Statistical properties of differences between low and high resolution CMAQ runs with matched initial and boundary conditions

Chae Young Lim*, Michael Stein†, Jason Ching‡, and Ruen Tang§

February 6, 2009

Abstract

Numerical simulations of air quality models provide unique and different outputs for different choices of grid size. Thus, an important task is to understand the characteristics of model outcomes as a function of grid size in order to assess the quality of the model as to its fitness for meeting a specifc design objective. This type of assessment is somewhat different than that of traditional operational performance and diagnostic type model evaluation. There, the objective is towards assessing errors in numerical models of air quality and utilize concentration measurements from monitors to provide the bases for guidance towards model improvement and for their assessment of ability to predict and retrospectively map air quality. However, observations used as "truth" to assess model performance have themselves properties unique to the data collection protocols, siting and spatial density of deployment. In the data assimilation community, the term "model error" is used for the difference between model output given perfect inputs and the "truth" (Kalnay, 2003). In this paper, we are concerned with one aspect of this "model error", the discrepancy due to discretization of space by choice of grid size in the model. To understand discrepancy due to discretization, outputs from the Community Multiscale Air Quality model (CMAQ) at two resolutions are studied. The lower resolution run is carried out so that its initial and boundary conditions are as similar as possible to those for the higher resolution run, thus minimizing this source of discrepancies and allowing us to isolate discrepancies due to discretization. Differences are analyzed from a statistical perspective by comparing marginal

*Department of Statistics and Probability, Michigan State University. A426 Wells Hall, East Lansing, MI 48824, Email: lim@stt.msu.edu, Phone: +1 517 353 7154, Fax: +1 517 432 1405.

†Department of Statistics, University of Chicago. 5734 S. University Avenue, Chicago, IL 60637, Email: stein@galton.uchicago.edu, Phone: +1 773 702 8326, Fax: +1 773 702 9810.

‡Atmospheric Modeling Division, National Exposure Research Laboratory, USEPA, RTP, NC 27709, Email: Ching.Jason@epa.gov

§CSC, RTP, NC 27709, E-mail: Tang.Ruen@epa.gov

distributions of two outputs and considering spatial variation of the differences. Results indicate sharp increases in spatial variation of the differences for the first few hours of running the model, followed by small increases thereafter. The spatial variation of the differences depends on the individual spatial structure of the original processes, which we show varies with the time of day. We also show that the spatial variations on sub-regions depend on whether the sub-region is in a rural or an urban area.
Keywords: CMAQ fine scale modeling; Sub-grid variability; Multiscale air-quality modeling; Spatial variogram

1 Introduction

Outputs from numerical models of geophysical phenomena differ from reality for a number of reasons, including problems with model inputs, imperfect and/or inadequate descriptions of processes in the model and errors from numerical simulation methods and discretizations. Distinguishing between these various sources of error for most applications would be difficult even if one had perfect observations to compare to model output; the problem is made much harder by the fact that observations are generally not sufficiently dense to compare with model outputs, have their own errors (Brohan et al., 2006; Hegerl et al., 2001) and often do not measure the same quantities that models produce due, for example, to differing levels of aggregation in space, time, and/or chemical composition (Gelfand et al., 2001; Gotway and Young, 2002).

Quantifying and attributing error to appropriate sources is critical to understanding and improving numerical models. One particularly important use of numerical models is forecasting, for which distinguishing errors due to the model formulation as opposed to model inputs can play a major role. For example, the Kalman filter, which is a popular method for data assimilation in geophysical models, requires a statistical specification (e.g., through its covariance structure) of "model error", or the error produced and propagated by a model over some period of time when given perfect input. Together with the observation error, the difference between the truth and the observation, the model error is an important ingredient to performing the data assimilation (for example, see Kalnay (2003)).

When forecasting weather or ocean currents, although its importance for analysis has been noted by Cohn (1993), Dee (1995) and Houtekamer et al. (2005), model error is often either assumed to be zero (i.e., the only source of error in model outputs is incorrect inputs) or assumed to have a simple structure. For example, Mitchell and Houtekamer (2000) and Houtekamer et al. (2005) assumed horizontal-vertical separability and horizontal isotropy for a model error covariance matrix in their Ensemble Kalman filter scheme. Evensen (2003) simulated model error using a relatively simple, time evolving equation. Their studies assumed simple spatial and temporal forms for the covariance structure of the model error. Learning more about the spatial and temporal structure of the model error and being able to specify it should lead to more efficient assimilation schemes and more realistic assessments of forecast uncertainties. Developing better approaches to quantifying model error should be of considerable value not only for improving data assimilation, but also for providing model developers with a better understanding of the strengths and weaknesses of their

models.

One big obstacle to studying model error is the unknown truth which we should compare with the model output. One could compare model output to observation, as a surrogate of the truth. However, the differences in these two quantities are due to many sources other than model error as we use the term here, including errors in emissions and initial and/or boundary conditions, which result in imperfect inputs for the model, and errors in observations. Even if one believes observation errors at monitoring sites were minimized, there is still incommensurability between volume-averaged model outputs and point wise observations (Park et al., 2006; Swall, 2008). Thus, our approach to study model error is to compare model outputs themselves instead of comparing model outputs to observations.

In this work, we mainly focus on discrepancy due to discretization of space in a numerical air quality model. Results of air quality models can depend strongly on discretization (for example, see Mousavi et al. (1999), Zoras et al. (2007) and references therein). We study the discrepancy due to discretization by comparing two model outputs at different resolutions with matched Initial and Boundary conditions (IC/BC) so that we minimize discrepancy due to IC/BC. Although we study only a part of the model error by restricting ourself to the discrepancy due to discretization, we believe that this component of the model error can be well understood by our approach and thus provides a meaningful lower bound to the total model error. Furthermore, we find this component of error is not small. We denote the difference between simulation outputs at two resolutions as the SubGridscale Discrepancy, or SGD and consider SGD as a surrogate for the actual discrepancy due to discretization in the model.

Model outputs from CMAQ at different resolutions were also studied by Ching et al. (2006) in which a conceptual framework for comparing observations and model outputs was proposed when the distribution of sub-grid variability is available. Its distribution was then estimated using model outputs at fine resolution within each coarse grid cell to illustrate their framework. Some characteristics of sub-grid variability were also studied by looking at the difference between coarse scale outputs and aggregation of fine scale outputs. In Ching et al. (2006), the model outputs were obtained from the usual practice of CMAQ simulation. The simulation outputs we study here are different from the usual practice in CMAQ simulation in that we ran CMAQ at various starting times with matched IC/BC for two grid resolutions to see propagation of the difference over time while minimizing possible errors due to IC/BC. Thus, one of the main differences between the work by Ching et al. (2006) and our work is that we investigate the difference between two model outputs from the specialized CMAQ simulation so that we are able to distinguish the discrepancy due to discretization in the model from discrepancies due to IC/BC.

The present work makes use of some specialized runs of a complex air quality modeling system, CMAQ (Community Multiscale Air Quality model). For regional scale simulations, CMAQ is typically run using grid sizes of order 36, 12 and as small as 4 km. As grid size decreases, smaller scale features become resolved and are certainly more evident. For example, Jiménez et al. (2006) argue that even for a secondary pollutant like O_3, 2 km runs capture small-scale features that do not appear with coarser grids. Due to many contributing factors, including chemical nonlinearities, and other processes working at sub-grid scales,

the simulated concentration at a coarse resolution is *not* in general equal to the simulated concentration of a fine resolution run aggregated to coarse resolution (Ching et al., 2006). These differences are due to the inability of coarse scale modeling to adequately incorporate processes operating at sub-grid scales. Examples of such processes include: inadequate spatial characterization of emissions sources within grid; the variable land-use structures and vegetation and its corresponding different degrees of pollutant deposition, dispersion parameters due to sub-grid variations in turbulent mixing, mixing height and underlying surface roughness; the variability of biogenic precursor emissions due to subgrid variability in vegetation; the subgrid variability in the incoming radiation due to subgrid cloud fields, thus inducing variability in heating rates, photochemically active radiation (PAR) and surface energy budgets. For photochemically reactive species, the model's chemical mechanism operates on the chemical mixtures where each precursor compound is based on the sum of the emission from all source within the grid. Thus, the potentially important sub-grid chemistry that must be occurring e.g., such as near source titration of ozone by NO_x is ignored at coarser scales and consequently, this error can be substantial for grids containing large and complex distribution of such sources.

The higher resolution simulation output will itself contain levels of SGD since it can not capture sub-grid scale behaviors at even higher resolution as well as variability of the original process. In addition, since both simulation runs use the same model, they may have a number of errors in common, which will tend to make their outputs similar. Thus, use of two simulations from the same model at different resolutions will undoubtedly lead to an underestimate of the total model error. However, we believe having even a lower bound that has a sound basis will be of considerable value, for example, in implementing assimilation schemes, especially since it turns out that this lower bound is not small.

