Application of Kolmogorov-Zurbenko Filter for the Dynamic Evaluation of a Regional Air Quality Model

Daiwen Kang[1], Christian Hogrefe[2], Kristen Foley[2], Sergey Napelenok[2], Rohit Mathur[2], and S. Trivikrama Rao[2]

[1] Computer Science Corporation, Research Triangle Park, 79 T.W. Alexander Drive, NC 27709, USA
[2]Atmospheric Modeling and Analysis Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

Corresponding Author: Daiwen Kang

Computer Science Corporation
79 T.W. Alexander Drive
Building 4401 Room 430A
Research Triangle Park, NC 27709

(919) 3616211
kang.daiwen@epa.gov

September, 2012

Abstract

Regional air quality models are being used in a policy-setting to estimate the response of air pollutant concentrations to changes in emissions and meteorology. Dynamic evaluation entails examination of a retrospective case(s) to assess whether an air quality model has properly predicted the air quality response to known changes in emissions and/or meteorology. In this study, the Kolmogorov-Zurbenko (KZ) filter has been used to spectrally decompose pollutant time series into different forcings that are controlled by different atmospheric processes affecting the predicted and observed pollutant concentrations. Through analyses of the different components influenced by different forcings in the dynamic evaluation, we can discern which of the component(s) or scale(s) of forcing are simulated well by the model and the component(s) or scale(s) of forcing needing further improvement in the model. The KZ filter has been applied to both the observed and Community Multiscale Air Quality (CMAQ) model simulated summertime ozone (O_3) time series in years 2002 and 2005. The 2002-2005 time period is a good candidate for the dynamic evaluation case study because of the large NOx emission changes that occurred as a result of the U.S. Environmental Protection Agency's (USEPA) NO_X State Implementation Plan (SIP) call along with a more gradual decreasing trend in mobile emissions. Results suggest that the CMAQ model performs similarly for both years in terms of capturing the observed synoptic forcing. However, the changes in the observed ozone baseline component (i.e. longer-term variations) are not properly captured by the model at some locations. The factors contributing to the ozone baseline include emissions, boundary conditions, and other parameters that vary slowly over time. Analysis using a reduced form model developed from the sensitivity coefficients calculated from the decoupled direct method in three dimensions (DDM-3D) reveals that ground-level NO_X emissions, especially those from mobile and area source

sectors, may be overestimated in 2005 as there is an increase in the mean bias for the 2005 ozone simulations from its 2002 level at the majority of AQS sites located within the states affected by the NO_X SIP Call.

1. Introduction

Regional-scale atmospheric chemical transport models are being widely used in the development and implementation of air pollution control rules and regulations (Bachmann, 2007; Gego et al., 2008; Cohan et al., 1994), air quality management (Frost et al, 2006), and short-term air quality forecasting (Otte et al., 2005; Kang et al., 2005; Eder et al., 2009). Due to the complex interrelationships among the variables that constitute the photochemical modeling systems, it can be very challenging to perform process-level model evaluations (Dennis et al., 2010). Using standard model evaluation metrics generated by pairing a modeled quantity with the corresponding observation, it is difficult to assess the respective impacts on model performance from the two primary inputs, i.e. meteorology and emissions, for photochemical air quality models (AQMs). Dynamic evaluations of regional air quality models relevant to air quality management and policies, assessing the model's response to changes in emission inputs or meteorology, have gained more attention in recent years (Dennis et al., 2010; Napelenok et al., 2011). For dynamic evaluation, retrospective cases are examined to evaluate whether the model has properly predicted the air quality response to known changes in emissions and/or meteorology (Dennis et al., 2010). A useful dynamic evaluation case study is the time period between 2002 and 2005 because of the large NO_X emission changes due to the implementation of U.S. Environmental Protection Agency's (USEPA) NO_X State Implementation Plan (SIP) call in addition to a more gradual decreasing trend in mobile emissions (Gilliland et. al, 2008; Godowitch et al., 2010; Pierce, et al., 2010; Napelenok et al, 2011). Previous dynamic evaluation studies (Gilliland et. al, 2008; Godowitch et al., 2010; Napelenok et al., 2011) that focused on this period found a somewhat smaller change in model predictions of daily maximum 8-hr (DM8HR) ozone (O_3) concentrations compared to the change seen in the observations. Specifically, these studies reported that the model predicts around 40-80% of the observed change and hypothesized that a lower magnitude for the modeled response may be attributable to errors in NO_X emission inputs used in the simulations. However, it should be noted that these changes in the pollutant levels are influenced both by changes in emissions as well as in meteorology, making it difficult to isolate the influence of either forcing. One approach for segregating the effects of emissions and meteorology on pollutant concentrations is to separate the various spectral components in the time series of atmospheric variables which represent different scales of forcing. Variations in ambient levels of pollutants can be thought of as the baseline of pollution that is created by emission sources and modulated by the prevailing weather conditions (Rao et al., 2011).

Several different techniques for separating the different scales of motion imbedded in atmospheric variables have been proposed (e.g. Goody et al., 1998; Salcedo et al., 1999; Hogrefe et al., 2003). The Kolmogorov-Zurbenko (KZ) filter can be used to separate time series data into the short-term component (influenced by the prevailing meteorological conditions) and baseline components (influenced by the emissions and boundary conditions) (Flaum et al., 1996; Milanchus et al., 1998). Because the KZ filter provides good separation of frequencies and does not require special treatment for missing data (Rao and Zurbeno, 1994; Eskridge et al., 1997), the KZ filter was applied to both observed and modeled ozone (O_3) concentration time series to examine the component characteristics in earlier studies (Rao et al., 1997; Vukovich, 1997; Hogrefe, et al., 2000). Since the purpose of this analysis is to evaluate the model performance from the dynamical evaluation perspective, the KZ filter is applied to the raw time series (i.e., original data in the ppb scale) for this study.

