} \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}_o)$ (Eq. 5), where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. We discuss here three estimators that we considered to obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and our selection of one of these estimators.

The three parameter estimators that we considered are described in table S1. They are the least square (LS), the maximum likelihood (ML), and the maximum a posteriori (MAP)

estimators. The LS estimator generally involves the least square fitting of parameterized models to relevant data, while the ML estimator seeks parameters that are most likely given the observed data, and the MAP estimator is an improvement upon the ML estimator if prior knowledge about the parameters is available.

Table S1
Definition of the least square (LS), maximum likelihood (ML) and maximum a posteriori (MAP) estimators of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ of Eq. 5

Parameter Estimator	Description	Resulting approx. error in simulated study
Least Square (LS)	Generalized least square fitting a polynomial mean trend model with parameters $\boldsymbol{\beta}$ on the observed values $\hat{z}_o$, Least square fitting of the covariance model with parameters $\boldsymbol{\theta}$ on experimental covariance values	$\lvert \hat{z}_k - \hat{z}_k^{(LS)} \rvert < 0.29 ppb$
Maximum Likelihood (ML)	$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ which maximizes the likelihood $f(\hat{z}_o \mid \boldsymbol{\beta}, \boldsymbol{\theta})$	$\lvert \hat{z}_k - \hat{z}_k^{(ML)} \rvert < 0.038 ppb$
Maximum a Posteriori (MAP)	$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ which maximizes $f(\hat{z}_o \mid \boldsymbol{\beta}, \boldsymbol{\theta})\ f(\boldsymbol{\beta}, \boldsymbol{\theta})$	$\lvert \hat{z}_k - \hat{z}_k^{(MAP)} \rvert < 0.0078 ppb$

When implementing the parameter estimators described in Table S1, we define the mean trend model as a polynomial function of the spatial and temporal coordinates. For example, a polynomial of order 1 corresponds to a linear drift with respect to space and time, and in that case $\boldsymbol{\beta}$ consists in the intercept and linear coefficients of that linear space/time drift. The likelihood $f(\hat{z}_o \mid \boldsymbol{\beta}, \boldsymbol{\theta})$ is calculated as a multivariate Gaussian PDF with a mean vector and covariance matrix obtained from the mean trend model $\mu(\boldsymbol{p};\boldsymbol{\beta})$ and covariance model $c_Z(\boldsymbol{p},\boldsymbol{p}';\boldsymbol{\theta})$ evaluated at the observation points $\boldsymbol{p}_o$. The prior PDF $f(\boldsymbol{\beta}, \boldsymbol{\theta})$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ should be selected based on prior knowledge of these parameters, though the selection of this prior is somewhat arbitrary in practice.

The BME mean estimate $\hat{z}_k$ of ozone at an unmonitored point is given by the mean of the BME PDF $f_K(z_k) = A^{-1} \int d\boldsymbol{z}_m f_S(\boldsymbol{z}_m) f_G(\boldsymbol{z})$ (Eq. 3) where the G–KB based PDF $f_G(\mathbf{z})$ is calculated using Eq. 4. If however $f_G(\mathbf{z})$ is approximated using $f_G(\boldsymbol{z}) \approx f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{z}}_o)$ (Eq. 5), where $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ is obtained using either the LS, ML or MAP parameter estimator, then we get approximate BME mean estimates referred to as $\hat{z}_k^{(LS)}$, $\hat{z}_k^{(ML)}$ or $\hat{z}_k^{(MAP)}$, respectively. We conducted a simulation study and found that the approximation is numerically almost exact when using the BME MAP estimate $\hat{z}_k^{(MAP)}$, and robust when using the LS estimate $\hat{z}_k^{(LS)}$ and the ML estimate $\hat{z}_k^{(ML)}$.

For conciseness sake and without loss of generality we present here results obtained when the parameter estimation is reduced to that of estimating the covariance parameters

$\boldsymbol{\theta}$ by assuming that the mean trend is known to be zero. In effect this corresponds to the estimation of mean trend removed ozone concentration, or in other words we redefine the S/TRF $Z(\boldsymbol{p})$ as ozone concentration minus a known mean trend (this simulation is relevant for our study as we found no evidence that using a polynomial mean trend model improves ozone prediction). We used the Choleski decomposition method to simulate realizations of observed mean trend removed ozone values $\hat{z}_o$ with a zero mean and an exponential covariance function $c_Z(r,\tau) = \sigma_Z^2\left(\exp(-3r/a_r)\right)\left(\exp(-3r/a_t)\right)$, where the covariance parameters $\boldsymbol{\theta}=(\sigma_Z^2, a_r, a_t)$ were set to σ_Z^2=0.000159ppm^2, a_r=10km, and a_t=20$hours$, which is representative of the short length/time scale variability of ozone mean trend removed concentrations for our study area (as described in the next section). The mean trend removed ozone concentrations simulated at one monitoring station are shown with filled circles in Fig. S1. These simulated values exhibit a variability that is similar to that of the mean trend removed ozone concentrations in our study area (see for e.g. Fig. S2).

