

## Examination of Model Predictions at Different Horizontal Grid Resolutions\*

EDITH GEGO<sup>a</sup>, CHRISTIAN HOGREFE<sup>b</sup>, GEORGE KALLOS<sup>c</sup>, ANTIGONI VOUDOURI<sup>c</sup>, JOHN S. IRWIN<sup>d,\*\*</sup> and S. TRIVIKRAMA RAO<sup>d</sup>

<sup>a</sup>*University Corporation for Atmospheric Research*

<sup>b</sup>*University at Albany, Albany, NY 12222, U.S.A.*

<sup>c</sup>*University of Athens, Athens, Greece*

<sup>d</sup>*NOAA/Atmospheric Sciences Modeling Division, Research Triangle Park, NC, 27711, U.S.A.*

Received 30 June 2003; accepted in revised form 13 February 2004

**Abstract.** While fluctuations in meteorological and air quality variables occur on a continuum of spatial scales, the horizontal grid spacing of coupled meteorological and photochemical models sets a lower limit on the spatial scales that they can resolve. However, both computational costs and data requirements increase significantly with increasing grid resolution. Therefore, it is important to examine, for any given application, whether the expected benefit of increased grid resolution justifies the extra costs. In this study, we examine temperature and ozone observations and model predictions for three high ozone episodes that occurred over the northeastern United States during the summer of 1995. In the first set of simulations, the meteorological model RAMS4a was run with three two-way nested grids of 108/36/12 km grid spacing covering the United States and the photochemical model UAM-V was run with two grids of 36/12 km grid spacing covering the eastern United States. In the second set of simulations, RAMS4a was run with four two-way nested grids of 108/36/12/4 km grid spacing and UAM-V was run with three grids of 36/12/4 km grid spacing with the finest resolution covering the northeastern United States. Our analysis focuses on the comparison of model predictions for the finest grid domain of the simulations, namely, the region overlapping the 12 km and 4 km domains. A comparison of 12 km *versus* 4 km fields shows that the increased grid resolution leads to finer texture in the model predictions; however, comparisons of model predictions with observations do not reveal the expected improvement in the predictions. While high-resolution modeling has scientific merit and potential uses, the currently available monitoring networks, in conjunction with the scarceness of highly resolved spatial input data and the limitations of model formulation, do not allow confirmation of the expected superiority of the high-resolution model predictions.

**Key words:** air quality, grid resolution, model evaluation

## 1. Introduction

Three-dimensional Eulerian models for the simulations of atmospheric dynamics or photochemistry, for example RAMS [1], MM5 [2], UAM-V [3], CAMx [4], CMAQ [5]), are formed from an impressive number of differential equations which describe the temporal and spatial evolution of the relevant physical and chemical processes. Solving the system of equations requires a finite difference approximation scheme. Accordingly, the domain studied is discretized in space and time into a finite number of regular cells within which model inputs are considered uniform. Until recently, coarse discretization schemes were used because of limited computational resources.

As the power of computational hardware improved, it became possible to discretize the model domain into finer and finer grids in an effort to improve the quality and 'realism' of the model predictions. In theory, higher resolution modeling is expected to yield better predictions because of better resolved model input fields (e.g., topography, land cover or emissions), and better mathematical characterization of physical and chemical processes. For instance, finer grids reduce the volume into which the ozone precursors emitted are artificially diluted and allow more refined modeling of ozone chemistry near emission sources.

The use of smaller grid sizes undeniably improves the appearance of simulation results, but quantitative statistical measures of performance show only slight, if any, improvement (e.g., [6]). Examining temperature and wind predictions obtained by the MM5 model, Mass *et al.* [7] noticed considerable improvements when grid size decreased from 36 km to 12 km but only minor skill improvement when grid size was further reduced to 4 km. Colle *et al.* [8], also examining MM5 temperature and wind predictions, arrived at the same conclusions as Mass *et al.* [7], but acknowledge the superiority of the 4 km wind estimates in coastal areas.

Furthermore, for any given model application, the potential benefits of higher-resolution modeling should be weighed against the increased demands on input fields, CPU time, and disk space requirements. While a 1 km grid cell may be relevant for local-scale air quality forecasting or exposure assessment modeling, for instance, it may not be rational to study regional-scale air pollution problems affecting domains extending over several thousand square kilometers. Additionally, model users may not be able to accurately specify the particulars of all input data at a fine grid resolution. If a model user cannot properly characterize local variations in emission patterns, land use, soil moisture and other input data, then modeling with a finer grid resolution may not be greatly advantageous.

A separate but related issue is the spatial coverage of observations available to evaluate model performance. While high resolution modeling will always yield spatial fields with more texture and higher gradients than low resolution mod-

\*The U.S. Government's right to retain a non-exclusive royalty-free licence in and to any copyright is acknowledged.

\*\*Corresponding author, E-mail: irwin.john@epamail.epa.gov

eling, and while there may be good scientific justifications for increasing model resolution, observational data may not be sufficient to prove the benefits of higher-resolution modeling efforts.

The preceding comments suggest that there can be no single answer as to the merit of high-resolution modeling for all applications. Rather, in this paper, we attempt to illustrate some of the issues raised above for a particular application: the simulation of meteorology and ozone concentrations for three episodes of high regional ozone concentrations that occurred in the northeastern U.S. during the summer of 1995. We compare model predictions obtained using a 4 km horizontal grid size to those obtained using a 12 km horizontal grid size as described in the next section. We also use observations from routine monitoring networks to evaluate both simulations. To the extent possible, our results are discussed in terms of their implications to the use of these modeling approaches for regulatory purposes.

## 2. Brief Description of the Modeling System and its Setting

To investigate the effect of the horizontal grid spacing on model predictions, two distinct model configurations were chosen for both the meteorological and photochemical simulations. Meteorological simulations were performed with the RAMS-4a model in both 3- and 4-nest configurations. For the 3-nest setup, horizontal grids of 108 km, 36 km, and 12 km resolution were chosen in a 2-way nesting mode, allowing feedback from the finer to the coarser grids. For the 4-nest setup, a 4 km resolution inner grid was added to the above configuration. Aside from the addition of the 4 km grid for the 4-nest setup, all other modeling options such as the number of vertical levels, the parameterization for the treatment of different physical processes, and the use of Four Dimensional Data Assimilation (FDDA) were common between the two simulations. FDDA was used to nudge RAMS4a predictions towards a blend of surface observations and analysis fields obtained from the European Center for Medium Range Weather Forecasting. Both the 3- and 4-nest RAMS4a simulations were performed in a 'dry' mode, i.e., without the simulation of explicit micro-physics, an option we considered valid since the focus of this study is high ozone days with relatively clear sky conditions. Therefore, the main expected improvement of the meteorological fields simulated by RAMS4a in the 4-nest (smallest grid 4 km) compared to the 3-nest (smallest grid 12 km) configuration is through a better representation of topographical features such as valleys, mountain ranges, and coastlines.

RAMS4a meteorological fields were then used to drive two photochemical simulations with the UAM-V system [3]. These simulations were performed using a 2-nest setup with 36 km and 12 km grids, and a 3-nest setup with 36 km, 12 km, and 4 km grids. The 3-nest RAMS4a simulations were used as inputs to the 2-nest UAM-V simulations, and the 4-nest RAMS4a simulations were used as inputs to the 3-nest UAM-V simulations. For this purpose, RAMS4a fields were transformed from their original polar-stereographic coordinate system to the longitude-

latitude grid system used by UAM-V. As stated above, the RAMS4a simulations were performed in a 'dry' mode, i.e., without the simulation of clouds. To include the radiative effects of clouds on biogenic emission rates and photochemical reaction rates, horizontal cloud cover fields were derived from observations and used as input to the emissions processor as well as the UAM-V photochemical model. Anthropogenic emissions were based on the OTAG inventory; biogenic emissions were estimated with the Biogenic Emission Inventory System 2 [9] using the meteorological fields generated by RAMS4a.

The county-level total emissions for area sources (and vehicle miles traveled for mobile sources) are identical in both runs. When processing the emissions, the county-level totals are converted to gridded emissions by making use of 12 km and 4 km spatial surrogate data, such as the distribution of population, roads, housing, agricultural area within that county. Since the same the county-level emissions are used in constructing the 4-km and 12-km emission inventories, total temperature-independent emissions in the nine 4 km cells within a 12 km cell are equal to the 12 km emissions. However, temperature-dependent 4 km and 12 km emissions, such as biogenic and mobile source emissions, may be slightly different, since the 4 km temperature predictions are not necessarily strictly identical to the corresponding 12 km temperature prediction.