The objective of this study is to perform a dynamic evaluation of the Community Multiscale Air Quality (CMAQ) modeling system (Byun and Schere, 2006) for the NO_x SIP Call period by utilizing the KZ filter to decompose both modeled and observed time series into different forcings. Specifically, the change in the modeled components between the summers of 2002 and 2005 in the eastern United States is evaluated against the change seen in the corresponding observed components to gauge how well the model has reproduced the observed changes. In addition, we quantify the level of uncertainty in the 2005 mobile and area source emissions that would need to be assumed in order to improve the agreement between observed and simulated changes.

2. MODEL DESCRITPTION AND OBSERVATIONS

Model simulations were performed with 12 km x 12km horizontal grid cells covering the eastern United States nested within the 36 km x 36km grid covering the entire continental United States, including parts of Canada and Mexico. The modeled period spanned the period from 1 June to 31 August in both 2002 and 2005 with a 10-day spin-up at the end of May. CMAQ version 4.7.1 (Foley et al., 2010), with decoupled direct method in three dimensions (DDM-3D) (Napelenok et al., 2008) and with the latest available version of the carbon bond mechanism (CB05) (Sarwar et al., 2008) for atmospheric chemistry, was used to perform the simulations. Meteorological inputs were supplied by MM5 version 3.6.3 (Grell et al., 1994) configured with the standard physics options as described by Godowitch et al. (2010). Boundary conditions for the continental domain were specified based on the global model GEOS-Chem (Bey et al., 2001) for both years. Emissions were developed using the Sparse Matrix Operator Kernel Emissions (SMOKE) processor (http://www.smoke-model.org) version 2.4 based on temporally and spatially resolved wildfire, electricity generating units, and mobile sources. Domain-wide NO_X emissions reflected reductions from the SIP Call as well as reductions in the mobile sector. On average, point source emissions of NO_X were reduced by 22% and mobile source emissions were reduced by 18% between the summer of 2002 and the summer 2005 over the entire model domain; NO_X emissions from all other sources, including biogenic NO remained relatively unchanged over these two years (0.2% reduction).

O_3 observations used in this study were extracted from EPA's Air Quality System (AQS, http://www.epa.gov/air/data/aqsdb.html). Data from over 700 monitoring stations were available within the 12 km modeling domain during the two modeling periods.

3. METHODS OF ANALYSIS

3.1 KZ Filter and Error Analysis

The KZ filter and its application in the spectral decomposition of O_3 time series is described in detail by Hogrefe et al. (2000) and Rao et al. (2011). A more generalized form of presentation of the KZ filter and the spectral components are presented in Appendix A and the error propagation analysis in terms of the decomposed components is presented in Appendix B.

3.2 Perturbation of Emissions

Calculated DDM-3D sensitivity coefficients were used to provide responses to the perturbations in the uncertain inputs of the NO_x emissions categories (area and mobile) through Taylor series expansion (Morgan and Henrion, 1990; Hakami et al., 2003; Napelenok et al., 2011). Generally, pollutant concentration as a function of any one perturbation can be reconstructed using the following:

$$C_j(\bar{\mathbf{x}},t) = C_0(\bar{\mathbf{x}},t) + \Delta\varepsilon_j S_j^{(1)}(\bar{\mathbf{x}},t) + \frac{1}{2}\Delta\varepsilon_j^2 S_{j,j}^{(2)}(\bar{\mathbf{x}},t) + h.o.t. \qquad (1)$$

where $C_j(\bar{\mathbf{x}},t)$ is the concentration due to a specific perturbation *j* at time *t* and location $\bar{\mathbf{x}}$; $C_0(\bar{\mathbf{x}},t)$ is base, unperturbed concentration; $\Delta\varepsilon_j$ is the fractional perturbation of the parameter *j*; $S_j^{(1)}(\bar{\mathbf{x}},t)$ and $S_{j,j}^{(2)}(\bar{\mathbf{x}},t)$ are the first and second order sensitivity coefficients, and *h.o.t.* are higher order terms with little impact on the approximation.

To account for perturbations of more than one emission sector at once, for example area and mobile emissions, and dropping the higher order terms, the Taylor Series were expanded as follows:

$$C_{a+m}(\bar{\mathbf{x}},t) \approx C_0(\bar{\mathbf{x}},t) + \Delta\varepsilon_a S_a^{(1)}(\bar{\mathbf{x}},t) + \frac{1}{2}\Delta\varepsilon_a^2 S_{a,a}^{(2)}(\bar{\mathbf{x}},t)$$

$$+ \Delta\varepsilon_m S_m^{(1)}(\bar{\mathbf{x}},t) + \frac{1}{2}\Delta\varepsilon_m^2 S_{m,m}^{(2)}(\bar{\mathbf{x}},t) + \Delta\varepsilon_a \Delta\varepsilon_m S_{a,m}^{(2)}(\bar{\mathbf{x}},t) \qquad (2)$$

where the subscripts a and m represent area and mobile NO_X emissions respectively.

3.3 Verification Statistics

The *Root Mean Square Error (RMSE)* is often used to evaluate model performance since it quantifies the magnitude of the error in the model (Fox, 1981).

The RMSE can be estimated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{(m,i)} - C_{(o,i)})^2} \qquad (3)$$

where i is the ith paired (model-observation) data point, N is the total number of paired data points, and $C_{(m,i)}$ and $C_{(o,i)}$ are the ith modeled and observed mixing ratios, respectively.

For independent component decomposition as in Equation (B1), the relationship between the total RMSE and its component RMSEs can be expressed as:

$$RMSE(f) = \sqrt{RMSE(A)^2 + RMSE(B)^2} \qquad (4)$$

Mean Bias (MB) is another statistical metric commonly used in model evaluation. *MB* is defined as:

$$MB = \frac{1}{N}\sum_{i=1}^{N}(C_{(m,i)} - C_{(o,i)}) \qquad (5)$$

The MB provides the information on overall overprediction/underprediction in the forecasts.