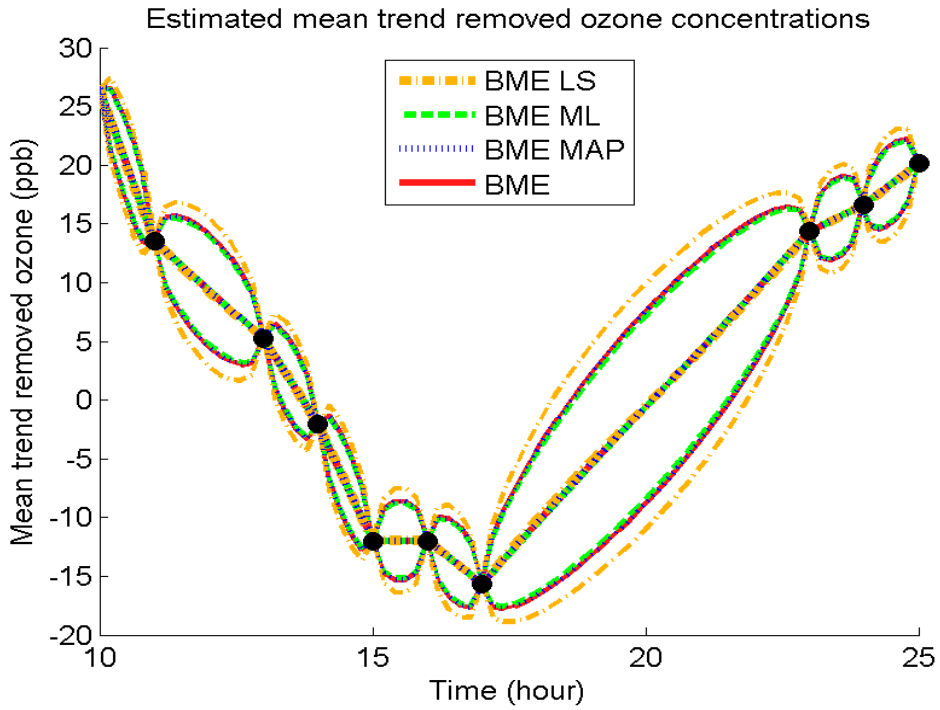


Figure S1. Simulated mean trend removed ozone observations (filled circle). The BME mean trend removed concentration estimates $\hat{z}_k^{(LS)}$, $\hat{z}_k^{(ML)}$, $\hat{z}_k^{(MAP)}$ and $\hat{z}_k$ are shown as approximately straight lines connecting the filled circles, while the round contours delineate the corresponding one standard-deviation lower and upper bounds for each BME estimate.

We then used the simulated $\hat{z}_o$ values to obtain the LS and ML estimates of the covariance parameters, while a prior of the covariance parameters needs to be specified in order to obtain the MAP estimates of the covariance parameters. For instance the LS and ML estimates for the temporal covariance range a_t are directly calculated from the simulated $\hat{z}_o$ shown in Fig. S1 as $\hat{a}_t^{(LS)}$=14.2$hours$ and $\hat{a}_t^{(ML)}$=28.0$hours$, respectively. However; we need a prior for a_t in order to calculate its MAP estimate. One possibility is to take a prior that is Gaussian distributed but truncated below zero, with mean equal to $\hat{a}_t^{(LS)}$, and a standard

deviation equal to 10*hours*. In that case we obtain $\hat{a}_t^{(MAP)}$=25.0*hours*., which indicates some Bayesian updating toward the ML estimate. However this example illustrates how arbitrary the choice of the prior can be in practice.

We finally calculated the mean trend removed BME estimate $\hat{z}_k$, and its approximate estimates $\hat{z}_k^{(LS)}$, $\hat{z}_k^{(ML)}$ or $\hat{z}_k^{(MAP)}$, which are shown in Fig S1 together with the lower and upper bounds of their corresponding standard deviation. As seen from that figure, the approximate estimates $\hat{z}_k^{(LS)}$, $\hat{z}_k^{(ML)}$ and $\hat{z}_k^{(MAP)}$ are in such good agreement with $\hat{z}_k$ that it is not visually possible to distinguish them. Indeed, as indicated in Table S1, the absolute difference between $\hat{z}_k$ and $\hat{z}_k^{(LS)}$ is at most 0.29*ppb* over the time interval shown in Fig. S1. Likewise this difference is at most 0.038*ppb* for the ML estimator, while it is at most 0.0078*ppb* for the MAP estimator. These approximation errors are very small compared to typical ozone concentrations (which mostly vary between about 10*ppb* to 100*ppb* in our study area), hence this simulated study provides quantitative support indicating that the approximation of Eq. 5 leads to an approximation of the BME estimate that is almost numerically exact when using the MAP estimate of the covariance parameters. Indeed Fig. S1 shows not only that $\hat{z}_k$ and $\hat{z}_k^{(MAP)}$ are almost exactly equal, but so are their standard deviation (i.e. their respective lower and upper bound interval are visually indistinguishable). This means that the estimation is robust to parameter uncertainty. This fact is actually illustrated by the fact that even though the LS and ML parameter estimates of a_t are somewhat different than that of the MAP estimator, their corresponding BME estimate of ozone are very similar to that obtained by accounting for parameter uncertainty.

Since we found that $\hat{z}_k^{(LS)}$, $\hat{z}_k^{(ML)}$ and $\hat{z}_k^{(MAP)}$ are good approximations of $\hat{z}_k$, which indicates that Eq. 5 is a reasonable approximation of Eq. 4, then we are free to use any of these estimates. In this study we select the LS parameter estimator because it does not require the arbitrary selection of a prior for the parameters, and it is based on a least square fitting that provides an opportunity to check whether the estimated parameters values are physically meaningful. For example, the covariance parameters are obtained by fitting a covariance model to experimental covariance values (see next section), which can be guided by expert knowledge gained from other ozone studies.

3. Description of the data and its mean trend and covariance functions

Figure S2 provides an example of the temporal trend of both monitored and modeled ozone concentrations at one of the monitoring stations from the NC regulatory monitoring network. Figure S3 illustrates the results of our mean trend analysis for a single station.