Conceptually, the 4 km UAM-V simulations might be expected to provide better ozone predictions than the 12 km simulations for several reasons. First, they utilize the higher-resolution RAMS4a simulations discussed above. Second, the treatment of chemistry near large emission sources could be expected to be more adequate because emissions no longer would be artificially diluted over a large grid area. Third, the 4 km simulations also utilize better resolved emission and land-use inputs.

Additional details on the setup of UAM-V can be found in Sistla *et al.* [10] who used the same UAM-V configuration (driven by a RAMS3b simulation) for their study.

RAMS4a and UAM-V were executed to simulate the three episodes when the 1-hour ozone standard (0.12 ppb) was exceeded in the northeastern U.S. during the summer of 1995. More specifically, these 3 episodes correspond to June 19–20, July 14–15, and August 1–2, 1995. Each simulated period included several spin-up days; the results pertaining to these spin-up days are not included in this investigation. More details about the RAMS4a simulations and the ozone episodes selection are provided in Sistla *et al.* [10], Biswas and Rao [11], and Lagouvardos *et al.* [12, 13].

We focus our analysis on the RAMS4a temperature and turbulent kinetic energy predictions from the 3-nest 12 km grid and the 4-nest 4 km grid (interpolated to the corresponding UAM-V grids), and the UAM-V ozone predictions from the 2-nest 12 km and 3-nest 4 km grids. In other words, only results from the finest grid in each simulation are used for analysis. The analysis domain is the common area between the 12 km and the 4 km UAM-V domains; it lies between  $-83^{\circ}$  W and

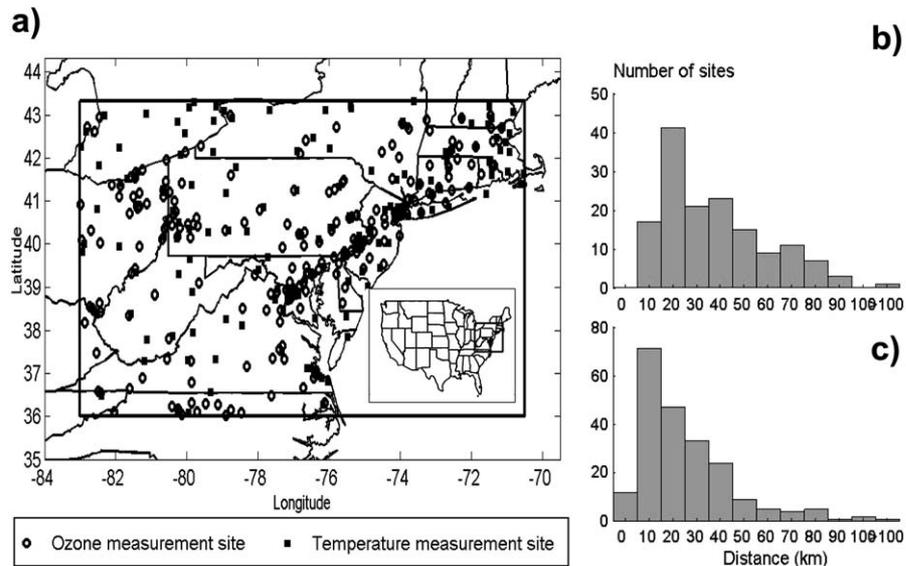


Figure 1. (a) location of the analysis domain and of the temperature and ozone measurement sites. (b) histogram of the distance between temperature measurement sites and their nearest neighbors. Panel (c) histogram of the distance between ozone monitoring sites and their nearest neighbors.

$-70.50^{\circ}$  W, and  $36^{\circ}$  N and  $43.33^{\circ}$  N (Figure 1a). This common area is decomposed into 44,550 or 4,550 surface cells depending on whether the 4 km or the 12 km grid is used. About 30% of these cells lie over the Atlantic Ocean or Lake Erie.

### 3. Observations

The hourly temperature observations used to evaluate the model predictions were retrieved from the Data Support Section at the National Center for Atmospheric Research (NCAR-DSS), whereas ozone observations were extracted from the USEPA's AIRS data base. All together, 148 temperature and 214 ozone sites were identified in the analysis domain. The locations of these sites are specified in Figure 1, along with the histograms summarizing the separation distances between each monitoring site and its nearest neighbor. Typical distances between neighboring temperature sites are on the order of 20 km. Ozone monitoring sites are more numerous in the domain and are therefore on average less separated. The modal distance between ozone sites is around 10 km. Yet, monitoring density is highly variable in both the ozone and the temperature network with at least 40 temperature sites and 22 ozone ones isolated by more than 50 km from their nearest neighbor.

Meteorological and air quality observations reflect variations occurring at various spatial scales. The smallest scale that can accurately be resolved is determined by the distance between adjacent sites. Due to the very uneven distribution of temperature and ozone sites throughout the domain, the spatial resolution of both

networks is highly variable, ranging from a couple of km for ozone in some urban area to tens of km in more remote areas.

#### 4. Methods

Differences between the 4 km and 12 km predictions were examined from two perspectives. First, we assessed the distinctiveness between corresponding spatial fields without considering their resemblance to measurements. In this step, hereafter referred to as ‘model to model comparison’, attention is paid to identifying the common features and dissimilarities between spatial fields, but no judgement concerning their ‘quality’ is passed. In the second step of our analysis, referred to as ‘model evaluation’, we calculate traditional evaluation statistics, such as biases and correlation coefficients, to measure the closeness of the relationship between the model predictions obtained with both grid sizes and the observations.

##### 4.1. MODEL TO MODEL COMPARISON

An informative way to compare the 4 km and 12 km results, although mostly qualitative, is by visually examining maps of the model outputs pertaining to each scheme. Other useful visual aides for this purpose are maps delineating the areas where extreme values have been modeled. Histograms summarizing the differences between the 4 km and the 12 km model predictions at each location are also informative. To detail information provided by such histograms, maps delineating areas with large local variability (as quantified by the standard deviation of the nine 4 km values within each 12 km cell) and where the use of the finest grid may therefore be the most relevant were produced.

##### 4.2. MODEL EVALUATION

A model’s ability to reproduce past events is typically quantified by different statistical parameters, most of which are a linear function of the differences between model predictions and observed data [14–18]. Whether these differences are considered in real or relative terms, in absolute value or with their signs, normalized or not, squared or not, one obtains different evaluation statistics that occasionally lead to conflicting conclusions. Our evaluation relies on the magnitude of the correlation coefficient between model predictions and observations ( $R$ ), the mean model bias error (MBE) and the mean gross error (MGE). Table I describes the calculations of these statistics. The USEPA has recommended the use of the normalized version of the two latter parameters for evaluation of ozone predictions for urban-scale models. The normalized bias error at each location and time is defined by the difference between the model prediction and observation, divided by the observation. Our preliminary research revealed that normalizing the evaluation parameters tended to exaggerate the significance of the results pertaining to the morning hours.

Table I. Relationship used for calculation of the evaluation parameters.

Mean Bias Error (MBE)	$\text{MBE} = \frac{1}{n} \sum_{i=1}^n P(x, t) - O(x, t)$
Mean Gross Error (MGE)	$\text{MGE} = \frac{1}{n} \sum_{i=1}^n  P(x, t) - O(x, t) $
Correlation coefficient (R)	$R = \frac{\sum_{i=1}^n \{P(x, t) - \overline{P(\cdot, t)}\} \times \{O(x, t) - \overline{O(\cdot, t)}\}}{\sqrt{\sum_{i=1}^n \{P(x, t) - \overline{P(\cdot, t)}\}^2} \times \sqrt{\sum_{i=1}^n \{O(x, t) - \overline{O(\cdot, t)}\}^2}}$
With: $P(x, t)$ : value predicted by the model for node $x$ and time $t$ , $O(x, t)$ : observed value at node $x$ and time $t$ , $\overline{P(\cdot, t)}$ : mean of the predicted values at time $t$ , $\overline{O(\cdot, t)}$ : mean of the observations at time $t$	

During these hours a slight bias does lead to high values of the normalized bias and normalized gross error, circumstances that should not matter since modeled and observed ozone concentrations are low. For this reason, we decided not to use the normalized evaluation metrics.