4. RESULTS AND DISCUSSION

4.1 Characteristics of decomposed components

To demonstrate how the decomposed components behave in both observations and model simulations, Figure 1 presents the time series of the mean observed and modeled components for the 12 worst-performing sites (left panel) and the 12 best-performing sites (right panel) within the eastern US during the summer of 2005. The worst and best performing sites were chosen by ranking all the monitoring sites within the modeling domain by Index of Agreement (IOA, Kang et al., 2008) (metrics are calculated based on paring the raw hourly observations and the model predictions in space and time). The 12 worst-performing sites are the 12 sites that have the smallest IOA values, while the 12 best-performing sites are the 12 sites that have the largest IOA values. As shown in Figure 1, the short-term components – SY (b, f), DU (c, g), and ID (d, h) – are reproduced by the model reasonably well for both the worst-performing sites (a-c) and best-performing sites (e-g), though the time series at the best-performing sites seem to be in better agreement than at the worst-performing sites. The largest difference between the worst-performing sites and the best-performing sites is evident in the baseline (BL) component (Figures 1a and 1e). At the worst-performing sites, the temporal variation in the modeled BL is similar to that in the observations; however, the modeled BL level is almost 20 ppb larger than that of the observed, resulting in the large difference for the total O_3 concentrations between modeled and observed values. In contrast, at the best-performing sites, the model reproduced the observed BL quite well, both in phase and magnitude, and consequently, the model reproduced the observed total O_3 concentrations at these sites. These figures demonstrate that ozone concentrations on any given day are strongly influenced by the prevailing baseline as discussed by Rao et al. (2011). Consequently, if the model cannot correctly reproduce the processes to capture the baseline signal, the model will fail to correctly predict pollutant levels on any given day.

To assess the correlation between the different components separated by the KZ filter, Table 1 presents the Relative Dependency (RD) values as defined in Appendix B. The hourly O_3 mixing ratios at all the AQS monitoring locations available for both the 2002 and 2005 summers are decomposed into ID, DU, SY and BL components using KZ filter. The RD values are then calculated both for the full decomposition and for multiple combinations of the individual components. The first Row of Table 1 shows that the RD values for the full decomposition are from 25.4% (2002 model) to 33.9% (difference between 2002 model outputs and observations), meaning that the covariance among the components is more than 25% of the sum of the individual errors (if they are treated as independent). In other words, the components cannot be separated cleanly because they are correlated. If only the separations between ID and DU and between DU and SY are examined, the second and third rows of Table 1 show that the RD values are from 23 to 29% indicating that KZ filter is not able to separate these adjacent short-term components. However, the RD values between SY and BL (forth row of Table 1) are all reduced to less than 5% indicating that KZ filter is able to effectively separate the long-term component BL from the short-term component SY. To cross-check the separation ability of KZ filter, Table 1 also presents the RD values for non-adjacent components in the separation spectrum (Row 5 to Row 7). The RD values for ID and SY (DU in between) are less than 10%, for DU and BL (SY in between) from 0.3% to 2%, for ID and BL (DU and SY in between) less than 0.5%. If the three short-term components (ID, DU, and SY) are combined into one component ST (ST = ID + DU + SY), and the RD value between ST and BL is less than 5%, which is well within the acceptable error range if the two components, ST and BL, are treated as independent (Hogrefe et al., 2003).

To justify the independent treatment of ST and BL, Figure 2 displays the spatial distribution of RMSE values between modeled and observed hourly O_3, BL, ST, and the combination of BL and ST for the summer of 2005 (similar results are found for the 2002 simulations). The RMSE values of BL (Figure 2a) are generally smaller than the RMSE values of ST (Figure 2b) (on average, of the total error (square of RMSE for O_3) ~30% is associated with BL and ~70% is attributed to ST). The spatial patterns of RMSE values for hourly O_3 (Figure 2c) and the combined components (BL+ST) calculated by Eq. 4 (Figure 2d) are almost identical suggesting that the independent treatment of the two components (BL and ST) as the decomposition of hourly O_3 time series is acceptable.

4.2 Dynamical assessment of model performance for O_3 simulations in 2002 and 2005

The evaluation of the base CMAQ simulations analyzed in this study was presented in Napelenok et al. (2011). It is motivated by some preliminary analysis to split the modeling domain and the corresponding AQS sites into two groups: the SIP sites that are located within the states where the NO_X SIP Call program was implemented by 2004 and non SIP (NSIP) sites where the NO_X SIP Call program was not implemented. Figure 3 presents the states that were required to implement the NO_X SIP Call program or that were significantly affected by this program for the purpose of this analysis (http://www.epa.gov/airmarkets/progsregs/nox/docs/NBPbasicinfo.pdf). Model performance for the base case simulation and the corresponding SIP and NSIP sites for the two years was assessed using the traditional statistical metrics (Kang et al., 2005) and is summarized in Table 2. Even though these statistics are generally comparable to the previous studies for this domain (Eder and Yu, 2006; Appel et al, 2007), subtle features can be seen between the 2002 and 2005 results. Both NME and NMB increased for SIP sites (from 41% to 48% for NME and from 17% to 30% for NMB) and slightly decreased for NSIP sites (from 18% to 17% for NME and from 20% to 17% for NMB). Note, the MB/NMB values in 2002 for SIP and NSIP sites were the same, but the MB/NMB values at SIP sites in 2005 were almost twice as large as those at NSIP sites.

To further investigate the differences in the error and bias patterns between the simulations of these two years, process level analysis (utilizing the KZ decomposed components) is employed. Since the independent assumption is acceptable when hourly O_3 time series is decomposed into short-term (ST) and longer-term (BL) components using the KZ filter, we examine the two-component decomposition for the assessment of model performance for O_3 simulations in the summers of 2002 and 2005 in the following section.