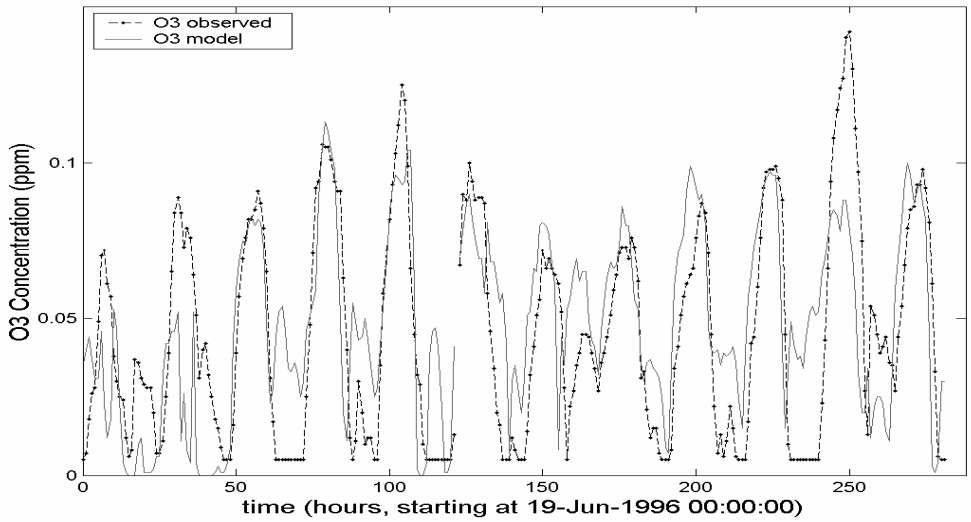


Figure S2. Plot of observed (dashed line) and modeled (plain line) ozone as a function of time at the station bearing the highest ozone concentration observed throughout the study period.

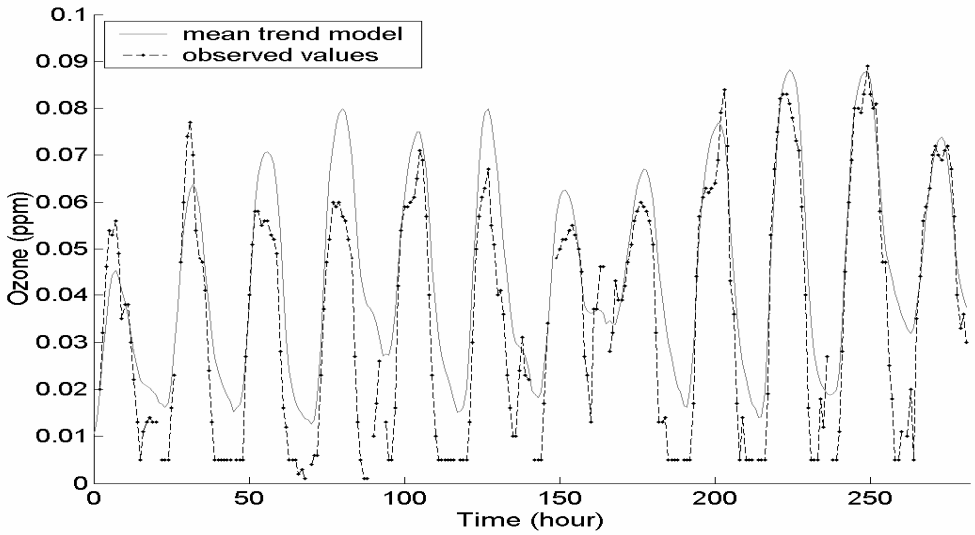


Figure S3. Plot showing ozone observed values $\hat{z}_j$ as a function of time at the Davie County station in NC piedmont (dots connected by a line), and the corresponding mean trend model $\mu(\boldsymbol{p};\boldsymbol{\beta})$ (Eq. 11) (solid line).

Figure S4 shows the results of covariance analysis. Experimental covariance values c_Z are shown with dotted lines as a function of spatial lag r for τ=0 in Fig.S4(a), and as a function of temporal lag τ for r =0 in Fig.S4(b).