The two sets of CMAQ simulations that we consider have grid cells of 1.33 km and 4 km on a side. The grid size of 1.33 km is the finest available resolution. There are some concerns whether increasing resolution produces more accurate prediction (Mass et al. (2002) and references therein). On the other hand, there are some works regarding model performance of MM5-CMAQ at fine resolutions ($1 \sim 2$ km) in an urban area, that show that the performance of the model at fine resolution is reasonably good. For example, Jiménez et al. (2006) show that the performance of various statistical parameters used for MM5-CMAQ modeling evaluation improves when decreasing grid spacing up to 2 km in the northwestern Iberian Peninsula, which is urban and industrial over a complex terrain. Sokhi et al. (2006) show that MM5-CMAQ modeling system at 1 km resolution reproduces the temporal trends well in London, UK, although it underpredicts the daily maximum level of O_3, especially nighttime concentrations. Also, they argue that statistical comparisons for the model performance is reasonable for the modeled and observed data.

In any case, for our approach to have merit, we do not require that the higher resolution output be closer to the actual pollution levels, only that the differences between model outputs have similar *statistical* properties as the actual discrepancy due to discretization (i.e., the errors we would get if discretization were the only problem with the model). Indeed, Mass et al. (2002) found that finer resolution model output explains spatial variation better although it may cause more timing and location errors in a numerical weather prediction

model. Since we mainly focus on the spatial variation in our study, comparing two runs with 1.33 km and 4 km resolutions will provide a useful statistical characterization of SGDs, thereby leading to an understanding of the discrepancy due to discretization.

In the usual simulations of CMAQ on a regional scale, the IC and BC are obtained by nesting within another model with coarser resolution. In other words, low resolution output provides the IC and BC to simulate high resolution output. More specifically, outputs with 4 km and 12 km resolutions were used in the simulation of 1.33 km and 4 km resolutions, respectively. 1.33 km output and 4 km output normally have different IC and BC as well as different resolution. The influence of IC/BC in air quality models is certainly another source of discrepancy when we look at the difference between model outputs at two resolutions (Jiménez et al., 2007; Liu et al., 2001). Since we want to study the discrepancy due to discretization, we simulated another set of specialized 4 km runs for which IC/BC are almost the same as those for the 1.33 km runs so that any differences in results cannot be attributed to differences in IC/BC.

Section 2 gives further details on our simulation. To compare runs at different resolutions, we need to either interpolate one or aggregate the other to match the resolutions. Our results (not shown) indicate not much difference between these two cases, so we will show results at high resolution. That is, we will show results about SGD at 1.33 km, the difference between 1.33 km runs and interpolation of the specialized 4 km runs. Section 3 presents our analysis and show how the differences between the outputs at the two resolutions depend on time of day, level of urbanization, species of pollutant and how long the model has been allowed to run at the two resolutions since starting at the same initial conditions.

2 Simulation Experiments

The CMAQ modeling system is characterized by its comprehensiveness and state-of-science description of atmospheric chemistry and physical processes applicable over a wide range of spatial scales. Briefly, CMAQ is comprised of processors and numerical solvers with complex linkages to the high level of model input information (emissions and meteorology) to drive the system. Because air pollution varies over a wide range of spatial scales and evolves in time, the CMAQ system utilizes model nesting to obtain model granularities at different scales, and numerical solvers with time discretization at short time intervals to increase accuracy. Data assimilation techniques in meteorological models are employed to constrain meteorological simulations for accurate predictions of transport, dispersion and photo and aqueous chemistry. More details concerning CMAQ can be found in Byun and Ching (1999) and Byun and Schere (2006). Given the myriad of processes and scales simulated, the sources of model errors are large in number, and generally difficult to assess. Many current efforts to understand, appreciate and treat model uncertainties and error have been reported in a special issue of *Atmospheric Environment* (Volume 40, Issue 26, 2006). These and other studies have addressed model uncertainties and errors in CMAQ and its preprocessors in particular.

We study the CMAQ simulations for 4 km and 1.33 km grid sizes for July 12, 1995 in

the Philadelphia metropolitan area. In particular, we focus on simulated concentrations of NO_x (the sum of NO_2 and NO), O_3 and CO. The Penn State/NCAR Mesoscale Model (MM5) provides meteorological fields for CMAQ. For fine-scale simulation (about 1 km grid spacing), an urban canopy parametrization (UCP) is included in MM5 to improve meteorological fields in urban areas (Otte et al., 2004). The CMAQ simulations were derived as nested runs of coarser grid sizes. Spatial resolution is gained by successively decreasing the grid and domain size from the base 36 km run covering the eastern United States. Standard practice for nested CMAQ runs is to perform a five day run at hourly intervals for the base 36 km grid size. The simulation for the 12 km grid size in the nested domain begins after a two-day spin up of the 36 km run. This 12 km run then provides the IC/BC for the 4 km run, which begins a day after the beginning of the 12 km run. The 1.33 km run is obtained similarly from the 4 km run. We thus have a single day for the 1.33 km set. For the 4 km and the 1.33 km set, the focus of our analysis is the last day of the original 5 day run.

We denote the two highest resolution runs by $Z_H(t)$ and $Z_L(t)$ where H stands for high resolution, i.e. 1.33 km grid size, L stands for low resolution, i.e. 4 km grid size, and t is time. We have hourly data from 12am to 8pm EDT. These hourly values are averaged values over 10 time steps with a 6 minute interval. Although the resolutions we consider here are higher than that of the typical runs done by CMAQ on a regional scale, the procedure to get IC/BC is standard. Because Z_H and Z_L use different IC and BC, we did some specialized low resolution runs to help us to better evaluate the SGDs. These additional runs were at 4 km. Specifically, for a given starting time, t_0, we used the aggregated value of high resolution output, $Z_H(t_0)$ as IC and the low resolution output, $Z_L(t)$ as BC for the new 4 km run, which we will call $W_L(t, t_0)$. That is, $W_L(t, t_0)$ is a simulated output at time t started from t_0 using $A(Z_H(t_0))$ as IC and $Z_L(t)$ as BC where $A(\cdot)$ represents aggregation. Thus, $Z_L(t)$ was used as the BC for simulation of both $Z_H(t)$ and $W_L(t, t_0)$. We have $W_L(t, t_0)$ for many t_0 (see Figure 1). The starting times, t_0, range from 8pm of the previous day to 4pm; ending times, t, range from t_0 to 7pm. Times are local times (EDT), i.e. $GMT - 4$.

In our simulation, meteorology and emission inputs are still different at the two resolutions in particular, UCP was implemented for the 1.33km simulation and these differences may also contribute to differences between results. Thus, the SGDs we are finding are not just due to differences in CMAQ at the two resolutions, but also include effects of differing meteorology and emissions. Since one always needs to run a meteorological and emissions model in one way or another to obtain inputs to CMAQ, it is reasonable to include these effects in the SGD. It does mean, though, that we cannot disentangle to what extent the discrepancies we observe are due to running CMAQ at different resolutions versus running MM5 or the emission model at different resolutions. Further specialized runs could in principle be done to investigate this matter.

3 Statistical analysis

We are interested in isolating the SGDs, the differences between model outputs due to differing resolutions, unconfounded as much as possible by differences in IC/BC. We are also interested in examining how the SGDs propagate in time. Thus, we consider the following procedure: for any given t_0 and $t \geq t_0$, we compare $Z_H(t)$ and $W_L(t, t_0)$ as high and low resolution runs with similar as possible IC/BC. Specifically, we define the difference at high resolution by $D_H(t, t_0) = Z_H(t) - B(W_L(t, t_0))$, where $B(\cdot)$ represents bilinear interpolation. Then, we can think of D_H as the discrepancy due to running the model at different resolutions, i.e. SGDs. Note that each value of $A(Z_H(t))$ is the average of $Z_H(t)$ over nine neighboring cells since the grid size of the lower resolution output is 3 times larger than that of the high resolution output. From now on, the ending time means output time, the time at which the simulation returns output, and t_0 means starting time, the time at which the low resolution simulation starts. We will call $t - t_0$ the lag time, which measures how long the models with two different resolutions have run since starting under the same conditions (see Figure 1 for illustration).

3.1 Marginal distributions of two simulated runs

We first look at the simulated outputs themselves (see Figure 2 for NO_x and Figure 3 for O_3). We compare the two simulated runs ($Z_H(t)$ and $W_L(t, t_0)$) in two different ways. Figures 2 and 3 display how the model outputs change for fixed ending time t as starting time t_0 varies. Each column of Figure 2 and 3 shows one high resolution output $Z_H(t)$ from the first simulation for ending times t and two low resolution outputs $W_L(t, t_0)$ from the second simulation for t, in each case for two different starting times t_0. Figures 2 and 3 also display how the model outputs change for fixed starting time as ending time varies. For example, both (c) and (f) in Figure 2 share the same starting time, 2am, but (f) is the output after the model runs for 5 hours while (c) is the output after the model runs only for 1 hour. We can see how these two outputs, (c) and (f), differ from high resolution outputs, (a) and (b). As one might expect, the low resolution output misses much of the local variability of the process, most notably, the high values for NO_x and the lower values for O_3 near major highways. Concentration of NO_x is usually very high near major highways while concentration of O_3 is opposite to that of NO_x since NO, a component of NO_x, will initially titrate the ozone so that O_3 is smoothly spread out over most of the region except where we have high values of NO_x. The simulated results show that for both pollutants, the differences between $Z_H(t)$ and $W_L(t, t_0)$ become larger as lag time, $t - t_0$, increases. This means that the longer the model runs, the larger the SGD will be, which is again what we would normally expect.

