THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT # 242

Recalibrating ozone chemical transport models

By

Zhong Liu

Nhu Le

James V Zidek

September 2008

Recalibrating ozone chemical transport models*

Zhong Liu†, Nhu D. Le‡, James V. Zidek§

September 1, 2008

*This work was partially supported by grants from the Pacific Institute of Mathematical Sciences and the Natural Science and Engineering Research Council of Canada

†Department of Statistics, U. of British Columbia. Email: zliu@stat.ubc.ca

‡British Columbia Cancer Research Centre. Email: nle@bccrc.ca

§Department of Statistics, U. of British Columbia. Email: jim@stat.ubc.ca

Abstract

This report concerns the recalibration of deterministic mesoscale computational models to align their outputs with measurements made on a microscale of the phenomena they model. Although the methods presented in the report are quite general, it focuses on chemical transport models (CTMs) for ground level ozone concentrations so that they can be empirically compared.

Such a model have been used in setting North American Air Quality Standards for ozone, a regulatory process that has required specification of ozone's policy relevant background (PRB) level, the level that would be observed if all anthropogenic sources in North America were eliminated. The level is needed to provide the baseline against which the impact of proposed new NAAQS can be assessed in environmental risk analysis. However, the PRB cannot be measured since few if any areas uncontaminated by anthropogenic sources now exist in North America. Moreover those that do exist may not represent very well the contaminated areas.

So instead, a deterministic chemical transportation models (CTM) has been used to infer the PRB by suppressing the anthropogenic sources in the computational model. To validate use of that model in that way, its output has been favorably compared with measured ozone concentrations with the anthropogenic sources turned on.

However CTMs generate their outputs on a mesoscale, making their outputs inherently incomparable with measurements, which are by their nature made on a micro-scale. Thus although comparisons of raw output do resemble their microscale counterparts, recalibrating them to adjust for their misalignment can greatly increase their resemblance. This report considers two very

different statistical methods for recalibrating CTMs and finds through an empirical investigation, a multi-stage multivariate regression method to be the best overall.

Keywords: Chemical transport models; Bayesian recalibration; policy related background levels; misaligned data; NAAQS for ozone; combining physical & statistical models.

INTRODUCTION

This report presents two general methods for recalibrating the outputs from deterministic mesoscale computational models to align them with microscale measurements of the environmental phenomena they represent. The former are not created to represent the latter ([4]). Ensuring their computational feasibility can mean ignoring things such as topography, turbulence, evaporation and friction that play an important role in determining the latter. Thus the model outputs and measurements are by their nature comparable.

Nevertheless in some situations it may be necessary to use these mesoscale models to simulate microscale measurements. One such situation of great interest in its own right will be the focus of this paper so that the methods presented here can be empirically assessed. That situation arises in setting North American Air Quality Standards (NAAQS) for ozone. The regulatory process involved there has required the specification of ozone's policy relevant background (PRB) level, the level that would be observed if all anthropogenic sources in North America were eliminated. The level is needed to provide the baseline against which the impact of proposed new NAAQS can be assessed in environmental risk analysis.

However the PRB cannot be measured since few areas (if any) uncontaminated by anthropogenic sources now exist in North America. For example, although one might have expected the Yellowstone National Park (WY) to be a pristine site, the monthly maximal hourly ozone concentration level there exceeds 50 ppb (parts per billion), a relatively high level even by urban standards, for most months from 1998 to 2001 (Figure 3-25a in [1]). Moreover those pristine areas that do exist may not represent the remainder very well.

Since the PRB level cannot be measured, it has to be imputed. The strategy adopted in the most recent NAAQS assessment of the ozone standard in the United States, uses a deterministic model in which anthropogenic sources were suppressed (i.e. "turned off") to get outputs representing PRB levels. However the PRB's vital role demands some sort of validation of these imputed PRB levels. A comparison with measurements being ruled out led instead to a comparison of those outputs with the anthropogenic sources "turned on" and measurements. Similarity in these two series over broad geographical and temporal domains supported use of the model outputs when they are turned off.