When calculating the evaluation statistics, one may attempt to involve all cells in the modeled domain or only those which contain one or more monitoring sites. If the first option is chosen, a spatially continuous map that constitutes the basis for comparison must first be built from existing observations. Such a map can be obtained by spatially interpolating existing observations, using a kriging technique or another one. The interpolated values can be fully trusted if the spatial density of monitoring is adequate and if the salient points (high and low) of the spatial fields are monitored. Given the relative isolation (more than 50 km) of some ozone and temperature monitoring sites, added to the fact that some salient points may be missing, we have no assurance that the monitoring networks are sufficiently dense to resolve what might be simulated by the models on the 12 km and the 4 km grids. The interpolation option was therefore discarded. As a result, the sole comparison pairs used for evaluating the models are formed by the observed values and the model predictions in grid cells containing the observation sites.

Simulations by photochemical modeling systems are known not to reproduce faithfully the morning hours after sunrise and the evening hours after sunset when the mixing height experiences rapid changes [10, 19]. Aware of these model sensitivities, we chose not to calculate the evaluation parameters for all 144 h (3 episodes  $\times$  2 days/episode  $\times$  24 h/day) considered in this study but to limit our investigation to periods during which meteorological changes are gradual and models are expected to perform well. Two 5-h periods, from 1 a.m. to 5 a.m. and from 1 p.m. to 5 p.m., have been selected for this purpose. The separate calculation of the eval-

uation statistics for these two periods allowed differentiating model performance for nighttime and daytime simulations.

The evaluation statistics were first calculated using all observed and predicted pairs within the analysis domain during the two time periods chosen for evaluation. However, as illustrated in the next section, 12 km and 4 km predictions are very similar in large portions of the domain and are, therefore, quite useless for determining the merits of the finer discretization. In an effort to better differentiate the performance results for the two grid sizes, the evaluation statistics were also calculated for a subset of monitoring sites that only includes sites located in areas with high local variability (i.e., areas where the standard deviation of the nine 4 km values within each 12 km cells is in the top 5%), and where differences between 12 km and 4 km predictions are, therefore, to be expected. This subset of monitoring sites is referred to hereafter as the 'high local variability subset'.

The strength of most evaluation statistics considered here is that they conveniently summarize the differences between model results and observations into a single number, usually obtained by some type of averaging of these differences. Yet, averaging is a double-edged sword. While limiting the quantity of figures to examine, the process may also hide potential model flaws. An insignificant mean model bias (MBE) that could be interpreted as a proof of model adequacy may result from the combination of sub-regions in the domain with large positive and negative bias. Indeed, the space averaging process utilized for calculating the evaluation statistics dissimulate extreme local discordances between observations and predictions and disallows any region-specific assessment of model quality. To circumvent these two limitations, we prepared scatter plots of the mean gross error and coefficient of correlation obtained with the 12 km and the 4 km grid sizes. We also constructed maps identifying location of monitoring sites where the largest improvement and deterioration of model performance (in terms of MGE and correlation coefficient) were encountered when using the 12 km or the 4 km grid size. The variables considered for this task are the afternoon temperature and ozone averages over the six episode days analyzed in this study. In contrast to the hourly predictions, these variables have the advantage of characterizing the entirety of the time period considered. All 149 temperature stations and 214 ozone stations were utilized to produce the scatter plots.

## **5. Results and discussion**

### **5.1. MODEL TO MODEL COMPARISON**

The model to model comparison of temperature and ozone spatial fields involved examination of multiple maps pertaining to different days and hours, during the 3 episodes simulated. In order to limit the size of this paper, only one example of each type of map considered useful for comparing the two grid sizes is presented in this paper.

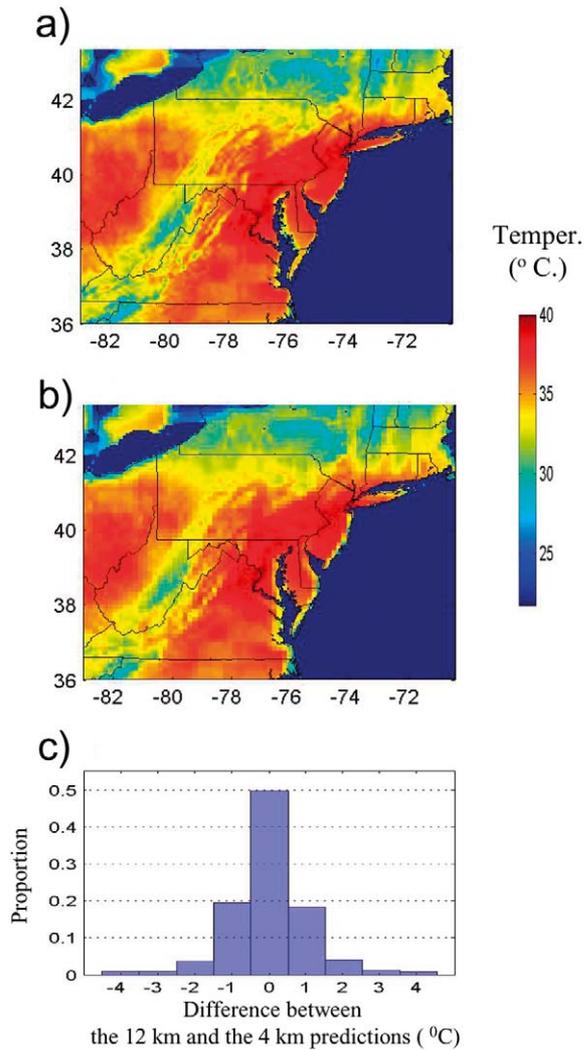


Figure 2. Comparison of 4 km and 12 km temperature predictions. (a) 4 km model predictions. (b) 12 km model predictions (c) histogram of the differences between the 4 km and 12 km predictions.

### 5.1.1. Simulation of Temperature

Figure 2 displays the 4 km and 12 km temperature fields estimated by RAMS4a (a and b) for July 15 1995, 5 p.m. (U.S. eastern standard time). The general aspect of both maps is very similar. Yet, the detailing allowed by the 4 km discretization gives the map a slightly more realistic look. The cumulative distribution function (CDF) of predictions pertaining to both grid sizes are indistinguishable (results not shown). Detailed evaluation of the differences between predictions at each location

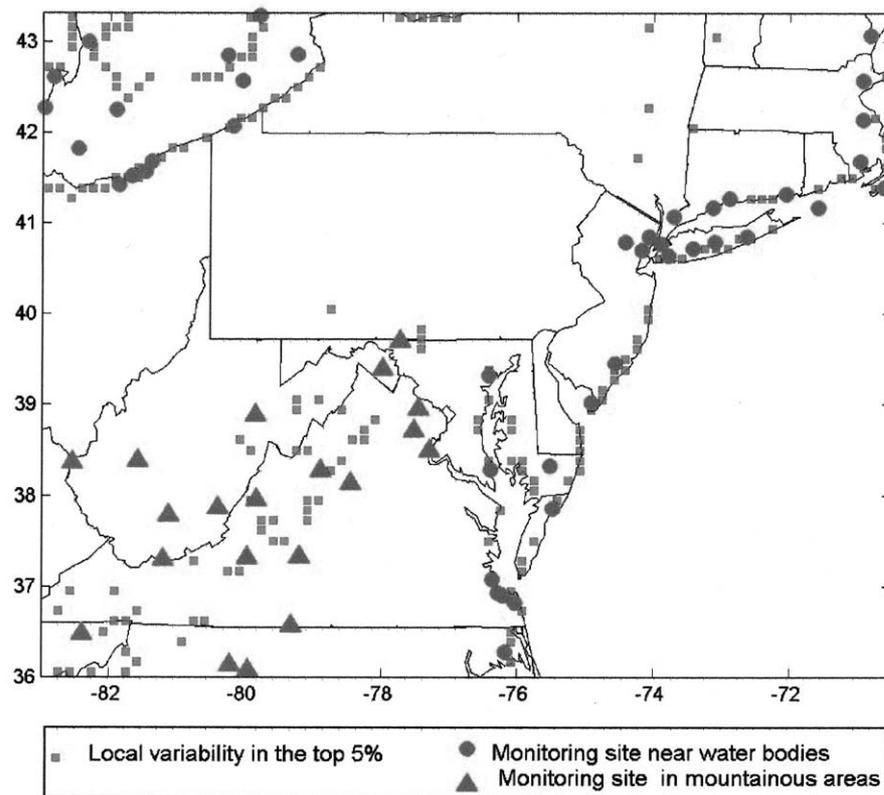


Figure 3. Location with the highest local variability (top 5%) in the 4 km predictions, where variability is expressed by the standard deviation of the nine 4 km predictions within each 12 km grid cell.

reveals that, in 50% of the domain, the 12 km and the 4 km predictions vary by more than  $0.5^{\circ}\text{C}$  and occasionally by more than  $2.0^{\circ}\text{C}$  (Figure 2c).