Next, we compare the marginal distributions of two runs by looking at the quantile-quantile, or QQ, plots between them. The QQ-plot shows if there is a difference in overall pattern of levels without regard to whether $W_L(t, t_0)$ has the right levels in the right locations. If the impact of different resolutions is small, the two distributions would be similar and this would be shown as a straight line at 45° in a QQ plot. The maximum concentration of the high resolution output is usually higher than that of the low resolution output for

a highly localized pollutant such as NO_x, which is to be expected, since aggregation will tend to lower the most extreme values (see the left plot of Figure 4). The more interesting finding is that $B(W_L)$ produces substantially higher values from about the 90th to 99th percentiles, the more so as $t - t_0$ increases. This result is at least in part due to $B(W_L)$ producing more total NO_x over the region than W_L. The right plot of Figure 4 for O_3 shows that the marginal distributions of the low resolution and high resolution output look similar. Similarity of the marginal distributions between the low resolution and high resolution output at 10am might be because O_3, which is a secondary pollutant, is smoothly spread out over the region, especially during the late morning and afternoon so that the low resolution simulation has little problem to describe the variability of O_3 at those times.

3.2 Variogram of the difference

Quantile plots allow comparisons of marginal distributions, but these results contain no spatial information. Since the concentration of the air pollutants we consider here show substantial spatial variation, it is informative to consider spatial information in the analysis for SGDs. One popular tool for describing spatial variation is the variogram (Cressie, 1993). The variograms of D_H can show how the spatial variability of the SGDs changes with spatial distance, ending time or lag time. A small value for the variogram of SGDs indicates that the low resolution output captures well the spatial fluctuations in the high resolution output. A estimated variogram $\hat{\gamma}$ for $Z(s)$ observed for s in some set of spatial location S is defined by

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{S} \{Z(s) - Z(s+h)\}^2,$$

where h is a two dimensional vector representing the direction and distance between a pair of points and the sum is over all points s in S such that both $Z(s)$ and $Z(s+h)$ are observed, and $N(h)$ is the number of such points. If Z is intrinsically stationary (Cressie, 1993), then $\gamma(h)$ is the mean of $\frac{1}{2}\{Z(s) - Z(s+h)\}^2$, independent of s, and $\hat{\gamma}(h)$ is an unbiased estimator of $\gamma(h)$. In the present setting, our processes are perhaps not well modeled as intrinsically stationary due to urban/rural differences and the effects of major highways. However, $\hat{\gamma}(h)$ can still be used as a measure of average variation over the region for pairs of points separated exactly by h (ignoring the slight curvature of the earth over this region of interest). $\hat{\gamma}(h)$ also cancels out any spatially constant errors, such as the bias seen in O_3 (see Figure 4). In this paper, we plot the square roots of the variograms, so that we have the same units as the original air pollutant processes.

At fixed ending time t, the spatial variability of $D_H(t, t_0)$ increases as the low resolution model has run longer, growing rapidly for smaller $t - t_0$ and then more gradually for larger $t - t_0$. This pattern holds for NO_x, O_3 and CO and each time t =midnight, $\cdots$, 7pm. Figure 5 shows the square root variograms of $D_H(10, t_0)$ for NO_x at various t_0: the variation of SGDs increases as the lag time $t - t_0$ increases, confirming what we found from Figure 2. Note that the square root variogram at $t - t_0 = 0$ is not zero. This is because the output at t_0 is the average over 10 time steps forward in time and thus is different from the input at t_0.

From Figure 5, it is difficult to see how the square root variogram changes over the lag time $t - t_0$ when the spatial lag is small. We can check this pattern directly by looking at the square root variogram as the lag time changes for fixed spatial lag and ending time. Figure 6 shows the square root variogram of D_H for NO_x at $t = 10$am for various lag times. Of course, the square root variogram increases with the lag time but we can see that it does not increase much after 2 hours running from the starting time, which suggests that most of the local spatial variation is created in the first few hours of running the model. Thus, for example, a data assimilation scheme, even if it incorporates data as frequently as hourly, which would be very frequent relative to present practice, should, ideally, take into account such spatial variation of SGD by including it in the covariance structure of the model error.

We are also interested in a pattern of the variation in SGDs along time of day, since a large variation in SGDs at certain hours of the day suggests the need to include a diurnal pattern in any model for discrepancy due to discretization. For example, we can see how spatial variation of $D_H(t, t_0)$ for NO_x and O_3 changes at different ending times t when $t - t_0 = 1$ in Figure 7. Figure 8 compares the square root variogram of $D_H(t, t-1)$ at distance 1.33 km for each hour of the day. If the mean levels of the original process are very different at two different times, we may expect to find the square root variogram tracks with the mean level of the pollutant at that time, since it is common for natural processes to vary more when the mean levels are high. Thus, we plot the estimated square root variogram of SGDs and the average of high resolution output together along time (local time) and their ratios. The first three plots in Figure 8 show $\sqrt{\hat{\gamma}(1)}$, the square root variogram of $D_H(t, t-1)$ at distance 1.33 km and $\bar{Z}_H$, the average of $Z_H(t)$ for NO_x, O_3 and CO and the last plot shows the ratio, $\sqrt{\hat{\gamma}(1)}/\bar{Z}_H$, for those three air pollutants. Both the square root variogram of $D_H(t, t-1)$ and the mean level of $Z_H(t)$ change with the hour of the day over local time. As expected, the square root variogram values generally, but not always, follow the ups and downs of mean level of the pollutant process.

The mean level and the square root variogram for NO_x in the early morning and in the afternoon roughly track each other, although there are some irregular movements of the square root variogram in the early morning, which results in fluctuations of $\sqrt{\hat{\gamma}(1)}/\bar{Z}_H$ at the corresponding period of time. On the other hand, both the mean level and the square root variogram decrease over the morning but with a different rate, which makes the ratio increase. This implies that the relative spatial variations of $D_H(t, t-1)$ in the afternoon are stronger than in the morning.

For O_3, the increases and decreases of the square root variogram go along with those of the mean level in the early morning and the afternoon, while they do not in the morning (5am to 9am). Thus, spatial variation of $D_H(t, t-1)$ in the morning is relatively strong since the square root variogram at those times is relatively high while the mean level of the high resolution output is relatively low. This is also shown by the small peak in the ratio around 7am for O_3. This result suggests that a low resolution model is less capable of describing spatial variation of O_3 in the morning.

CO is a highly localized pollutant, so most grid cells have very small values in our simulated outputs. The overall maximum CO level, 5438.43 ppbV, occurred at 7am and

most values are a small fraction of this (see Figure 9). Indeed, at 7am, only 6.37% of the grid cells are within 80% of this maximum and at other hours, this percentage is even lower. Both the square root variogram of $D_H(t, t-1)$ and the mean level of high resolution output is bigger around that time. However, the increase of the square root variogram is larger than that of the mean level and this is shown by the peak in the ratio around 7am for CO. This reflects the fact that the large variation around 7 am is not captured well by the lower resolution model.

It is useful to explore the behavior of the SGDs for various sub-regions within our study area. In particular, differences between urban and more rural areas may help to decide in which area SGDs are likely to be larger. Consider dividing the whole domain into 9 sub-regions of size 40 km $\times$40 km. Thus, the number of grid cells in a sub-region is 30×30 at the high resolution. Among the 9 sub-regions, we focus on two sub-regions. One is in the center and the other is the sub-region directly east of the center. We can consider the center sub-region as an urban area and the east sub-region as a rural area. In Figure 10, we plot the square root variogram of $D_H(t, t-1)$ and the average of $Z_H(t)$ for each sub-region and for NO_x, O_3 and CO. The last row in Figure 10 shows the corresponding ratio, $\sqrt{\hat{\gamma}(1)}/\bar{Z}_H$. First of all, the square root variogram, $\sqrt{\hat{\gamma}(1)}$, and the average, $\bar{Z}_H$, in the center sub-region look similar to those in the whole region (see Figure 8 and the left column of Figure 10). Since the whole region (Philadelphia metropolitan area) is an urban area by itself, the similarity in the behavior of SGDs in both cases is expected. We can also see the clear difference in the magnitudes of $\sqrt{\hat{\gamma}(1)}$ and $\bar{Z}_H$ between the center sub-region and the east sub-region for all three species. This indicates that the spatial variations of SGDs are larger in the urban area than in the rural area, despite the fact that O_3 levels are generally at least as high in the east sub-region as in the center sub-region. However, there are some differences in the comparison of SGDs between the center sub-region and the east sub-region for each species when we consider the square root variogram and the mean level together.