The deterministic model selected for this analysis is an ozone chemical transport model (CTM) called GEOS-CHEM (Goddard Earth Observing System-Chemistry). [Details about the various versions of GEOS-CHEM models can be found at http://www-as.harvard.edu/chemistry/trop/geos/.] Use of that model led to an inferred PRB level of between 15 and 35 ppb ([1]), the level adopted in the review of ozone standards.

Clearly, validating the model with sources turned on does not in and of itself validate use of the model with sources turned off. However, a great deal of physical knowledge is built into these CTMs and that gives vital added sup-

port. Nevertheless, a logical gap remains: the misalignment of model outputs and measurements, pointing to the need to recalibrate the former for this application. This report presents two very different statistical methods for doing just that.

Since the authors did not have access to GEOS - CHEM simulated data, they used instead outputs from the MAQSIP (Multiscale Air Quality Simulation Platform) as described by [10], whose outputs are thought to be similar to those of the former. [Its spatial resolution is on a grid scale of 6×6 km^2, its temporal resolution, one hour.] We use both methods to recalibrate MAQSIP's simulated data using measurements obtained from the EPA's Air Quality System (AQS) monitoring network which measures ozone concentration levels over the eastern and central part of USA. Validation data obtained from the same source, but not used in the recalibration itself, enables a comparison of these methods.

We organize the report as follows. Section introduces the data. Section presents two statistical models to calibrate the deterministic model outputs. We show the calibration results in Section . Finally in Section we summarize our findings and give recommendations.

DATA DESCRIPTION

The data comes from two sources: regional surface ozone concentration measurements and model outputs from a deterministic model AQM (air quality model), a non-hydrostatic version of the MAQSIP (Multiscale Air Quality Simulation Platform) model. This AQM system has been described in detail by [11]. The AQM model outputs are based on grid cells with resolution 6×6 km^2. The measurements are from the Air Quality System (AQS) monitoring network.

Both the measurements and model outputs are hourly concentrations starting from May 15 to September 11, 1995 over a 120-day period. The dataset represents 375 monitoring stations in the AQS network and 307 grid cells in the AQM output. Besides the fact that measurements and model outputs are based on different supports, the data represent different time standards, model outputs being based on the GMT (Greenwich Mean Time) time standard, the measurements on local time. Ignoring this time difference would result in poor correlation between measurements and model outputs.

The measurement series, unlike that for the model outputs, have missing values. For example, all the measurements from station 550730005 (in Wisconsin state) are missing. To deal with the missing values, we first choose those stations that have no more than 100 hours of missing measurements. Second, we use the 24 hour mean to fill in the missing values. For example, if the missing value occurs at 10 AM, then we use the average of the available values at 10 AM every day to fill in this missing value. After adjusting for different time standards and ignoring the stations with more than 100 missing measurements, we have measurements at 81 stations and model outputs on 375 grid cells of 2856 hours (119 days). In these 375 grid cells, there are 78 grid cells which contain one and only one station. To enable us understand better the role of model-to-measurement correlation, from now on the data always will focus on the 78 grid cells with 78 stations inside the grid cells. The calibration will focus on the 8-hour (10AM-17PM) daytime average because the measurements and model outputs are more correlated during those hours. Although we have 119 days from May to September we only focus on the days from July 1 to July 30 because the ozone concentration is at peak in the summer due to the high temperature. We use the measurements at 48 stations to fit the models and

the measurements at the rest 30 stations are used as validation.

METHODOLOGY

This section presents two statistical models which can be used to calibrate the deterministic model outputs. First, we introduce the Bayesian melding model proposed in [4] then we present an alternative spatial-temporal implemented in [7].

BAYESIAN MELDING MODEL

The Bayesian melding model assume the existence of an underlying process to connect the measurements and model outputs. The mathematical forms of the model is in the following.