Figure 4 presents the spatial map of the difference between 2005 and 2002 in RMSE values of BL (Figure 4a) and ST (Figure 4b) components for the summers of 2005 and 2002.

As shown in Figure 4b, the RMSE values associated with the ST components displayed only small changes (generally within ± 2.5 ppb) across the modeling domain between 2002 and 2005, indicating that the model performed similarly in simulating the short-term forcing (i.e., weather-induced variations).On the other hand, the BL RMSE values, shown in Figure 4a, significantly increased from 2002 to 2005 in the East, especially in the southeastern US (in the area surrounding Virginia, North Carolina, South Carolina, and Georgia), while the values decreased or changed little over the rest of the modeling domain.

Figure 5 displays the mean (spatially averaged over all SIP or NSIP sites) time series of the differences between the modeled and observed O_3 baseline for SIP sites and NSIP sites for 2002 and 2005, respectively. As shown in Figure 5, these time series of differences in the baseline have comparable magnitudes of about 6 ppb for 2002 at both the SIP sites and NSIP sites and for 2005 at the NSIP sites, indicating the presence of an overall positive mean bias (MB) in the modeled baseline. However, this bias becomes much more pronounced when considering the baseline differences between model predictions and observations for 2005 at the SIP sites, especially after July 1st. Since BL reflects the forcings induced by emissions and boundary conditions, the presence of BL errors can be a strong indication of errors in these key inputs. Since the MB values (Figure 5) at NSIP sites are comparable for the two years, the large differences in MB values for the two years at SIP sites are *unlikely* to be due to errors introduced by boundary conditions since such errors should have led to differences in BL model performance between 2002 and 2005 at the NSIP sites as well which are located closer to the boundary. Thus, Figure 5 strongly suggests that there were problems with emission inputs for the 2005 simulation.

To further investigate the impact of emissions on the MB values associated with BL components, the reduced form model (RFM) defined in Equations 1 and 2 was employed to estimate the effects of anthropogenic emission perturbations on the BL component. This analysis focuses on 2005 since the BL MB was particularly large at SIP sites during that year as discussed above. As mentioned earlier, the largest known changes in emissions between 2002 and 2005 are the reductions in NO_X emissions resulting from U.S. EPA's NO_X SIP Call program in the selected states (Figure 3). Additionally, a more gradual decreasing trend in mobile and area emissions was observed within the modeling domain. Since the CEM (Continuous Emissions Monitoring) data were used for the majority of the point source emissions in the CMAQ simulations (the point source inventory also contains other minor sources and the uncertainty associated with these sources may not be negligible), emissions from the utility sector are assumed to be accurate, and only mobile and area source emissions were perturbed in this study. The perturbation of emissions was implemented as follows: the fractional perturbation factor $\Delta\varepsilon$ (Equations 1 & 2) changed from 0.5 to -0.5 (corresponding to the ±50% emission uncertainty as stated in Napelenok et al. (2011)) with an increment of 0.05 for both 2005 mobile and area emissions, and the error metric (MB of 2005 BL in this case) was then calculated for each $\Delta\varepsilon$ at each site. At the end of the iteration, the minimum error (the smallest absolute value in the case of 2005 BL MB) and the corresponding $\Delta\varepsilon$ (referred to as optimal $\Delta\varepsilon$) is identified for each site. Note that limiting the perturbations to ±50% ensures that the RFM is applied within the range for which the sensitivity coefficients calculated by CMAQ-DDM have been shown to compare well with the results from brute force emission perturbation simulations (Cohan et al., 2005).

The optimal $\Delta\varepsilon$ values at each site for 2005 over the modeling domain are shown in Figure 6a and the time series of BL bias for the SIP and NSIP sites are displayed in Figure 6b for the CMAQ simulations and the RFM calculations using the optimal $\Delta\varepsilon$. As shown in Figure 6a, there is a mixture of positive and negative optimal $\Delta\varepsilon$ values within the modeling domain; however, the majority of the optimal $\Delta\varepsilon$ are negative values, especially within the SIP states, where some of the optimal $\Delta\varepsilon$ values reached the lower limit -0.5. As evidenced by Figure 6b, all MB values for both SIP sites and NSIP sites were significantly reduced when using the optimal $\Delta\varepsilon$ values to recalculate the O_3 BL using the RFM.

Since the major difference between the 2002 and 2005 simulations is the BL performance at the SIP sites, rather than calculating the optimal $\Delta\varepsilon$ that minimizes the BL bias for 2005, one may instead ask how much emission perturbation is needed to keep the same bias level for the 2005 simulation as that seen for the 2002 simulation. Figure 7 shows a map of the $\Delta\varepsilon$ values that are needed for the 2005 simulation to have the same bias as the 2002 simulation at any given site. For the results shown in Figure 7a, the 2005 target bias was the MB of the 2002 O_3 BL baseline component, while for Figure 7b the 2005 target bias was the MB of the 2002 DM8HR O_3 concentrations. In the following analysis, the $\Delta\varepsilon$ values displayed in Figures 7a and 7b are referred to as "$\Delta\varepsilon$ constant BL bias" and "$\Delta\varepsilon$ constant DM8HR bias", respectively. The main difference between these two approaches is that the BL O_3 bias is influenced by both daytime and nighttime values since the BL component is calculated from hourly values while the DM8HR O_3 bias typically is influenced by daytime values only. Figures 7a-b show that the patterns of the "$\Delta\varepsilon$ constant BL bias" and "$\Delta\varepsilon$

constant DM8HR bias" are very similar, however, the "$\Delta\varepsilon$ constant DM8HR bias" values tend to be smaller (typically more negative) than the "$\Delta\varepsilon$ constant BL bias" values. This implies that there is a greater difference in model performance between 2002 and 2005 when comparing daytime vs. nighttime simulations. As shown in Figure 7, the majority of $\Delta\varepsilon$ values in the SIP states are negative except for some isolated locations which may be influenced by some local VOC emission sources. In the Carolinas, the $\Delta\varepsilon$ values are predominantly negative (< -0.3, especially in Figure 7b), suggesting that the mobile and area emissions may be overestimated by more than 30% in 2005 assuming that emissions were unbiased in 2002 and that errors in 2005 emissions are driving the deterioration in model performance for 2005. However, when moving from east to west, the $\Delta\varepsilon$ values gradually increase, and in the Midwest states, the positive values dominate suggesting an underestimation of emissions for these locations in 2005, again assuming that emissions were unbiased in 2002 and that errors in 2005 emissions are driving the difference in model performance between 2002 and 2005.

