Published in final edited form as:
Environ Sci Technol. 2010 August 1; 44(15): 5707–5713. doi:10.1021/es100228w.

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], and **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

Abstract

States in the USA are required to demonstrate future compliance of criteria air pollutant standards by using both air quality monitors and model outputs. In the case of ozone, the demonstration tests aim at relying heavily on measured values, due to their perceived objectivity and enforceable quality. Weight given to numerical models is diminished by integrating them in the calculations only in a relative sense. For unmonitored locations, the EPA has suggested the use of a spatial interpolation technique to assign current values. We demonstrate that this approach may lead to erroneous assignments of non-attainment and may make it difficult for States to establish future compliance. We propose a method that combines different sources of information to map air pollution, using the Bayesian Maximum Entropy (BME) Framework. The approach gives precedence to measured values and integrates modeled data as a function of model performance. We demonstrate this approach in North Carolina, using the State's ozone monitoring network in combination with outputs from the Multiscale Air Quality Simulation Platform (MAQSIP) modeling system. We show that the BME data integration approach, compared to a spatial interpolation of measured data, improves the accuracy and the precision of ozone estimations across the State.

1. Introduction

Ozone is the main component of smog and a powerful oxidizing agent, associated with a wide range of adverse health outcomes including premature mortality, asthma exacerbation and hospital admissions for respiratory causes [1]. The Clean Air Act charged the U.S. Environmental Protection Agency (EPA) with establishing standards and implementation rules for six "criteria" pollutants, including ozone. Consistent with recommendations by its panel of science advisors (the Clean Air Scientific Advisory Committee) and based on its review of scientific evidence, the EPA is currently revising the National Ambient Air Quality Standard (NAAQS) for ozone established in 2008, deemed insufficiently protective of people, trees, and plants [2]. The EPA proposes to strengthen the "primary" ozone

[*] corresponding author: marc_serre@unc.edu, Tel: (919) 966 7014, Fax: (919) 966 7911.

Supporting Information
Further details on the data, methods, variance and cross-validation results, and attainment demonstration analyses can be found in the online supplementary material.

NAAQS, from 0.075 parts per million (ppm) to a level within the range of 0.060–0.070 ppm.

The form of the primary standard, established in 1997, is not disputed. To meet the standard, the 3-year average of the fourth highest 8-hour value at a particular monitor, known as the design value (DV), is not to exceed the established standard (0.080 ppm in 1997, 0.075 ppm in 2008). An important and unchallenged element of compliance demonstration instituted in 1997 is the reliance on observed ozone values for determining current and future attainment. As detailed later, the attainment guidance states how ozone model predictions are to be used in a relative rather than absolute sense, and in concert with measured values. That is, the ratios of the model's future to current (baseline) predictions at monitors, called the *relative response factors,* are multiplied with the observation-base "baseline" design value to predict the "future concentrations" which are then compared to the standard [3]. This attainment test in principle makes the process more forgiving of the potential inaccuracies of model outputs because it is based on "real" ambient concentration.

Reliance on observed data, however, poses a challenge for determining compliance in areas where no monitoring stations exist. The EPA suggested using interpolated measured values such as through Kriging [3], which naturally raises concerns since the method cannot account for chemistry and emission information. Un-monitored areas away from polluting sources - where monitoring stations are predominantly located - or rural and suburban areas where the monitors are usually sparsely located, may thus not reliably be assessed.

We propose an approach within a spatiotemporal epistemic knowledge processing framework that integrates observations and model predictions to obtain baseline design value, which we apply to North Carolina. Specifically, instead of using measured or modeled values by default, our approach, while giving preference to measured data, also uses model outputs as a function of model performance, thus making the integration more robust.

Examples exist of data integration efforts for air quality mapping making the best use of available information [4–9]. EPA guidance suggested combining modeled and monitored data to assess un-monitored areas. The recommended adjusted Voronoi Neighbor Averaging (eVNA) method interpolates neighbouring monitored values with a simple inverse-distance weighing scheme, and scales upwards or downwards given relative model predictions at the measured sites and the model prediction at location of interest [3]. This approach, while accounting in a limited way for the rich information contained in air quality models, suffers from the classic drawbacks of distance-weighing interpolation techniques, such as inability to provide error estimates, and exaggerated variability [10].

Statistical approaches provide a more rigorous framework to integrate observations and model predictions. An important member of these approaches is the Bayesian melding framework developed by Fuentes and Raftery [4]. That approach basically consists in taking a weighted average of the linear kriging estimator in a way that accounts for parameter uncertainty. However; the limitation of Bayesian melding, and the limitation of linear Bayesian hierarchical modeling approaches in general, is their reliance on Gaussian assumptions and linear models, which limits their estimation accuracy.

By contrast we use in this work the Bayesian Maximum Entropy (BME) method of modern Geostatistics. BME is a spatiotemporal epistemic knowledge processing framework that can integrate a wide variety of non-linear, non-Gaussian knowledge bases [11,12]. BME has been applied to map criteria pollutants previously [13–15]. In particular, spatiotemporal ozone maps were developed in California using 15 years of monthly measured values from the state's large monitoring network[15]. Importantly for exposure assessment, the BME maps were shown to be more accurate than linear kriging methods and provided important insights into the spatial distribution of seasonal patterns. However, to the best of our knowledge, none of the previous example of BME criteria pollutant mapping dealt with the problem of integrating ozone observations and model predictions, and this work is the first BME analysis to do so.

In our data integration approach, rather than rely on a Gaussian, linearized model to account for the change of support (as in Fuentes and Raftery[4]), we use a more transparent non-linear, non-Gaussian data-driven approach based on the direct comparison between model predictions and observations, which might be more acceptable within the contentious policy context. Our aim is to give precedence to monitoring data, thus we only weigh the model prediction according to how well it reproduces the observed values. A stochastic analysis is used to generate soft data from air quality model predictions, which is then processed by the BME method together with the hard observed data to provide a representation of space-time ozone distribution.

We begin by presenting the EPA compliance test, then our data, and finally present results comparing our approach to an analysis that only accounts for observed data. We conclude on the policy relevance of this work for ozone attainment demonstration.

2. Materials and Methods

2.1. Ozone attainment demonstration

The ozone compliance demonstration test accompanying the 8-hour primary standard formulation makes use of photochemical models to calculate *relative response factors* (RRF) that compare future-scenario to baseline modeled ozone levels [3]. Where monitoring stations exist, the RRF is applied to the measured current design value (DVC) to calculate the future design value (DVF) that is to be compared to the standard:

$$\mathrm{DVF}=\mathrm{RRF} * \mathrm{DVC}, \tag{1}$$

where RRF = (mean projected 8-hr daily max “near” monitor “x”)$_{\text{future}}$/(mean projected 8-hr daily max “near” monitor “x”)$_{\text{present}}$, and DVC is the 3-year average of the fourth highest 8-hour value observed at the monitor.

