A Comparison of Statistical Techniques for Combining Modeled and Observed Concentrations to Create High-Resolution Ozone Air Quality Surfaces

Valerie C. Garcia
U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

Kristen M. Foley
U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

Edith Gego
Gego Associates, Idaho Falls, ID

David M. Holland
U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

S. Trivikrama Rao
U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

ABSTRACT

Air quality surfaces representing pollutant concentrations across space and time are needed for a multitude of applications, including tracking trends and relating air quality to human and ecosystem health. The spatial and temporal characteristics of these surfaces may reveal new information about the associations between emissions, pollution levels, and human exposure and health outcomes that may not have been discernable before. This paper presents four techniques, ranging from simple to complex, to statistically combine observed and modeled daily maximum 8-hr ozone concentrations for a domain covering the greater New York State

area for the summer of 2001. Cross-validation results indicate that, for the domain and time period studied, the simpler techniques (additive and multiplicative bias adjustment) perform as well as or better than the more complex techniques. However, the spatial analyses of the resulting ozone concentration surfaces revealed some problems with these simpler techniques in limited areas where the model exhibits difficulty in simulating the complex features such as those observed in the New York City area.

IMPLICATIONS

Linking emission control actions to human health impacts is important in determining whether the regulations that have been implemented are reducing air pollution as intended. Measurements of pollutant concentration levels are often spatially sparse, and modeled outputs are only an estimate of the "true" pollutant concentration levels, hampering our ability to detect a relatively small signal of change embedded in ambient concentrations. This paper assesses four techniques to combine the strengths of modeled and observed data to provide high-resolution ozone concentration surface maps for use in human health studies and assessing whether regulatory control actions have had the intended impact.

INTRODUCTION

Air pollutant concentrations across space and time are used in a multitude of applications, including tracking trends and relating air quality to human and ecosystem health. Often, changes in air quality attributable to emission reductions stemming from control policies are weak signals within the overall changes in observed or modeled concentrations. This signal can be further confounded when investigating the impacts of emission reductions on human health. Although air quality observations taken at various locations represent the "ground truth", these observations are often limited in terms of spatial and temporal coverage. Air quality models can predict pollutant concentrations over a given spatial domain, but the modeled values are uncertain due to model input errors and the model's inability to perfectly simulate the various physical and chemical processes occurring in the atmosphere. To alleviate these problems, four techniques that statistically combine air quality measurements with model output to produce high-resolution ambient pollutant concentrations are considered here to better characterize air quality in a study area encompassing New York State for the 2001 ozone season. These

63 improved air quality surfaces may reveal associations between pollution levels and health
64 outcomes not discernable before.[1]
65 Several investigators [2-7] have applied techniques, such as the Kalman Filter and ensemble
66 approaches, to adjust forecasted meteorological variables and air pollution concentrations using
67 observations. Gego et al.[8] and Hogrefe et al.[9] applied bias-adjustment approaches to predict air
68 pollutant concentrations for use in health studies. Hirst et al.[10] used a hierarchical modeling
69 approach to combine measured and modeled deposition values to estimate long-range transport
70 of air pollutants in Europe by modeling the underlying (unobserved) "true" deposition process as
71 a function of two stochastic components; one non-stationary and correlated over long distances,
72 and the other describing variation within a grid square. Fuentes and Raftery [11] used a Bayesian
73 approach to combine observed and modeled dry deposition pollutant concentrations to improve
74 spatial predictions, evaluate the model and remove the model bias. Similar to Hirst et al. [10],
75 Fuentes and Raftery[11] assumed the model output and available observations can be represented
76 as a function of an unobserved ground truth plus error and bias terms. McMillan et al.[12] defined
77 an approach used in this study (discussed later in the paper) that again applies a hierarchical
78 Bayesian approach to predict ozone concentrations, but extends the model to include a temporal
79 dimension using an autoregressive structure.
80 Although varying in their approach, the intent of these statistical data combination
81 techniques is to retain the strongest components of each data type (i.e., observations and model
82 outputs) to best represent the pollutant of concern. For example, observations are the best
83 representation of the "true" pollutant concentration value at a given site and time, however,
84 depending on the network and the chemical being measured, the spatial and temporal extent may
85 be limited. Classical interpolation of observed concentrations helps to "fill-in" the spatial and
86 temporal gaps, but tends to produce overly smoothed results. Three-dimensional deterministic
87 air quality models, such as the Community Multiscale Air Quality (CMAQ) model[13], can
88 estimate pollutant concentrations across a uniform spatial and temporal scale. However, the
89 accuracy of these estimates is based on our uncertain understanding of the physical and chemical
90 processes underlying the formation, interaction and fate of atmospheric pollutants, and errors in
91 the model input (e.g., emissions, meteorology, boundary conditions). Even a perfect model with
92 perfect model input cannot reproduce the observations exactly since random variations
93 embedded in the observations taken at individual monitoring locations are not explicitly

94 estimated in current regional-scale models. Thus, an observation reflects a single event out of a
95 population, whereas, the modeled concentration represents the population average. Moreover,
96 the predictions from a deterministic model represents the volume-averaged concentration for the
97 grid cell while observations at a given monitoring location reflect point measurements. [14]
98 Figure 1 illustrates some of the strengths and weaknesses inherent in the modeled and
99 observed data used in this study by displaying the modeled, observed and interpolated (kriged)
100 surfaces for the domain encompassing New York State for June 13, 2001. On this day, the
101 interpolated observations appear to be overly smooth and miss the "hot spots" generated by
102 emissions from significant sources of ozone precursor chemicals captured by the model. The
103 modeled surface shows the effects of titration in the area of New York City (NYC) where high
104 nitrogen oxide (NO) emissions "scavenge" ozone creating areas of low ozone concentrations
105 around the emission source. If these local features are measured, kriging tends to smooth them
106 out. The model captures this important feature, but may overestimate the extent of the titration
107 effect. Thus, capturing "hot spots" and other spatial gradients that may be smoothed out by
108 kriging, yet correcting the bias that may exist in model estimates may produce improved air
109 quality surface concentrations critical to detecting health impacts. In this study, we apply
110 statistical combination techniques that integrate the observed concentrations with the model
111 estimates to optimize the strengths of each dataset. We focus on the summer of 2001 for the
112 New York State (NYS) domain as a pilot study to demonstrate the use of these enriched air
113 quality data in an epidemiological health study and risk assessment for NYS.
114
115 **APPROACH**
116 Four techniques ranging from relatively simple to complex, are investigated here for
117 combining observed and modeled ozone concentrations for a domain in the greater NYS area
118 from June 1 through August 31, 2001 (Figure 2). The focus of this investigation is on providing
119 the daily maximum 8-hr ozone concentrations to state health assessors for their use in
120 investigating relationships between air quality and human health endpoints (e.g., respiratory
121 related hospital admissions) across multiple years. The daily maximum 8-hr ozone
122 concentrations used in the study were calculated from observations and from CMAQ model
123 predictions. Four statistical combination techniques were applied to the observed and modeled
124 data to produce combined daily maximum 8-hr ozone concentration surfaces on a 12 km x 12 km

