5.3 On the use of Principal Component and Spectral Density Analyses to evaluate the Community Multiscale Air Quality (CMAQ) Model

Brian Eder, Wyat Appel, Thomas Pierce

Atmospheric Modeling and Analysis Division, National Exposure Research Laboratory, United States Environmental Protection Agency; Research Triangle Park, NC 27711

1. Motivation

The scope of air quality model evaluation has expanded greatly recently, in both methodology development as well as the amount of modeled data that needs to be evaluated. As recently as five years ago, modelers would have at most a short simulation, perhaps a month in length to evaluate. As computing capabilities increased, so did the modeler's ability to simulate over greater spatial and temporal scales, with annual and even decadal, continental-scale simulations becoming possible. While such large scale simulations can be advantageous, they can also provide new challenges, as the amount of output that needs to be evaluated can become overwhelming. Rudimentary techniques and performance statistics become less elucidating, even when broken down into seasonal or monthly periods (Appel et al., 2008; Eder and Yu, 2006).

Accordingly, the purpose of this paper is to introduce and demonstrate a multivariate statistical technique called Principal Component Analysis (PCA), with the hope of motivating the evaluation community in its use. Though infrequent, the use of PCA and similar analyses in the evaluation of air quality models is not unprecedented. For example, Li et al. (1994) use PCA to evaluate the Eulerian Acid Deposition and Oxidant Model (ADOM) against aerosol and gas observations in Ontario, Canada. With Li's and other similar applications, the PCA was performed separately on model output and observations, and then subsequently compared. The approach demonstrated here is different, in that we apply PCA directly to a measure of the Community Multi-scale Air Quality (CMAQ) model's performance, namely the modeled SO_4^{2-} bias (CMAQ concentration – CASTNet concentration). Keep in mind that the analysis could be applied to any measure of performance (i.e. root mean square error) and for any specie concentration or deposition simulated by CMAQ.

The advantage of using such an approach is that it will identify any systematic patterns of model bias across a myriad of spatial and temporal scales (i.e. not constrained to geopolitical boundaries nor monthly/seasonal time periods). Such analysis is useful in that it: (1) provides "weight of evidence" concerning the regional-scale nature of any CMAQ bias patterns; (2) facilitates understanding of the probable mechanisms responsible for the statistically unique behavior among bias patterns; and (3) identifies stations that can be used as indicators for more diagnostic evaluation.

2. Data

2.1 CMAQ Simulation

This evaluation used a five-year (2002-2006) simulation of CMAQ (Version 4.7) released in December 2008 (Byun and Schere, 2006). The modeling domain covered the contiguous United States using a 36 km x 36 km horizontal grid resolution (148 (columns) x 112 (rows) = 16,576 grid cells) and a 24-layer logarithmic vertical structure, extending from the surface to ~ 100 hPa. The meteorological fields were provided from MM5, the Fifth-Generation Pennsylvania State University/National Center for Atmospheric Research (NCAR) Meso-scale Model and were processed using the Meteorological-Chemistry Interface Program (MCIP). This 5-year simulation used the CB05 gas-phase chemistry mechanism. The emissions, which were processed using the

Sparse Matrix Operator Kernel Emissions (SMOKE) processor, were based on EPA's 2002 National Emissions Inventory, with year-specific fire, mobile (from MOBILE6), biogenic (from Biogenic Emission Inventory System (BEIS) v. 3.13) and major point source Electrical Generating Unit (EGU) data.

2.2 CASTNet

Currently operated by EPA's Clean Air Markets Division (CAMD), CASTNet is a long-term, predominately rural monitoring network designed to measure and characterize broad-scale spatial and temporal trends of pollutants contributing to acidic deposition. While the primary purpose of the Network is to assess the efficacy of emission control strategies established by EPA (i.e. the NO_x "State Implementation Plan" Call), its long-term, high-quality data also makes it ideal for use in evaluation of deterministic air quality models like CMAQ. Ambient air concentrations are measured weekly (Tuesday to Tuesday) from a height of 10 m using an open faced, three-stage filter pack. The CASTNet data used in demonstration of this evaluation technique consist of ambient air concentrations of particulate SO_4^{2-} ($\mu g\ m^{-3}$).

The CMAQ modeled concentrations were post-processed in order to achieve spatial (monitor to grid cell, with no interpolation) and temporal (weekly) compatibility with the CASTNet observations. A total of 45 stations were selected, focusing on the Eastern United States (locations of the CASTNet sites are shown in the left panel of Fig. 1). Given CMAQ's five year simulation period and CASTNet's weekly sampling schedule, a total of 256 weekly observations were available (Tuesday, Jan. 1, 2002 through Tuesday, Dec. 26, 2006), resulting in a total of 11,520 data pairs. Missing observation data were imputed by using a spline interpolation scheme, across time.

