Appendix A

**Regionalized PM2.5 Community Multiscale Air Quality model performance evaluation across a continuous spatiotemporal domain**

Jeanette M. Reyes,a Yadong Xu,a William Vizuete,a Marc L. Serrea\*

a: Department of Environmental Sciences and Engineering, UNC, 135 Dauer Drive, Chapel Hill, NC 27599-7431

\*Corresponding author: email: marc\_serre@unc.edu; telephone: 919 966 7014

**1. Model Performance Metrics**

In the model performance literature typical definitions exist regarding nomenclature of certain metrics, namely, “bias” and “error”. In the Statistics realm, metrics such as “bias” and “error” are defined differently. This work utilizes naming schemes of metrics that is consistent with Statistical language. In Table A1 denotes a modeled value at space/time point , is its paired observed value (i.e. observed at the same space/time location) and is the corresponding error. In Table A2 denotes an estimate of the error standard deviation, which in our work is obtained using .

**Table A1. Table of commonly used model performance evaluation statistics used in the CMAQ literature. The left column displays the typical nomenclature used and the right column displays the nomenclature used in this work. The third column states whether the metrics quantifies systematic or random error. The second half of the table displays metrics less commonly used in the literature.**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric name used in the CMAQ literature | Definition | Systematic / Random | Metric Name used in this Work |
| Regulatory Performance Metrics Used in Air Quality Modeling | | | |
| # of data pairs |  | NA | **# of data pairs** |
| Mean observation value |  | NA | **Mean observation value** |
| Mean simulation value |  | NA | **Mean modeled value** |
| Mean bias |  | Systematic | **Mean error** |
| Normalized bias |  | Systematic | **Mean normalized Error** |
| Normalized mean bias |  | Systematic | **Normalized mean Error** |
| Fractional bias |  | Systematic | **Fractional error** |
| Mean Error |  | Systematic / Random | **Mean absolute error** |
| Normalized Error |  | Systematic / Random | **Mean normalized absolute error** |
| Normalized Mean Error |  | Systematic / Random | **Normalized mean absolute error** |
| Fractional error |  | Systematic / Random | **Fractional absolute error** |
| Correlation |  | Random | **Correlation** |
| Less Commonly Used Regulatory Performance Metrics | | | |
| Correlation squared |  | Random | **Correlation squared** |
| Standard bias |  | Random | **Standard error** |
| Mean squared bias |  | Systematic / Random | **Mean squared error** |
| Root mean squared bias |  | Systematic / Random | **Root mean squared error** |
| Normalized root mean squared bias |  | Systematic / Random | **Normalized root mean squared error** |
| Mean bias/standard bias |  | Systematic | **Mean error/standard error** |
| Mean bias squared/mean squared bias |  | Proportion of Systematic | **Mean error squared/mean squared error** |
| Variance of bias/mean squared bias |  | Proportion of Random | **Variance of errors/mean squared error** |

**Table A2. Table of model performance evaluation statistics used when the estimate has a corresponding variance . These metrics are used in the validation statistics.**

|  |  |
| --- | --- |
| BME Metric | Definition |
| Variance standardized error |  |
| Root mean squared standardized error |  |
| Mean root variance |  |

|  |  |  |
| --- | --- | --- |
| ME2 🡹, VE🡻 | ME2 🡻, VE 🡹 | ME2 🡹, VE 🡹 |
| (a) | (b) | (c) |
|  |  |  |
| (d) | (e) | (f) |
|  |  |  |
| (g) | (h) | (i) |
|  |  |  |

**Figure A1. Visual representations of systematic and random error. Panels (a)-(i) show different scenarios of high and low systematic and random errors. The left column (plots (a), (d), (g)) displays three different visual representations of estimates with large systematic error (i.e. high ME2) and low random error (i.e. low VE). The middle column (plots (b), (e), (h)) displays representations of estimates with low systematic error and large random error. The right column (plots (c), (f), (i)) displays representations of estimates with large systematic error and large random error. The top row displays error using a target analogy, where estimates should ideally land on the target. The middle row displays the distribution of error via a PDF. The bottom row displays a group of paired modeled and observed concentrations within a space/time region . The modeled values are displayed on the independent axis as and the observed values are displayed on the dependent axis as . The solid line is the one-to-one line.**

**2. Data**

*2.1 Observed Data*

The daily PM2.5 concentration for each monitoring site/day during 2000-2002 were constructed based on raw monitoring data from monitoring stations measuring either hourly or daily PM2.5 concentrations using the procedure described below.