125 spatial grid for a total of 92 days in the summer of 2001. These combined surfaces were
126 quantitatively compared through cross-validation and qualitatively compared through analysis of
127 the spatial surfaces produced by each of the methods.
128
129 **Observations**
130 Hourly ozone observations for June 1 through August 31, 2001 were obtained from: (1)
131 the Environmental Protection Agency's (EPA) Air Quality System (AQS) database
132 (http://www.epa.gov/oar/data/aqsdb.html) and Clean Air Status and Trends Network
133 (CASTNET) (http://www.epa.gov/castnet/); and (2) the Canadian Environmental Assessment
134 Agency's National Air Pollution Surveillance Network (NAPS) (http://www.etc-
135 cte.ec.gc.ca/publications/napsreports_e.html). All monitoring networks provided hourly
136 concentrations for each day in the summer with a total of 200 sites (139 AQS sites, 52 NAPS
137 sites and 9 CASTNET sites). Information on the quality assurance conducted for each network
138 can be found on the websites provided above. The daily maximum 8-hr ozone concentrations
139 were calculated by applying an 8-hr moving window to the hourly time series and selecting the
140 8-hr time window with the highest ozone concentration value (referred to as the 8-hr maximum
141 daily average (MDA) throughout the paper). Only those days having greater than 20 hours of
142 data were used for computing the MDA.
143
144 **Model Output**
145 Ozone 8-hr MDAs were calculated from the hourly concentration values simulated by
146 EPA's CMAQ model, version 4.5. Specifically, the simulated concentrations for June 1 through
147 August 31 were extracted from the 2001 annual simulation with 12 km x 12 km horizontal grid
148 cells. The meteorology and emissions inputs for this simulation were from the Fifth-Generation
149 NCAR / Penn State Mesoscale Model (MM5) and EPA's 2001 National Emissions Inventory,
150 respectively. The 12-km simulation encompassed most of the eastern United States (U.S.) and
151 was nested within a 36 km x 36 km horizontal grid simulation covering the contiguous U.S.
152 using the same model configuration as the 12-km nested simulation. Boundary conditions for
153 the 36-km simulation were provided by a global chemical transport model (GEOS-CHEM).[15]
154 Several evaluations[16-21] have been done of the CMAQ model. In particular, Appel et al.[22]
155 evaluated the CMAQ annual simulation used in this study and found that the median observed

156 and predicted ozone concentrations correspond well between 10 a.m. and 6 p.m. which are the
157 hours that typically makeup the 8-hr MDA. Appel et al.[22] also report that the model consistently
158 over-estimated low concentrations and under-estimated high concentrations of ozone. Additional
159 details on the CMAQ simulation used in this study and the evaluation of the simulation results
160 can be found in Byun and Schere[13] and Appel et al.[22], respectively.
161
162 **Statistical Combination Techniques**
163 The four techniques investigated for combining the observed and modeled values are:
164 additive bias adjustment, multiplicative bias adjustment, weighted average, and hierarchical
165 Bayesian. The application of each of these techniques resulted in a spatial surface (12 km
166 horizontal resolution) of 8-hr MDA ozone concentrations (ppb) for each day of the summer in
167 2001. In general, the bias adjustment approaches modify the modeled surface concentration
168 values by accounting for bias in the modeled ozone concentrations. The weighted-average
169 approach uses weights for the interpolated observations and modeled surfaces based on a relative
170 accuracy of each surface at each grid cell, and the hierarchical Bayesian approach treats both the
171 observations and the model concentrations as representing a true surface and calculates model
172 parameters using a Markov Chain/Monte Carlo technique.
173 *Interpolated Observations.* In addition, 8-hr MDA ozone observations were interpolated
174 with a kriging approach that applies a Matern spatial correlation function[23] to produce daily 12
175 km x 12 km maps. For each day of the study, the parameters for the Matern covariance function
176 were estimated using the restricted maximum likelihood estimation technique. Kriging was
177 performed to estimate the concentration at the grid center-point rather than block kriging since
178 the results of the two techniques were similar.[14] Differences in the spatial correlation structure
179 along different directions (anisotropy) was also accounted for in the model. This same kriging
180 approach was used in the bias-adjustment approaches and the weighted-average technique
181 discussed below.
182 *Additive Bias Adjustment.* The additive bias was calculated by subtracting the modeled
183 value from the observed value at each observation site for each day. This bias was then
184 interpolated to a 12 km x 12 km grid structure using the kriging method explained above to
185 match the model grid structure. To derive the final corrected ozone concentration value, the
186 interpolated bias fields were added to the modeled values as follows:

$$O_j^{corr} = \hat{C}_j + O_j^{\text{mod}} \quad (1)$$

$\hat{C}_j$ is C_i kriged to estimate the grid-cell center bias values. C_i is calculated as follows:

$$C_i = O_i^{obs} - O_i^{\text{mod}} \quad (2)$$

Where i refers to ozone monitor i, j refers to the grid-cell center point of a 12 km x 12 km horizontal grid structure (variables are also indexed by day, but this designation is omitted for simplification); $\hat{C}$ refers to the estimated kriged bias; O^{obs} refers to observed ozone concentration value; O^{mod} refers to modeled ozone concentration value; and O^{corr} refers to the corrected ozone concentration value.