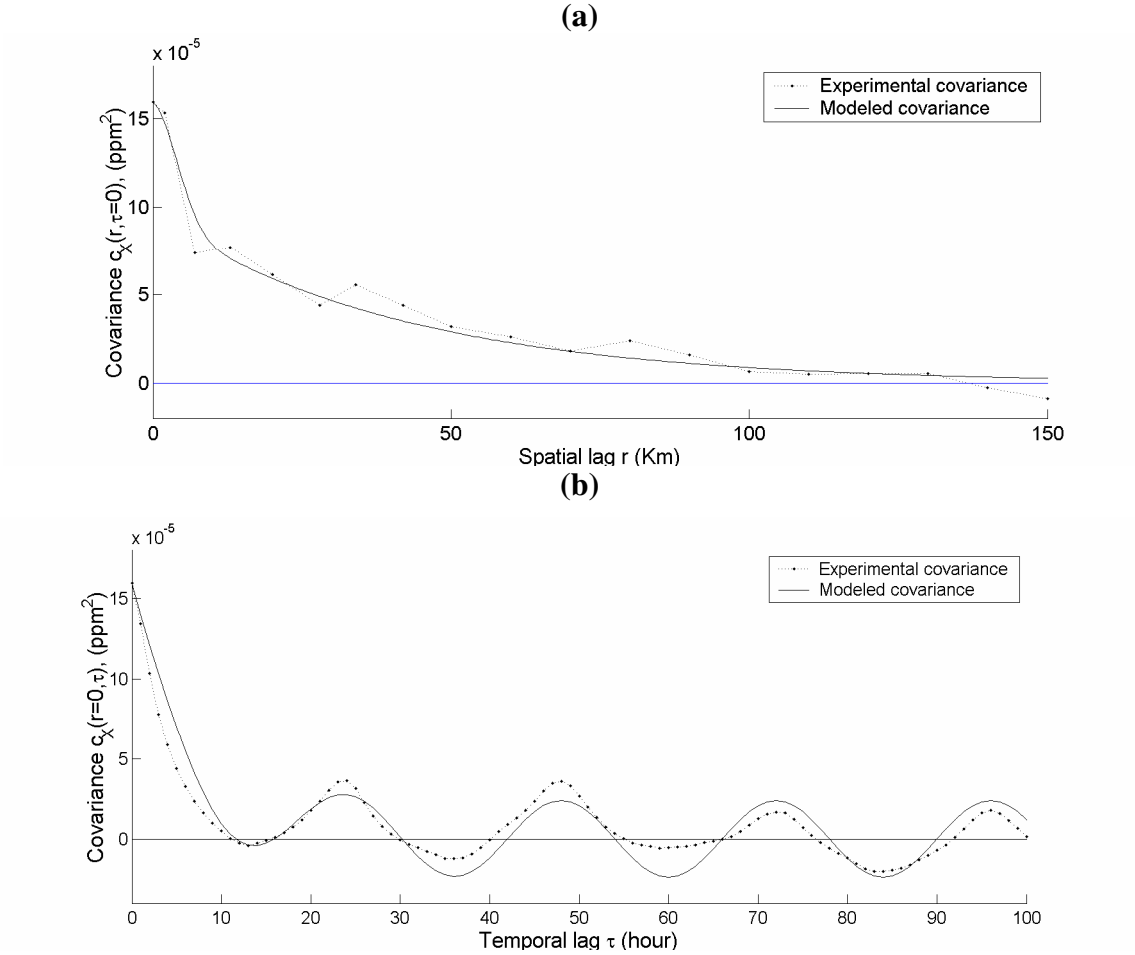


Figure S4. Experimental (dots) and modeled (solid line) covariance shown as a function of **(a)** the spatial lag r for τ=0*hr*. and **(b)** the temporal τ for r=0*km*.

We then fit to these experimental covariance values the following space/time separable covariance model:

$$c_Z(r,\tau) = \sigma_Z^2\left(0.4\exp(\frac{-3r^2}{a_{r1}^2}) + 0.6\exp(\frac{-3r}{a_{r2}})\right)\left(0.85\exp(\frac{-3\tau}{a_{t1}}) + 0.15\cos(\frac{2\pi\tau}{a_{t2}})\right) \quad \text{(S10)}$$

where the ozone concentration variance is σ_Z^2=0.000159ppm^2, the Gaussian spatial range is a_{r1}=10*km*, the exponential spatial range is a_{r2}=125*km*, the temporal exponential range is a_{t1}=20*hours*, and the periodicity for the cosinusoidal temporal range is a_{t2}=24*hours*. Plain lines in Fig.S4 show the resulting fitted covariance model.

Figure S5 shows the results of covariance analysis in a 3D plot.

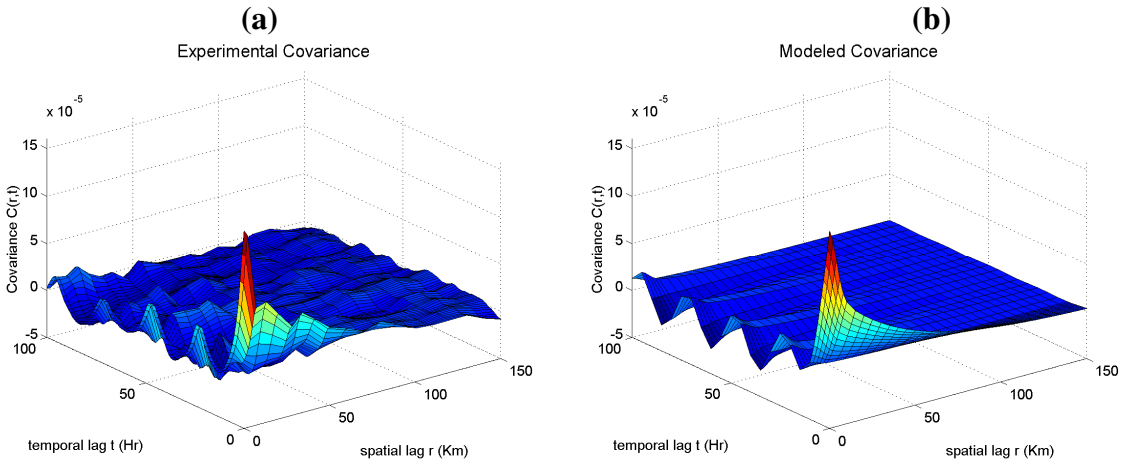


Figure S5. Experimental (a) and modeled (b) space time covariance of the ozone field in North Carolina. The temporal lag (0 to 100*hr*) is shown on the left axis, the spatial lag (0 to 150*km*) is shown on the right axis, and the space time covariance on the vertical axis.

4. Error variance of model performance assessment

Stars in Figure S6 represent the error variances $\hat{\lambda}_2(\tilde{z}_i)\}$ (Eq. 8) calculated within each bin *i,* and plotted against the average modeled predictions $\hat{\lambda}_1(\tilde{z}_i)\}$ (x-axis) for the corresponding bin, and the line connecting the stars is the interpolation model used to assign a soft distribution variance var[Z|$\tilde{z}$] to each modeled data point $\tilde{z}$. Note that higher predicted ozone levels (above approximately 0.07ppm) have lower error variance, so that greater confidence can be assigned to these values.