PM2.5 monitoring data (raw data) sampled during the study period (2000-2002) were obtained from the Air Quality Systems (AQS) database maintained by the EPA, a repository of the monitoring data collected across various monitoring networks. PM2.5 data are available in a few data files on AQS depending on the source of data. These files are described in AQS as follows: 1) daily PM2.5 local conditions, 2) daily PM fine speciation from the Chemical Speciation Network (CSN) monitoring network and 3) daily PM fine speciation from the Interagency Monitoring of Protected Visual Environments (IMPROVE) monitoring network. Within each data file, the methodologies used to measure PM2.5 are defined using a parameter code which takes the following values: 1) 88101 for daily and hourly PM2.5 concentrations measured using a Federally Referenced Method (FRM) and 2) 88502 for PM2.5 Air Quality Index (AQI) values that provide acceptable measurements of PM2.5 concentrations in that they are comparable to FRM measurements. Data from the parameter 88502 are also known as Tapered Element Oscillating Microbalance (TEOM) data.

Hourly PM2.5 data were averaged into daily PM2.5 if at least 18 out of 24 hours were measured for a given day/monitor. Otherwise, a daily average was not constructed. More than 99.9% of hourly records were reported every hour on the hour. However, there were several records not reported on the hour. These hourly records were removed before constructing daily concentrations. All observations sampled at monitors whose measurement scale was “Microscale” were removed.

At each monitoring site with multiple monitors, the collocated daily concentrations recorded at any given day were combined using the following procedure to produce a constructed daily concentration for that site/day. First, priority rank scores were assigned to each collocated daily concentration based on its data source and type as follows:

**Table A3. Ranking scores used for averaging collocated PM2.5 values for a given site/day**

|  |  |
| --- | --- |
| Rank 1 | FRM daily PM2.5 |
| Rank 2 | TEOM daily PM2.5 from CSN |
| Rank 3 | TEOM daily PM2.5 from IMPROVE |
| Rank 4 | TEOM hourly PM2.5 |

If the collocated concentrations for a given site/day had varying priority ranks, then only the concentration with the highest rank (i.e. the smallest priority score) was retained. If there were more than one collocated daily concentrations with the highest priority rank, then these daily concentrations were averaged to produce a single daily concentration at that site/day.

*2.2 Modeled Data*

Daily concentrations for PM2.5 were also constructed from modeled CMAQ data. CMAQ inputs emissions and meteorological data which are then translated into complex chemical processes to estimate ambient air pollution over gridded geographical boundaries for different time steps. The modeled data used for this work were available at a 36km resolution every hour for the years 2001 and 2002 across the continental US. Data are projected using a Lambert Conic Conformal (LCC) projection.

**Table A4. Description of available CMAQ modeling data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | Model | Domain | Resolution | Source | Received date |
| 2001 | CMAQ v4.5 | The contiguous US | 36km | EPA | 08-26-2011 |
| 2002 | CMAQ  v4.5 | The contiguous US | 36km | CENRAP | 08-02-2011 |

The CMAQ data have full spatial and temporal coverage for the continental US (Table A4). All modeled runs were done using hindcasting. Daily modeled values were constructed by averaging the 24 hourly modeled values for a given grid location/day. Each modeled concentration over an area (i.e. the modeled concentration over a grid cell) is assigned the location defined by the centroid of that grid cell.

**3. Choice of S-curve Parameters for the RAMP analysis**

We conducted a visual analysis to select the parameters used to construct the and S-curves in the RAMP analysis. In the CAMP analysis and are calculated across the domain . In the CAMP analysis the S-curves do not capture the variability of model performance at a fine spatial scale. To address this issue the RAMP analysis regionalizes the calculation to a space/time region that consists of paired modeled/observed values at the closest stations within days of . Our aim is to capture fine scale spatial variability in the S-curves. In order to achieve the finest spatial resolution possible we choose . Using proximal stations define the smallest region possible near any space/time location of interest, which provides a description of model performance at the finest spatial resolution possible for any given location of interest. The time window was thus increased to days to allow approximately 150 paired values needed in the creation of the S-curve. To test whether days and leads to a stable analysis we performed a validation and stochastic simulation analysis that demonstrates that RAMP produces and values that outperforms CAMP. We conducted a visual sensitivity analysis by increasing and inspecting whether the and maps change appreciatively as increases (Fig. A2). As seen in this figure, the and maps do not change appreciatively as increases from 3 to 6. The same result is obtained using other values of between 3 and 6 (results not shown). This demonstrates that the parameters chosen (180 days and ) are as spatially specific as possible while still maintaining a stable estimation of and .