For NO_x, the increases and decreases of $\sqrt{\hat{\gamma}(1)}$ and $\bar{Z}_H$ are roughly similar for the two sub-regions except around 2am and 3pm, although the magnitudes are different and $\bar{Z}_H$ has a peak at a different time. So, the relative strength of spatial variations tends to decrease and then increase again for both sub-regions. This may indicate that the strength of spatial variations between the two sub-regions is not too different for NO_x in the morning through the early afternoon.

For O_3, it is interesting that $\bar{Z}_H$ is high in the east sub-region but the square root variogram is not, indicating that the low resolution model describes spatial variation fairly well in the rural sub-region for O_3 despite high mean levels. For CO, although $\bar{Z}_H$ in the east sub-region is roughly half that in the center sub-region, the square root variogram is much smaller. This makes the strength of spatial variation of SGDs in the center sub-region relatively stronger than in the east sub-region.

The patterns of the square root variogram of the difference between outputs at the two resolutions and the average of high resolution output tell us that the SGDs depend on how the original process behaves over time. NO_x has relatively high spatial variation of SGDs in the afternoon after a sharp increase over the morning. O_3 has relatively high spatial

variation of SGDs when the average of the original process is low. Meanwhile, CO has relatively high spatial variation of SGDs in the morning when the original process has a high average. We also learn that those patterns differ between rural and urban sub-regions. As we expected, the relative size of spatial variation of SGDs is larger in the urban area in general.

4 Conclusion and Discussion

In our study, we looked at a component of model error in an air quality model (CMAQ) that is actively used in practice. Specifically, we considered the discrepancy due to discretization in the model for short-range forecast by investigating SGD, the difference between two model outputs at different resolutions, recognizing that lack of resolution is only one source of the discrepancy, so that we are most likely underestimating total model error. Another issue here is the different meteorology and emission inputs at different resolutions in our simulation study. To investigate whether inputs from MM5 with UCP for CMAQ at high resolution can cause more discrepancy in results than inputs from MM5 without UCP, we compared two high resolution outputs with/without UCP and found that the spatial variation as measured by variogram is generally smaller than that of $D_H(t, t_0)$, particularly at the shortest spatial lags. For example, for NO_x, the average ratios of variograms of the difference between outputs with/without UCP and variograms of $D_H(t, t_0)$ are 0.493, 0.698 and 0.774 when t_0 = midnight and the spatial distances are 1×1.33 km, 2×1.33 km and 3×1.33 km, respectively (see Table 1). Thus, if we had done the specialized CMAQ runs at 4 km using IC/BC from CMAQ at 1.33 km resolution without UCP, we might expect to have found somewhat less SGD. However, it is not clear that these SGDs would give a better approximation to the statistical characteristics of the discrepancy due to discretization, since the lack of urban landscape features in MM5 is a source of error in MM5 runs without UCP.

Spatial variograms (or their square roots) of differences between two resolution outputs show substantial spatial variation in SGD. For example, the square root variograms of the SGD for NO_x (the sum of NO_2 and NO) in the East-West direction with spatial distance 1.33 km and 2×1.33 km at 10am after running the model for an hour are around 9 ppbV and 11.5 ppbV, respectively (see Figure 6), which is a nonnegligible fraction of the mean

Table 1: Average of $\frac{\text{variogram of } Z_H(t) - \tilde{Z}_H(t)}{\text{variogram of } D_H(t,t_0)}$ over time

spatial distance	NO_x	O_3	CO
1×1.33 km	0.493	0.715	0.654
2×1.33 km	0.698	0.821	0.807
3×1.33 km	0.774	0.840	0.851

$\tilde{Z}_H(t)$ is CMAQ output when UCP is not included in MM5. t is from 12am to 5pm and t_0 = 12am.

NO_x level at 10am of 19.21 ppbV (see Figure 9). We can also see that the SGD occurs mostly because low resolution output has problems describing small-scale variation in space and with nonlinear chemistry. Results show that the spatial variation of SGDs (as measured by the square root variogram) increases mostly in the first few hours of simulation and it has a diurnal pattern along output time for fixed lag time. The SGD also depends on urbanization.

These results for the discrepancy due to discretization as a component of model error suggest that we need to consider the characteristics of model error in data assimilation for air quality numerical models. For example, we could include parameters for a diurnal pattern in time for the model error equation in the Kalman filter. Our results may also provide some helpful information to modelers or model users when they need to decide the resolution of the model outputs or to interpret model outputs at different resolutions. For example, an urban area may need a fine scale model to have better description of spatial fluctuations. Further analysis regarding the spatial and temporal aspects of the discrepancy in the model together with identifying its sources to better characterize its nature is needed in the future.

Acknowledgments

The authors thank Dr. V. Rao Kotamarthi in Argonne National Laboratory for his helpful and valuable comments on this paper. We also thank to an editor and reviewers for their careful comments and suggestions which greatly help us to improve the paper. The United States Environmental Protection Agency through its Office of Research and Development funded and collaborated in the research described here under STAR Cooperative Agreement $\#R-82940201$ to the University of Chicago. It has been subjected to Agency review and approved for publication.

References

P. Brohan, J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, 111,D12106:doi:10.1029/2005JD006548, 2006.

D. W. Byun and J. K. S. Ching. *Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System.* EPA, Available at http://www.epa.gov/asmdnerl/CMAQ/CMAQscienceDoc.html, 1999.

D. W. Byun and K. L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied Mechanics Reviews*, 59:51–77, 2006.

J. Ching, J. Herwehe, and J. Swall. On joint deterministic grid modeling and sub-grid variability conceptual framework for model evaluation. *Atmospheric Environment*, 40: 4935–4945, 2006.

S.E. Cohn. Dyanmics of short-term univariate forecast error covariances. *Monthly Weather Review*, 121:3123–3149, 1993.

N. Cressie. *Statistics for Spatial Data.* Wiley, 1993.

D.P. Dee. On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, 123:1128–1145, 1995.

G Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

A. E. Gelfand, L. Zhu, and B. P. Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2,1:31–45, 2001.

C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97:632–648, 2002.

G. C. Hegerl, P. D. Jones, and T. P. Barnett. Effect of observational sampling error on the detection of anthropogenic climate change. *Journal of Climate*, 14:198–207, 2001.

P. L. Houtekamer, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen. Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly Weather Review*, 133:604–620, 2005.

P. Jiménez, O. Jorba, R. Parra, and J. M. Baldasano. Evaluation of mm5-emicat2000-cmaq performance and sensitivity in complex terrain: High-resolution application to the northeastern iberian peninsula. *Atmospheric Environment*, 40:5056–5072, 2006.

P. Jiménez, R. Parra, and J. M. Baldasano. Influence of initial and boundary conditions for ozone modeling in very complex terrains: A case study in the northeastern iberian peninsula. *Environmental Modelling & Software*, 22:1294–1306, 2007.

E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press, 2003.

T.-H. Liu, F.-T. Jeng, H.-C. Huang, E. Berge, and J.S. Chang. Influences of initial conditions and boundary conditions on regional and urban scale eulerian air quality transport model simulations. *Chemosphere-Global Change Science*, 3:175–183, 2001.

C. F. Mass, D. Ovens, K. Westrick, and B. A. Colle. Does increasing horizontal resolution produce more skillful forecasts? the results of two years of real-time numerical weather prediction over the pacific northwest. *Bulletin of the American Meteorological Society*, 83:407–430, 2002.

H. L. Mitchell and P. L. Houtekamer. An adaptive ensemble kalman filter. *Monthly Weather Review*, 128:416–433, 2000.

M. Mousavi, M. Soltanieh, and A. Badakhshan. Influence of turbulence and atmospheric chemistry on grid size with respect to location in modeling and simulation of photochemical smog formation and transport. *Environmental Modelling & Software*, 14:657–663, 1999.

T.L. Otte, A. Lacser, S. Dupont, and J.K.S. Ching. Implementation of an urban canopy parameterization in a mesoscale meteorological model. *Journal of Applied Meteorology*, 43:1648–1665, 2004.

S. K. Park, C. E. Cobb, K. Wade, J. Mulholland, Y. Hu, and A. G. Russell. Uncertainty in air quality model evaluation for particulate matter due to spatial variations in pollutant concentrations. *Atmospheric Environment*, 40:S563–S573, 2006.

R.S. Sokhi, R. San José, N. Kitwiroon, E. Fragkoua, J.L. Pérez, and D.R. Middleton. Prediction of ozone levels in london using the mm5cmaq modelling system. *Environmental Modelling & Software*, 21:566–576, 2006.

J. L. Swall. The impact of spatial correlation and incommensurability on model evaluation. *Atmospheric Environment*, page doi: 10.1016/j.atmosenv.2008.10.057, 2008.

S. Zoras, A.G. Triantafyllou, and P.J. Hurley. Grid sensitivity analysis for the calibration of a prognostic meteorological model in complex terrain by a screening experiment. *Environmental Modelling & Software*, 22:33–39, 2007.