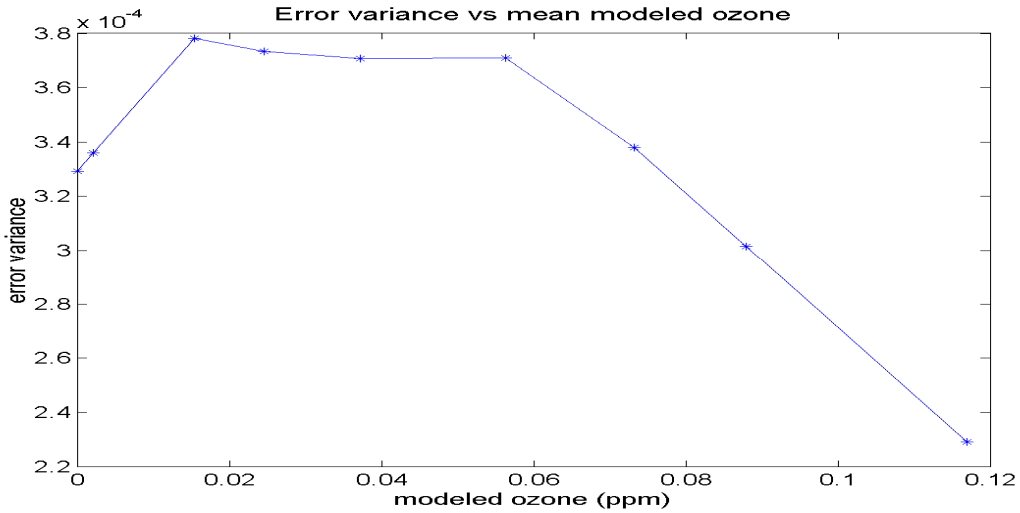


Figure S6. Photochemical model error variance as a function of modeled ozone values: the stars represent the error variances found within each class of ozone modeled values plotted against the mean of modeled ozone within that class; the line describes the interpolation between each point so that each photochemical model output is assigned an error variance.

5. Truncated Gaussian distributions

The choice of the parametric function ϕ(.) depends on the statistical distribution of prediction errors for the air quality model. Without loss of generality, we will consider the

Gaussian distribution truncated below zero. Examples of this distribution are shown in Fig. S7(a) and Fig. S7(b). This distribution is completely defined by two parameters, which are its expected value μ_1, and its standard deviation μ_2, hence we denote it as $\phi(u\ ;\ \mu_1,\mu_2)$. Note that since this parameterized PDF is truncated below zero, its mean μ_1 is greater than its mode.

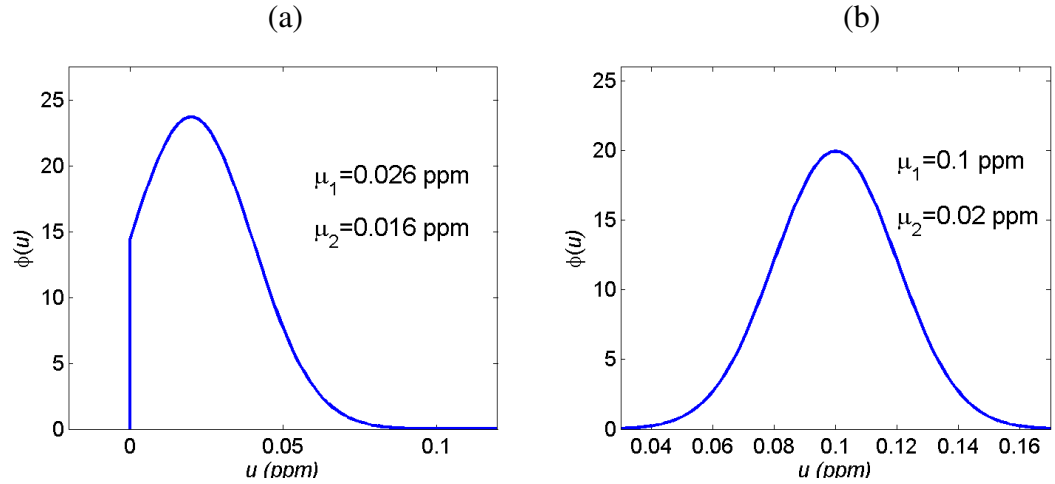
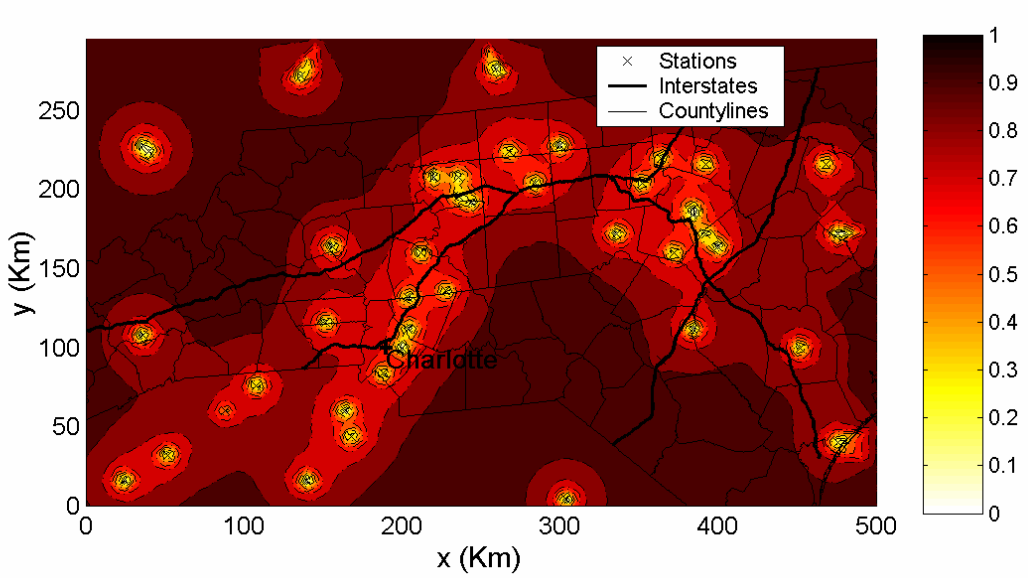


Figure S7: Gaussian distributions $\phi(u\ ;\ \mu_1,\mu_2)$ truncated below zero; for (a) μ_1=0.026*ppm* and μ_2=0.016*ppm* and (b) μ_1=0.1*ppm* and μ_2=0.02*ppm*

4. BME error variance estimates

Fig.S8 compares the BME error variance estimates corresponding to the main text Figure 3 representing hourly mean estimates. In these figures, the error variance σ_k^{*2} is normalized by the covariance sill σ_Z^2 defined in Eq. S10, so that the normalized variance has a value between 0 and 1.