The standard deviation of the nine 4 km predictions within each 12 km grid cell, a statistic we used to measure local variability, was found to fluctuate from  $0.02^{\circ}$  to  $3.32^{\circ}\text{C}$  with a mean of  $0.42^{\circ}\text{C}$ . Figure 3 shows pixels where this standard deviation exceeds  $1.12^{\circ}$  (top 5% of CDF). It appears that the areas with the largest local variability mostly correspond to Appalachian mountain ridges and the shores of water bodies. The more detailed topographic information input in the model for the 4 km simulation is thought to be the cause for this behavior. Illustrating another aspect of the particularities of both discretization schemes, Figure 4 shows areas where model predictions of temperature exceed the 90th CDF percentile of all temperature predictions in the analysis domain (90th percentile of temperature CDF =  $37.4^{\circ}\text{C}$ ). It appears that the bulk of the areas so-delineated for both grid resolutions coincide. It is noted that the 4 km simulation displaced slightly the

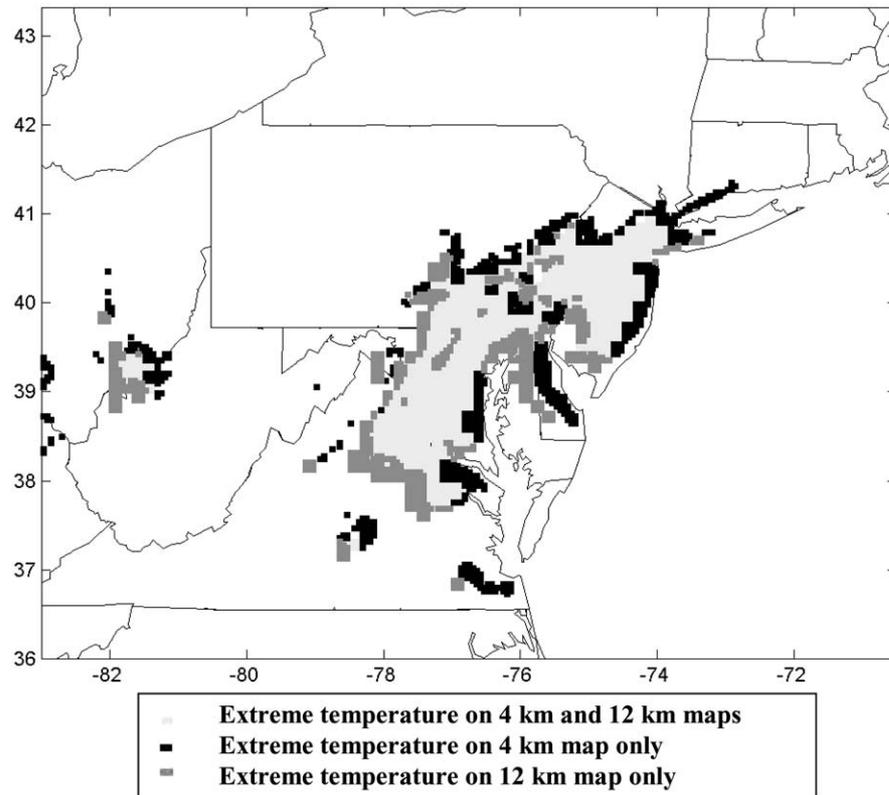


Figure 4. Comparison of the zones associated with extreme temperature (> 90th CDF percentile) in the 4 km and 12 km predictions.

warm front coming from the southwest towards the northeast, in comparison with the location of this front in the 12 km map.

#### 5.1.2. Simulation of TKE

The height of the mixing layer, the upper boundary for vertical dispersion of pollutants, is an essential parameter for air pollution studies. One method for its determination relies on examination of turbulent kinetic energy (TKE) isolines; the upper boundary of the mixing layer being the elevation at which TKE subsides below a threshold value, often fixed at  $0.1 \text{ m}^2/\text{s}^2$ . Figures 5 and 6 display isolines of TKE along a vertical cross-section, approximately parallel to the  $40^\circ \text{ N}$  latitude line (from south of Pittsburgh to Philadelphia). Figure 5 characterizes daytime simulations. It appears that the detailed topographic representation in the 4 km simulations results in a slightly higher mixing layer with small undulations (see segment 'a' in Figure 5) that follow the relief. The smaller grid size also allowed intensification of local gradients of TKE in the vicinity of topographic features (see segment 'b' in Figure 5, which is along the Susquehanna River valley). Night-

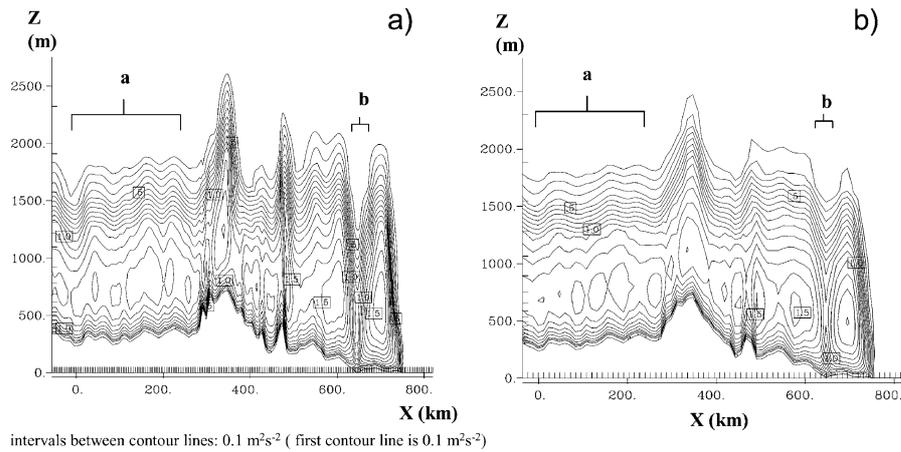


Figure 5. Vertical cross-section of TKE on 12 July 1995 at 1 p.m. (eastern standard time). (a) model estimates on the 4 km grid, (b) model estimates on the 12 km grid.

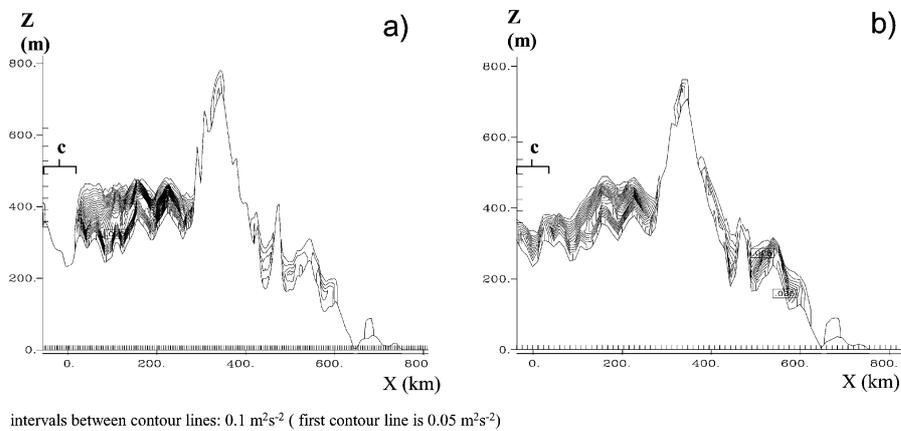


Figure 6. Vertical cross-section of TKE on 13 July 1995 at 1 a.m. (eastern standard time). (a) model estimates on the 4 km grid, (b) model estimates on the 12 km grid.

time results presented in Figure 6 indicated an overall reduction of turbulence and lowering of the height of the mixing layer. Differences between the 4 km and the 12 km TKE simulation results are not very noticeable, except around the Ohio River (see segment 'c' on Figure 6), where the 4 km results show a deeper mixing layer east of the river and no mixing layer west of it. TKE is also somewhat weaker on the 4 km map in the valleys east of the Appalachian ridge. Globally, differences between the 4 km and the 12 km TKE simulation results are more pronounced during daytime than nighttime, suggesting that the effects of a finer discretization on meteorological variables may be more pronounced for daytime predictions.