3. Methodology

3.1. Spatial analysis

The rotated principal component analysis begins with the calculation of a square, symmetrical correlation matrix **R** (having dimensions of 45 x 45) from CMAQ's SO_4^{2-} bias matrix having dimensions of 45 (CASTNet stations) x 256 (weekly biases) and containing 11,520 observations. By using **R** and the identity matrix (**I**) of the same dimensions, 45 eigenvalues (λ) were derived that satisfy the polynomial equation:

$$\det\left[{}_{45}R_{45} - \lambda_{45}I_{45}\right] = 0 \qquad (1)$$

For each root of (1), which is called the *characteristic equation*, a non-zero vector (**e**) can be derived such that:

$${}_{45}R_{45}\, e_1 = \lambda_{45} e_1 \qquad (2)$$

where **e** is the eigenvector of **R**, associated with its corresponding eigenvalue (λ). The eigenvectors represent the mutually orthogonal linear combinations or "modes of variation" of the bias matrix, while their respective eigenvalues represent the amount of variance explained by each of the eigenvectors. By retaining only the first few eigenvector-eigenvalue pairs, collectively called the *principal components*, a substantial amount of the total variance of each pattern of bias can be explained while ignoring the higher order principal components which explain smaller amounts of the variance. The exact number of components that should be retained was determined by examination of a Scree plot and revealed that a 10 component solution was most appropriate.

When the elements of each eigenvector are multiplied by the square root of the associated eigenvalue ($\lambda^{0.5}$), one obtains the principal component loading (L), which represents the correlation between the bias component and the CASTNet station. The retained principal components were then "rotated", to facilitate spatial interpretation,

using an orthogonal rotation. This rotation technique increases the segregation between component loadings, which in turn better defines the areas of homogenous bias. The station(s) with the highest loading in each of the bias regions can then be designated as the "bias pattern indicator", subsequently allowing more focused diagnostic analysis.

3.2. Temporal analysis

Having identified the spatial patterns of bias, examination of their time series was then achieved through the calculation of the rotated *principal component scores* (PC_s). The PC_s for week w on component i are weighted, summed values, whose magnitudes are dependent upon the weekly bias (B_{wj}) for week w at station j and the L_{ji} is the loading of station j on component i as seen below:

$$\left(PC_s\right)_{wi} = \sum_j B_{wj} L_{ji} \qquad (3)$$

The PCs are standardized, so they have a mean of zero and standard deviation of one. Positive scores correspond to positive CMAQ bias, while negative scores represent negative CMAQ bias. When plotted as a time series, the weekly PC_s provide excellent insight into the spectrum of temporal variance experienced by each bias pattern. A four week moving average was applied to each time series to aid in interpretation.

Spectral Density Analysis (SDA) using the finite Fourier transformation was also applied to each of the time series. This analysis decomposes each time series into a sum of cosine and sine waves of varying amplitudes and wavelengths, yielding a measure of the distribution of the bias variance over a continuous spectrum of all possible wavelengths. The abscissa on the spectral density plots ranges from 0 to 262 weeks, which corresponds to cycles or periodicities from as little as 2 weeks to as long as trends.

4. Results

4.1. Spatial analysis

A total of 10 principal components were deemed significant by the Scree test, which together explained 69.8% of CMAQ's SO_4^{2-} bias. For the sake of brevity, we present the results for just one of the components (PC_3), which, with an eigenvalue (λ_3) of 3.94, explained 8.8% (3.94/45) of CMAQ's total SO_4^{2-} bias. Examination of the left of panel Fig. 1 reveals this bias component's unique spatial characteristic, as the component loadings (which again represent the correlation between the pattern and the CASTNet stations) clearly indentify five high-elevation locations (Lye Brook (Loading: 0.74), elev.: 730 m; Claryville (0.70), 765 m; Horton Station (0.66), 920 m; Shenandoah (0.60), 1073 m and Cranberry (0.50), 1219 m. Of these five stations, four are classified by CASTNet as mountain top, with the fifth (Claryville) classified as a complex terrain site.

4.2. Temporal analysis

Examination of this component's principal component scores time series (right panel of Fig. 1) is equally compelling in that a strong, systematic pattern of bias is revealed in which CMAQ tends to over-predict SO_4^{2-} concentrations at these locations during the first six months of the year, under-predict during the months of July, August and September, then over predict from October through December. The strength of this pattern, which is strongest in 2002, 2003 and 2005, is affirmed in Fig. 2, which depicts a "typical year" constructed by using the median of each of the five years of simulations (left panel) and a SDA plot revealing a statistically significant cycle at 52 weeks (right panel.) Note the periodicity at 26 weeks may be a spurious alias of the 52 week and will be investigated. Having identified and characterized this statistically unique pattern of SO_4^{2-} bias, the next step in the evaluation approach (not discussed) will involve a diagnostic analysis of its most representative site (Lye Brook), which should facilitate

understanding into the process(es) responsible for this systematic bias. The other principal components characterized nine equally compelling systematic patterns of bias, each providing "indicator stations" that will be the focus of future diagnostic evaluation.