Multiplicative Bias Adjustment. The multiplicative bias was calculated by dividing the observed value by the modeled value at each observation site for each day. Similar to the additive bias approach, the bias ratio was interpolated to a 12 km x 12 km grid structure using the kriging technique described above. However, because large ratios can result in those cases where the modeled value is small in comparison to the observed value, the ratios were log-transformed before interpolating and then back-transformed before multiplying the ratios by the model surface. The corrected ozone concentration values were calculated as follows:

$$O_j^{corr} = \hat{C}_j \times O_j^{\text{mod}} \quad (3)$$

$\hat{C}_j$ is C_i kriged to estimate the grid-cell center bias values. C_i is calculated as follows:

$$C_i = \frac{O_i^{obs}}{O_i^{\text{mod}}} \quad (4)$$

Neither the additive or multiplicative bias adjustment approaches calculate an error estimate.

Weighted-average. This technique used the kriging interpolation method described above to calculate a gridded surface based on the observed values. The final estimated ozone concentration was calculated using a weighted average of the observation-based estimate, O_j^{krige}, and the CMAQ output value, O_j^{mod}. The following statistical model was used to combine these two sources of information about the true (unknown) ozone concentration, O_j^{true}, at grid cell j:

$$O_j^{krig} = O_j^{true} + \varepsilon_j; \qquad \varepsilon_j \sim (0, \sigma_{\varepsilon_j}^2) \quad (5)$$

213 $$O_j^{\text{mod}} = O_j^{true} + \eta_j; \qquad \eta_j \sim (0, \sigma_{\eta_j}^2) \tag{6}$$

214

215 This statistical model does not rely on any assumptions about the distribution of the errors, ε_j and
216 η_j, except that they each have mean zero and known variances, $\sigma_{\varepsilon_j}^2$ and $\sigma_{\eta_j}^2$, respectively. The
217 final weighted average estimate at each grid cell is:

218

219 $$O_j^{corr} = w_j O_j^{krig} + (1 - w_j) O_j^{\text{mod}} \tag{7}$$

220 The weights are determined by minimizing the mean square error of O_j^{corr}. Using this least
221 squares approach, the weight factor is defined as: $w_j = \dfrac{\sigma_{\eta_j}^2}{\sigma_{\varepsilon_j}^2 + \sigma_{\eta_j}^2}$.

222 The kriging analysis provides an estimate for the error variance, $\sigma_{\varepsilon_j}^2$, for each grid cell
223 based on the covariance structure of the observed ozone. This estimated error variance accounts
224 for measurement error in the observations and for uncertainty in the kriging prediction, due to
225 sparseness of the monitoring network and the heterogeneity of the underlying unobservable
226 ozone field. Since the uncertainty in the model output is more difficult to characterize, the error
227 variance for the model values is held constant across all N grid cells and is set equal to the
228 maximum kriging error variance for a given day: $\sigma_\eta^2 = \max_j(\sigma_{\varepsilon_j}^2)$ for j = 1, …, N. This choice
229 for the model error variance was made because it produced the following properties for the final
230 estimate. At locations where the kriging error variance is large (e.g. in regions of very few
231 monitors), the kriging estimate will be given less weight than at other locations, but its weight is
232 never less than half that of the model. In grid cells that contain a monitoring site, the kriging
233 estimate will be weighted more heavily as compared to the modeled value. Thus this approach
234 uses the error variances to quantify the relative quality or accuracy of the observation-based
235 gridded concentrations compared to the model output. In addition, the error variance of the final
236 estimate is the ratio: $\dfrac{\sigma_{\varepsilon_j}^2 \sigma_\eta^2}{\sigma_{\varepsilon_j}^2 + \sigma_\eta^2}$.

237

238 *Hierarchical Bayesian Modeling (HBM).* A Bayesian hierarchical space-time fusion
239 modeling approach[12] has been developed for integrating various sources of air quality data. This
240 flexible model was developed to provide daily pollutant predictions over the continental U.S. for
241 multiple years. In this application, the HBM model was applied to estimate ozone concentration
242 values for the greater NYS domain (Figure 1) for the summer of 2001. Bayesian analysis
243 decomposes the modeling problem into linked stages: 1) air quality monitoring data; 2) CMAQ
244 output; 3) measurement errors and CMAQ bias; and 4) the underlying "true" concentration
245 surface. A Bayesian approach incorporates 'prior knowledge' of the unknown parameters which
246 results in improved estimation of the uncertainty of the 'true' pollutant surface at any location in
247 space and time. This model assumes that both monitoring data and CMAQ output provide good
248 information about the same underlying pollutant surface, but with different measurement error
249 structures. Discussion of the choice of parameters used and additional details on the overall
250 HBM approach can be found in McMillan et al.[12]
251
252 **Comparison of different techniques**
253 The interpolated observations and combined surfaces resulting from each of the
254 techniques described above were evaluated using cross-validation. Selection of cross-validation
255 sites used in an evaluation can present many challenges. In this study, the number and location
256 of the ozone monitors was relatively dense. However, because monitors are sometimes placed to
257 determine compliance with regulatory exceedance thresholds, the monitors tend to be clustered
258 around urban areas (Figure 2). As a result, random selection of monitors can result in a
259 relatively large number of urban sites. This tendency can bias the results of the evaluation to
260 favor the interpolated observations as interpolation will always perform best in those areas where
261 there are many monitors. In addition, the clustering of monitoring sites in urban areas can result
262 in under-representation of rural areas. In order to ensure that rural areas as well as urban areas
263 were represented in the selection, the observation sites were overlaid on the 2000 Census Bureau
264 urban metropolitan area boundaries using a Geographical Information System to determine
265 whether the observations were in a rural or urban environment. Cross-validation sites were then
266 selected in two steps: (1) 7 rural CASTNET monitoring sites were used for cross-validation; and
267 (2) 20 sites were randomly selected from the AQS and NAPS networks for a total of 27 sites and
268 2,454 observations (Figure 3). Four of the 27 randomly selected AQS and NAPS sites were

designated as rural, resulting in a total of 11 rural sites (7 CASTNET sites + 4 AQS/NAPS sites) and 16 urban sites. These observations were set aside for cross-validation and all methods utilized the remaining observations to generate the combined surfaces.

The coefficient of determination (R^2), mean bias, and root mean square error (RMSE) were calculated between the observed value at the cross-validation site and the modeled output, interpolated observation and each of the three combination techniques for all 92 days of the summer. In addition, time series plots were generated to evaluate the error for each day at all sites. R^2, mean bias and RMSE were also compared for urban sites versus rural sites, and by network (AQS, NAPS and CASTNET), however, these latter analyses are not shown as they did not result in substantial differences among the different data combination techniques.