(a)



(b)

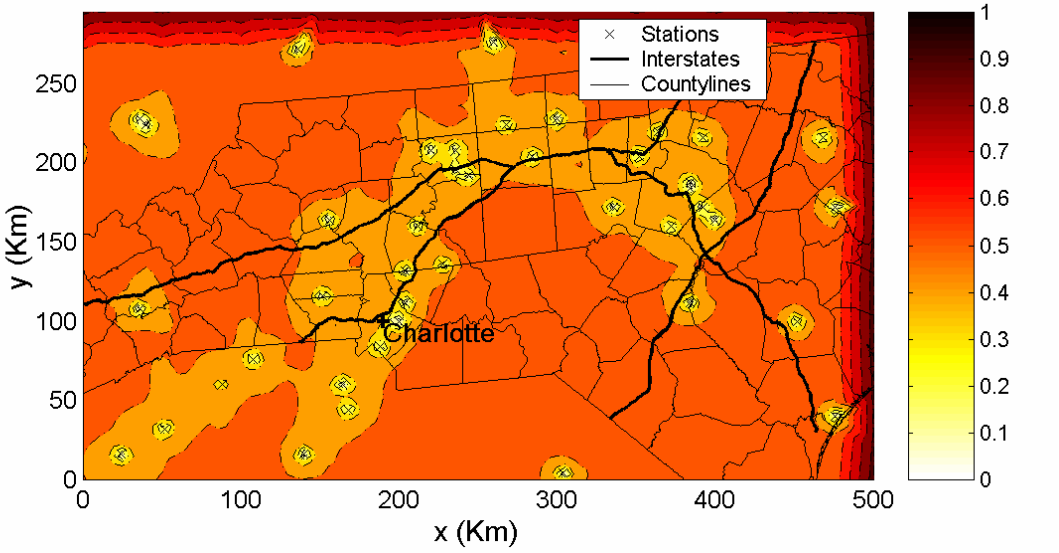


Figure S8. BME estimate of normalized error variance σ_k^{*2}/σ_Z^2 (unitless), obtained when **(a)** using ozone observations only and **(b)** using both ozone observations and model predictions, for hour 250

6. Testing the statistical significance of a reduction in mean square error

We may test whether any findings that our proposed approach did lead to a decrease in mean square error is statistically significant, i.e. whether any findings that *MSEs−MSEh* is found to be negative can be said to be a statistically significant result. We apply the test for the difference of means *d=MSEs−MSEh*, where *MSEs* is the mean of squared prediction error $e_{j,s}^2(r_v)=(z_{j,s}^*(r_V)-\hat{z}_j)^2$ for the proposed approach accounting for the air quality model data treated as soft, and *MSEh* is the mean of squared prediction error $e_{j,h}^2(r_v)=(z_{j,h}^*(r_V)-\hat{z}_j)^2$ for the approach that does not account for the air quality model data. Under this test, the variance of the difference of means is

$$var_d = [\text{variance}(e_{1,s}^2(r_v)) + \text{variance}(e_{1,h}^2(r_v))]/n_o, \quad \text{(S11)}$$

Under the null hypothesis we assume that the difference *d* in mean square error is normally distributed with mean zero and variance var_d. Since we test the alternative hypothesis that *d<0*, we use the one tail *p*-value formulae to calculate the probability of obtaining a specific *MSEs−MSEh* if the null hypothesis were true,

$$p = \int_{-\infty}^{MSEs-MSEh} du\ \phi(u;0,\text{var}_d), \quad \text{(S12)}$$

where $\phi(u\ ;0,var_d)$ is the Gaussian PDF with mean 0 and variance var_d, and *MSEs−MSEh* is the difference in mean square error obtained in the cross validation analysis. Traditionally a *p*-value less than 0.05 indicates that the null hypothesis may be rejected, leading to the conclusion that the decrease in mean square error obtained in the cross validation is statistically significant. The *p*-value may be calculated for the ensemble of monitoring

stations, or it may be calculated for individual monitoring stations. In the later case, we can calculate the percentage of stations where a statistically significant reduction in mean square error is found at different levels of significance (i.e. with *p*-values less than 0.05, 0.01, and 0.1).

We first performed a cross-validation analysis that showed that, to begin with, advancing the analysis from a purely spatial to a spatial-temporal analysis reduces the MSE by almost an order of magnitude, from MSE=$2.1*10^{-4} ppm^2$ to MSE=$2.5*10^{-5} ppm^2$. This is not surprising as observations are hourly, and ozone levels from one hour to the next are highly correlated.

Table 1 in the main text shows cross-validation statistics for the comparison of the hard and soft vs. hard data only methods for five different cross-validation radii r_v of excluded monitoring station data. The results show that the larger the cross-validation radius, the better the soft data methodology performs compared to using hard data only. Only a 1.5% reduction in error is found when a radius of 10*km* is used, however a 28% improvement is observed for a radius of 100km, i.e. in the situation where no monitoring stations exist within 100km of the point of interest. The overall mean square error reduction is not statistically significant for the 10*km* cross-validation radius, but is highly significant (p-value < 0.0001) for radii above 30*km*. At the 95% significance level, only 14% of validation points show a statistically significant improvement in mean square error reduction for the 10*km* radius, while that number reaches 90% for a 100*km* cross-validation radius.

7. Implications for non-attainment status

The method used to estimate ozone maps has implications for determining ozone attainment areas based on a design value (DV), such as the DVC or the DVF defined in Eq. (1). For illustration purposes, we define a pseudo-standard where the DV consists of the second highest 8-hour average, and where the non-attainment area is the collection of points where the DV is above 0.08ppm.