$$
\begin{aligned}
&\hat{Z}(\boldsymbol{s}) = Z(\boldsymbol{s}) + e(\boldsymbol{s}) \\
&\boldsymbol{s} \in \Re^D\} D \in \{1,2,3\} \\
&Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \\
&Z(B) = \frac{1}{|B|}\int_B Z(\boldsymbol{s})d\boldsymbol{s} \\
&\tilde{Z}(\boldsymbol{s}) = a(\boldsymbol{s}) + b(\boldsymbol{s})Z(\boldsymbol{s}) + \delta(\boldsymbol{s}) \\
&\tilde{Z}(B) = \frac{1}{|B|}\int_B a(\boldsymbol{s})d\boldsymbol{s} + \frac{1}{|B|}\int_B b(\boldsymbol{s})Z(\boldsymbol{s})d\boldsymbol{s} + \frac{1}{|B|}\int_B \delta(\boldsymbol{s})d\boldsymbol{s} \\
&\mu(\boldsymbol{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 \\
&\boldsymbol{s} = (s_1, s_2)^{\mathrm{T}}
\end{aligned}
\tag{1}
$$

The subscript $\boldsymbol{s}$ stands for the station and B stand for the grid cell. In the above model, Melding links processes with responses on mismatched scales through an

underlying true process $\{Z(\boldsymbol{s})$. Denote the measurement process by $\{\hat{Z}(\boldsymbol{s})$, and the deterministic model output process by $\{\tilde{Z}(B)\}$. To match $Z(\boldsymbol{s})$, we also hypothesize the existence of deterministic model output process $\{\tilde{Z}(\boldsymbol{s})$. The measurements error and model output error are independent of each other. The measurement errors, $e(\boldsymbol{s})$, are independent and identically distributed, having a normal distribution $N(0, \sigma_e^2)$. The model output errors, $\delta(\boldsymbol{s})$, are independent and identically distributed with a normal distribution $N(0, \sigma_\delta^2)$. The spatially correlated residuals, $\epsilon(\boldsymbol{s})$, have zero mean and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the covariance parameter vector. The mean of $\boldsymbol{Z}(\boldsymbol{s})$ is a linear function of the longitude (s_1) and latitude (s_2) at station $\boldsymbol{s}$. The number of locations is n. $Z(B)$ and $\tilde{Z}(B)$ are integrals of $Z(\boldsymbol{s})$ and $\tilde{Z}(\boldsymbol{s})$ over grid cell B. We only observe realizations of process $\hat{Z}(\boldsymbol{s})$ and $\tilde{Z}(B)$ at measured stations and grid cells for model outputs. We use the Gibbs sampling algorithm([5]) to fit this Bayesian melding model. More details can be found in [9].

The calibrated model output at grid cell B by using Bayesian melding model is

$$\left(\tilde{Z}(B) - \frac{1}{|B|}\int_B a(\boldsymbol{s})d\boldsymbol{s}\right)/b. \tag{2}$$

The multiplicative calibration parameter b is assumed constant across space and $a(\boldsymbol{s}) = a_0 + a_1 s_1 + a_2 s_2$. One can plug-in the posterior mean of $a(\boldsymbol{s})$ and b into (2) to obtain the calibrated model output ([4]). Departing the plug-in approach used in [4], we use formula (2) to calculate the calibrated model outputs at each iteration of the Gibbs sampling so we have a distribution for the calibrated model outputs. We use the mean of that distribution as the final calibration results.

SPATIAL - TEMPORAL MODEL

This section presents a Bayesian hierarchical spatial-temporal model. At each station, we assume there is a linear relationship between the measurements and model outputs. We model this relationship by a linear regression with temporally correlated residuals. To incorporate the spatial correlation, we assume both the coefficients and residuals are also spatially correlated. Thus we have a spatial-temporal model with the following form.