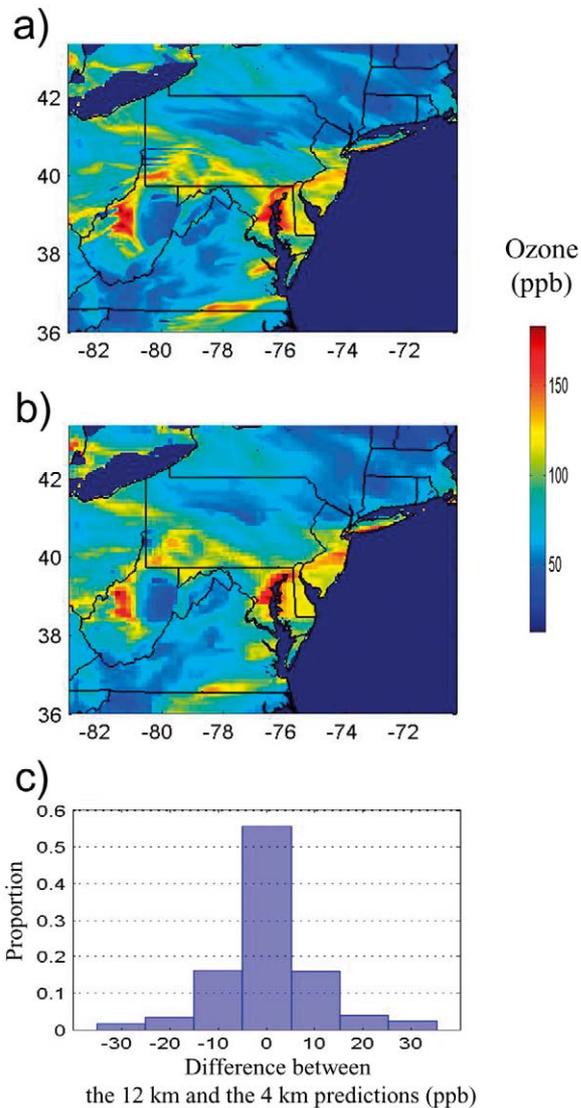


Figure 7. Comparison of 4 km and 12 km ozone predictions. Panel (a) 4 km model predictions. (b) 12 km model predictions. (c) histogram of the differences between the 4 km and 12 km predictions.

### 5.1.3. Simulation of Ozone

The 4 km and 12 km ozone spatial fields estimated by UAM – V for July 15<sup>th</sup> 1995 at 5 p.m. (U.S. eastern standard time) are presented in Figure 7a,b. The global features of these two maps are quite similar with the zones of high (e.g., Baltimore area) and low (e.g., upstate New York) concentrations almost identically located. That observation is not surprising since both models use the same county-wide

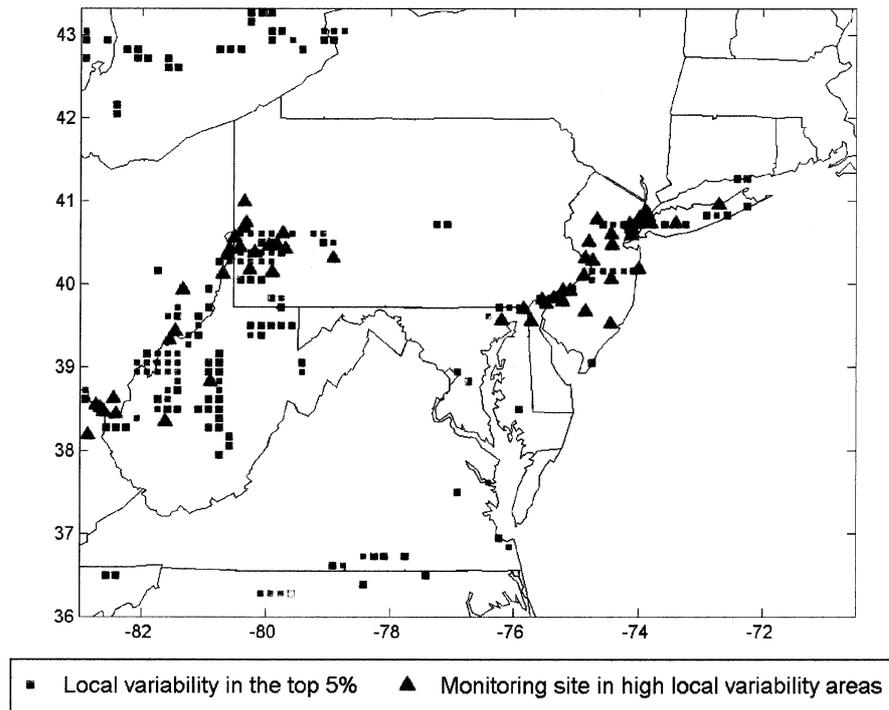


Figure 8. Locations of highest local variability (top 5%) in the 4 km ozone predictions, where variability is expressed by the standard deviation of the nine 4 km predictions within each 12 km grid cell.

emission fields. However, the emissions are more resolved in the 4 km map than in the 12 km one and the corresponding ozone maps exhibit more texture with finer features. For example, a thin stripe of high concentration (exceeding 120 ppb) can be distinguished around the 40° N – 77° W area (Pittsburgh, Pennsylvania) in the 4 km results, a feature that is not evident in the 12 km results. The cumulative distribution functions of 4 km and 12 km model predictions are very similar (results not shown). A detailed examination of the differences between 4 km and 12 km ozone concentration predictions at each location reveals that they are less than 5 ppb for about 50% of the domain (Figure 7). Differences larger than 25 ppb are encountered in 4% of the domain. The areas with large differences also correspond with high predicted concentrations (results not shown).

The standard deviation of the nine 4 km predictions within each 12 km grid cell varies from 0.10 ppb to 43.2 ppb with a mean of 4.2 ppb. The local coefficient of variation of ozone predictions is substantially greater than that of temperature predictions (mean coefficient of variation = 5.6% for ozone and = 1.3% for temperature), suggesting that the effect of the finer discretization scheme is more pronounced for ozone than for temperature. Figure 8 displays pixels with the largest local variability (top 5%). Some of these pixels are along the shores of

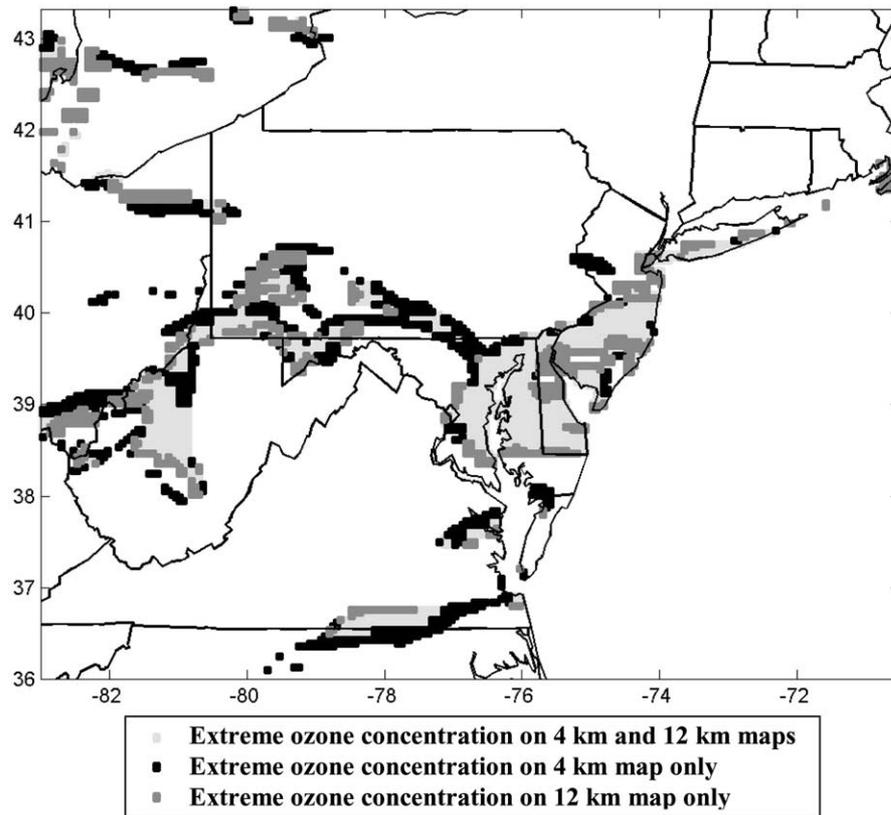


Figure 9. Comparison of the zones associated with extreme ozone concentrations (> 90th CDF percentile) in the 4 km and 12 km predictions.