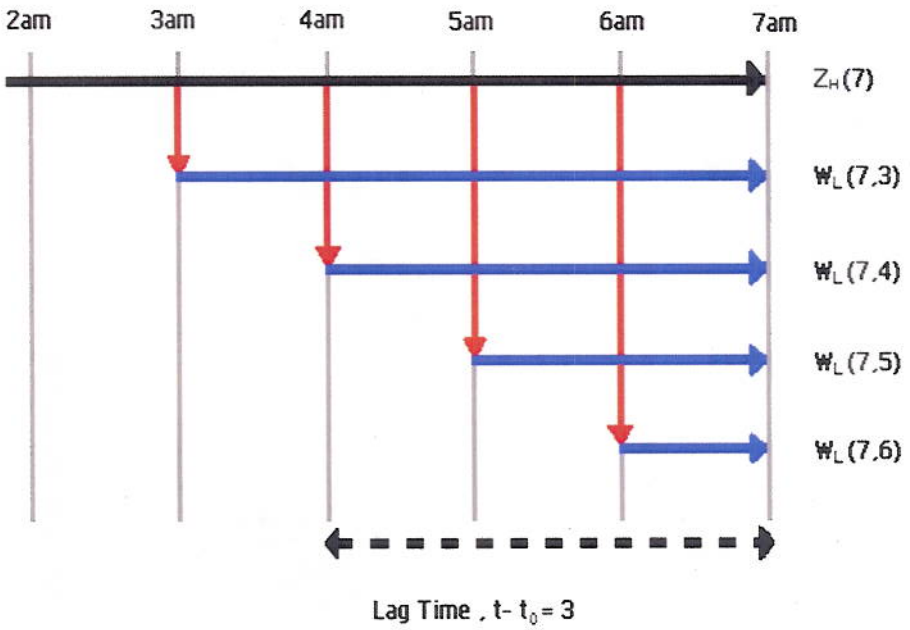

Figure 1: Diagram for specialized low resolution (4 km) runs with various starting times. For example, $W_L(7,3)$ is a simulated output at $t = 7$am started from $t_0 = 3$am using the aggregated value of $Z_H(3)$ as IC and $Z_L(t)$ as BC. We call the difference between the starting time and output time, $t - t_0$, the lag time.

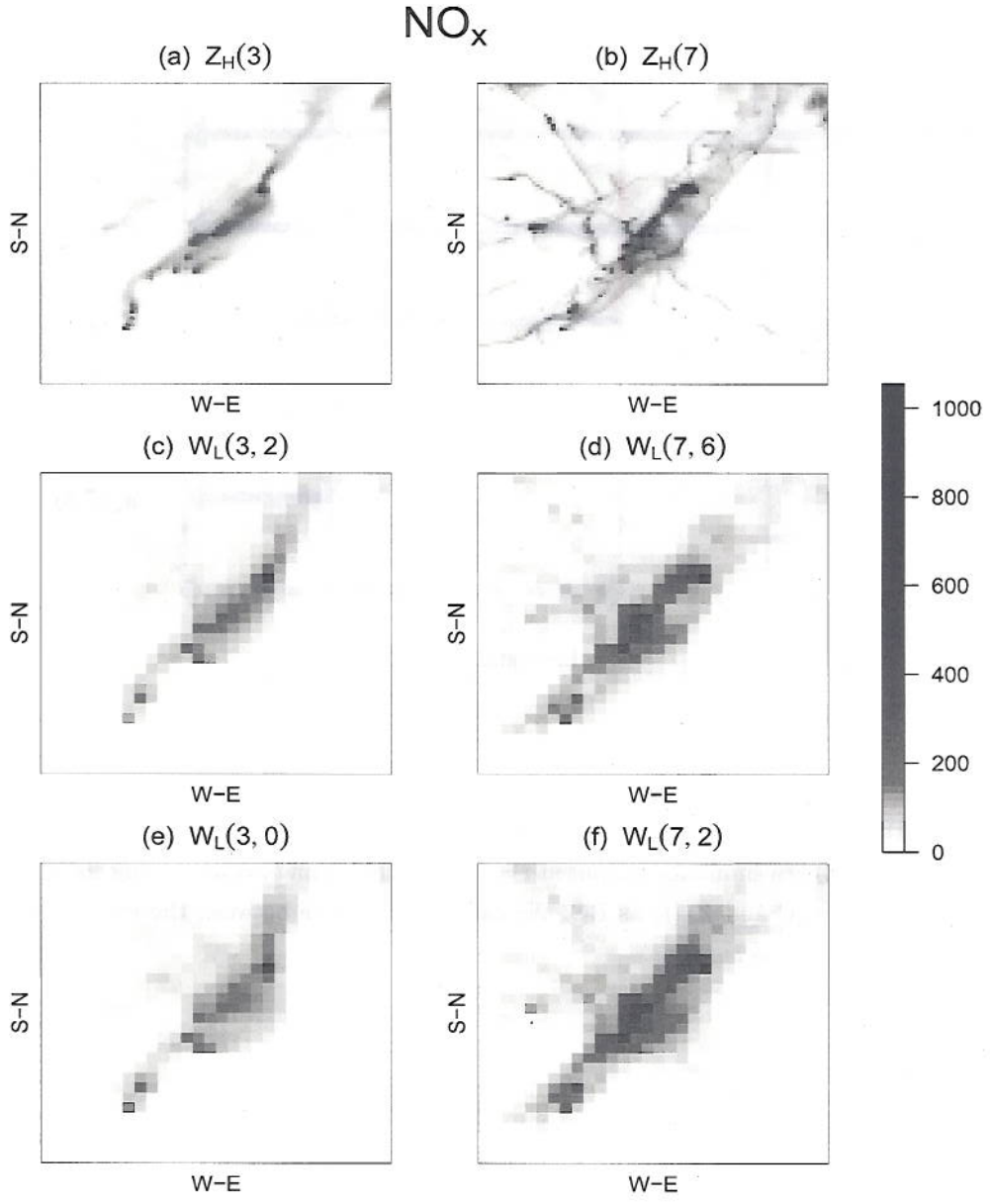

Figure 2: NO_x, Simulation output on July 12, 1995 in Philadelphia metropolitan area (120 km $\times$120 km). (a) $Z_H(3)$ and (b) $Z_H(7)$ are high resolution outputs from the first run at $t = 3$am and $t = 7$am. (c) $W_L(3,2)$, (d) $W_L(7,6)$, (e) $W_L(3,0)$ and (f) $W_L(7,2)$ are low resolution outputs from the second run that use the high resolution outputs from the first run as initial conditions. The unit is ppbV.