2.1. Ozone monitoring data

Ozone observations used in this work, retrieved from the U.S. EPA Aerometric Information Retrieval System (AIRS), were collected hourly at 46 sites across NC June 19th to June 30th 1996 (Fig. 1 and Fig. S1 in Supporting Information). This period represents one of the high 8-h ozone episodes extensively studied for NAAQS attainment demonstration in NC [16].

2.2 Air quality model predictions

The model data consists of ozone concentrations predicted by the Multiscale Air Quality Simulation Platform (MAQSIP) on a 4×4 km grid resolution covering North Carolina, for every hour of our study period. MAQSIP is a comprehensive urban- to intercontinental-scale atmospheric chemistry-transport model, developed in collaboration with the U.S. EPA; it has served as a prototype for the U.S. EPA’s Community Multiscale Air Quality (CMAQ) modeling system [17,18], and is under continuing development. For this study, we modeled only gas phase chemistry, and used CBM-IV with updated isoprene chemistry [19]. We modeled horizontal and vertical advection using the Bott scheme, horizontal diffusion using a constant K_h, and vertical diffusion based upon K-theory. We used one-way nesting

(36/12/4-km) in MAQSIP for all the episodes, and did not use plume-in-grid technique for any of the three resolutions. Detailed model specifications and evaluation are discussed elsewhere [16,20].

2.3 The proposed BME approach

As described in section 1 of Supporting Information (SI) the traditional Bayesian melding approach relies on linear and Gaussian assumptions. By contrast, the BME method is based on a knowledge processing framework grounded on sound epistemic principles that can process a wide variety of knowledge bases (stochastic physical laws, high order multi-point statistical moments, empirical relationships, non-Gaussian distributions, hard and soft data, etc.) that are beyond the reach of linear Geostatistics and classical spatial statistics. We only provide here the fundamental BME equations for the ozone mapping problem; the reader is referred to other works for descriptions of theoretical underpinnings used to derive these equations[11, 12] and of the numerical implementation in *BMElib* [21, 22].

The BME framework is based on a Space/Time Random Field (S/TRF) representation of ozone concentration $Z(\boldsymbol{p})$, where $\boldsymbol{p}=(\boldsymbol{s},t)$ is the space/time coordinate, $\boldsymbol{s}$ is the spatial location, and t is time. 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. $\boldsymbol{Z}=[Z_1,Z_2,\ldots]^T$ and $\boldsymbol{z}=[z_1,z_2,\ldots]^T$. The knowledge available is organized in the general knowledge base (G-KB) about the S/TRF (e.g. describing its space/time variability, physical laws, high order statistical moments, etc.), and the site-specific knowledge base (S-KB) corresponding to the hard and soft data available at a set of specific space/time data points $\boldsymbol{p}_d$. The BME fundamental set of equations is [23,24]

$$\left.\begin{array}{l}\int d\mathbf{z}(\mathbf{g}(\mathbf{z})-E[\mathbf{g}])e^{\mu^T\mathbf{g}(\mathbf{z})}=0\\ \int d\boldsymbol{\xi}_S(\mathbf{z})e^{\mu^T\mathbf{g}(\mathbf{z})}-Af_K(z_k)=0\end{array}\right\}, \tag{2}$$

where $\boldsymbol{z}$ is a vector of ozone concentrations at mapping points $\boldsymbol{p}$ consisting of the union of the data points $\boldsymbol{p}_d$ and the estimation point $\boldsymbol{p}_k$, $\boldsymbol{g}$ is a vector of functions selected such that their expected values E[$\boldsymbol{g}$] is known from the G –KB, $\xi_S[.]$ is an operator representing the S-KB, A is a normalization constant, and f_K is the BME posterior probability density function (PDF) describing ozone concentration z_k at the estimation point $\boldsymbol{p}_k$, where the subscript K = G U S means that f_K is based on the blending of the G– and S–KB.

The G–KB for the S/TRF $Z(\boldsymbol{p})$ corresponds to that obtained by extending the process model (eq S3) used in Fuentes and Raftery (see SI section1) to the space/time continuum, i.e. it consists in the space/time mean trend function $\mu(\boldsymbol{p;\beta})=E[Z(\boldsymbol{p})]$ parameterized on $\boldsymbol{\beta}$ and the space/time covariance function $c_Z(\boldsymbol{p},\boldsymbol{p}';\boldsymbol{\theta})=\mathrm{cov}(Z(\boldsymbol{p}),Z(\boldsymbol{p}'))$ parameterized on $\boldsymbol{\theta}$. The key conceptual difference between our work and that of Fuentes and Raftery [4] is how we treat the data models (S1–2) to obtain the S–KB. First, we restrict our application to a regulatory context, and as a result, because of the widely recognized legal precedent of using observations as enforceable evidence to calculate the DVC (Eq. 1), we treat the vector of observations $\hat{\boldsymbol{z}}_o$ as an exact value for the vector of random variables $\boldsymbol{Z}_o$ representing ozone $Z(\boldsymbol{p})$ at the set of points $\boldsymbol{p}_o$ where the observations were taken, i.e. we have $Prob.[\boldsymbol{Z}_o=\hat{\boldsymbol{z}}_o]=1$, and we will thereafter refer to $\hat{\boldsymbol{z}}_o$ as the "hard" data. Second, we do not restrict the complex stochastic relationship between model prediction $\tilde{Z}(B)$ and ozone concentration $Z(\boldsymbol{p})$ to the Gaussian linearized model (S2). Instead, we refer to the vector of model prediction values $\tilde{\boldsymbol{z}}_m$ as soft data for the vector of random variables $\boldsymbol{Z}_m$ representing ozone $Z(\boldsymbol{p})$ at the set of points $\boldsymbol{p}_m$ corresponding to the centroid of the computational nodes $\{B\}$ for which the model predictions were obtained. In other words, we do not make any restrictive assumptions about the PDF $f_S(\boldsymbol{z}_m)$ characterizing $\boldsymbol{Z}_m$ given knowledge provided by the model predictions $\tilde{\boldsymbol{z}}_m$.