To further investigate how the BL errors caused by uncertainties in area and mobile emissions impact the hourly and DM8HR O_3 concentrations, the emission perturbation factors in Figures 7a and 7b were plugged back into the RFM to calculate the hourly and DM8HR O_3 concentrations. The operational performance statistics for RFM (emission factors in Figure 7a) as listed in Table 2 (values in parentheses) indicates that compared to the base model, the RFM performed better than the base CMAQ simulation for all the metrics, especially at SIP sites. To evaluate how the observed changes from 2002 to 2005 are simulated by the base CMAQ model and the RFM using the "$\Delta\varepsilon$ constant BL bias" and "$\Delta\varepsilon$ constant DM8HR bias" emission factors, Figure 8 displays the scatter plots between the modeled change and the observed change for the mean DM8HR O_3 concentrations and the mean of all the DM8HR O_3 concentrations $\geq 95^{th}$% at each site, respectively. As shown in all panels of Figure 8, the RFM using adjusted emissions model performed better than the base model with smaller intercepts and significantly better (closer to one) slopes than the base CMAQ model. To further demonstrate the difference at SIP sites and NSIP sites, the average change (2005-2002) in the mean of the DM8HR O_3 over the summer and $\geq 95^{th}$% of the distribution across SIP sites and NSIP sites and the corresponding ratios of the modeled and observed change are listed in Table 3 (the RFM used the "$\Delta\varepsilon$ constant DM8HR bias" emission factors shown in Figure 7b). The change (both the mean value and the standard deviation σ) simulated by the RFM is much better than the base model to match the observed change for both the mean and the $\geq 95^{th}$% DM8HR O_3. At SIP sites, the ratio of modeled change to observed change increased from 48% in the base model to 71% in the RFM for the seasonal mean of the DM8HR O_3 and from 61% in the base model to 79% in the RFM for the $\geq 95^{th}$% DM8HR O_3. On the other hand, at the NSIP sites, there were only small changes between 2002 and 2005 for these two metrics and both the base model and the RFM performed similarly. For the $\geq 95^{th}$% maximum daily 8-hr O_3 at the NSIP sites, the base model overestimated the change, while the RFM reduced the overestimation from 37% to 15%. Note that the RFM not only increased the ratios of the mean values of the change, but the standard deviations associated with the RFM matched the observations much better than the base model (the ratios of the modeled standard deviation to the observed standard deviation changed from 53%-79% for the base model to 73%-95% for the RFM). The base model results are comparable to those reported in Gilliland et al. (2008) for the modeled change with CB4 taking into consideration that this earlier study included both the AQS sites and CASTNet sites, while only AQS sites were included in this analysis.

The above analysis is based on the assumption that only errors in mobile and area source emissions have caused the differences in the MB patterns for the two years considered. As Napelenok et al. (2011) pointed out, other sources of uncertainty including emissions of Volatile Organic Compounds (VOC), and other emission sources besides mobile and area sources not in the inventory, may also have influenced the model's ability to accurately predict a change in the ozone baseline during this time period. As noted earlier, boundary conditions likely played an insignificant role in explaining the differences in BL performance for the two years, but the influence of VOC emissions may not be ignored at some locations. As also pointed out by Napelenok et al. (2011), at some locations during certain time periods, the relative change of VOC/NO_X ratio may lead to possible O_3 formation regime change by transitioning some NO_X-limited regions to become VOC-limited due to the fact that NO_X and VOC emissions were changed in 2005 from their 2002 levels. Most noticeably, even though negative $\Delta\varepsilon$ values dominate the northeast region, some large positive values greater than 0.3 occurred at sites in the northeastern urban corridor along the interstate 95 in Figure 7, and it is likely that the chemical regime for the formation of O_3 changed at these locations. Nevertheless, it is demonstrated that uncertainties in low level NO_X emissions likely are one of the major factors that caused the increased bias of simulated O_3 baseline in 2005 from the 2002 levels. This study reveals that it is difficult for a regional air quality model to reproduce the observed change in air quality induced by the NOx SIP Call given large uncertainties in the mobile and area source emission inventories.

5. SUMMARY

The ability to separate the impacts of meteorology and emissions, the primary inputs to air quality models, on the formation, destruction, and transport of atmospheric pollutants for performance evaluations of air quality models is extremely valuable. The time series of an atmospheric pollutant can be decomposed into different temporal components. In hourly time series data, the short-term (ST) component is considered to be the combined effect of intra-day forcing (stochastic processes), diurnal forcing (solar variation, heating or cooling cycle, mixing height change, etc.), and synoptic weather evolutions. The baseline (BL) component represents slow changing emission inputs, boundary conditions, and trends of climate change. It is very challenging to find a technique that can be used to cleanly decompose the hourly time series of air pollutants into different components (spectra), particularly for small temporal scales. However, the KZ filter is a good method to separate O_3 time series effectively into the various spectral components that are of most interest to policy-making. The error propagation analysis for both the observed and modeled O_3 time series indicates that it is acceptable to treat the KZ decomposed ST and BL components as independent components.