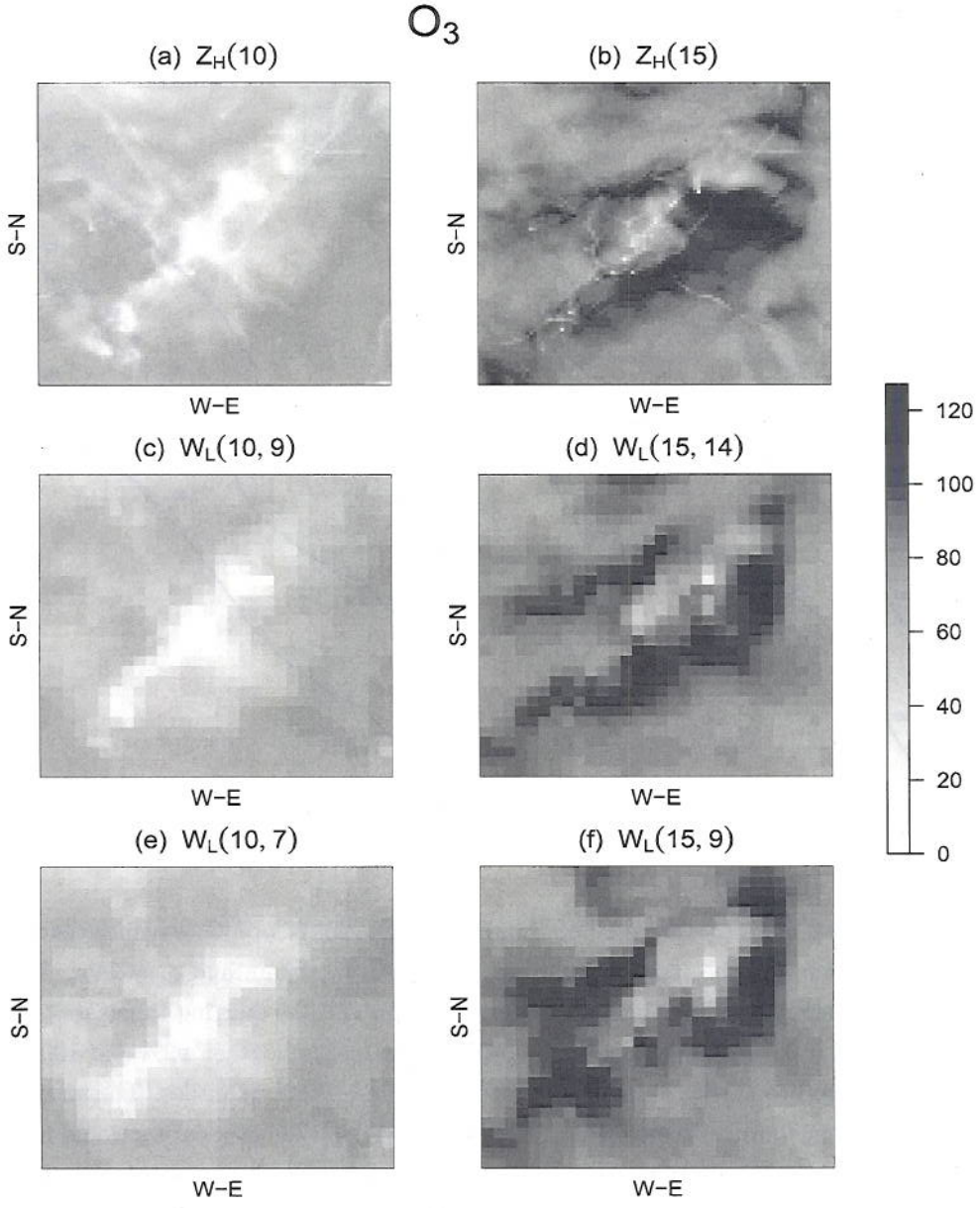

Figure 3: O_3, Simulation output on July 12, 1995 in Philadelphia metropolitan area (120 km $\times$120 km). (a) $Z_H(10)$ and (b) $Z_H(15)$ are high resolution outputs from the first run at $t = 10$am and $t = 3$pm. (c) $W_L(10,9)$, (d) $W_L(15,14)$, (e) $W_L(10,7)$ and (f) $W_L(15,9)$ are low resolution outputs from the second run that use the high resolution outputs from the first run as initial conditions. The unit is ppbV.

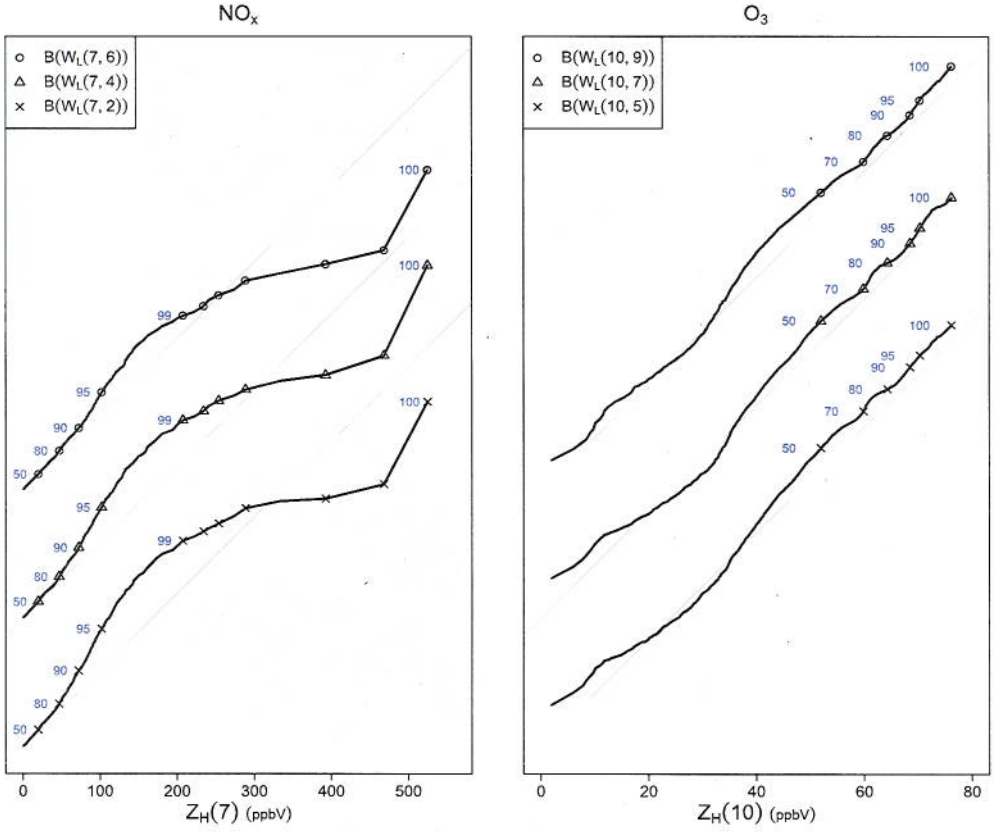

Figure 4: QQ-plots of $Z_H(t)$ versus $B(W_L(t, t_0))$. The left plot is for NO_x when the ending time is fixed at $t = 7$am and the starting times are $t_0 = 6$am, $t_0 = 4$am and $t_0 = 2$am. The right plot is for O_3 when the ending time is fixed at $t = 10$am and the starting times are $t_0 = 9$am, $t_0 = 7$am and $t_0 = 5$am. The marks are quantiles. The marks between 99% and 100% in the left plot correspond to the $99.2\%, 99.4\%, 99.6\%, 99.8\%$ and 99.9% quantiles. The straight gray lines are one-to-one lines between $Z_H(t)$ and $B(W_L(t, t_0))$, so the black curves would follow the gray lines if the two distributions were the same. The left plot clearly shows that for values of NO_x between about 50 and 250 ppbV, $B(W_L(7, 2))$ has worse agreement with $Z_H(7)$ than $B(W_L(7, 6))$. For O_3, the low resolution model gives higher values at nearly all quantiles and the disagreement is only slightly worse for earlier t_0. Note that we have put 3 QQ-plots in each box to save space and ease comparisons across different starting times.

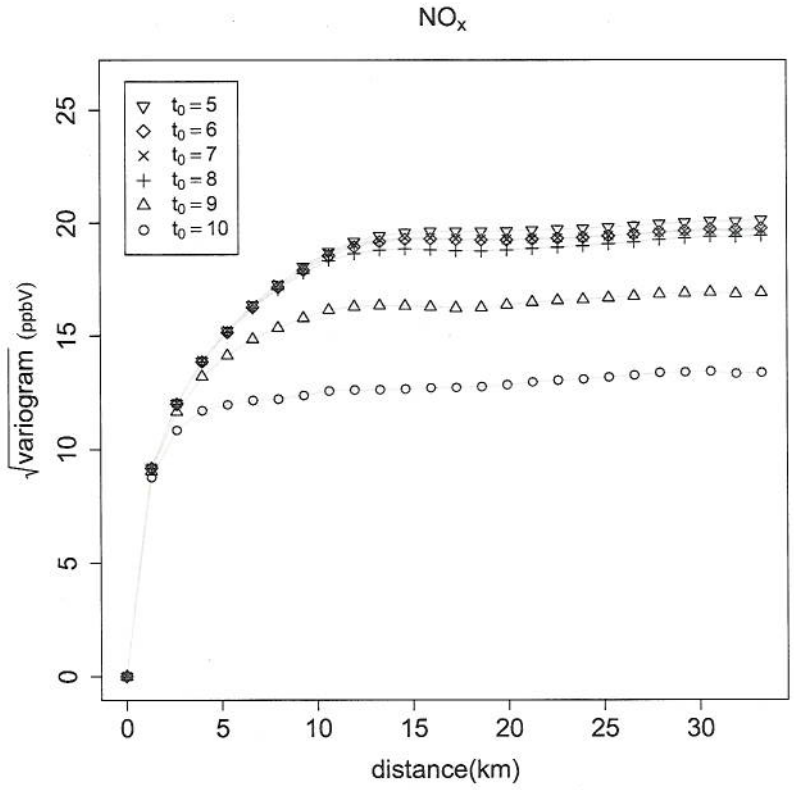

Figure 5: NO_x, $\sqrt{\text{Variogram}}$ of $D_H(t, t_0)$ at ending time $t = 10$am. The East-West variogram. Each line in the plot corresponds to the square root variogram for a different starting time.

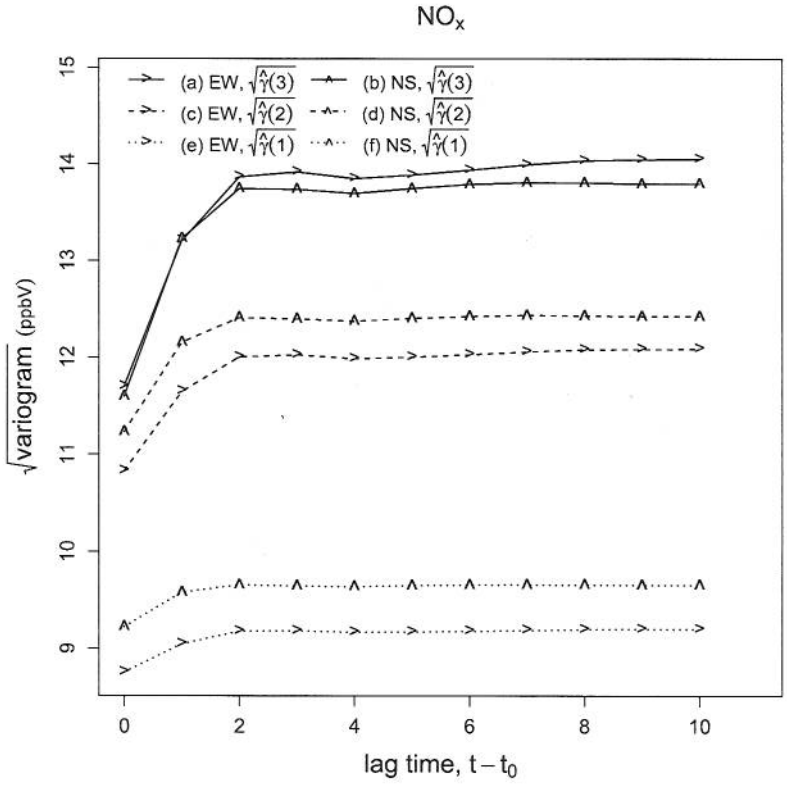

Figure 6: NO_x, $\sqrt{\text{Variogram}}$ of $D_H(t, t_0)$ at t = 10am. The horizontal axis represents the running hours of the model from the starting time. The spatial distances are fixed at 1×grid size = 1.33 km ((e) and (f)), at 2×grid size = 2×1.33 km ((c) and (d)) and at 3×grid size = 3×1.33 km ((a) and (b)).