Hence, overall, the knowledge bases considered consist of G = $\{\mu(.,\boldsymbol{\beta}), c_Z(.,\boldsymbol{\theta})\}$ and S = $\{\hat{\boldsymbol{z}}_o, f_S(.)\}$. In this case the BME fundamental set [25] of equations (2) reduces to

$$f_K(z_k)=A^{-1}\int dz_m f_S(z_m) f_G(z), \tag{3}$$

where $f_G(\boldsymbol{z}) = e^{\mu^T \boldsymbol{g}(z)}$ is the multivariate Gaussian PDF for $(Z_k,\boldsymbol{Z}_o,\boldsymbol{Z}_m)$ obtained from the G–KB, $\boldsymbol{z} = (z_k, \hat{\boldsymbol{z}}_o, \boldsymbol{z}_m)$ is a realization of $(Z_k,\boldsymbol{Z}_o,\boldsymbol{Z}_m)$, $\hat{\boldsymbol{z}}_o$ is the hard observed data, z_k and $\boldsymbol{z}_m$ are dummy variables that can take any value, $f_S(\boldsymbol{z}_m)$ is the (generally non-Gaussian) PDF of $\boldsymbol{Z}_m$ given knowledge of the soft data $\tilde{\boldsymbol{z}}_m$, and the normalization constant is $A = \int dz_k \int d\boldsymbol{z}_m\, f_S(\boldsymbol{z}_m)\, f_G(\boldsymbol{z})$. When the parameters $(\boldsymbol{\beta},\boldsymbol{\theta})$ are known, $f_G(\mathbf{z})$ is simply equal to a multivariate normal PDF with mean and covariance given by $(\boldsymbol{\beta},\boldsymbol{\theta})$, i.e. $f_G(\mathbf{z}) = f(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\mathbf{z}}_o)$, where the conditioning on $\hat{\boldsymbol{z}}_o$ was noted to reflect that these values are known. When $(\boldsymbol{\beta},\boldsymbol{\theta})$ are not known, we can remove the conditionalization on these parameters by taking the marginal PDF of $f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \hat{\mathbf{z}}_o) = f(\mathbf{z} \mid \hat{\mathbf{z}}_o, \boldsymbol{\beta}, \boldsymbol{\theta})\, f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \hat{\mathbf{z}}_o)$ with respect to $\boldsymbol{z}$, which leads to

$$f_G(\mathbf{z})=\int d\beta \int d\theta f(\mathbf{z}|\beta,\theta,\widehat{\mathbf{z}}_o) f(\beta,\theta|\widehat{\mathbf{z}}_o), \tag{4}$$

where $f(\boldsymbol{\beta}, \boldsymbol{\theta}| \hat{\boldsymbol{z}}_o)$ is fully defined by the likelihood $f(\hat{z}_o \mid \boldsymbol{\beta}, \boldsymbol{\theta})$ and some prior PDF $f(\boldsymbol{\beta}, \boldsymbol{\theta})$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. However we found through simulations that the approximation

$$f_G(z) \approx f(z|\widehat{\beta},\widehat{\theta},\widehat{z}_o), \tag{5}$$

where $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})$ are estimates of $(\boldsymbol{\beta},\boldsymbol{\theta})$, is reasonable (SI section 2). Since there is no noticeable numerical difference between Eqs. (4) and (5), we use Eq. (5) with least square estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ that are found to be physically meaningful from the exploratory analysis of the observed data $\hat{\boldsymbol{z}}_o$.

Eqs. (3) and (5) provide a solution to our problem once we specify how we obtain $f_S(\boldsymbol{z}_m)$ from the soft data $\tilde{\boldsymbol{z}}_m$, which we present next.

2.4 Generation of the PDF characterizing the performance of the air quality model

The PDF $f_S(\boldsymbol{z}_m)$ essentially characterizes how well the air quality model predicts observed ozone values. As explained above, $f_S(\boldsymbol{z}_m)$ is the PDF for $\boldsymbol{Z}_m$ conditional solely on the knowledge of the model predictions $\tilde{\boldsymbol{z}}_m$, hence, we may write

$$f_S(\mathbf{z}_m)=\prod_i^{n_m} f(z_i|\tilde{z}_i, \mathbf{p}_i), \tag{6}$$

where n_m is the number of computational nodes, $f(z_i \mid \tilde{z}_i, \boldsymbol{p}_i)$ s the PDF for the ozone concentration Z_i at the centroid $\boldsymbol{p}_i \in \boldsymbol{p}_m$ of a particular computational node $B_i \in \{B\}$, $\boldsymbol{z}_m$ is a vector of size n_m and with value z_i at the i-th computational node, $\tilde{z}_i$ is the computer model prediction at the i-th node, and the conditioning is made explicit on the space/time location $\boldsymbol{p}_i$ to indicate that the performance of the air quality model may vary across space or time.

As explained above, a key conceptual aspect of our work is that our proposed framework does not impose any restriction on the form of the PDFs $f(z_i \mid \tilde{z}_i, \boldsymbol{p}_i)$. This is an important advantage over approaches restricted to using normal distributions, which as noted by Riccio *et al.* [6], can be a problem because they may allow negative ozone concentrations that cannot occur in reality.

Our approach is to use a parameterized statistical distribution $f(z_i \mid \tilde{z}_i, \boldsymbol{p}_i) = \Phi(z_i; \boldsymbol{\lambda}(\tilde{z}_i, \boldsymbol{p}_i))$, where $\Phi(.)$ may be an exponential, truncated normal, Johnson S_{BB}, etc. distribution with parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$. Then, the problem becomes that of estimating $\boldsymbol{\lambda}(\tilde{z}_i, \boldsymbol{p}_i)$ rather than $f(z_i \mid \tilde{z}_i, \boldsymbol{p}_i)$.

An analysis of observations and their model predictions in our datasets indicated that a good choice for $\Phi(.)$ is the normal distribution truncated below zero (Fig. S7) with an expected value λ_1 and variance λ_2 that are only a function of model prediction (i.e. there was no evidence that the prediction errors changed across space/time). Hence without loss of generality, we can assume that

$$f(z_i|\tilde{z}_i, \mathbf{p}_i) = \Phi(z_i; \lambda_1(\tilde{z}_i), \lambda_2(\tilde{z}_i)), \tag{7}$$

where $\Phi(.)$ will thereafter refer to the normal distribution truncated below zero.