In order to assess how the various combination techniques compared across different percentiles, the observed concentrations used in the cross-validation were ranked by concentration level, and then binned by non-uniform percentiles (0-50th, 50th-75th, 75th-90th, 90th-95th, and 95th-100). The matching cross-validation surface results for each technique were also binned, and the binned values were compared to the binned observations through scatterplots. In addition, the error (predicted – observed) for each technique was calculated and averaged for each bin, allowing for all techniques to be shown on one line plot. Finally, the spatial features of the combined surfaces were qualitatively assessed by comparing the spatial maps (i.e., concentration estimates at each 12 km x 12 km grid cell) produced by the three combination techniques, the model and the interpolated observations. For the spatial analysis, maps displaying the mean, median, various percentiles, standard deviation, coefficient of variation (standard deviation/mean) were examined. Only the most relevant of these maps are included in the paper.

RESULTS AND DISCUSSION

Of the four combination techniques, the additive and multiplicative bias adjustment approaches were the easiest methods to use. The HBM approach was the most complex model, requiring specification of prior distributions for all model parameters. Related to this requirement, estimating the model error parameter for both the weighted-average and HBM approaches was problematic as this value is unknown. In the near future, however, use of ensemble runs may improve our ability to estimate the model variance. It should also be noted

300 that the weighted-average and HBM approaches were the only combination techniques compared
301 in the study that provided an estimate of predicted error. This estimate of predicted error can be
302 important for some applications.

303 For the standard metrics examined, the cross-validation produced similar results for all
304 four combination techniques and the kriged observations (Table 1). The RMSE for the four
305 combination techniques and kriged observations was within 1.4 ppb of each other, and the mean
306 bias was within 1.05 ppb of each other. R^2 ranged from 85 to 88 percent. The metrics also
307 indicated that all four combination techniques substantially improved the modeled surface (R^2 of
308 0.66).

309 The percentile rank-ordered analysis, however, revealed interesting differences among
310 the combination techniques and the interpolated observations. Figure 4 displays scatterplots
311 (predicted versus observed), highlighted by color code to depict the percentile range. Note that
312 the weighted- average and HBM techniques correct the model bias fairly well at the lower
313 percentiles, but follow the scatterplot pattern of the raw modeled output at the higher percentiles.
314 The additive and multiplicative bias techniques follow the one-to-one line closely, indicating that
315 a simple correction of the model bias may be effective for improving the spatial characterization
316 of ozone concentrations.

317 Although cross-validation often favors kriging of the ozone observations (due to the
318 concept discussed earlier of randomly selecting cross-validation data from clustered monitoring
319 sites), the bias-adjustment techniques produce slightly better results than kriging at the higher
320 percentiles. This same difference is evident in the error plots in Figure 5. The tendency of the
321 CMAQ model to overestimate low ozone values and underestimate high ozone values can be
322 clearly seen. Similar to the scatterplots, the weighted-average and HBM techniques appear to
323 correct this overestimation at the lower ozone concentrations, but do not do as well at
324 reproducing the observed ozone concentration levels at the higher percentiles (it should be noted
325 that the HBM technique was designed to provide Bayesian predictions over large national spatial
326 scales rather than the small regional domain of this study.) The additive and multiplicative bias
327 adjustment approaches appear to perform best across all percentiles, slightly out-performing the
328 kriging of the observations at the higher percentiles as noted earlier. Although it is recognized
329 that inferences from the smaller sample sizes in the higher percentile bins must be done with
330 caution (sample size ranges from 370 to 123 site-days for the three highest percentile bins), the

sample sizes range from 5% - 15% of the total sample size of 2,454 site-days, providing credence to the results discussed above.

In addition, the spatial texture seen with the combined surfaces indicates that model-based spatial information seen in CMAQ is retained with the combined surfaces. Figure 6 shows 8-hr MDA ozone values for the (a), CMAQ model (b) kriged observations and (c) multiplicative bias adjusted values for two representative days on June 7, 2001 and July 19, 2001. Note the overly smooth surface inherent in the interpolated observations and the influence of the model in the spatial texture of the combined surface. Figure 7 shows the mean concentration values for each grid cell. The model estimates of the ozone titration effect near NYC and Boston can clearly be seen in the modeled surface. The measurements in these same areas also show low ozone concentration values; however, these low values are averaged out by the interpolation. In addition, high ozone values are predicted by the model over the Great Lakes and the Atlantic Ocean due to lower planetary boundary layer heights, stable atmospheric conditions, and reduced turbulence and deposition; all physical processes known to exist over large waterbodies. Ozone measurements taken over Lake Michigan and aircraft observations over the coastal areas of the Northeast indicate the presence of high ozone concentrations over large waterbodies which the model seems to capture [24; 25]. Similar to titration in the urban core, the interpolated observations do not show this physical phenomenon. Since the purpose of this study is to provide improved air quality data for health studies, the difference in the estimation of titration is particularly relevant as this physical phenomenon can occur in highly-populated areas.

The coefficient of variation (Figure 8) calculated for all days at each grid cell reveals a problem introduced by the large standard deviation values (relative to the mean) produced by the titration effect in both the additive and multiplicative bias adjustment approaches. For example, the coefficient of variation is high for the multiplicative bias surface near Staten Island and Boston for which the model predicts low ozone concentration values due to titration. These high coefficient of variation values are the result of large differences between the modeled and observed concentrations that result in very large observed-to-modeled ratios (Eq 4; Figure 9a). Kriging the ratios and multiplying them by the model surface creates high ozone concentrations in the non-titrated area surrounding the titrated area (Figure 9b). While selection of a cross-validation site in one of the impacted areas may have changed the results, this effect occurs for less than 0.03% of the total concentrations and for only 3 days during the 92-day time period. In

addition, the high observed-to-modeled ratios do not produce excessively high ozone concentrations except for over the ocean outside of Staten Island where the population is low.