In terms of model simulations, the impact of most parameterization of physical and chemical processes is also filtered into the BL component. Thus, the effective separation of the ST and BL components can greatly facilitate dynamical model evaluations in terms of a model's separate responses to meteorology, emissions, and change in model configurations. The application of this technique in conjunction with the DDM-3D method to the 2002 and 2005 simulations suggests that the NO_X emission inputs for the 2005 simulations in the states where the NO_X SIP Call program was implemented were not represented as well in emission inventories for ground-level sources as in 2002. Specifically, the 2005 NO_X emissions, especially those from mobile and area sectors, may have been overestimated at majority of the AQS sites within the NO_X SIP Call states by more than 50% at some locations such as the Carolinas.

A time series of atmospheric pollutant can be viewed as the time series of its BL component being superimposed on the time series of its ST component. The BL component forms the underlying forcing of the time series, while the ST component represents the influence of weather conditions on the baseline. If an air quality model fails to reproduce the BL component of an air pollutant, no matter how well it simulates the ST component, it is impossible for the model to reproduce the observed air pollutant concentration levels. This analysis demonstrates that model performance for these types of applications will be driven by the ability of modeled BL component to reproduce the observed BL. In most of the regulatory applications of air quality models, such as assessment of emission control strategies, the primary concern is the models' ability to simulate the mean levels of air pollutants; when the mean level is reduced, we can expect the entire distribution will be affected.

Appendix A. KZ filter and spectral decomposition

A time-series of hourly species (S) data can be presented by:

$$S(t) = ID(t) + DU(t) + SY(t) + BL(t) \qquad (A1)$$

Where *S(t)*: original time-series, ID(t): intra-day component, DU(t): diurnal component, SY(t): synoptic component, and BL(t): baseline component.

The KZ(*m*, *p*) filter (a filter with window length *m* and *p* iterations) can be defined by:

$$KZ_{m,p} = \Re_{i=1}^{p}\{\Im_{k=1}^{W_i}[\frac{1}{m}\sum_{j=-(m-1)/2}^{(m-1)/2} S(t_i)_{k,j}]\} \quad \{\begin{matrix}\Im: running window \\ \Re: iteration\end{matrix} \qquad (A2)$$

$$W_i = L_i - m + 1 \qquad L_i : length\ of\ S(t_i)$$

The components of interest in this study are estimated as follows:

$$ID(t) = S(t) - KZ_{3,3} \qquad (A3)$$
$$DU(t) = KZ_{3,3} - KZ_{13,5} \qquad (A4)$$
$$SY(t) = KZ_{13,5} - KZ_{103,5} \qquad (A5)$$
$$BL(t) = KZ_{103,5} \qquad (A6)$$

Appendix B. Error Propagation and Relative Dependency

To investigate how components (A3) – (A6) relate to each other, we examine the characteristics of the error distribution. According to error propagation theory, if

$$f = A + B \qquad (B1)$$

then the relationship of the errors (from repeat measurements or reproduction) is:

$$\sigma_f^2 = \sigma_A^2 + \sigma_B^2 \pm 2COV_{AB} \qquad (B2)$$

where COV_{AB} is the covariance between A and B and its magnitude is a measure of the independence of A and B. If $COV_{AB} = 0$ then A and B are completely independent. In other words, A and B can be independently measured or reproduced and the total error in the combined quantity (f) is simply the sum of errors of the individual components (A and B). To measure the degree of independency between components, we define a metrics referred to as relative dependency (RD):

$$RD = \left| \pm \frac{2COV_{AB}}{\sigma_A^2 + \sigma_B^2} \right| = \left| \frac{\sigma_f^2}{\sigma_A^2 + \sigma_B^2} - 1 \right| \times 100\% \qquad (B3)$$

RD is simply the ratio of the covariance to the sum of the individual component errors. This can be applied to more than two components, and the covariance term would be the sum of covariance between any two components and the denominator would still be the sum of all the individual component errors. The smaller the RD values are, the more independent of the components are.

In terms of the model evaluation with hourly observations, there are two paired time series at each site: S(obs, t) and S(mod, t). To apply KZ filter to each of the time series, we have the following decomposed relationships:

$$S(obs,t) = ID(obs,t) + DU(obs,t) + SY(obs,t) + BL(obs,t) \qquad (B4)$$

$$S(\mathrm{mod},t) = ID(\mathrm{mod},t) + DU(\mathrm{mod},t) + SY(\mathrm{mod},t) + BL(\mathrm{mod},t) \qquad (B5)$$

The two time series and the corresponding components can be treated as two separate measurements or reproduction of the same quantity. Even though the RD metric is developed using two components (A and B) above, it can be applied to more than two components. However, when more than two components (for example, f = A+B+C) are involved, cautions shall be exercised under some special occasions. For instance, if A and B are highly positively correlated and A and C are highly negatively correlated, then the RD for A+B and for A+C is very high but the RD for A+B+C is much lower. Therefore, when more than two components are involved, in addition to the RD values of total sum, it is also suggested to compute the RD values for each of the pair-wise sums of the components (such as A+B, B+C, and A+C) to avoid the misinterpretation of the RD values resulting from these special situations.

Acknowledgements and Disclaimer

The authors thank Brian Eder and Shawn Roselle for their helpful comments on the manuscript. Although this manuscript has been reviewed and approved for publication, it does not necessarily reflect the policy or views of the US Environmental Protection Agency.

6. REFERENCES

Appel, K. W., Gilliland, A. B., Sarwar, G., Gilliam, R. C., 2007. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: sensitivities impacting model performance Part I – ozone. *Atmos. Environ.* 41 (40), 9603-9615.

Bachmann, J., 2007. Will the circle be unbroken: A history of the U.S. national ambient air quality standards. J. Air Waste Manage. Assoc. 57, 652-697.

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., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *J. Geophys. Res.-Atmos.* 106(D19), 23073-23095.

Byun, D.W., and Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Appl. Mech. Rev*., 59, 51-77.