Estimators for $\lambda_1(\tilde{z}_i)$ and $\lambda_2(\tilde{z}_i)$ are

$$\left.\begin{array}{l} \widehat{\lambda}_1(\tilde{z}_i) = \frac{1}{n_o(\tilde{z}_i)} \sum_{j=1}^{n_o(\tilde{z}_i)} \widehat{z}_j \\ \widehat{\lambda}_2(\tilde{z}_i) = \frac{1}{n_o(\tilde{z}_i)} \sum_{j=1}^{n_o(\tilde{z}_i)} \left(\widehat{z}_j - \widehat{\lambda}_1(\tilde{z}_i)\right)^2 \end{array}\right\}, \tag{8}$$

where only co-located observed and model ($\hat{z}_i$ and $\tilde{z}_i$) values are considered, $n_o(\tilde{z}_i)$ is the number of observed values $\hat{z}_j$ throughout the study area for which the co-located model prediction value $\tilde{z}_j$ was within some tolerance Δz of $\tilde{z}_i$, i.e. such that $\tilde{z}_i - \Delta z \le \tilde{z}_j \le \tilde{z}_i + \Delta z$. A small tolerance Δz was selected and progressively increased so as to eliminate the noise in the $\hat{\lambda}_1(\tilde{z}_i)$ and $\hat{\lambda}_2(\tilde{z}_i)$ relationships.

Hence Eqs. (6–8) provides the soft data PDF $f_S(\boldsymbol{z}_m)$ needed to calculate the BME posterior PDF $f_K(z_k)$ (Eq. 3 and 5) of the ozone concentration Z_k at any estimation point $\boldsymbol{s}_k$ of interest, from which the expected value z_k^* provides a BME estimate of ozone concentration that is relevant for the calculation of the DVC (Eq. 1) in a regulatory spatiotemporal context, and the standard deviation σ_k^* provides an assessment of associated estimation uncertainty.

2.5 Cross-Validation

Cross-validation is used to compare the accuracy of different BME estimation scenarios. Each observed value $\hat{z}_j$ at space/time point $\boldsymbol{p}_j = (\boldsymbol{s}_j, t_j)$ is considered one at a time, and the corresponding ozone concentration Z_j is re-estimated using only non collocated data outside of a radius r_V of $\boldsymbol{s}_j$, i.e. none of the data collected any time within a radius of $\boldsymbol{s}_j$ is considered in the estimation of Z_j. The cross-validation estimate $z_j^*(r_V)$ obtained is a function of the cross-validation radius r_V. The cross-validation errors $e_j^*(r_V) = z_j^*(r_V) - \widehat{z}_j$ can then be summarized in terms of their mean square error $MSE(r_V)$ equal to the arithmetic average of estimation errors squared, i.e.

$$MSE(r_V)=\frac{1}{n_o}\sum_{j=1}^{n_o}\left(z_j^*(r_V)-\widehat{z}_j\right)^2 \tag{9}$$

where n_o is the number of observed ozone values $\hat{z}_j$ throughout the study domain. $MSE(r_V)$ varies as a function of the cross-validation radius r_V considered. We are particularly interested to learn how an estimation scenario performs when estimating the ozone concentration away from any monitoring station, which we may do by calculating $MSE(r_V)$ for large cross-validation radii r_V.

Using the cross-validation MSE we compare two estimation scenarios consisting in scenario (a) using ozone observations only and scenario (b) using both observations and model predictions. We let *MSEh* and *MSEs* be the mean square error for scenario (a) (hard data only) and scenario (b) (hard and soft data), and we define the percent change in mean square error *PCMSE* as

$$PCMSE(r_V)=100\frac{MSEs(r_V)-MSEh(r_V)}{MSEh(r_V)} \tag{10}$$

A negative *PCMSE* indicates a decrease in *MSE*, which corresponds to the percent improvement in estimation accuracy resulting in integrating model predictions in the estimation of the ozone concentration. (See SI section6 for significance test derivations).

3. Results

3.1. Mean trend model

As described above, the mean trend model $\mu(\boldsymbol{p};\boldsymbol{\beta})$=E[$Z(\boldsymbol{p})$] is a function of space and time parameterized on $\boldsymbol{\beta}$ that captures systematic trends in ozone concentration. A simple approach is to model this mean trend as a function of some covariates, e.g. a polynomial function of latitude, longitude and time. However, similarly to Riccio *et al.*[6], we found that that simple mean trend model was not adequate because it failed to capture large area spatial and temporal patterns. We found that a better model for the mean trend is

$$\mu(\boldsymbol{p};\beta)=\mu_s(\boldsymbol{s};\beta_s)+\mu_t(t;\beta_t)+\mu^{\boldsymbol{\prime}}(\boldsymbol{p};\beta^{\boldsymbol{\prime}}) \tag{11}$$

where μ_s (***s***; $\boldsymbol{\beta}_s$) captures consistent geographical trends in ozone concentration, μ_t (t; $\boldsymbol{\beta}_t$) captures the strong statewide daily cyclic pattern in ozone concentration, and μ' (***p***; $\boldsymbol{\beta}'$) is a polynomial function of space and time. We obtain μ_s (s; $\boldsymbol{\beta}_s$) and μ_t (t; $\boldsymbol{\beta}_t$) using an exponential kernel smoothing of the time-averaged and spatially-averaged data. Following Akita et al.[26], an exploratory analysis was used to select kernel smoothing parameters resulting in a physically meaningful mean trend model (see for e.g. Fig. S3). We then found no evidence that adding a space/time polynomial function μ' (***p***; $\boldsymbol{\beta}'$) decreases the cross validation MSE, so we set μ' (***p***; $\boldsymbol{\beta}'$) =0.

Three stations located in eastern NC mountainous areas were shown to be governed by very specific and localized inversed ozone cyclic patterns, and were subsequently excluded from the analysis. Once these areas were removed, the mean trend model selected, illustrated in Fig. S3 for a single station, was found to provide a good representation of systematic trends for our study domain.

3.2. Space-time covariance model

According to the “process” model (Eq. S3) extended to the space/time continuum, $Z'(\boldsymbol{p})=Z(\boldsymbol{p})-\mu(\boldsymbol{p};\boldsymbol{\beta})$ is a zero-mean residual S/TRF with space/time covariance function $c_Z(\boldsymbol{p},\boldsymbol{p}';\boldsymbol{\theta})=\mathrm{cov}(Z(\boldsymbol{p}),Z(\boldsymbol{p}'))$ parameterized on $\boldsymbol{\theta}$. Since $\mu(\boldsymbol{p};\boldsymbol{\beta})$ captures the non-homogenous/non-stationary trend in ozone, it is reasonable to assume that the covariance is homogeneous/stationary, i.e. that the covariance between points $\boldsymbol{p}=(\boldsymbol{s}, t)$ and $\boldsymbol{p}' =(\boldsymbol{s}',t')$ is only a function of the spatial lag $r=\|\boldsymbol{s}-\boldsymbol{s}'\|$ and the temporal lag $\tau =|t-t'|$;

$$c_Z(\boldsymbol{p},\boldsymbol{p}';\theta)=c_Z(r=\left\|\boldsymbol{s}-\boldsymbol{s}'\right\|,\tau=|t-t'|;\theta). \tag{12}$$

Examination of the residual data $\hat{z}_j-\mu(\boldsymbol{p}_j,\boldsymbol{\beta})$, $j=1,\ldots,n_o$ did not provide any evidence rejecting this covariance model.