SUMMARY

The cross-validation results of this pair-wise comparison using standard statistical metrics did not reveal a large difference among the four combination techniques, but did reveal that all techniques provide improved estimates of 8-hr MDA ozone concentrations as compared to the model surface alone. The percentile analysis of the cross-validation results revealed interesting results not discerned by the all-days/all-sites metrics alone. The percentile analysis indicated that the additive and multiplicative bias adjustment techniques tended to improve the combined 8-hr MDA ozone concentrations at the higher percentiles as compared to the other techniques, including kriging the observations. Further analysis of the resulting spatial surfaces, however, revealed problems with the additive and multiplicative bias adjustment approaches introduced by the modeled titration effect, yielding artificially high ozone concentration values in adjacent cells. This problem, though, occurred for less than 0.03% of the total concentrations on only 3 days of the total summer, and primarily over waterbodies where the population is low. The qualitative spatial analysis performed supported that the combination techniques added spatial information from the model as compared to kriging the observations alone. In the case of this study, the intended application of the combination approach is to provide improved air quality surface maps for conducting epidemiology studies in NYS. The additive and multiplicative bias adjustment approaches are considered appropriate for this application because; (1) accurately representing days of high-ozone concentrations is important for the health study of interest and the additive and multiplicative bias adjustment approaches out-performed the other methods at the higher ozone concentration percentiles, (2) the additive and multiplicative bias adjustment approaches are relatively simple and can readily be applied by the state health community, and (3) to date, estimates of predicted error produced by the HBM and weighted average approaches are not generally used in health studies. However, as epidemiology studies move towards the use of predictive distributions, more complex approaches such as HBM may be needed to estimate prediction error. Finally, these results are limited in applicability to the domain, pollutant and time period studied.

ACKNOWLEDGEMENTS

The authors thank Drs. Steven Porter, Christian Hogerefe, and Jenise Swall for many helpful discussions on the data fusion techniques. Although this paper has been reviewed and approved for publication, it does not necessarily reflect the views and policies of the U.S. Environmental Protection Agency.

398

399

400 **REFERENCES**

401 1. Bell, M.L. The Use of Ambient Air Quality Modeling to Estimate Individual and
402 Population Exposure for Human Health Research: A Case Study of Ozone in the
403 Northern Georgia Region of the United States;*Environment International* 2006, 32(5),
404 586-593.
405
406 2. Delle Monache, L.;Wilczak, J.;McKeen, S.;Grell, G.;Pagowski, M.;Peckham, S.;Stull,
407 R.;McHenry, J.;McQueen, J. A Kalman-Filter Bias Correction Method Applied to
408 Deterministic, Ensemble Averaged and Probabilistic Forecasts of Surface Ozone;*Tellus*
409 *Ser. B-Chem. Phys. Meteorol.* 2008, 60(2), 238-249.
410
411 3. Kang, D.;Mathur, R.;Rao, S.T.;Yu, S. Bias-Adjustment Techniques for Improving Ozone
412 Air Quality Forecasts;*Journal of Geophysical Research* 2008.
413
414 4. Wilczak, J.;McKeen, S.;Djalalova, I.;Grell, G.;Peckham, S.;Gong, W.;Bouchet,
415 V.;Moffet, R.;McHenry, J.;McQueen, J.;Lee, P.;Tang, Y.;Carmichael, G.R. Bias-
416 Corrected Ensemble and Probabilistic Forecasts of Surface Ozone over Eastern North
417 America During the Summer of 2004;*Journal of Geophysical Research-Atmospheres*
418 2006, 111(D23), 15.
419
420 5. Houtekamer, P.L.;Mitchell, H.L.;Pellerin, G.;Buehner, M.;Charron, M.;Spacek,
421 L.;Hansen, M. Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results
422 with Real Observations;*Monthly Weather Review* 2005, 133(3), 604-620.
423
424 6. McKeen, S.;Wilczak, J.;Grell, G.;Djalalova, I.;Peckham, S.;Hsie, E.Y.;Gong,
425 W.;Bouchet, V.;Menard, S.;Moffet, R.;McHenry, J.;McQueen, J.;Tang, Y.;Carmichael,
426 G.R.;Pagowski, M.;Chan, A.;Dye, T.;Frost, G.;Lee, P.;Mathur, R. Assessment of an
427 Ensemble of Seven Real-Time Ozone Forecasts over Eastern North America During the
428 Summer of 2004;*Journal of Geophysical Research-Atmospheres* 2005, 110(D21); doi:
429 10.1029/2005jd005858.
430
431 7. Whitaker, J.S.;Compo, G.P.;Wei, X.;Hamill, T.M. Reanalysis without Radiosondes
432 Using Ensemble Data Assimilation;*Monthly Weather Review* 2004, 132(5), 1190-1200.
433
434 8. Gego, E.;Porter, P.S.;Garcia, V.C.;Hogrefe, C.;Rao, S.T. Fusing Observations and Model
435 Results for Creation of Enhanced Ozone Spatial Fields: Comparison of Three
436 Techniques. In: Borrego C., and Miranda A.I. (Eds.). *Air Pollution Modeling and Its*
437 *Application Xix*. Springer Science + Business Media B.V., 2008, pp 341-348.
438
439 9. Hogrefe, C.;Hao, W.;Civerolo, K.;Ku, J.Y.;Sistla, G.;Gaza, R.S.;Sedefian, L.;Schere,
440 K.;Gilliland, A.;Mathur, R. Daily Simulation of Ozone and Fine Particulates over New
441 York State: Findings and Challenges;*J. Appl. Meteorol. Climatol.* 2007, 46(7), 961-979;
442 doi: 10.1175/jam2520.1.

10. Hirst, D.;Storvik, G.;Syversveen, A.R. A Hierarchical Modelling Approach to Combining Environmental Data at Different Scales;*J. R. Stat. Soc. Ser. C-Appl. Stat.* 2003, 52, 377-390.

11. 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.

12. McMillan, N.;Holland, D.M.;Morara, M.;Feng, J. Combining Numerical Model Output and Particulate Data Using Bayesian Space-Time Modeling;*Environmetrics* 2010, 21, 48-65; doi: DOI: 10.1002/env.984.

13. Byun, D.W.;Schere, K.L. Review of the Governing Equations, Computational Algorithms, and Other Componenets of the Models-3 Community Multiscale Air Qualtiy (Cmaq) Modeling System;*Applied Mechanics Reviews* 2006, 59, 51-77.

14. Swall, J.L.;Foley, K.M. The Impact of Spatial Correlation and Incommensurability on Model Evaluation;*Atmospheric Environment* 2009, 43(6), 1204-1217.

15. Bey, I.;Jacob, D.J.;Yantosca, R.M.;Logan, J.A.;Field, B.D.;Fiore, A.M.;Li, Q.B.;Liu, H.G.Y.;Mickley, L.J.;Schultz, M.G. Global Modeling of Tropospheric Chemistry with Assimilated Meteorology: Model Description and Evaluation;*Journal of Geophysical Research-Atmospheres* 2001, 106(D19), 23073-23095.