Usually non-attainment areas are defined as the set of points for which the *estimated* DV is greater than the standard value, which essentially corresponds to an area that is more likely than not to be in attainment. However, the methods presented in this work offer additional information on the level of confidence we can attribute to each estimated DV. Hence we may define the probability of violation (PV) of the standard at a point $\boldsymbol{s}$ as

$$PV(\boldsymbol{s})=\text{Prob}[DV(\boldsymbol{s})>NA], \qquad \text{(S13)}$$

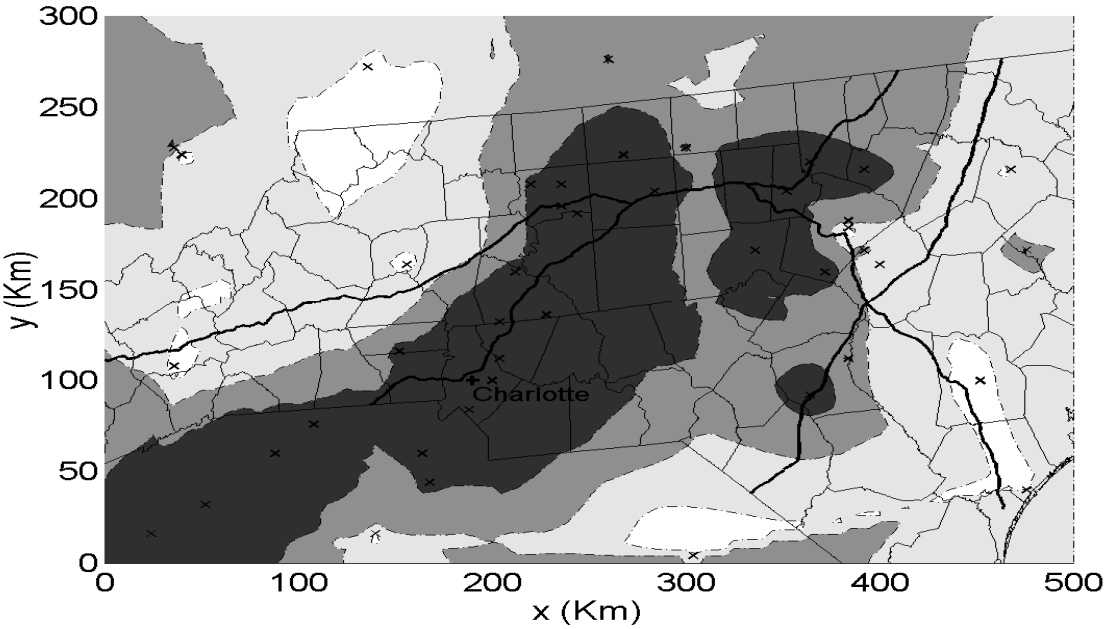
where *NA* is the non-attainment standard value. Fig.S9 illustrates non-attainment areas corresponding to four levels of the PV for the pseudo standard defined above, from the darkest to the lightest shade of grey: PV>0.9 (highly likely in non-attainment), 0.5<PV<0.9; (likely in non-attainment), 0.1<PV<0.5 (near non-attainment), and PV<0.1 (highly likely in attainment). We are thus able to define areas that depict different confidence in attaining the standard, which may be useful for policy purposes in a regulatory context.

Fig.S9 shows the non-attainment areas obtained with estimation scenario (a) using only ozone observations, and estimation scenario (b) using both observations and model-predictions. The difference between Fig.S9(a) and (b) is quite striking, The estimation scenario (b) not only yields different sizes of areas estimated to be in non-attainment, but also allows a more precise delineation of non-attainment boundaries because it has substantially lower estimation variance (e.g. compare Fig.S8a with S8b). Yet, it is quite clear

that the two estimation scenarios show the same core as being in non-attainment, and only disagree on the extent of the area covered under different conditions of uncertainty. Therefore the two approaches are shown to be consistent, but estimation scenario (b) increases the precision of the estimate. The following analysis provides a comparative analysis with the EPA proposal for classification of non-attainment areas.

To allow comparison with EPA results, Figure S10 reproduces estimates and classifications of non attainment areas proposed by the EPA for the year 2004 using the correct design value calculation, based on 2001-2003 data. The figure depicts the county and partial-county violation classifications (marked by an X, source: US EPA 2004a), and also shows the result of their spatial interpolation analysis using the ordinary Kriging methodology (grey area, source: US EPA 2004b). The appropriate comparison with our results is with the 50% confidence level boundaries, as mean estimates are used in the regulatory framework. The boundaries of the EPA's spatial interpolation are similar to the 50% confidence spatial-temporal analysis using hard data only shown in figure S9a, but generally broader with some narrower parts in the south western part of the map. We would expect our analysis to provide more precise estimates because of incorporation of the temporal component, however the results are not fully comparable because of our data limitation (the use of a single ozone episode as opposed to 3 years worth of episodes). We compare the EPA's determination of nonattainment counties and partial counties to the 50% confidence boundary of the BME hard and soft methodology nonattainment estimate. Adapting EPA's designation rule, we count an entire county as being in nonattainment if an area of that county is found to be in nonattainment by the continuous BME estimate. We find general agreement for 80% of the North Carolina counties for which we have data. Seven counties or partial counties designated by the EPA as nonattainment were not found to be in violation of the pseudo standard used for our analysis by the BME estimation method. Three of these had no monitoring stations, one had design value just below the standard, but they were determined to contribute to nearby violating monitoring stations by the EPA [2]. Seven counties were shown to violate our pseudo standard, but were not designated as nonattainment by the EPA. Only one of these had a monitoring station (which did not violate the real standard), the six others could not have a design value computed since they do not have monitoring stations[2]. One of these six counties (Anson) has an area that falls into the "highly likely in nonattainment" category in our analysis.