Using Eq. (12) we obtained experimental covariance values $c_Z(r, \tau)$ which we then used to fit a space/time separable covariance model, detailed in SI section3.

3.3. Stochastic analysis of the air quality model performance

The model performance analysis (Fig. 2) shows that the photochemical air quality model on average underestimates the measured values for very low ozone concentrations, and on average overestimates values for levels above approximately 0.03*ppm*. However at low levels fewer data points exist and they are widely dispersed. Fig. 2 illustrates the steps described in the methods section (Eqs. 7–8) to obtain the PDF $f\,(z\mid\tilde{z})$ describing ozone concentration z given a model prediction value $\tilde{z}$. (see Fig. S6 for the corresponding variance var[Z| $\tilde{z}$]). These PDFs are then used in Eq. (6) to construct the soft data PDF $f_S(\boldsymbol{z}_m)$. We note that higher predicted ozone levels (above approximately 0.07*ppm*) have lower error variance, so that greater confidence can be assigned to the soft data generated using these values.

3.4. Ozone estimates

Maps of the BME mean estimate of hourly ozone concentration z_k^* are obtained for the two estimation scenarios: (a) using observations only and (b) using both observations and model predictions, as shown in Fig. 3 for hour 250. It can be seen from this figure that ozone levels in the immediate proximity of the monitoring stations (‘x’ marks) are the same in both maps, but then the concentration gradient is steeper moving away from the stations for estimation scenario (b), resulting in levels dropping lower far away from any stations.

Corresponding BME error variance maps are shown in Fig. S8. We find that the two scenarios provide quite different results, with error variances remaining relatively low in areas where no stations exist for the estimation scenario (b) (except on the map borders where the model predictions are not available). Thus, estimates of ozone away from monitoring stations are found to be more accurate (and substantially different) when integrating both observations and model predictions.

3.5. Cross-validation

Fig. 4 and Table 1 show the percent change in MSE (Eq. 10) as a function of cross-validation radius r_V. Let's consider first the Fig. 4 curve labeled "cut=0*ppm*", which includes all observations in the cross-validation dataset. As can be seen, the *PCMSE* is consistently negative, indicating that integrating both observations and model predictions is consistently more accurate than relying solely on observations. Furthermore, the percent reduction in MSE consistently improves as r_V increases: there is only a 1.5% reduction in error for r_V=1.5*km*, while there is at least a 28% reduction in error at locations more than 100 km away from a monitoring station. The MSE reduction is highly significant (p-value<0.0001) for $r_V \geq 30$ *km* (Table 1).

We found that the reduction in MSE was even greater when restricting our cross-validation to high ozone observations, as illustrated in Fig. 4 for observations greater than 0.04*ppm* and 0.07*ppm*. This was expected since the air quality model was shown to perform better for higher concentrations.

3. Discussion

We have presented an ozone mapping approach that integrates predictions weighed according to model performance and observations treated as an error-free proxy for ozone concentration. This framework uses observations as hard data, and uses model predictions to construct soft data (Eqs. 6–8). We thus produce estimates that put priority on observations and take advantage of model prediction only to the extent that they are deemed accurate with respect to observed values. Importantly in this regulatory context, spatial fields generated provide an observation-driven representation of ozone across space/time that is more accurate and precise than those produced using observation data only, especially for locations distant from monitoring stations and at higher concentrations. A strength of our approach is that the mapping error variance can reflect various forms of uncertainty –such as distance to monitoring stations and variance in model performance as a function of ozone prediction levels (Fig. S5), so that the mapping accuracy is meaningful everywhere. The better performance of the model is explained by the historical purpose of grid-based air quality models developed to study extreme events, i.e. periods of high air pollution. In fact, guidelines for evaluating air quality model performance for ozone typically used a cutoff value of 60 ppb in the past[27]. Since this application the model has been updated to perform better during low observed concentrations[28], and could be integrated in future work.

To assess implications in determining ozone attainment areas, we performed additional analyses (SI section7) in which we find clear consistency between the two estimation scenarios in determining the core of non-attainment areas, and disagreements on the extent of the area in non-attainment. Our results validate concerns regarding an observed-ozone spatial interpolation methodology to assess current design values (DVC). In un-monitored areas where interpolation gives an artificially high DVC, if no major sources are present and thus predicted ozone reductions are relatively insignificant the *relative response factors* are necessarily close to 1. This would make future compliance difficult to demonstrate if the DVC is above the standard, since, applying Eq (1), the future design value would be close to the DVC.

Further, the BME methodology offers information on the precision of ozone estimates, providing opportunities for more tailored approaches to ozone compliance (see illustrations of probability of attainment scenarios in SI section7). The EPA could reward greater precision, useful in particular for developing targeted health-protective policies, by establishing criteria for granting reclassification requests based on the level of confidence associated with estimates. In addition, the error variance assessment can be particularly useful for designing the spatial distribution of monitoring networks. Such considerations are particularly relevant in the context of the current ozone NAAQS review process which may include modifications in the monitoring network design [2].

Finally, our data integration framework provides a novel approach to account for the complex non-Gaussian, non-linear stochastic relationship between observed values and model predictions. One limitation of our current application is the 4-km grid resolution of the air quality model, too low to assess exposures near scavenging sources. To study ozone impacts near roadways and high traffic densities using models at this resolution, one can use sophisticated techniques such as variable-grid resolutions or adaptive grid resolutions in air quality models, where the grid resolution can be a lot finer than 1-km, perhaps in the hundreds of meters range. There is no limitation in the type of information the BME framework may integrate, hence it would be an interesting future application to develop higher resolution ozone maps using such models. Similarly, the methodology could usefully be applied for health assessments or regulatory purposes to other criteria pollutants that are routinely monitored, extensively modeled, and display spatial-temporal patterns. Particulate matter, for example, is of great health concern and is strongly influenced by emission sources, and as such would benefit from our mapping approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge Harvey Jeffries, and are thankful for his feedback. This work was supported by grants from the National Institute of Environmental Health Sciences (grants P30ES10126 and 5-P42-ES005948). The MAQSIP modeling was performed under funding from the North Carolina Department of Environment and Natural Resources under Modeling Assistance Contracts EA2012, EA03017 and EA05009 to the UNC Institute for the Environment.