16. Gego, E.;Gilliland, A.;Godowitch, J.;Rao, S.T.;Porter, P.S.;Hogrefe, C. Modeling Analyses of the Effects of Changes in Nitrogen Oxides Emissions from the Electric Power Sector on Ozone Levels in the Eastern United States;J. Air Waste Manage. Assoc. 2008, 58(4), 580-588; doi: 10.3155/1047-3289.58.4.580.

17. Gilliland, A.B.;Hogrefe, C.;Pinder, R.W.;Godowitch, J.M.;Foley, K.L.;Rao, S.T. Dynamic Evaluation of Regional Air Quality Models: Assessing Changes in O-3 Stemming from Changes in Emissions and Meteorology;*Atmospheric Environment* 2008, 42(20), 5110-5123; doi: 10.1016/j.atmosenv.2008.02.018.

18. Godowitch, J.M.;Gilliland, A.B.;Draxler, R.R.;Rao, S.T. Modeling Assessment of Point Source Nox Emission Reductions on Ozone Air Quality in the Eastern United States;*Atmospheric Environment* 2008, 42(1), 87-100. doi: 10.1016/j.atmosenv.2007.09.032.

19. Godowitch, J.M.;Hogrefe, C.;Rao, S.T. Diagnostic Analyses of a Regional Air Quality Model: Changes in Modeled Processes Affecting Ozone and Chemical-Transport Indicators from Nox Point Source Emission Reductions;*Journal of Geophysical Research-Atmospheres* 2008, 113(D19); doi: 10.1029/2007jd009537.

20. Eder, B.;Yu, S.C. A Performance Evaluation of the 2004 Release of Models-3 Cmaq;*Atmospheric Environment* 2006, 40(26), 4811-4824; doi: 10.1016/j.atmosenv.2005.08.045.

21. Hogrefe, C.;Porter, P.S.;Gego, E.;Gilliland, A.;Gilliam, R.;Swall, J.;Irwin, J.;Rao, S.T. Temporal Features in Observed and Simulated Meteorology and Air Quality over the Eastern United States;*Atmospheric Environment* 2006, 40(26), 5041-5055; doi: 10.1016/j.atmonsenv.2005.12.056.

22. Appel, K.W.;Gilliland, A.B.;G, S.;Gilliam, R.C. Evaluation of the Community Multiscale Air Quality (CMAQ) Model Version 4.5: Sensitivities Impacting Model Performance; Part I--Ozone;*Atmospheric Environment* 2007; doi: 10.1016/j.atmosenv.2007.08.044.

23. Stein, M. *Interpolation of Spatial Data*. Springer, New York, 1999.

24. Dye, T.S.;Roberts, P.T.;Korc, M.E. Observations of Transport Processes for Ozone and Ozone Precursors During the 1991 Lake Michigan Ozone Study;*J. Appl. Meteorol. Climatol.* 1994, 34, 1877 - 1889.

25. Winkler, P. Surface Ozone over the Atlantic Ocean;*Journal of Atmospheric Chemistry* 1988, 7(1), 73-91.

About the Authors

Valerie Garcia (garcia.val@epa.gov) is the Chief of the Atmospheric Exposure Integration Branch, Atmospheric Modeling and Analysis Division (AMAD), within the U.S. Environmental Protection Agency (EPA). Dr. Kristen Foley is a statistician within AMAD, Dr. David Holland is a Senior Statistician with the National Exposure Research Laboratory within EPA, and Dr. S.Trivikrama Rao is the Director of AMAD. Edith Gego is a consultant with Gego Associates in Idaho Falls, ID.

Table 1. Cross-Validation Results

Table 1: Cross-Validation Results
Ozone 8-hour Maximum Daily Averages

n = 2,454	RMSE (ppb)	Mean Bias (ppb)	R^2
CMAQ Only	11.70	0.88	0.60
Kriged Observations	6.40	0.49	0.88
Additive Bias Adjustment	6.80	0.33	0.86
Multiplicative Bias Adjustment	6.80	0.05	0.86
Weighted average	6.70	0.50	0.88
Hierarchical Bayesian	7.80	1.10	0.85

Figure 1: Ozone 8-hr MDA (ppb) for June 13, 2001: (a) modeled, (b) observed and (c) kriged observations. Kriging smoothes the hotspots whereas, the model may over-predict titration in NYC.

Figure 2: Domain and observations used in study. Solid circles show location of AQS sites, triangles show NAPS sites and squares show CASTNET sites. Open circles highlight sites clustered in urban areas.

Figure 3: Location of cross-validation sites. Circles denote NAPS sites, triangles denote CASTNet sites and diamonds denote AQS sites.

Figure 4: Observed (y-axis) and predicted (x-axis) 8-hr MDA ozone (ppb) binned by percentile for cross-validation sites. Black = 0-50%, red = 51-75%, orange = 76-90%, blue = 91-95%, green = 96-100%.

Figure 5: Mean error (binned by percentile) between modeled, kriged and the four combination techniques versus mean observed concentrations for all cross-validation sites. Circles represent average mean for each binned percentile; lines are for identification of technique but do not represent linear relationships between averaged points.

Figure 6: Contribution of spatial information from model in combined surface of 8 hr MDA. Column (a) is modeled surface, column (b) is interpolated observations and column (c) is multiplicative adjusted bias. Top panels are for June 7, 2001 and bottom panels are for July 19, 2001.

Figure 7: Mean 8-hr MDA ozone concentrations across all days for (a) modeled, (b) kriged observations, (c) multiplicative bias, (d) additive bias, (e) weighted average, and (f) HBM.

556

557 Figure 8: Coefficient of Variation calculated for 8-hr MDA ozone concentrations across all days
558 for (a) modeled, (b) kriged observations, (c) multiplicative bias, (d) additive bias, (e) weighted
559 average, and (f) HBM.

560

561 Figure 9: Maximum values of (a) observed-to-modeled ratios (eq 4) and (b) 8-hr MDA ozone
562 concentrations resulting from the multiplicative adjusted bias approach.