5. Conclusions

Examination of only one of the ten components has demonstrated the advantage of using principal component and spectral density analyses, in that they have identified systematic patterns of CMAQ SO_4^{2-} bias across spatial and temporal scales unconstrained by geopolitical boundaries and calendar periods. Such analysis has provided "weight of evidence" concerning the regional-scale nature of these patterns while identifying stations that should be used for diagnostic evaluation, which will lead to a better understanding of the mechanisms responsible for the modeled bias.

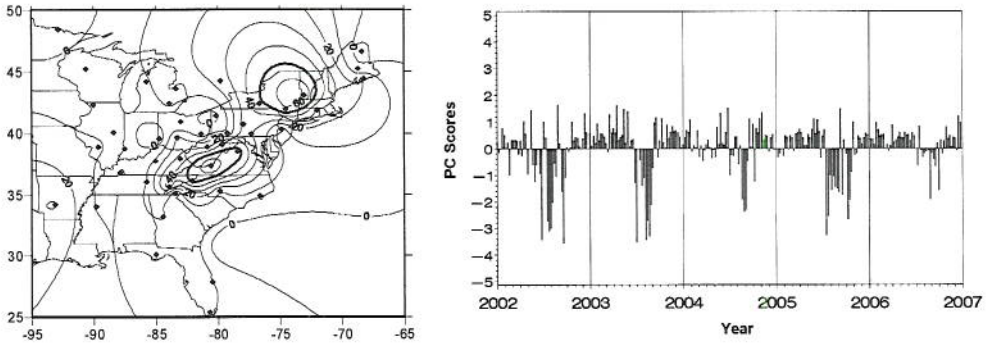

Fig. 1. Component loadings (x10) at each of the 45 paired CASTNet – CMAQ locations (left panel) and time series of the principal component scores (right panel) associated with the third rotated principal component.

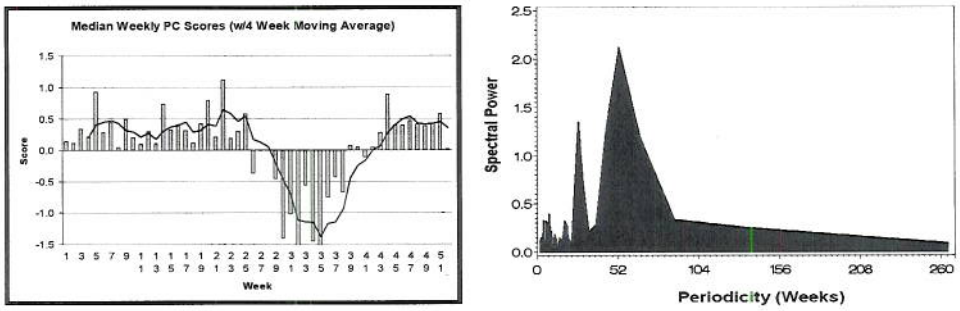

Fig. 2. Weekly median values of the principal component scores (with a four week moving average - left panel) and SDA of the raw time series (right panel) associated with the third rotated principal component.

References

Appel, W., P. Bhave, A. Gilliland, G. Sarwar and S. Roselle, 2008. Evaluation of the CMAQ model *v* 4.5: Sensitivities impacting model performance; Part II–particulate matter. *Atmos. Environ.*, **42**, 6057-6066.

Byun, D. and K. Schere, 2006. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality modeling system. *Applied Mechanics Review*, **59**, 51-77.

Eder, B. and S. Yu, 2006. A performance evaluation of the 2004 release of Models-3 CMAQ. *Atmos. Environ.*, **40**, 4811-4824.

Li, S., K. Anlauf, H. Wiebe, J. Bottenheim and K. Pucket, 1994. Evaluation of a comprehensive Eulerian air quality model with multiple chemical species measurements using principal component analysis. *Atmos. Environ.*, **28**, 3449-3461.

6. Questions and Answers

Question: How do you deal with missing data? (*Jeremy Silver*)

Answer: Principal component analysis should not be applied to a dataset that contains missing data. Accordingly, all missing data were imputed using a spline interpolating scheme. To limit the amount of interpolation, stations with less than 90% capture were excluded from the analysis.

Question: If the PCA was based on model error at points, how were the spatial maps made? (*Bruce Denby*)

Answer: The maps were produced by Kriging the principal component loadings. It is noted that care must be used when interpreting such results in areas where data are sparse.