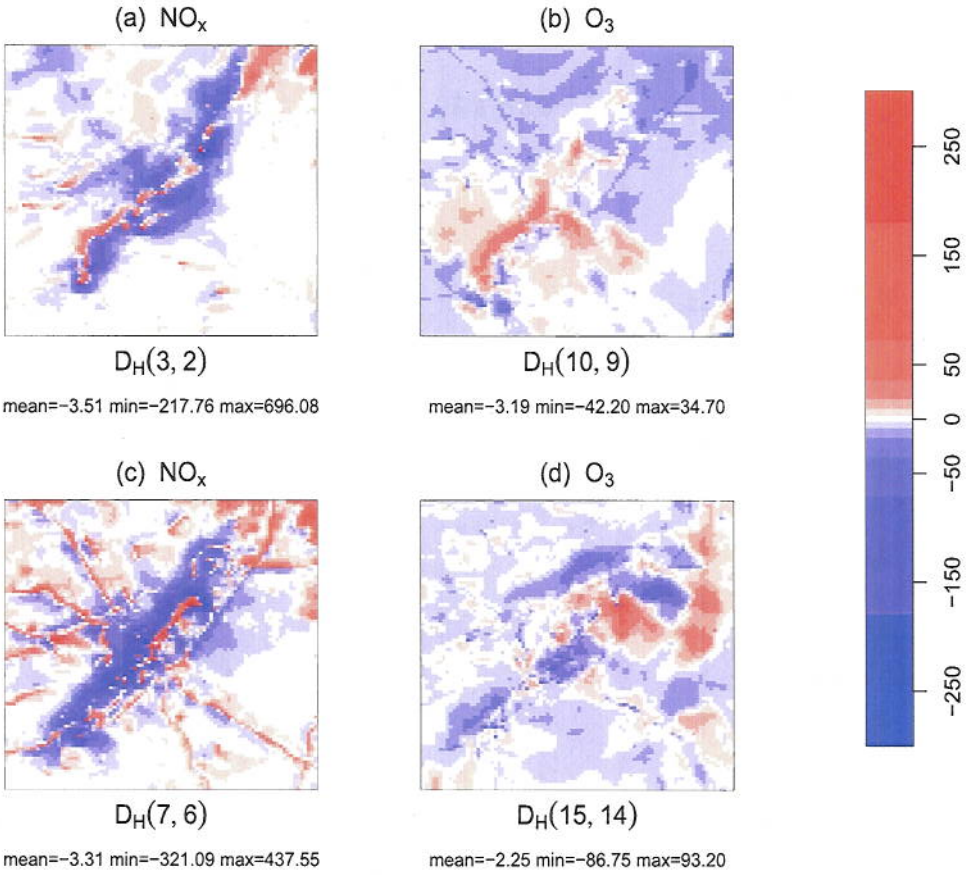

Figure 7: $D_H(t, t_0) = Z_H(t) - B(W_L(t, t_0))$, the difference between a high resolution output at t and an interpolated low resolution output at t from the second run started from t_0. $(t, t_0) = (15, 14)$ in (d) means $t = 3$pm and $t_0 = 2$pm. The left column is for NO_x. The right column is for O_3. The unit is ppbV for each.

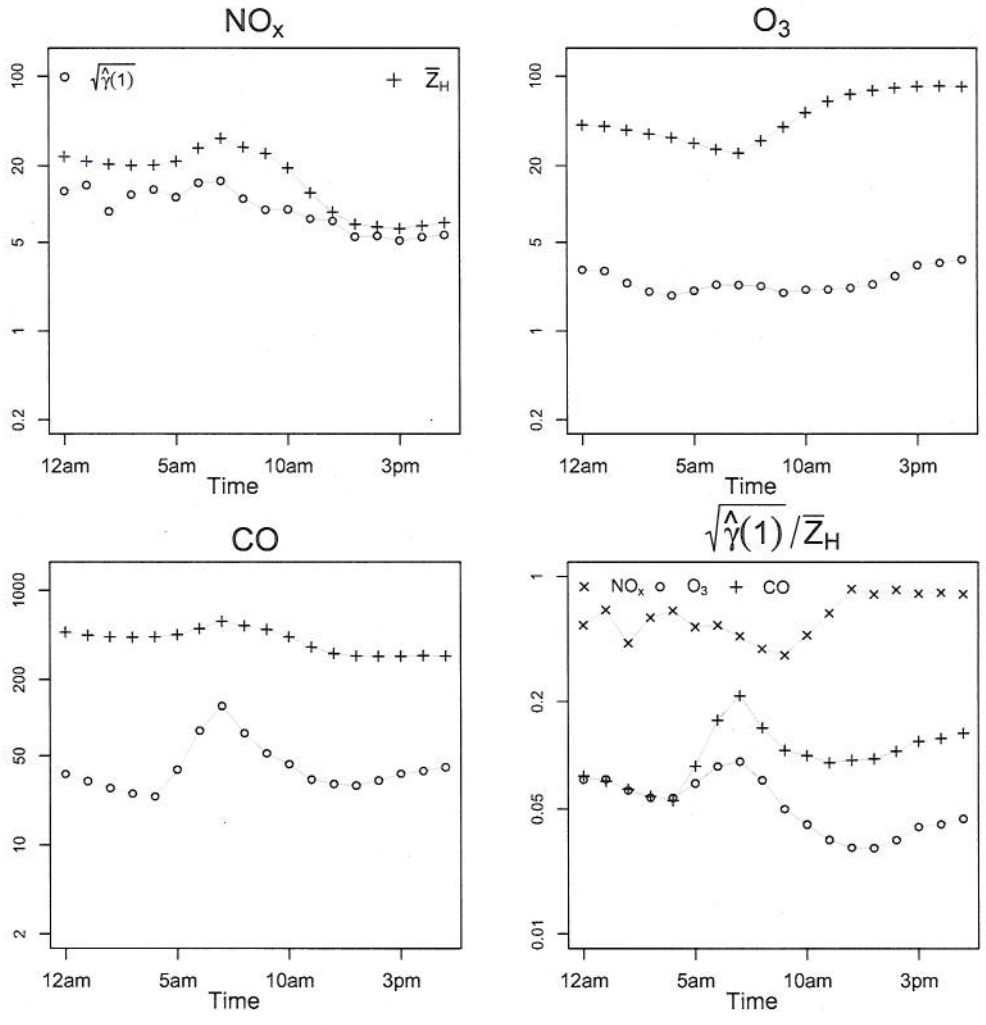

Figure 8: $\sqrt{\hat{\gamma}(1)}$, the square root variogram of $D_H(t, t_0)$ in the East-West direction at distance = 1.33 km when $t - t_0 = 1$ and $\bar{Z}_H$, the average of $Z_H(t)$ for the first three plots. $\sqrt{\hat{\gamma}(1)}/\bar{Z}_H$ in the lower right plot. The horizontal axis represents the ending time (local time). The vertical axis in the first three plots is scaled for $\sqrt{\hat{\gamma}(1)}$ of $D_H(t, t_0)$ and $\bar{Z}_H$. The unit in the vertical axes is ppbV.