Cohan, D., Stafford, R., Scheraga, J., Herrod, S., 1994. The global climate policy evaluation framework, in the Proceedings of the 1994 A&WMA Global Climate Change Conference, Phoenix, April 5–8, Air & Waste Management Association, Pittsburgh.

Cohan, D., Hakami, A., Hu, Y. T., Russell, A. G., 2005. Nonlinear response of ozone to emissions: Source apportionment and sensitivity analysis. *Environmental Science & Technology*, 39, 6739-6748.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D., Venkatram, A. , 2010. A framework for evaluating regional-scale numerical photochemical modeling systems. *Environ. Fluid Mech.* 10(4): 471-489.

Eder, B., Kang, D., Mathur, R., Pleim, J., Yu, S., Otte, T., Pouliot, G., 2009. A performance evaluation of the national air quality forecast capability for the summer of 2007, *Atmos. Environ.*, doi:10.1016/j.atmosenv.2009.01.033.

Eder, B., Yu, S., 2006. A performance evaluation of the 2004 release of model-3 CMAQ. Atmos. Environ. 40 (26), 4811-4824.

Eskridge, R. E., Ku, J.-Y., Rao, S. T., Porter, P. S., and Zurbenko, I. G., 1997. Separating different scales of motion in time series of meteorological variables, *Bull. Amer. Meteor. Soc.*, 78, 1473-1483.

Foley, K.M., Roselle, S.J, Appel, K.W., Bhave, P.V., Pleim, J.E., Otte, T.L., Mathur, R., Sarwar, G., Young, J.O., Gilliam, R.C., Nolte, C.G., Kelly, J.T., Gilliland, A.B, Bash, J.O., 2010. Incremental testing of the Community Multiscale Air Quailty (CMAQ) modeling system version 4.7. *Geosci. Model Dev.*, 3, 205 – 226.

Fox, D. G., 1981. Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Performance. *Bull. Amer. Meteor. Soc.*, 62, 599-609.

Flaum, J. B., Rao, S. T., and Zurbenko, I. G., 1996. Moderating the influence of meteorological conditions on ambient ozone concentrations. J. Air Waste Manag. Assoc., 46, 35-46.

Frost, G. J., McKeen, S. A., Trainer, M., Ryerson, T. B., Neuman, J. A., Boberts, J. M., Swanson, A., Holloway, J. S., Sueper, D. T., Fortin, T., Parrish, D. D., Fehsenfeld, F. C., Flocke, F., Peckham, S. E., Grell, G. A., Kowal, D., Cartwright, J., Auerbach, N., and Habermann, T., 2006. Effects of changing power plant NOX emissions on O3 in the eastern United States: proof of concept. J. Geophys. Res., 111:D12306, doi:10.1029/2005JD006354.

Gego, E., Gilliland, A., Godowitch, J., Rao, S. T., Porter, P. S., and Hogrefe, C., 2008. 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 Manag. Assoc., 58, 580-588.

Gilliland, A. B., Hogrefe, C., Pinder, R. W., Godowitch, J. M., Foley, K. L., Rao, S. T., 2008. Dynamic evaluation of regional air quality models: Assessing changes in O-3 stemming from changes in emissions and meteorology. *Atmos. Environ.,* 42(20), 5110-5123.

Godowitch, J. M., Hogrefe, C., and Rao, S. T., 2008. Diagnostic analyses of a regional air quality model: Changes in modeled processes affecting ozone and chemical-transport indicators from NO_X point source emission reductions. J. Geophys. Res., 113:D19303, doi:10.1029/2007JD009537.

Godowitch, J. M., Pouliot, G., Rao, S. T., 2010. Assessing multi-year changes in modeled and observed NOx concentrations from a dynamic evaluation perspective. *Atmos. Environ.*, 44(24), 2894-2901.

Goody, R., Anderson, J., and North, G., 1998. Testing climate models: An approach. *Bull. Amer. Meteor. Soc.*, 79, 2541-2549.

Grell, G. A., Dudhia, J., Stauffer, D. R., 1994. A description of the 5th generation Penn State/NCAR Mesoscale Model (MM5). NCAR Technical Note. NCAR/TN-398+STR.

Hakami, A., Odman, M. T., Russell, A. G., 2003. High-order, direct sensitivity analysis of multidimensional air quality models. *Environ. Sci. Technol*. 37 (11), 2442-2452.

Hogrefe, C., Rao, S. T., Zurbenko, I. G., and Porter, P. S., 2000. Interpreting the information in ozone observations and model predictions relevant to regulatory policies in the eastern United States, *Bull. Amer. Meteor. Soc., 81*, 2083-2106.

Hogrefe, C., Vempaty, S., Rao, S. T., and Porter, P. S., 2003. A comparison of four techniques for separating different time scales in atmospheric variables, *Atmos. Environ.*, 37, 313-325.

Kang, D., Eder, B.K., Stein, A.F., Grell, G.A., Peckham, S.E., McHenry, J., 2005. The New England air quality forecasting pilot program: Development of an evaluation protocol and performance benchmark. *J. Air Waste Manage. Assoc.*, *55*, 1782-1796.

Kang, D., R. Mathur, S.T. Rao, and S. Yu, 2008. Bias adjustment techniques for improving ozone air quality forecasts, *J. Geophys. Res.*, *113*, D23308, doi: 10.1029/2008JD010151.

Milanchus, M. L., Rao, S. T., and Zurbenko, I. G., 1998. Evaluating the effectiveness of ozone management efforts in the presence of meteorologica variability. J. Air Waste Manag. Assoc., 48, 201-215.

Morgan, M. G., Henrion, M., 1990. Uncertainty: A guide to dealing with uncertainty in quantitative risk and polic analysis. Cambridge University Press, pp. 172-220.

Otte, T.L., and coauthors, 2005. Linking the Eta Model with the Community Multiscale Air Quality (CMAQ) modeling system to build a national air quality forecasting system. *Wea. Forecasting*, *20*, 367-384.