Another choice for the number of station would be to use a fixed radius , and select all stations within of the location of interest. We found that in order to achieve a stable estimate of and , the radius has to be set to a long distance such that at least 3 stations are included in the most sparsely monitored area of the continental US. When moving to densely monitored areas, the number of station within the fixed radius becomes so large that essentially the RAMP method becomes equivalent to the CAMP method in these areas, and as a result this approach fails to assess model performance at a fine spatial resolution.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |

**Figure A2. Maps of and across the US on July 1, 2001 calculated using the RAMP method with two sets of S-curve parameters. is displayed in (a) and (c). is displayed in (b) and (d). (a) and (b) are obtained using the closest stations within days of , while (c) and (d) are obtained using the closest stations within days of . No appreciable difference can be seen by comparing (a) and (b) against (c) and (d).**

**4. Model Performance Metrics for Different Fixed Modeled Values**

The RAMP method allows for construction of model performance metrics as a function of both space/time region and arbitrary modeled values. Through the equation and an analogous equation for , one can visualize how mean error and variance of error changes across the United States for a given day for different increasing modeled values. Mean error for PM2.5 decreases consistently across the United States from to . CMAQ has consistent problems estimating PM2.5 at smaller concentrations and high concentrations. CMAQ performs best at mid-range values of PM2.5 (Fig. A3). Standard error increases as a function of modeled value (Fig. A4). Change in the standard deviation of errors demonstrates that model performance is non-homoscedastic. The RAMP method allows the flexibility of allowing model performance to be both non-linear and non-homoscedastic. Non-homoscedastic behavior is most clearly observed in the Great Lakes region of the county.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |

**Figure A3. Systematic error () for fixed modeled values for (a) and (b) across the US on July 1, 2001 using the RAMP method as interpolated by S-curves. Areas where are not displayed are outside the range of modeled values for the corresponding S-curve and are therefore not interpolated.**

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |

**Figure A4. Random error () for fixed modeled values (a) and (b) across the US on July 1, 2001 using the RAMP method as interpolated by S-curves. Areas where are not displayed are outside the range of modeled values for the corresponding S-curve and are therefore not interpolated. The boxed area in green corresponds to the Great Lakes region.**

**5. Maps of Other Model Performance Metrics**

All metrics can be visualized for each CMAQ grid cell. When looking at systematic error and random error as a proportion of total error, most error coming from CMAQ can be defined as random (Fig. A5a-b). Because such a large proportion of total error is coming from random error, visually, maps of random error and total error look similar (Fig. A5c). MNAE (Fig. A5e) is lowest in the Eastern part of the US where overall performance of CMAQ is known to be better. MNE has a large range illustrating the large potential normalized errors when observed values are small (Fig. A5d). Visually, the maps of MNE look similar to .

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | |
|  | |
| (d) | (e) |
|  |  |

**Figure A5. Maps of daily PM2.5 across the continental United States on July 1, 2001 displaying (a) , (b) , (c) , (d) and (e) .**

**6. and for Different Fixed Modeled Values**

Patterns of for increasing modeled values follow along similar lines of . That is, performance is most consistent with mid-levels of PM2.5. Most systematic error is seen with low and high concentrations. The corresponding error correction of shows the lowest levels of variation for , and . Areas that show the highest levels of error correction in the US start in the Eastern US and move predominately to the Appalachian Mountain region of the country. Patterns of for increasing modeled values follow along similar lines of . That is, performance is most homogenous for lower levels of PM2.5. Most random error is seen with mid to high level concentrations.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |
| (e) | (f) |
|  |  |

**Figure A6. Maps of PM2.5 concentrations in increments of across the continental United States on July 1, 2001 with (a) being and (f) being . Units for all figures are .**

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |
| (e) | (f) |
|  |  |

**Figure A7. Maps of PM2.5 concentrations in increments of across the continental United States on July 1, 2001 with (a) being and (f) being . Units for all figures are .**

**7. RAMP Stochastic Simulation**

We do not know the true values of and because they are not directly measured. As a result it is not possible to validate the RAMP method based on true or measured and values. However, we can use stochastic simulation to create a set of simulated values for , , and that have the same statistical properties as the true values. These simulated values are taken as the simulated truth, which can then be used to validate the RAMP method.