$$
\begin{aligned}
&O_{\boldsymbol{s},t} = a_{\boldsymbol{s}} + c_{\boldsymbol{s}} M_{\boldsymbol{s},t} + N_{\boldsymbol{s},t} \\
&N_{\boldsymbol{s},t} = \rho N_{\boldsymbol{s},t-1} + \epsilon_{\boldsymbol{s},t} \\
&\boldsymbol{a} = (a_1, \cdots, a_n)^{\mathrm{T}} \sim \mathrm{MVN}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) \\
&\boldsymbol{c} = (c_1, \cdots, c_n)^{\mathrm{T}} \sim \mathrm{MVN}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\
&(\epsilon_{\mathbf{1},t}, \cdots, \epsilon_{\boldsymbol{n},t})^{\mathrm{T}} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon) \text{ independently and identically} \\
&\text{times } t = 1, \cdots, T, \text{ and} \\
&\text{sites } \boldsymbol{s} = s_1, \cdots, s_n \\
&\boldsymbol{\mu}_a = (\mu_a, \cdots, \mu_a)^{\mathrm{T}} \\
&\boldsymbol{\mu}_c = (\mu_c, \cdots, \mu_c)^{\mathrm{T}} \\
&\boldsymbol{\Sigma}_a = \sigma_a^2 \exp(-\boldsymbol{D}/\lambda_a) \\
&\boldsymbol{\Sigma}_c = \sigma_c^2 \exp(-\boldsymbol{D}/\lambda_c) \\
&\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2 \exp(-\boldsymbol{D}/\lambda_\epsilon).
\end{aligned} \tag{3}
$$

As described in the previous section, each grid cell has one and only one station inside. So each station has two time series: measurements and model outputs on the grid cell which contains that station. After assigning proper prior dis-

tributions to the parameter, we have a Bayesian hierarchical spatial-temporal model. Details on how fit this model can be found in [8]. The calibration formula is

$$a + c\tilde{Z}(B). \tag{4}$$

Similar to the calibration with Bayesian melding model, we use the average of calibrated model outputs at each Gibbs sampling as the final calibration results.

CALIBRATION RESULTS

This section compare how the measurements are predicted by model outputs, calibrated model outputs with Bayesian melding and spatial-temporal model. The Bayesian melding model is a pure spatial model because it does not incorporate the temporal correlation within the measurements. So we apply Bayesian melding model to the data on each of the 30 days as if the data are temporally independent. The RMSPE (root mean square prediction error) measures the predictive performance. At day t, we define the RMSPE by

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \hat{O}_i)^2},$$

n being the number of stations to be predicted, O_i, the true measurement at station i and $\hat{O}_i$, the prediction. Table 1 summarizes the RMSPE of the model outputs, calibrated model outputs with Bayesian melding and spatial-temporal model. Figure 1 shows the scatter plots of BM model calibrated model outputs versus uncalibrated model outputs and Figure 2 shows scatter plots of

Bayesian spatial-temporal model calibrated model outputs versus uncalibrated model outputs.

We have the following conclusions from the calibration results. model output, we have the following conclusions.

- Figures 1 and 2 reveal marked differences between calibrated and uncalibrated model outputs because most points deviate from the solid line with intercept 0 and slope 1. In fact, both of the two approaches proposed in this paper, although very different in nature agree in suggesting that MAQSIP overestimates measurements at the high end of the scale and underestimates measurements at the low end.

- Table 1 shows that on average the spatial-temporal model calibrated model outputs have the smallest RMSPE than the uncalibrated ones and Bayesian melding calibrated model outputs also have smaller RMSPE than uncalibrated ones. After calibration, the mean RMSPE has been reduced by 16.07% for Bayesian melding model, 21.48% for spatial-temporal model. Out of all the 30 days, both BM and Bayesian spatial-temporal model calibrated model outputs have smaller RMSPEs than uncalibrated ones for 25 days. So it is beneficial to calibrate the model outputs by using either Bayesian melding or spatial-temporal model.

- Figure 3 shows that both BM and Bayesian spatial-temporal calibrated model outputs are closer to the measurements than the uncalibrated ones at most stations.