Pierce, T., Hogrefe, C., Trivikrama Rao, S., Porter, P. S., Ku, J.-Y. , 2010. Dynamic evaluation of a regional air quality model: Assessing the emissions-induced weekly ozone cycle. *Atmos. Environ.* , 44(29), 3583-3596.

Napelenok, S. L., Cohan, D. S., Odman, M. T., Tonse, S., 2008. Extension and evaluation of sensitivity analysis capabilities in a photochemical model. *Envion. Modell. Softw.*, 23(8), 994-999.

Napelenok, S. L., Foley, K. M., Kang, D., Mathur, R., Pierce, T., Rao, S. T., 2011. Dynamic evaluation of regional air quality model’s response to emissions in the presence of uncertain emission inventories. *Atmos. Eniron.*, 45, 4091-4098.

Rao, S. T., Porter, P. S., Mobley, J. D., and Hurley, F., 2011. Understanding the spatio-temporal variability in air pollution concentrations. EM, November 2011, 6-7.

Rao, S.T., Zurbenko, I.G. , Neagu, R., Porter, P.S., Ku, J.Y., and Henry, R.F., 1997. Space and time scales in ambient ozone data. *Bull. Amer. Met. Soc.*, 78, 2153-2166.

Rao, S.T., and Zurbenko, I.G., 1994, Detecting and tracking changes in ozone air quality, *J. Air Waste Manage. Assoc.*, 44, 1089-1092

Salcedo, R. L. R., Alvim Ferraz, M. C. M., Alves, C. A., and Martins, F. G., 1999. Time series analysis of air pollution data. *Atmos. Environ.*, 33, 2361-2372.

Sarwar, G., Luecken, D., Yarwood, G., Whitten, G. Z., Carter, W. P. L., 2008. Impact of an updated carbon bond mechanism on predictions from the CMAQ modeling system: Preliminary assessment. *J. Appl. Meteorol. Clim.,* 47(1), 3-14.

Vukovich, F. M., 1997. Time scales of surface ozone variations in the regional, non-urban environment. *Atmos. Environ.*, 31, 1513-1530.

Table 1. Mean RD (%) values of component signals for hourly O_3 time series during summers of 2002 and 2005

Seq	Signal Components	2002		2005	
		OBS	MOD	OBS	MOD
1	O3 = ID+DU+SY+BL	28.6	25.4	30.9	33.3
2	ID + DU	23.2	23.4	22.9	23.4
3	DU + SY	22.9	29.0	21.2	26.5
4	SY + BL	4.8	4.2	9.1	7.5
5	ID + SY	8.8	9.3	10.3	9.1
6	DU + BL	1.2	1.6	1.8	2.0
7	ID + BL	0.1	0.2	0.3	0.04
8	O3 = ST + BL (ST=ID+DU+SY)	3.2	3.9	4.7	4.7

Table 2. Operational model evaluation for hourly ozone concentrations in 2002 and 2005 for both SIP sites (located within the states where the State NO_X SIP Call program was implemented) and NSIP sites (located within the states where the State NO_X SIP Call program was not implemented). The values in the parentheses for 2005 correspond to the reduced form model (RFM) performance using emission factors as shown in Figure 7a.

	$Mean_{obs}$ (ppb)	$Mean_{model}$ (ppb)	RMSE (ppb)	NME (%)	MB (ppb)	NMB (ppb)	r
2002 All	**35.4**	**41.8**	**19.4**	**41.9**	**6.4**	**18.2**	**0.64**
SIP Sites	**38.7**	**45.1**	**20.6**	**40.9**	**6.4**	**16.6**	**0.64**
NSIP Sites	**32.0**	**38.4**	**18.0**	**43.2**	**6.4**	**20.2**	**0.63**
2005 All	**32.7**	**40.1 (38.9)**	**18.8 (17.3)**	**44.5 (40.5)**	**7.4 (6.2)**	**23.4 (18.9)**	**0.61 (0.66)**
SIP Sites	**32.8**	**42.6 (40.1)**	**20.2 (18.2)**	**47.9 (42.8)**	**9.8 (7.3)**	**30.1 (22.0)**	**0.61 (0.66)**
NSIP Sites	**32.5**	**37.9 (37.7)**	**17.3 (16.3)**	**41.1 (38.1)**	**5.4 (5.2)**	**16.7 (15.7)**	**0.62 (0.66)**

Table 3 Average change (2005-2002) in the seasonal mean of the daily maximum 8-hr ozone and ≥ 95th% of the distribution across sites.

	Observed Change	**Modeled change**	**RFM change**
All Sites (N=782)			
Average ±σ	-4.5±6.2 ppb	-2.4±3.2 ppb	-3.1±4.9 ppb
Ratio (mean, std. dev)		53%,53%	69%,79%
≥95th%			
Average ±σ	-8.6±9.6 ppb	-6.2±6.4 ppb	-7.3±8.2 ppb
Ratio (mean, std. dev)		72%,67%	85%,85%
SIP Sites (N=408)			
Average ±σ	-8.3±4.9 ppb	-4.0±2.6 ppb	-5.9±4.3 ppb
Ratio (mean, std. dev)		48%,53%	71%,88%
≥95th%			
Average ±σ	-14.0±7.4 ppb	-8.5±5.4 ppb	-11.1±7.0 ppb
Ratio (mean, std. dev)		61%,73%	79%,95%
NSIP Sites (N=374)			
Average ±σ	-0.4±4.7 ppb	-0.6±2.9 ppb	-0.05±3.5 ppb
***Ratio (std. dev)**		±62%	±74%
≥95th%			
Average ±σ	-2.7±8.1 ppb	-3.7±6.4 ppb	-3.1±7.3 ppb
Ratio (mean, std. dev)		137%,79%	115%,90%

***The mean value of the change is too small to take the statistically meaningful ratio.**