Lake Erie or the Atlantic Ocean, but most of them are in the area located along and downwind of the Ohio River valley, where a major portion of the point source emissions is located. Not surprisingly, most areas with large local variability are also areas where the model has predicted high ozone concentrations.

Also useful for illustrating subtle differences between the 4 km and 12 km ozone predictions are maps delineating zones of extreme concentrations. For example, Figure 9 depicts zones where model predictions are larger than the 90th percentile of ozone predictions CDF (115 ppb) on the 4 km and the 12 km grid. Only about half the areas selected this way coincide. In contrast to temperature, the zones of extreme ozone concentrations do not constitute a unified area comparable to a warm front but are quite dispersed throughout the domain, a pattern probably resulting from the spatial distribution of the NO<sub>x</sub> emissions.

## 5.2. MODEL EVALUATION

Tables II and III summarize the evaluation parameters (mean bias error, mean gross error, correlation coefficient) for temperature and ozone, estimated episode by episode for both grid sizes. The first set of statistics presented in each table were obtained using all the observation sites (148 temperature sites, 214 ozone sites) present in the domain (Tables IIa and IIIa). The other sets characterize the areas with high local variability only (Tables IIb, IIc and IIIb). Figure 3 depicts the locations of the temperature monitoring sites included in the high local variability subset. Forty-seven of these sites are near water bodies and twenty are located in the Appalachian Mountains. Suspecting different model performances for these two types of environments, the evaluation statistics for mountainous locations versus nearby water bodies were summarized separately. For ozone, the high local variability subset consists of 57 sites whose locations are defined in Figure 8. Since differences between gridded emissions, not natural environmental factors (e.g., meteorology), are thought to be the dominant cause for differences between the 4 km and 12 km model predictions, all 57 ozone sites were considered simultaneously.

### 5.2.1. *Simulation of Temperature*

An inspection of the results pertaining to all sites (Table IIa) suggests that refining the grid size from 12 km to 4 km led to a slight improvement in the afternoon temperature estimates, as evidenced by a decrease in the overall mean bias from  $-0.57$  to  $-0.18$ , a decrease in the overall mean gross error from 2.25 to 1.97, and an increase in the mean correlation coefficient from 0.57 to 0.64. The slight superiority of the afternoon 4 km estimates is consistently observed for all 3 episodes considered. Decreasing grid size seems to have no impact on the nighttime results (mean bias equals 0.32 for the 12 km grid and 0.35 for the 4 km grid).

Focusing on the subset of sites with high variability in the vicinity of water bodies (Table IIb), we see that refining the grid leads to smaller biases ( $-1.1$  for the 4 km grid and 1.67 for the 12 km grid) and gross errors (2.17 for the 4 km grid and 2.87 for the 12 km grid) for simulation of the afternoon hours. However, refining the grid does not clearly improve the quality of the nighttime (1 to 5 a.m.) estimates (mean bias and gross error less for the 4 km grid size than the 12 km grid size but average correlation for the 4 km grid less than for 12 km).

For the mountainous area sites (Table IIc), the 4 km afternoon estimates tend to be slightly more accurate than the 12 km ones, a tendency that cannot be confirmed for the nighttime estimates.

We also computed the evaluation statistics for the afternoon average temperature over the six episode days at all 148 temperature stations and prepared a scatter plot of the statistics obtained for the 12 km simulation compared with those obtained for the 4 km simulation. These results are presented in the plots and maps shown in Figure 10. As evident from the symmetric scatter around the

Table II. Evaluation statistics for temperature.

a. All observation sites considered.

	Episode 1		Episode 2		Episode 3		All episodes	
	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours
MBE (°C)								
4 × 4 grid	0.36	-0.04	-0.04	-0.27	0.75	-0.24	0.35	-0.18
12 × 12 grid	0.49	-0.28	-0.20	-0.63	0.66	-0.63	0.32	-0.57
MGE (°C)								
4 × 4 grid	2.09	2.12	1.87	1.72	2.12	2.08	2.02	1.97
12 × 12 grid	2.10	2.41	1.90	2.05	2.02	2.26	2.01	2.25
R								
4 × 4 grid	0.29	0.64	0.23	0.65	0.29	0.61	0.27	0.64
12 × 12 grid	0.24	0.57	0.21	0.56	0.31	0.56	0.26	0.57

b. High local variability subset, sites near water.

	Episode 1		Episode 2		Episode 3		All episodes	
	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours
MBE (°C)								
4 × 4 grid	-0.87	-0.78	-1.33	-0.70	-1.10	-1.82	-1.10	-1.10
12 × 12 grid	-0.84	-1.14	-1.67	-1.48	-1.18	-2.40	-1.23	-1.67
MGE (°C)								
4 × 4 grid	1.85	2.19	1.82	1.98	1.84	2.33	1.84	2.17
12 × 12 grid	1.81	2.70	2.01	2.88	1.83	3.04	1.88	2.87
R								
4 × 4 grid	0.58	0.73	0.61	0.74	0.44	0.78	0.54	0.75
12 × 12 grid	0.55	0.66	0.66	0.59	0.48	0.65	0.57	0.63

c. High local variability subset, sites in mountainous areas.

	Episode 1		Episode 2		Episode 3		All episodes	
	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours
Bias Error (°C)								
4 × 4 grid	2.39	2.28	1.96	1.06	2.56	0.60	2.90	1.31
12 × 12 grid	2.59	2.44	1.75	1.22	2.04	0.44	2.12	1.37
Gross Error (°C)								
4 × 4 grid	2.88	2.61	2.27	1.60	2.49	1.32	2.54	1.84
12 × 12 grid	2.89	2.84	2.14	1.81	2.18	1.40	2.40	2.02
Correl. Coef.								
4 × 4 grid	-0.02	0.80	0.34	0.56	0.33	0.65	0.21	0.65
12 × 12 grid	0.04	0.75	0.32	0.49	0.40	0.56	0.25	0.60

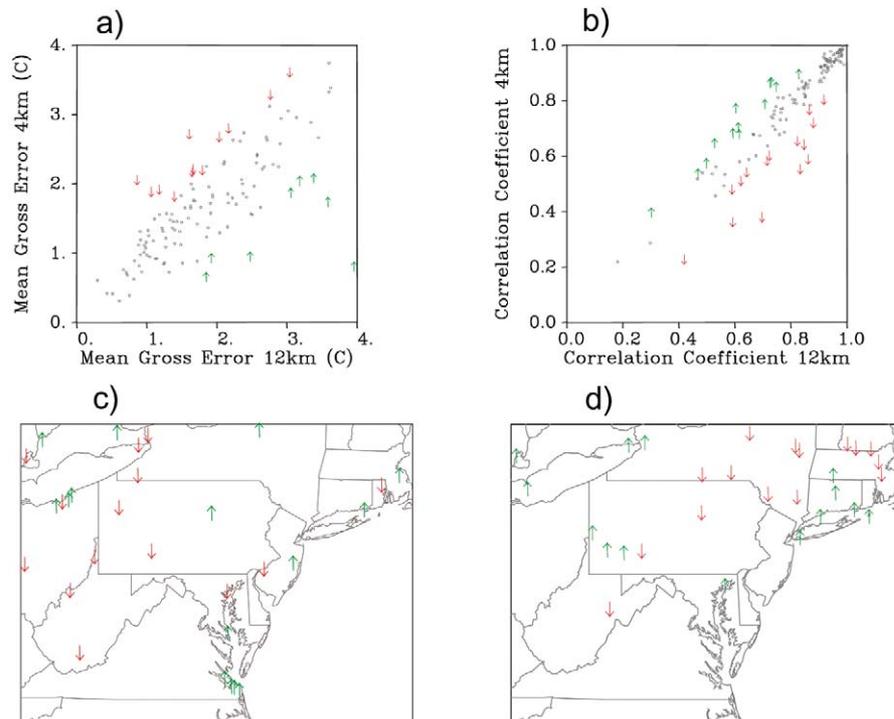


Figure 10. Comparison of mean gross errors (a) and correlation coefficients (b) calculated for both 12 km and 4 km simulations at all temperature monitors. The 10% of stations having the largest improvement of model performance for the 4 km predictions as measured by a particular metric are marked with upward-pointing arrows, while the 10% of stations having the largest deterioration of model performance for the 4 km simulations are marked with downward-facing arrows. (c) and (d) depict the location of the stations that were marked by arrows in panels (a) and (b), respectively.