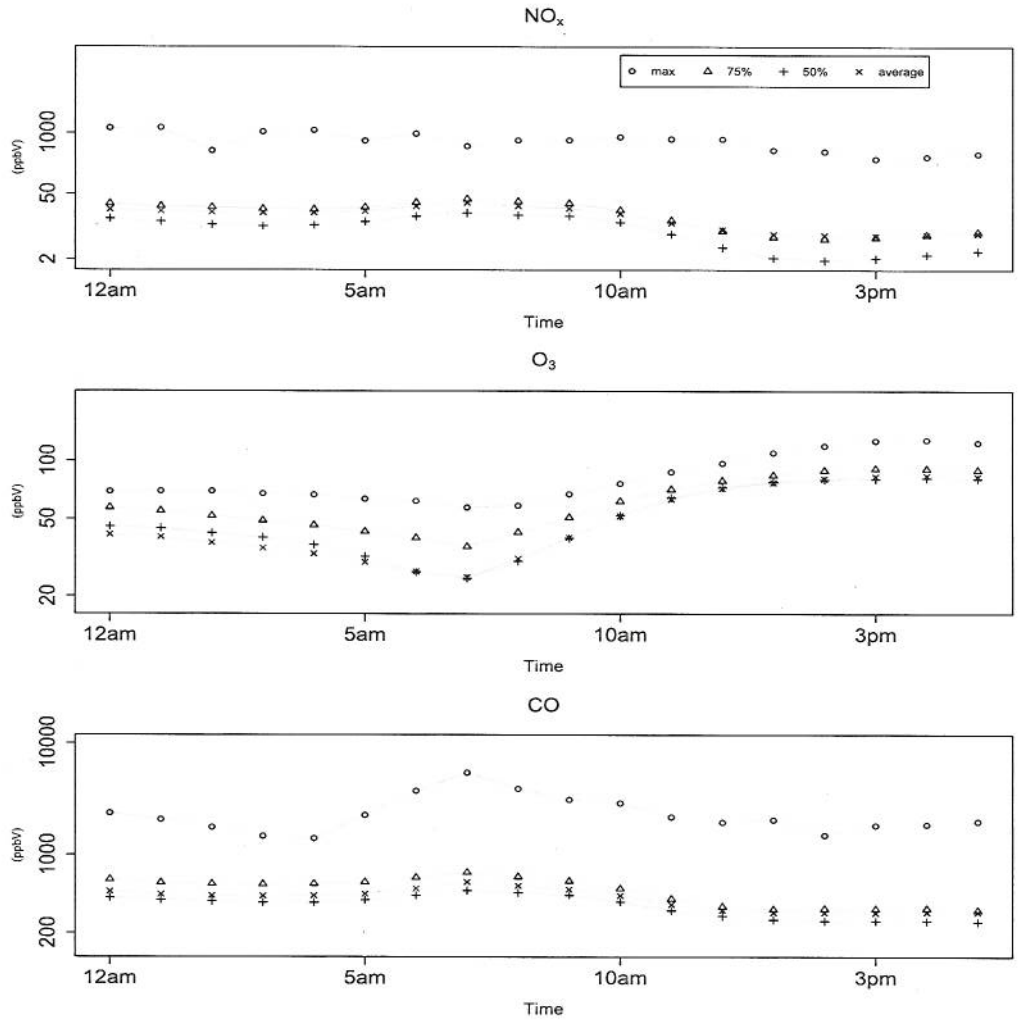

Figure 9: $Z_H(t)$, Maximum, 75% and 50% quantiles and average value at each time for species, NO_x, O_3 and CO.

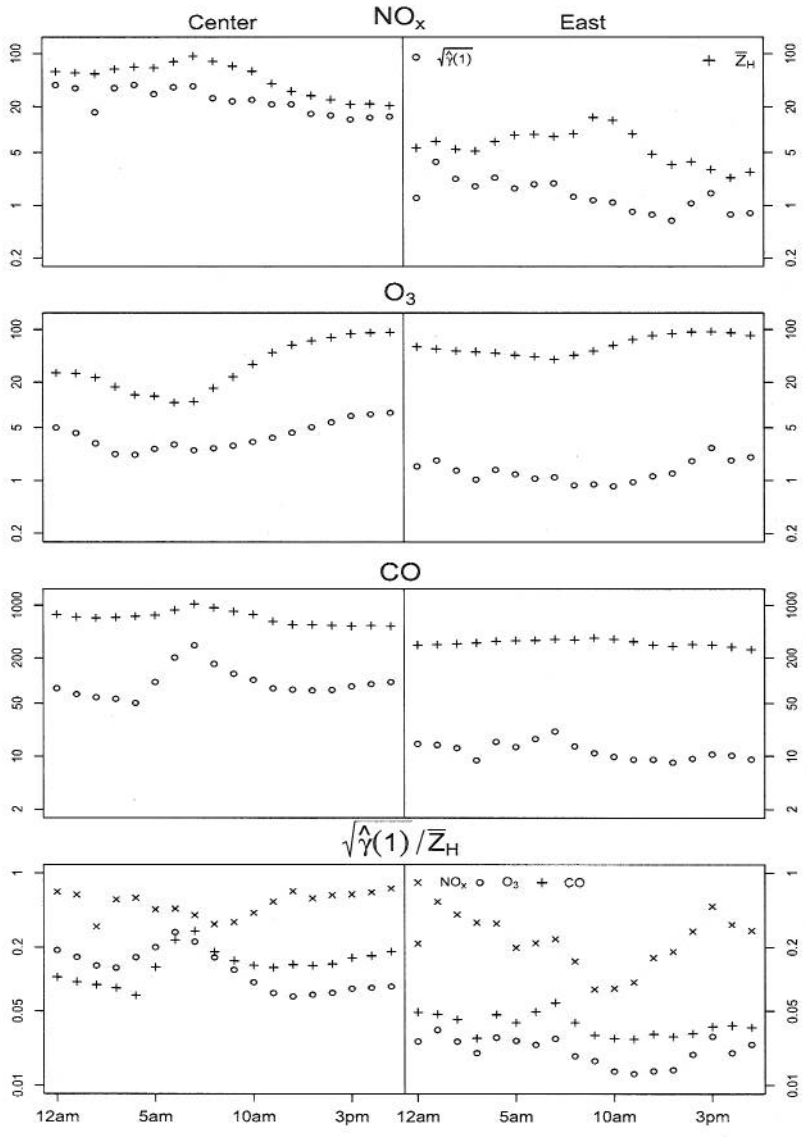

Figure 10: Same plots as in Figure 8, except separated by sub-region.

- A no-practices scenario, consisting of the results of revised model runs where the P-factor was set equal to 1.0 and the practice code was set such that the NRCS curve number represented conditions without conservation practices. All other model settings were the same as in the conservation-practices baseline scenario, including slopes and slope lengths and tillage practices.

Outputs from the no-practices scenario model runs were aggregated in the same manner as for the conservation practice baseline model runs. The two scenarios represent the same acreage. To determine the effects of the conservation practices, outputs for the URUs with practices were compared to the same set of URUs simulated without practices. Since the P-factor is not part of the wind erosion equation, the effects of the three practices on wind erosion was not assessed.

Representing irrigation in the model

Irrigation was simulated for URUs representing NRI sample points with irrigation. Irrigated land, as defined for NRI purposes, is land that shows physical evidence of being irrigated during the year of the inventory (presence of ditches, pipes, or other conduits) or having been irrigated during two or more of the four years preceding the inventory (USDA NRCS 1997b). Three types of irrigation are recorded in the NRI: gravity irrigated, pressure irrigated, or gravity and pressure irrigated.

For EPIC modeling, sprinkler irrigation was used to simulate pressure systems and furrow/flood irrigation was used to simulate gravity systems. The gravity pressure irrigation type was defined in the NRI as cases where water was delivered to the field by gravity flow and then applied through a pressurized sprinkler system (USDA NRCS 1997b); this was modeled in EPIC as a sprinkler system. When simulating no-till, however, a sprinkler system was always used. For rice, flood/furrow irrigation was always used. For URUs with average slopes greater than 3 percent, only sprinkler irrigation was used for non-hay crops.

Since information about the timing and amount of irrigation water used was not available, a generic irrigation schedule was simulated. A manual irrigation of 75 millimeters (3 in) for gravity and 50 millimeters (2 in) for sprinkler systems was applied prior to planting to ensure adequate moisture for seed germination. Subsequent irrigation events were simulated using the automatic irrigation feature of EPIC to irrigate during the growing season. The plant growth stress factor in this routine was set at 0.85, which caused the model to irrigate on any day that plant growth was less than 85 percent of potential growth if all other parameter conditions were met. Other parameters were set to: only irrigate to field capacity when irrigation was triggered; never irrigate more frequently than once in 5 days; irrigate with volumes between 25 and 75 millimeters (1–3 in); never irrigate more than 900 millimeters annually (35 in); limit irrigation volumes at each application so that no more than 5 percent is lost to runoff for sprinkler systems and no more than 20 percent is lost to runoff for gravity systems.

Overall, about 13 percent of the acres included in the study were irrigated (table 14). In the West, however, 79 percent of the acres were irrigated. The Southern Great Plains and South Central regions also had significant irrigation; 28 percent and 21 percent of the cropland acres included in the study were irrigated in these two regions, respectively. About 15 percent of the acres in the Northern Great Plains region were irrigated. Irrigated acres in the Southeast region represented 6 percent of the cropland acres included in the study. The Northeast and Upper Midwest regions had very few irrigated acres.

Representing commercial fertilizer applications in the model

Commercial fertilizer application is a critical factor for determining the amount of nitrogen and phosphorus loss from farm fields. The timing of application, the method of application (whether the materials are incorporated into the soil at application or not), and the amount applied all have significant influences on EPIC model results. Farmer surveys typically collect information on the number of applications, the timing of application, the amount applied at each application, and the method of application for both nitrogen and phosphorus. However, reports published by NASS and ERS seldom include summary statistics with this much detail because sample sizes from farmer surveys are usually too small to report these results on an annual basis.