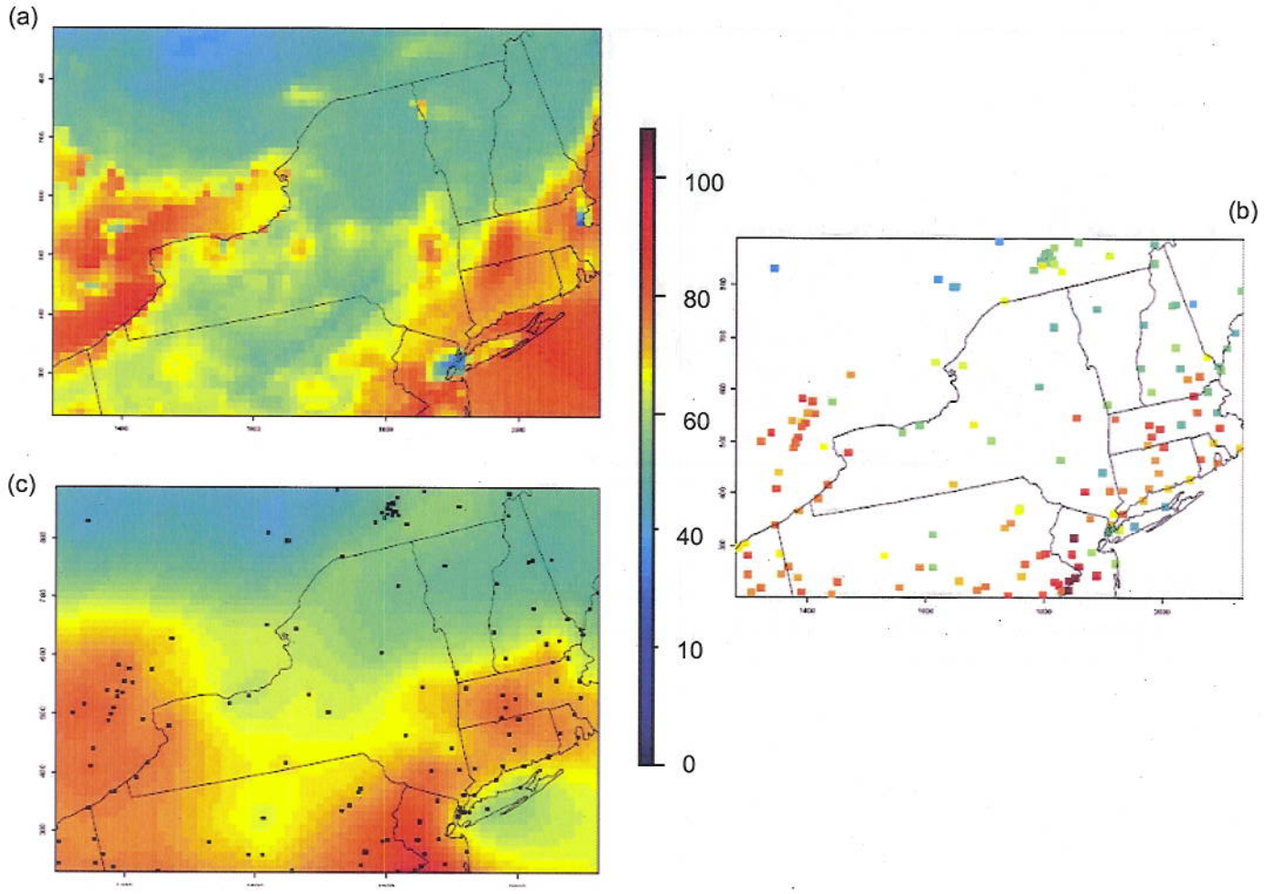

Figure 1: Ozone 8-hr MDA (ppb) for June 13, 2001: (a) modeled, (b) observed and (c) kriged observations. Kriging smoothes the hotspots whereas, the model may over-predict titration in NYC.

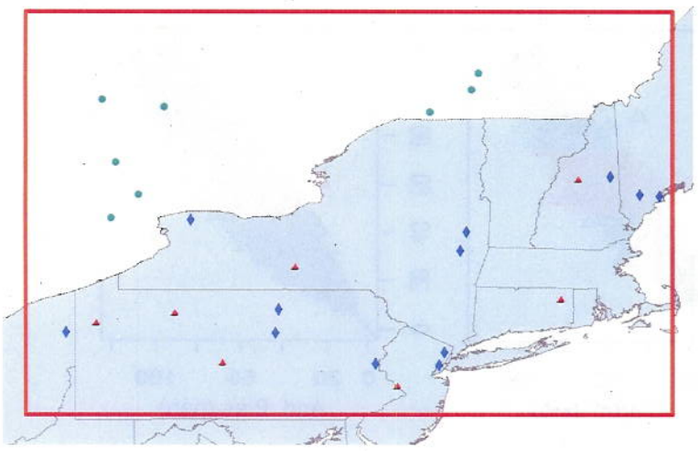

Figure 3: Location of cross-validation sites. Circles denote NAPS sites, triangles denote CASTNet sites and diamonds denote AQS sites.

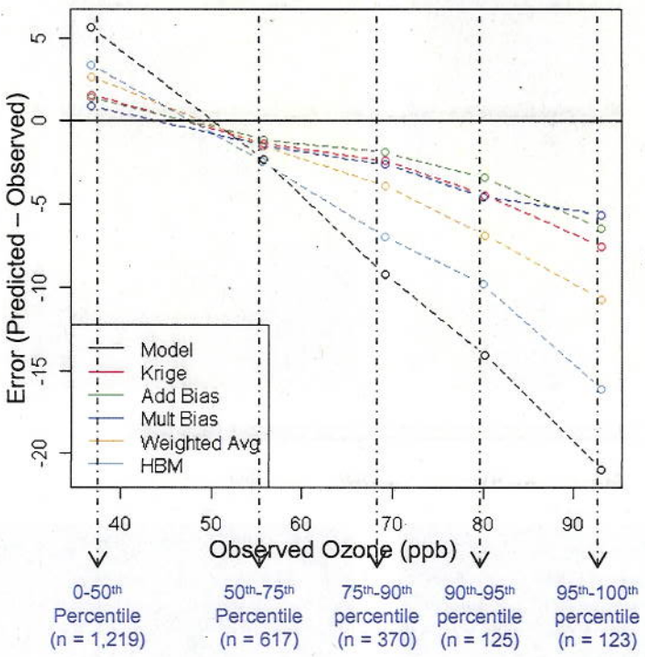

Figure 5: Mean error (binned by percentile) between modeled, kriged and the four combination techniques versus mean observed concentrations for allcross-validation sites. Circles represent average mean for each binned percentile; lines are for identification of technique but do not represent linear relationships between averaged points.