To do this, and that were obtained in this work were selected as the “true” values. From the selected and and the CMAQ modeled values used in this work, and were obtained by substituting with in and , respectively. Observed data were statistically simulated by randomly generating a stochastic realization . The set of , , and values represent the simulated truth. The validation then consists of using the Constant, CAMP and RAMP methods to obtain and based on a re-estimation that uses *only* the and . The re-estimates and can be compared with the selected “true” and to evaluate the estimation accuracy of each method. Fig. A8a-b show and , respectively. Fig. A8c-d show and , respectively, for the day of interest. The results of the re-estimation for the RAMP methods are shown in Fig. A8e-f. As can been seen in the figure, the maps of and are very similar to those of and , respectively, which demonstrates that the RAMP method is able to correctly estimate and across the space/time study domain based only on paired and values.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |
| (e) | (f) |
|  |  |

**Figure A8. Maps of stochastic simulation of daily PM2.5 across the continental United States on July 1, 2001. (a) Selected true , (b) selected true , (c) CMAQ modeled values for that day, (d) stochastic realization , (e) and (f) re-estimated by the RAMP method based only on and .**

In order to assess how well the RAMP method evaluates systematic errors compared to other methods, we show in Fig. A9a that was selected as the “truth” across the continental United States on July 1, 2001. We show in Fig. A9b-d values that were obtained using the Constant, CAMP and RAMP methods. The Pearson Correlation coefficient between and is for the Constant, for the CAMP and for the RAMP methods. These correlation values and the corresponding figures demonstrate that the Constant method is not able to capture any of the spatial variability in systematic errors and the CAMP method captures only some of the spatial variability of systematic errors. By comparison the RAMP method captures the spatial variability of systematic errors well.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |

**Figure A9. (a) Map of the selected true for daily PM2.5 across the continental United States on July 1, 2001, and maps of the corresponding re-estimated obtained using the (b) the Constant method, (c) the CAMP method and (d) the RAMP method.**

In order to assess how well the RAMP method is able to assess systematic errors compared to other methods, we show in Fig. A10a that was selected as the “truth” across the continental United States on July 1, 2001, and we show in Fig. A10b-d values that were obtained using the Constant, CAMP and RAMP methods. These maps indicates that the Constant method is unable to capture the spatial variability in random errors, the CAMP method captures only some of the spatial variability of random errors and, by comparison, the RAMP method is able to capture areas high and low random errors well.

|  |  |
| --- | --- |
| (a) | (b) |
|  |  |
| (c) | (d) |
|  |  |

**Figure A10. (a) Map of the selected true for daily PM2.5 across the continental United States on July 1, 2001, and maps of the corresponding re-estimated obtained using the (b) the Constant method, (c) the CAMP method and (d) the RAMP method.**

It should be noted that the and values were obtained using *only* the paired modeled and randomly generated observed values and yet the RAMP method is able to capture areas of high and low systematic and random errors across the continuous mapping domain. This demonstrates that the RAMP method is able to assess model performance at unsampled locations, i.e. at locations where observations are *not* available.

**8. S-curves across the US**

|  |  |
| --- | --- |
| (a) | (b) |
| \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x-170km_y-1200km.png | \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x-1800km_y-400km.png |
| (c) | (d) |
| \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x-1950km_y-400km.png | \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x-800km_y-940km.png |
| (e) | (f) |
| \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x1100km_y-780km.png | \\Aa.ad.epa.gov\ord\RTP\Users\E-J\jreyes02\Net MyDocuments\ToORISE\Publications\Reyes_AE_RAMP_US\current_stuff\SCurvesKeep\Scurve_20010701_x975km_y-700km.png |

**Figure A11. S-curves for selected CMAQ grids on July 1, 2001 across the continental US. Vertical cyan line marks and the horizontal cyan line marks .**