References

1. Smith KR, Jerrett M, Anderson HR, Burnett RT, Stone V, Derwent R, Atkinson RW, Cohen A, Shonkoff SB, Krewski D, Pope CA 3rd, Thun MJ, Thurston G. Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants. Lancet. 2009
2. Agency, EP., editor. National Ambient Air Quality Standard for Ozone: Proposed Rule. Vol. 75. 2010. Federal Register.
3. U.S. EPA . Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, PM2.5, and Regional Haze; EPA-454/B-07-002. Office of Air Quality, Planning and Standards; Research Triangle Park, NC: 2007.
4. Fuentes M, Raftery AE. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 2005;61(1):36–45. [PubMed: 15737076]
5. Van de Kassteele J, Stein A. A Model for External Drift Kriging with Uncertain Covariates Applied to Air Quality Measurements and Dispersion Model Output. Environmetrics 2005;17(4):309–322.
6. Riccio A, Barone G, Chianese E, Giunta G. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. Atmospheric Environment 2006;40(3):554–566.
7. Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall; New York: 2004.
8. Foley KM, Fuentes M. A Statistical Framework to Combine Multivariate Spatial Data and Physical Models for Hurricane Surface Wind Prediction. Journal of Agricultural, Biological, and Environmental Statistics 2008;13(1):37–59.

9. Wikle CK, Berliner LM, Cressie N. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 1998;5(2):117–154.

10. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Analysis and Environmental Epidemiology 2005;15(2):185–204. [PubMed: 15292906]

11. Christakos G. A Bayesian/Maximum-Entropy View To The Spatial Estimation Problem. Mathematical Geology 1990;22(7):763–776.

12. Christakos, G.; Hristopulos, DT. Spatiotemporal Environmental Health Modelling: A Tractatus Stochasticus. Kluwer Academic Publ; Boston, MA: 1998.

13. Christakos G, Serre ML. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmospheric Environment 2000;34(20):3393.

14. Yu HL, Chen JC, Christakos G, Jerrett M. BME estimation of residential exposure to ambient PM10 and ozone at multiple time scales. Environ Health Perspect 2009;117(4):537–44. [PubMed: 19440491]

15. Bogaert P, Christakos G, Jerrett M, Yu HL. Spatiotemporal modelling of ozone distribution in the State of California. Atmospheric Environment 2009;43(15):2471–2480.

16. Arunachalam S, Holland A, Do B, Abraczinskas M. A Quantitative Assessment of the Influence of Grid Resolution on Predictions of Future-Year Air Quality in North Carolina, USA. Atmospheric Environment' 2006;40(26):5010–5016.

17. Byun DW, Schere KL. Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. J Applied Mechanics Reviews 2006;59(51)

18. Byun, DW.; Ching, JKSE. Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. 1999.

19. Carter WPL, Atkinson R. Development and evaluation of a detailed mechanism for the atmospheric reactions of isoprene and NOx. Int J Chem Kinet 1996;28:497–530.

20. Arunachalam, SR.; Mathur, Z.; Adelman, D.; Olerud, J.; Holland, A. A Comparison of Models-3/CMAQ and the MAQSIP modeling systems for Ozone Modeling in North Carolina. 94th Annual Meeting of the A&WMA; Orlando, FL. 2001.

21. Serre ML, Christakos G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. Stochastic Environmental Research and Risk Assessment 1999;13(1–2):1–26.

22. Christakos, G.; Bogaert, P.; Serre, ML. Temporal GIS: Advanced Functions for Field-Based Applications. Springer-Verlag; New York, N.Y: 2002. p. 217

23. Christakos, G. Modern Spatiotemporal Geostatistics. Oxford university press; 2000.

24. Christakos, G. Chapter 6: Bayesian Maximum Entropy. In: Kanevski, M., editor. Advanced Mapping of Environmental Data: Geostatistics, Machine Learning, and Bayesian Maximum Entropy. J. Wiley & Sons; New York, NY: 2008. p. 247-306.

25. Christakos G, Serre ML. Spatiotemporal analysis of environmental exposure-health effect associations. J Expo Anal Environ Epidemiol 2000;10(2):168–87. [PubMed: 10791598]

26. Akita Y, Carter G, Serre ML. Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. J Environ Qual 2007;36(2):508–20. [PubMed: 17332255]

27. Russell A, Dennis R. NARSTO critical review of photochemical models and modeling. Atmospheric Environment 2000;34(12–14):2283–2324.

28. Appel KW, Gilliland AB, Sarwar G, Gilliam RC. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance: Part I--Ozone. Atmospheric Environment 2007;41(40):9603–9615.

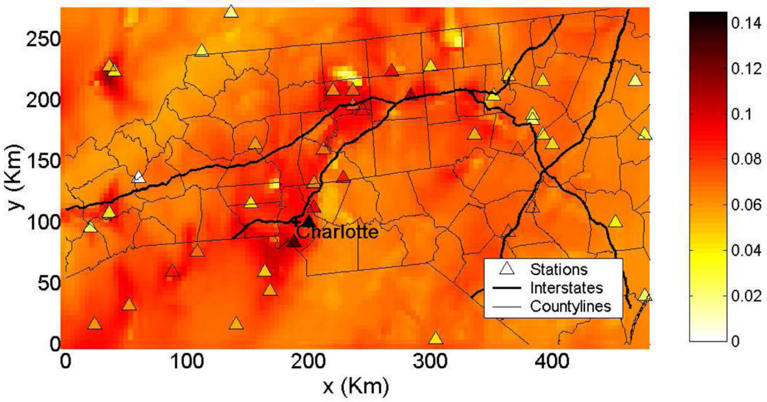

Figure 1.

Observed and model prediction of ozone concentrations (*ppm*) across North Carolina at hour 250 of the study period. The background color depicts model predictions at a 4×4*km* grid resolution. Triangles represent concentrations observed at monitoring stations.