(a)



(b)

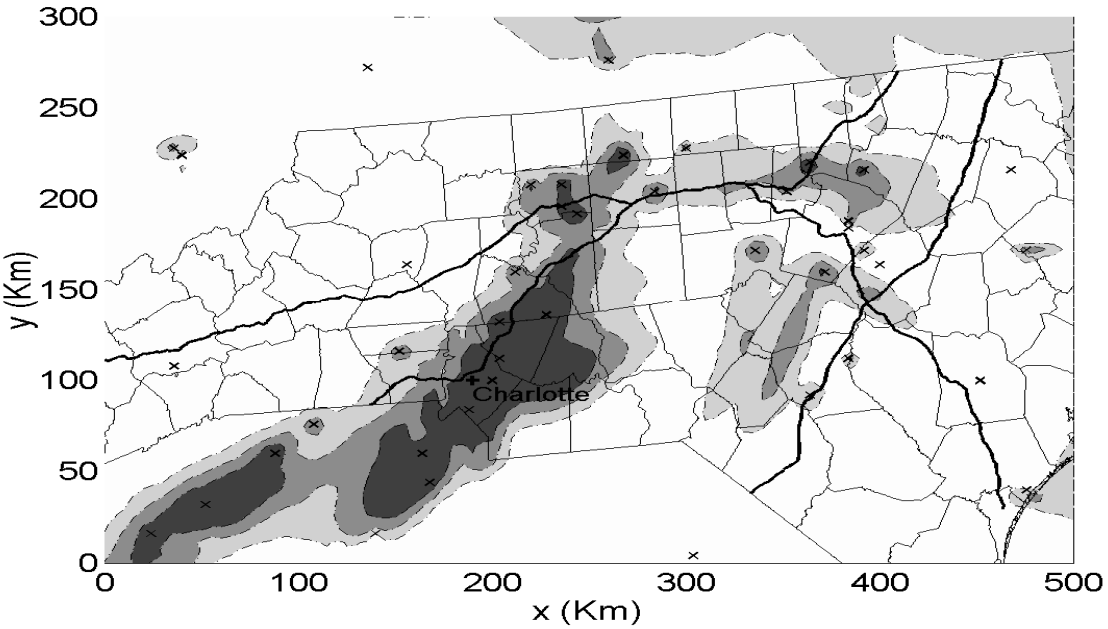


Figure S9. Non-attainment areas for the pseudo-standard of 2nd highest ozone level being above 0.08*ppm* obtained using estimation scenario (a) using only ozone observations, and estimation scenario (b) using both observations and model-predictions. The probability of violation (PV) of this standard are, from the darkest to the lightest shade of grey, PV>0.9 (highly likely in non-attainment), 0.5<Prob<0.9 (likely in non-attainment), 0.1<Prob<0.5 (near non-attainment), and Prob<0.1 (most likely in attainment)

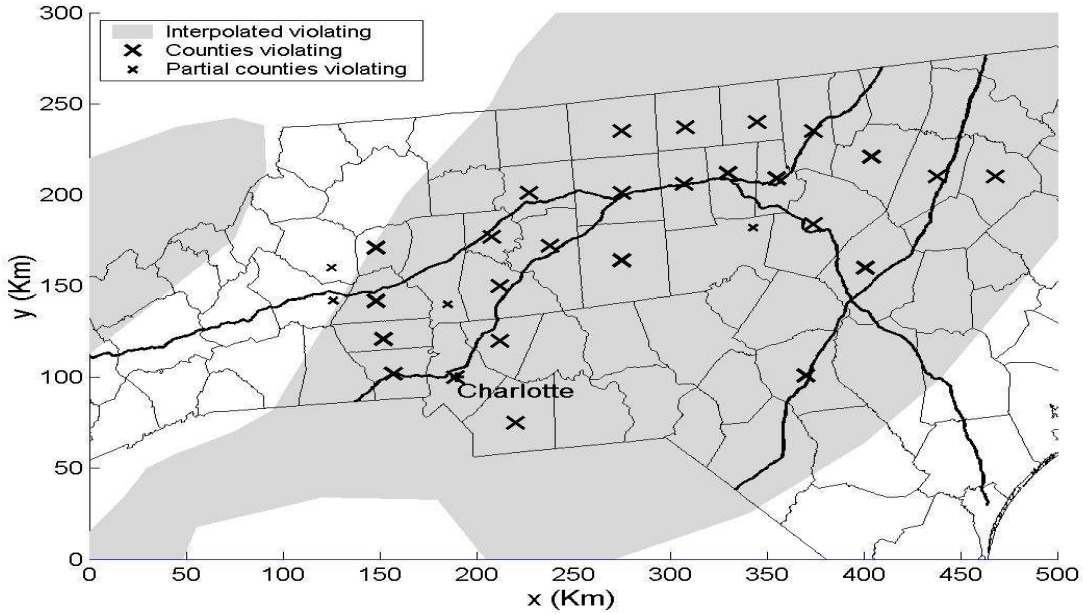


Figure S10. Reproduction of EPA maps of 2004 North Carolina 8-hour ozone nonattainment areas. The X marks indicate the counties classified in nonattainment by the EPA in 2004, smaller x marks representing partial counties classified in nonattainment [2]. The grey shade represents areas violating the standard estimated by the EPA by the Kriging methodology using 2001-2003 data [2].

References

1. Fuentes, M.; Raftery, A. E., Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **2005,** *61*, (1), 36-45.

2. US Environmental Protection Agency *8-Hour Ozone Nonattainment Areas* Office of Air Quality Planning and Standards, http://www.epa.gov/oar/oaqps/greenbk/o8index.html: 2004.