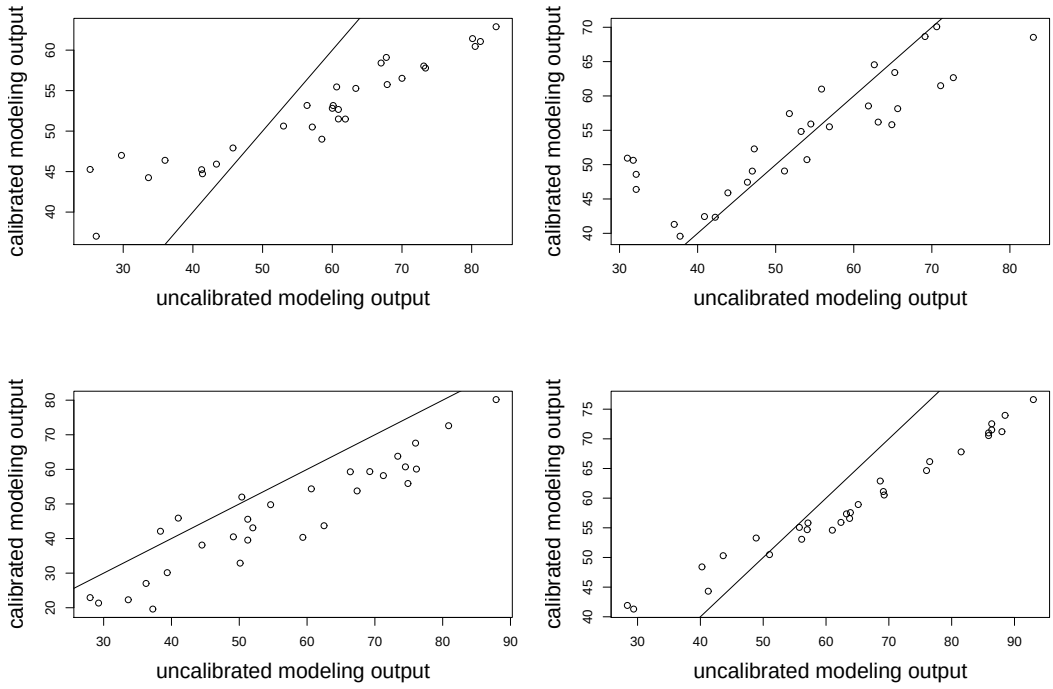

Figure 1: Scatter plots of uncalibrated model outputs versus BM calibrated model outputs. The solid lines have intercept 0 and slope 1. The plots are for days 2, 7, 8 and 26 from the upper left to the lower right in left-to-right sequence.

CONCLUSIONS

Results reported in the previous section, lead us to conclude that the spatial temporal model recalibrates the model outputs better than Bayesian melding, because unlike the latter, the former "borrows strength" across time as well as space. More specifically, the recalibrated outputs come closer to their measured counterparts for the former than the latter. Nevertheless Bayesian Melding does have appeal since it addresses the misalignment problem directly unlike its purely statistical competitor. In fact, its approach resembles Reynold's averaging where sub - grid cell (micro - scale) processes are averaged out through