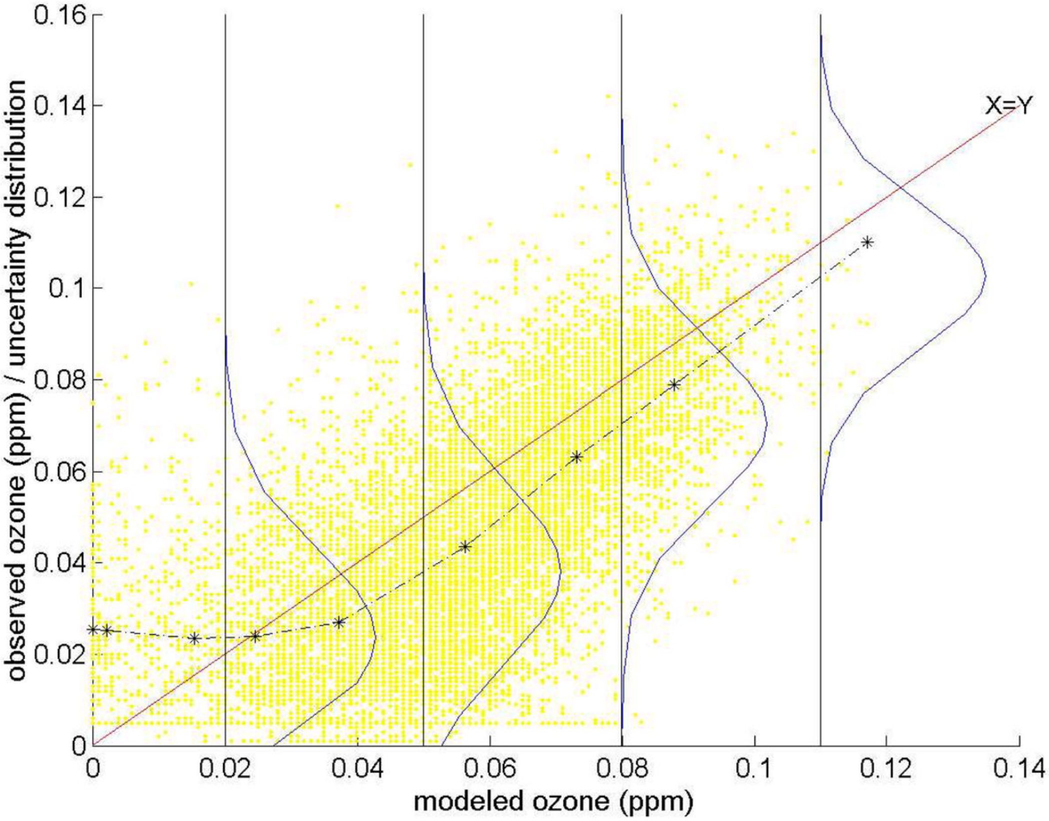

Figure 2.

Dots show ozone observed values $\hat{z}_j$ against their colocated model prediction values $\tilde{z}$. The stars-and-dashed lines show the estimator $\hat{\lambda}_1(\tilde{z})$ (Eq. 17) of the expected value E[Z| $\tilde{z}$] of observed ozone concentration given a model prediction $\tilde{z}$. The four PDFs shown are truncated normal PDF representing $f(z \mid \tilde{z})$ (Eq. 16) for $\tilde{z}$ =(0.02, 0.05, 0.08 and 0.11*ppm*).

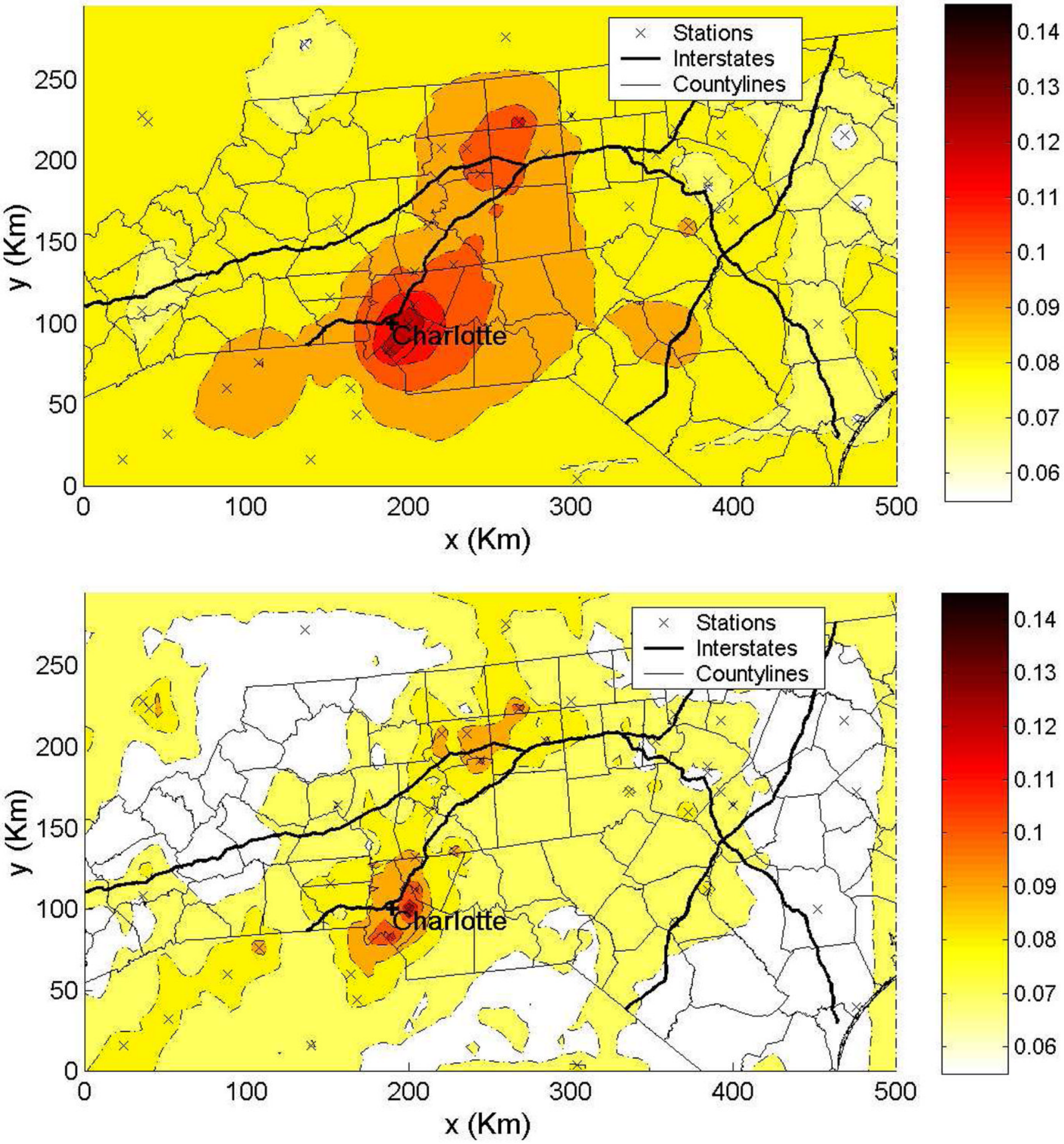

Figure 3.