1:1 line, there is no systematic improvement of the 4 km predictions compared to the 12 km predictions. However, by focusing on the data points located the furthest away from the 1:1 line and determining the corresponding monitoring stations, we investigated if there are regions in which there is a systematic improvement or deterioration for the 4 km simulation compared to the 12 km simulation. For this purpose, the 10% of the data points that show the largest improvement and the largest deterioration in MGE and correlation coefficient were highlighted on the scatter plots, and the corresponding monitors were plotted and highlighted on the maps displayed in Figure 10. For the MGE of temperature predictions, a cluster of stations that experienced the largest improvement (reduction of MGE) in the 4 km simulations can be seen from Baltimore south along the Chesapeake. The cluster of stations for which the 4 km simulation led to the largest deterioration of model performance as measured by MGE is located from eastern New York through

Table III. Evaluation statistics for ozone.

a. All observation sites considered.

	Episode 1		Episode 2		Episode 3		All episodes	
	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours
MBE (ppb)								
4 × 4 grid	-11.57	-9.22	-8.34	-0.12	-2.90	-3.43	-7.60	-4.25
12 × 12 grid	-10.56	-2.94	-7.64	4.29	-1.93	2.47	-6.70	1.27
MGE (ppb)								
4 × 4 grid	22.38	22.13	20.10	16.81	19.61	20.73	20.70	19.89
12 × 12 grid	22.43	20.78	19.62	16.76	19.15	18.65	20.40	18.74
R								
4 × 4 grid	0.17	0.58	0.16	0.58	0.18	0.37	0.17	0.51
12 × 12 grid	0.13	0.63	0.17	0.64	0.19	0.50	0.16	0.59

b. High local variability subset.

	Episode 1		Episode 2		Episode 3		All episodes	
	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours	a.m. hours	p.m. hours
MBE (ppb)								
4 × 4 grid	-14.0	-16.0	-12.6	-2.1	-11.7	-14.5	-12.8	-10.9
12 × 12 grid	-11.5	-6.8	-11.0	5.35	-8.3	-2.35	-10.3	-1.3
MGE (ppb)								
4 × 4 grid	30.7	29.4	24.0	19.5	20.4	26.3	25.0	25.0
12 × 12 grid	31.1	25.4	22.8	16.9	20.21	21.5	24.7	21.3
R								
4 × 4 grid	-0.34	0.31	0.06	0.39	0.16	0.12	-0.04	0.27
12 × 12 grid	-0.36	0.37	0.07	0.55	0.16	0.20	-0.05	0.37

West Virginia. When considering the locations of stations at which the correlation coefficient either showed the largest increase or decrease, no clear pattern emerges.

### 5.2.2. Simulation of Ozone

The results shown in Table IIIa for all observations and in Table IIIb for the high local variability sites suggest that, in the case of ozone, the 4 km grid did not lead to better results than the 12 km one, neither for nighttime nor daytime simulations. It even appears that refining the grid size from 12 to 4 km led to deterioration of the quality of ozone estimates. For all observations (Table IIIa), the overall afternoon (from 1 p.m. to 5 p.m.) bias increased from 1.27 to -4.25 when grid size decreased

from 12 km to 4 km. Similarly, the overall gross error increased from 18.74 to 19.89 and the correlation coefficient decreased from 0.59 to 0.51 for the afternoon ozone values. For nighttime simulations (from 1 to 5 a.m.), it appears that both the 4 km and the 12 km grid sizes led to less accurate results than for daytime simulations (12 km grid overall bias equals  $-6.7$  for nighttime and  $1.27$  for daytime, 12 km grid overall correlation coefficient equals  $0.16$  for nighttime and  $0.59$  for daytime).

The preceding comments concerning the comparative quality of the 12 km and the 4 km grid sizes also apply to the high local variability subset (Table IIIb). For this subset of 57 sites, the bias and gross error results are generally lower for the 12 km results than for the 4 km results, both for the nighttime and daytime estimates, (overall daytime bias equals  $-1.3$  for the 12 km grid and  $-10.9$  for the 4 km grid, overall nighttime bias equals  $-10.3$  for the 12 km grid and  $-12.3$  for the 4 km grid). Interestingly, model estimates are generally less accurate in the high local variability areas than on average in all the domain, as indicated by larger biases and smaller correlation.

In addition, by analogy to our analysis of temperature predictions, we computed the evaluation statistics for the afternoon average ozone predictions over the six episode days at all 214 ozone monitors and prepared a scatter plot of the statistics obtained for the 12 km simulation compared to the statistics obtained for the 4 km simulation. These results are presented in plots and maps in Figure 11. It is seen that the sites for which the 4 km and 12 km evaluation statistics differ the most, both in a positive and negative manner, are located in urban areas or near large emission sources. In other words, ozone levels in urban areas are most sensitive to changes in grid resolution. The maps also indicate that the attempt of the 4 km simulation to more accurately describe the chemical formation of ozone in those urban areas was not generally successful. While a reduction of MGE is evident in the Baltimore/Washington and Boston areas, an increase can be seen in the New Jersey/New York region. This points to underlying problems in either input fields or model formulations that cannot be alleviated by high-resolution modeling.

## 6. Summary

It is commonly believed that refining a model grid size will improve its ability to properly simulate all pertinent physical and chemical processes. However, decreasing grid size is a costly option in terms of computation resources and data requirement. It is therefore important to examine if, for the model application on hand, the expected benefits of a smaller grid size justify the necessary expenses. For simulation of secondary pollutants in the regulatory context, for instance, the uncertainties in the underlying emissions are such that a very fine grid resolution may not be justified.

In this paper, we examined RAMS4a temperature and UAM-V ozone predictions obtained using 12 km and 4 km grid sizes for three high ozone episodes that occurred over the northeastern United States during the summer of 1995. Compar-

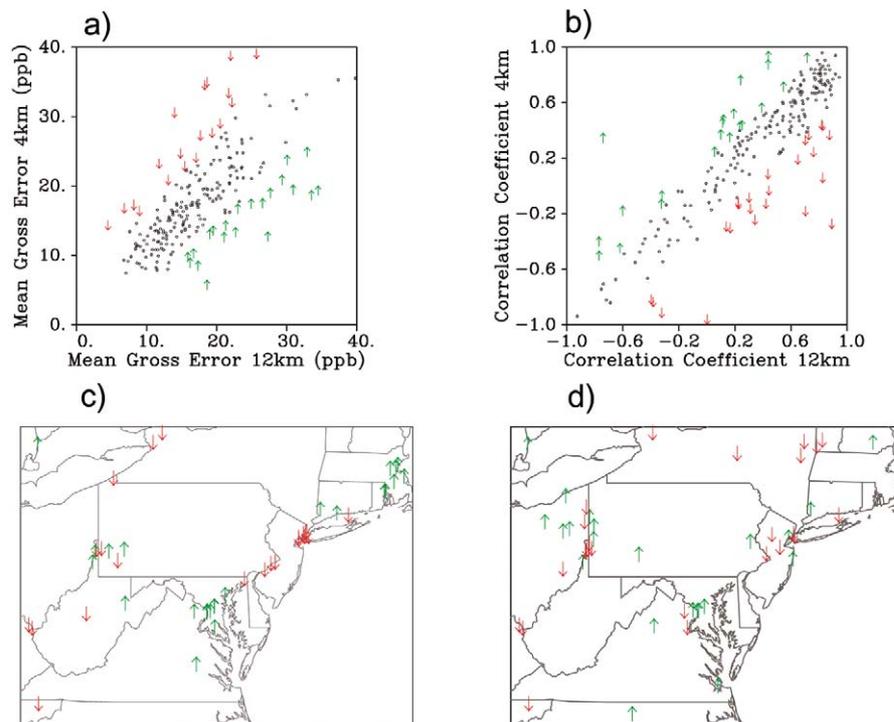


Figure 11. Comparison of mean gross errors (a) and correlation coefficients (b) calculated for both 12 km and 4 km simulations at all ozone monitors. The 10% of stations having the largest improvement of model performance for the 4 km predictions as measured by a particular metric are marked with upward-pointing arrows, while the 10% of stations having the largest deterioration of model performance for the 4 km simulations are marked with downward-facing arrows. Panels (c) and (d) depict the location of the stations that were marked by arrows in panels (a) and (b), respectively.