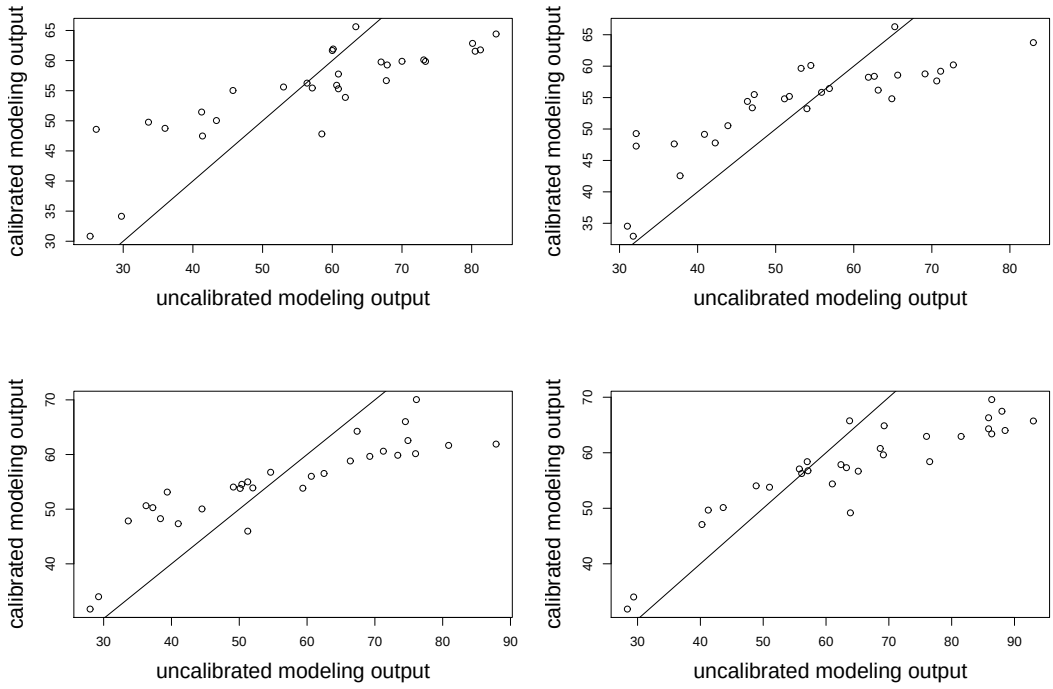

Figure 2: Scatter plots of uncalibrated model outputs versus Bayesian spatial-temporal model calibrated model outputs. The solid lines have intercept 0 and slope 1. The plots are for days 2, 7, 8 and 26 from the upper left to the lower right in left-to-right sequence.

integration.

However, both agree that MAQSIP overestimates measured ozone concentrations at the high end and underestimates them at the low end. It is not clear if this finding carries over to other CTMs such as GEOS - CHEM, nor whether result would apply once the anthropogenic sources are suppressed in the model. These results seems generally in accord with findings reported in [2] and [3] which report amongst other things that GEOS - CHEM outputs are about 10 ppbv too high in the southeastern United States in summer and

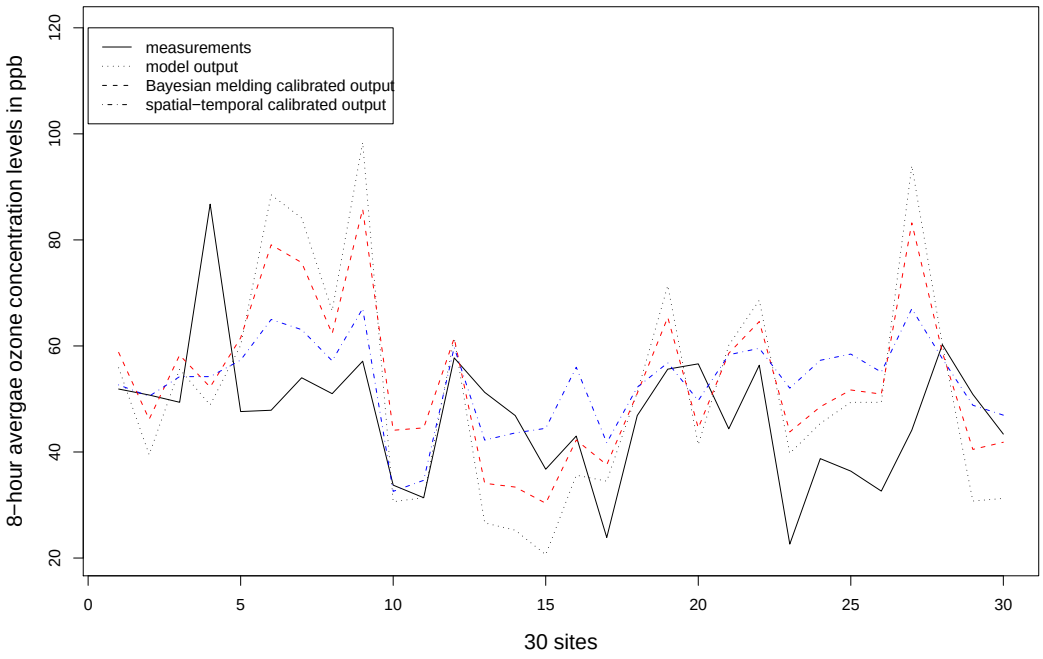

Figure 3: Measurements versus the uncalibrated and calibrated model output of 30 grid cells on day 2. The solid line shows the measurements inside the grid cells. The dotted and dashed lines are the uncalibrated and calibrated model outputs respectively.

in excess of measured levels in highly populated coastal areas. The second of these papers finds the O_3 to be in the range of 15 - 35 (ppbv) with "occasional incidences of 40 - 50 ppbv at high altitude western sites in spring".