BME mean estimate of ozone (*ppm*) at hour 250 obtained for two estimation scenarios consisting in **(a)** using ozone observations only and **(b)** using both ozone observations and model predictions, for hour 250. Monitoring stations are represented by “x” marks.

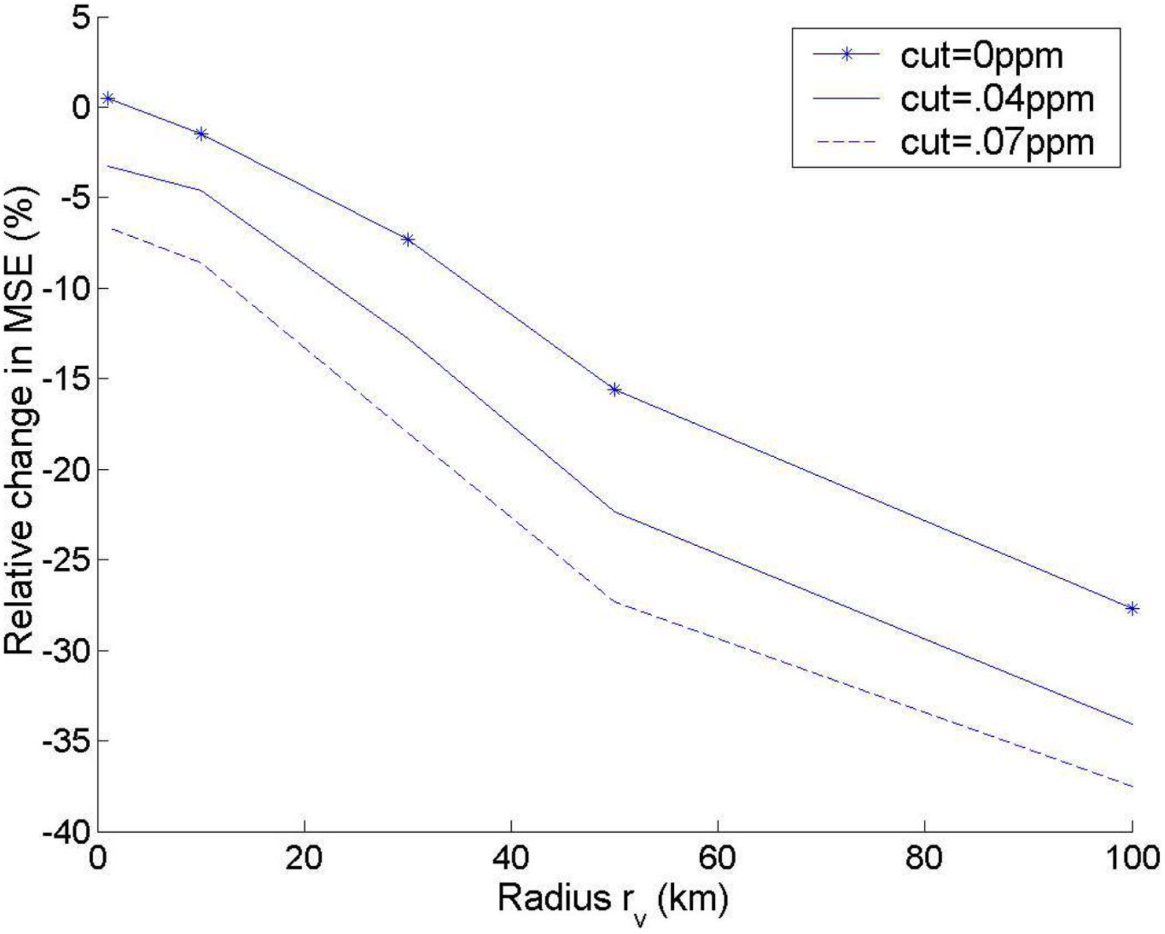

Figure 4.

Percent change in mean square error *PCMSE* (Eq. 19) shown as a function of cross-validation radius r_V. Each curve corresponds to the *PCMSE* calculated using only observations above a given cutoff (0*ppm*, 0.04*ppm*, and 0.07*ppm*) of all observations values.

Table 1

Statistics to compare estimation methods using hard and soft data versus using hard data only. The analysis uses a constant mean trend, and compares statistics for different cross-validation radii of exclusion points

Stat	**r_v=10**	**r_v=30**	**r_v=50**	**r_v=100**
MEs	−1.1407E-03	−1.0634E-03	−9.1051E-04	−8.9985E-04
MEh	−4.5917E-04	−2.7240E-04	5.0473E-05	1.6704E-04
MSEs	2.5634E-04	2.8936E-04	3.3093E-04	4.1262E-04
MSEh	2.6021E-04	3.1216E-04	3.9202E-04	5.7070E-04
PCMSE	−1.4863	−7.3057	−15.5833	−27.6988
MaxRE	−27.2217	−27.2217	−33.8706	−44.2374
Pval	0.2117	<0.0001	<0.0001	<0.0001
P01	4.6512	13.9535	18.6047	74.4186
P05	13.9535	23.2558	44.1861	90.6977
P1	16.2791	27.9070	53.4884	100.0000

r_V is the cross-validation radius (*km*) around monitoring stations within which all observation points are excluded in the cross-validation estimation

MEs is the mean error using soft data

MSEs is the mean square error using soft data

MEh is the mean error using hard data only

MSEh is the mean square error using hard data only

PCMSE is the percent change in mean square error = (MSEs-MSEh)/MSEh*100;

MaxRE is the maximum achieved relative reduction error at a station (soft compared to hard)

Pval is the p-value testing the hypothesis that there is no difference between the mean square errors using soft vs. hard data

P01 is the percentage of stations with a significant reduction in error with a p value less or equal to 0.01

P05 is the percentage of stations with a significant reduction in error with a p value less or equal to 0.05

P1 is the percentage of stations with a significant reduction in error with a p value less or equal to 0.1