isons of the 12 km versus 4 km spatial fields predicted by the models show that these fields have many common features. The zones of high and low temperature or the zones of high and low ozone concentrations are almost identically located with both grid sizes. The cumulative distribution functions of model predictions in 4 km cells or 12 km cells are very similar. The increased grid resolution led to finer texture of the model predictions, conferring the output maps a more realistic appearance. Yet, due to the limited monitoring density we were precluded from verifying the authenticity of the spatial details present in the finely discretized fields. Since augmenting the number of monitoring sites to increase the density of spatial coverage is probably beyond current national resources, uncertainty about the veracity of the model details present in finely discretized fields will not be lifted in the near future.

Traditional evaluation statistics such as mean bias, mean gross bias and correlation coefficient were used to compare observations and predictions. Given the

current monitoring networks, it was shown that, in the case of temperature, the 4 km predictions tend to be slightly more accurate than their 12 km counterpart. However, the superiority of the finest grid could not be confirmed for ozone simulations; in fact, the 4 km ozone predictions appear to be less accurate than the 12 km predictions.

When using simulation models as tools for assisting the regulatory decision making process, the parsimony principle calls for use of grid cell sizes no larger than the general size of the States involved but no smaller than necessary to obtain adequate model performance. For the three episodes we investigated, the 12 km ozone predictions appeared to be as good or better than the 4 km predictions.

### Acknowledgements

The authors would like to thank Winston Hao (Department of Environmental Conservation, State of New York) for performing the UAM-V Simulations and Steve Porter (University of Idaho) for his constructive remarks throughout the course of the editing of this paper. The authors also wish to express their gratitude to the reviewers whose comments allowed improvement of the paper. One of the authors (C. Hogrefe) would like to acknowledge support by the U.S. Environmental Protection Agency under STAR grant R-82873301. Additional funding was provided by the New York State Energy and Development Authority under project numbers 4914 and 6085. This work was also funded in part by the United States Environmental Protection Agency through Interagency Agreements (DW 13938634 and DW 13938483) with the National Oceanic and Atmospheric Administration. This paper has been subjected to agency review for approval for presentation and publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

### References

1. Walko, R.L., Tremback, C.J. and Hertenstein, R.F.A.: 1995, *RAMS – The Regional Atmospheric Modeling System*, Version 3b, *User's Guide*, ASTER Division, Mission Research Corporation, Fort Collins, CO.
2. Grell, G.A., Dudhia, J. and Stauffer, D.: 1994, *A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5)*. NCAR Technical Note, NCAR/TN-398 + STR.
3. Systems Applications International: 1995, *Users Guide to the Variable Grid Urban Airshed Model (UAM-V)*, Systems Applications International, San Rafael, CA, 131 pp. [Available from Systems Applications International, 101 Lucas Valley Rd., San Rafael, CA 94903.]
4. ENVIRON: 1997, *User's Guide to the Comprehensive Air Quality Model with Extensions (CAMx)*. Available from ENVIRON International Corporation, 101 Rowland Way, Novato, CA 94945.
5. Byun, D.W. and Ching, J.K.S. (eds.): 1999, *Science algorithms of the EPA Models-3 Community Multiscale Air Quality Model (CMAQ) modeling system*. EPA/600/R-99/030, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC 20460.

6. Kunz, R. and Moussiopoulos, N.: 1997, Implementation and assessment of a one-way nesting technique for high resolution wind flow simulations, *Atmos. Environ.* **31**, 3167–3176.
7. Mass, C., Ovens, D., Albright, M. and Westrick, K.: 2002, Does Increasing Horizontal Resolution Produce Better Forecasts? The Results of Two Years of Real-Time Numerical Weather Prediction in the Pacific Northwest, *Bull. Amer. Meteorol. Soc.* **83**, 407–430.
8. Colle, B.A., Olson, J.B. and Tongue, J.S.: 2001, Verification of the Eta and Real-Time MM5 over the Eastern U.S., paper presented at Eleventh PSU/NCAR MM5 Users' Workshop, Foothills Laboratory, NCAR, Boulder, Colorado, June 25–June 27 2001. (Available at <http://www.mmm.ucar.edu/mm5/workshop/ws01/>).
9. Geron, C., Guenther, A. and Pierce, T.: 1994, An Improved Model for Estimating Emissions of Volatile Organic Compounds from Forests in the Eastern United States, *J. Geophys. Res.* **99**, 12773–12791.
10. Sistla, G., Hao, W., Ku, J.-Y., Kallos, G., Zhang, K., Mao, H. and Rao, S.T.: 2001, An Operational Evaluation of Two Regional-Scale Ozone Air Quality Modeling Systems over the Eastern United States, *Bull. Amer. Meteorol. Soc.* **82**, 945–964.
11. Biswas, J. and Rao, S.T.: 2001, Uncertainties in Episodic Ozone Modeling Stemming from Uncertainties in the Meteorological Fields. *J. Appl. Meteorol.* **40**, 117–136.
12. Lagouvardos, K., Kallos, G. and Kotroni, V.: 1997a, *Modeling and Analysis of Ozone and its Precursors in the Northeast U.S.A. (Atmospheric Model Simulations)*, Final report to Electric Power Research Institute, Palo Alto, CA, 46 pp. [Available from Office of Science and Technology, NYSDEC, 50 Wolf Rd., Albany, NY 12233–3259.
13. Lagouvardos, K., Kotroni, V., Kallos, G. and Rao, S.T.: 1999, An Analysis of the Meteorological and Air Quality Conditions during an Extreme Ozone Episode over the Northeastern USA, *Intern. J. Environ. Poll.* **14**, 581–587.
14. Willmott, C.J.: 1982, Some Comments on the Evaluation of Model Performance, *Bull. Amer. Meteorol. Soc.* **63**, 1309–1313.
15. Hanna, S.R.: 1994, Mesoscale Meteorological Model Evaluation Techniques with Emphasis on Needs of Air Quality Models. In: R.A. Pielke and R.P. Pearce (eds.), *Mesoscale Modeling of the Atmosphere*, pp. 47–58, Meteorological monographs 25, American Meteorological Society, 45 Beacon Str., Boston, MA 02108.
16. Olerud, D. and Wheeler, N.: 1997, *Investigation of the Regional Atmospheric Modeling System (RAMS) for the Ozone Transport Assessment Group (OTAG)*, Report to the Southeast Modeling Center, MCNC Environmental Programs, Research Triangle Park, NC 27709.
17. McNally, D. and Tesche, T.W.: 1993, *MAPS Sample Products*, Alpine Geophysics, 16225 W. 74th Dr., Golden, CO 80403.
18. American Society for Testing and Materials D 6589: 2000, *tandard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589)* (available at <http://www.astm.org>), 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428, 17 p.
19. Sistla, G., Zhou, N., Hao, W., Ku, J.-Y., Rao, S.T., Bornstein, R., Freedman, F. and Thunis, P.: 1996, Effects of Uncertainties in Meteorological Inputs on Urban Airshed Model Predictions and Ozone Control Strategies, *Atmos. Environ.* **30**, 2011–2025.