Since the PRB inferred from GEOS - CHEM lies at the low end of the scale, our results based on MAQSIP leads to concerns that the imputed level of PRB is too high. Were this true for GEOS - CHEM, the excess disease outcome counts (above the PRB baseline) due to ozone would be underestimated pointing to a need for even more stringent regulations than those proclaimed on March

12, 2008 by the EPA Administrator. Clearly this is an issue that needs to be addressed more carefully prior to the next review of the US ozone standards.

ACKNOWLEDGEMENTS

We thank Prasad Kasibhatla for providing the measurements and AQM model outputs of ozone air pollution for the application addressed in this paper. We also thank him for stimulating conversations about its topic while the third author was visiting the Statistical and Applied Mathematical Sciences Institute.

References.

1. Garner, J.; Lewis, T.; Hogsett, W.; Andersen, C. Air quality criteria for ozone and related photochemical oxidants (Second external review draft) Volume III, *U.S. Environmental Protection Agency*, 2005.

2. Fiore, A.M.; Jacob,I.B.; Yantosca, R. M.; Field, B.D.; Fusco, A.C.; Wilkinson, J.G. Background ozone over the United States in summer: Origin, trend, and contribution to pollution episodes, *J. Geophys. Res.*, 2002, 107(D15), 4275.

3. Fiore, A., D. J. Jacob, H. Liu, R. M. Yantosca, T. D. Fairlie, and Q. Li, Variability in surface ozone background over the United States: Implications for air quality policy, *J. Geophys. Res.*, 2003, 108(D24), 4787.

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

5. Gelfand, A.E.; Smith, A.F.M. Sampling-based approaches to calculating marginal densities, *J Amer Statistic Assoc*, 1990, 85, 398-409.

6. Kalnay, E. Data assimilation and predictability, *Cambridge: Cambridge University Press*, 2003.

7. Liu, Z. Bayeisan Melding Model Software in R. In preparation, 2007.

8. Liu, Z.; Le, N.D.; Zidek, J.V. Calibrating deterministic modeling output with application of ozone fields, *TR 232 Department of Statistics, The University of British Columbia*, 2007.

9. Liu, Z.; Le, N.D.; Zidek, J.V. An appraisal of Bayesian melding for physical-statistical modeling, *TR 233,Department of Statistics, The University of British Columbia*, 2007.

10. Odman, M.T.;Ingram, C.L. Multiscale air quality simulation platform (MAQSIP) source sode documentation and validation, *Environmental Programs, MCNC-North Carolina Supercomputing Center*,1996.

11. Wheeler, N.; Houyoux, M. Development and implementation of a seasonal model for regional air quality, *Proc. Air and Waste Manage Association, Paper 98 - A739, A&WMA, Pittsburgh, PA*, 1998.

Table 1: RMSPE of uncalibrated and calibrated model outputs for the prediction of measurements at the 30 stations. column 1: day 1-30; column 2: RMSPE of uncalibrated model output; column 3: RMSPE of BM calibrated model output; column 4: RMSPE of Bayesian spatial-temporal model calibrated model output; The number with a * indicates the "winner" in that row.

Day	model output	BM	spatial-temporal
1	20.43	16.54	13.39*
2	19.88	10.89*	12.17
3	10.41	7.60*	8.78
4	14.82	13.79	12.00*
5	22.72	16.05	15.94*
6	23.67	15.09	14.02*
7	16.14	13.75	13.12*
8	15.36	14.42	11.30*
9	14.66	14.61*	15.06
10	12.23	11.19*	12.42
11	12.96	10.01	9.78*
12	17.41	16.19	15.00*
13	15.07	18.45	15.25*
14	20.83	15.77*	22.28
15	23.19	20.14*	22.79
16	16.49	17.28	16.22*
17	15.68	11.47	11.09*
18	15.89	15.35	13.83*
19	10.30	10.44	7.29*
20	14.20	11.59*	13.14
21	19.78	26.13	14.37*
22	17.96	11.07	8.98*
23	12.29	11.37	8.29*
24	15.87	10.02	9.53*
25	22.67	12.35*	13.92
26	18.41	14.06 *	14.12
27	16.59	9.99*	10.80
28	19.70	18.75	13.15*
29	13.44	11.33*	15.92
30	15.06	17.41	11.69*
mean	16.80	14.10	13.19*