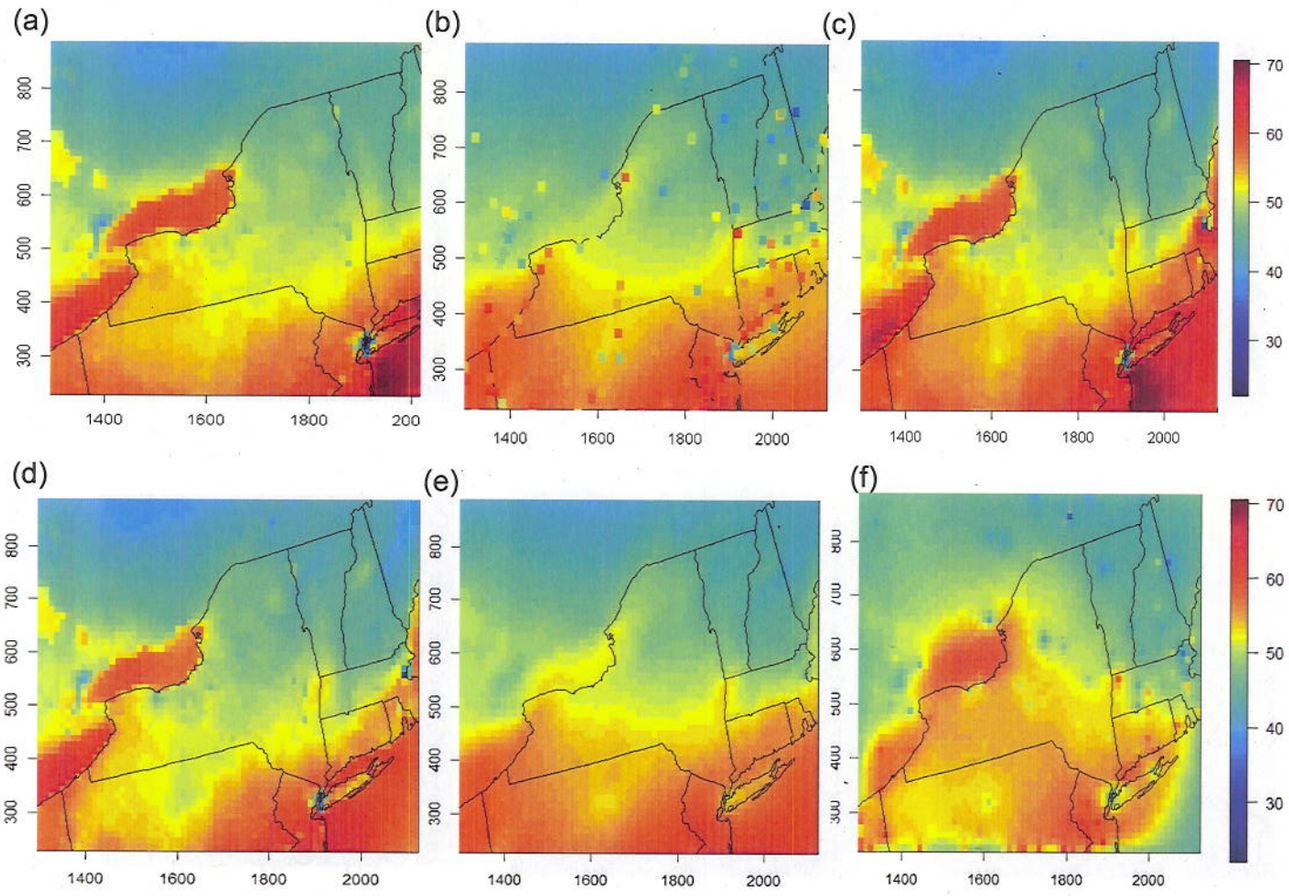

Figure 7: Mean 8-hr MDA ozone concentrations across all days for (a) modeled, (b) kriged observations, (c) multiplicative bias, (d) additive bias, (e) weighted average, and (f) HBM.

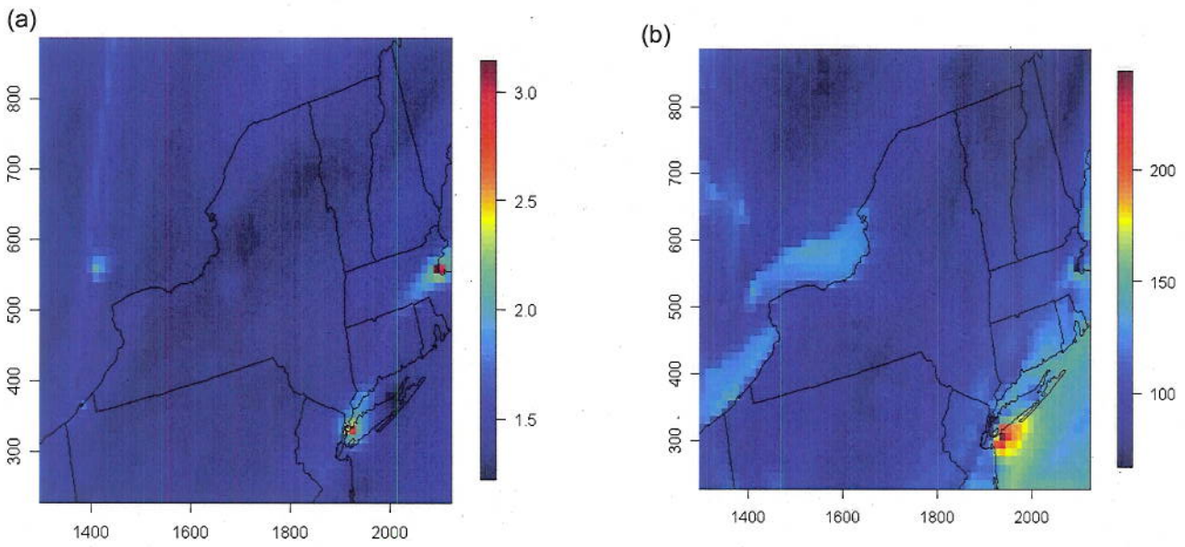

Figure 9: Maximum values of (a) observed-to-modeled ratios (eq 4) and (b) 8-hr MDA ozone concentrations resulting from the multiplicative adjusted bias approach.