New Categorical Metrics for Air Quality Model Evaluation

DAIWEN KANG*

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, Research Triangle Park, North Carolina

ROHIT MATHUR AND KENNETH SCHERE

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, and National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

SHAOCAI YU*

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, Research Triangle Park, North Carolina

BRIAN EDER

Atmospheric Sciences Modeling Division, Air Resources Laboratory, National Oceanic and Atmospheric Administration, and National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

(Manuscript received 15 February 2006, in final form 10 August 2006)

ABSTRACT

Traditional categorical metrics used in model evaluations are "clear cut" measures in that the model's ability to predict an "exceedance" is defined by a fixed threshold concentration and the metrics are defined by observation–forecast sets that are paired both in space and time. These metrics are informative but limited in evaluating the performance of air quality forecast (AQF) systems because AQF generally examines exceedances on a regional scale rather than a single monitor. New categorical metrics—the weighted success index (WSI), area hit (aH), and area false-alarm ratio (aFAR)—are developed. In the calculation of WSI, credits are given to the observation–forecast pairs within the observed exceedance region (missed forecast) or the forecast exceedance region (false alarm), depending on the distance of the points from the central line (perfect observation–forecast match line or 1:1 line on scatterplot). The aH and aFAR are defined by matching observed and forecast exceedances within an area (i.e., model grid cells) surrounding the observation location. The concept of aH and aFAR resembles the manner in which forecasts are usually issued. In practice, a warning is issued for a region of interest, such as a metropolitan area, if an exceedance is forecast to occur anywhere within the region. The application of these new categorical metrics, which are supplemental to the traditional counterparts (critical success index, hit rate, and false-alarm ratio), to the Eta Model–Community Multiscale Air Quality (CMAQ) forecast system has demonstrated further insight into evaluating the forecasting capability of the system (e.g., the new metrics can provide information about how the AQF system captures the spatial variations of pollutant concentrations).

* Additional affiliation: Science and Technology Corporation, Hampton, Virginia.

Corresponding author address: Daiwen Kang, Atmospheric Modeling Division, U.S. EPA, Mail Drop E243-03, Research Triangle Park, NC 27711.
E-mail: kang.daiwen@epa.gov

DOI: 10.1175/JAM2479.1

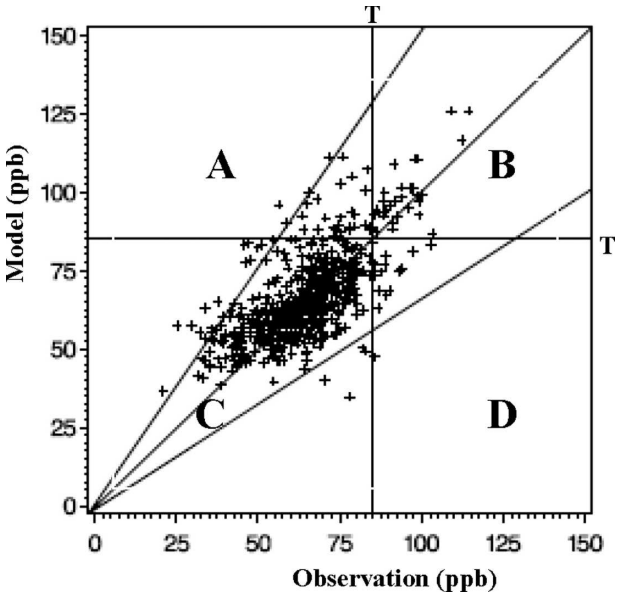

FIG. 1. Example scatterplot for the definition of traditional categorical metrics.

1. Introduction

The skill of an air quality forecast system is gauged by how well the modeling system predicts species concentrations in relation to threshold values. These events, which are referred to as "exceedances" and "nonexceedances," can be evaluated using categorical metrics. Current categorical metrics used in model evaluations (e.g., Kang et al. 2005) measure the model's ability to predict an exceedance using a fixed threshold mixing ratio (or "clear cut"), and the metrics are calculated using direct observation–forecast pairs in space and time. However, both the observations and the model forecasts represent different spatial and temporal scales. For example, observations are point measurements at fixed locations, whereas model forecasts represent volume-average mixing ratios. The direct matching of the point observations to the volume-mean model forecast may result in misleading conclusions in model performance evaluations, especially when exceedances are very sparse spatially.

To avoid the so-called clear-cut effect, three new categorical metrics—weighted success index (WSI), area hit rate (aH), and area false-alarm ratio (aFAR)—are developed as a supplement to the traditional metrics. As shown in Fig. 1, an observation–forecast scatterplot is generally classified into four quadrants by threshold value that marks exceedances. Quadrant A contains forecast exceedances that were not observed, quadrant B presents both forecast and observed exceedances, in quadrant C reside the data points that are both forecast and observed nonexceedances, and quadrant D holds the observed exceedances that were not forecast. In the definition of WSI, credits are given to the observation–forecast pairs within the observed exceedance region (but missed forecast) or the forecast exceedance region (but false alarm) depending on the distance of the data points from the perfect observation–forecast match line (i.e., the 1:1 line on a scatterplot). Metrics aH and aFAR are defined to resemble the manner in which forecasts are usually issued. In practice, a warning is issued for a region of interest (such as a metropolitan area) if an exceedance is forecast to occur anywhere within the region. So, the aH is defined as a hit if an observed exceedance is matched within an area (adjacent grid cells) surrounding the observation location (central cell) and the aFAR is defined to reflect number of false-alarm ratios of the forecast if a forecast exceedance is not paired with an observed one within the area. In this paper, the use of these metrics is demonstrated using model guidance from the national air quality forecasting system.

2. Existing categorical metrics

For the categorical forecast evaluation, the model's accuracy, bias, hit rate H, false-alarm ratio (FAR), and critical success index (CSI) are typically examined using the National Ambient Air Quality Standards (NAAQS) as the thresholds to define exceedance values. As presented in Fig. 1 [using the NAAQS for ozone (O_3)], a represents the number of forecast 8-h exceedances (O_3 mixing ratios ≥ 85 ppb) that were not observed (i.e., data pairs in quadrant A), b represents the number of correctly forecast 8-h exceedances (data pairs in quadrant B), c represents the number of correctly forecast 8-h nonexceedances (data pairs in quadrant C), and d represents the number of observed 8-h exceedances that were not forecast (data pairs in quadrant D). The number of model–observation pairs in A, B, C, and D form the basis for calculating the categorical evaluation metrics. If the denominator in any of the definitions is 0, then it is considered to be not applicable.

The FAR measures the percentage of times an exceedance was forecast and did not occur:

$$\mathrm{FAR} = \left(\frac{a}{a+b}\right) \times 100\%. \tag{1}$$

Smaller values of FAR are desirable: for example, a FAR of 0% indicates no false alarms and a FAR of 50% indicates that one-half of the forecast exceedances were not observed.

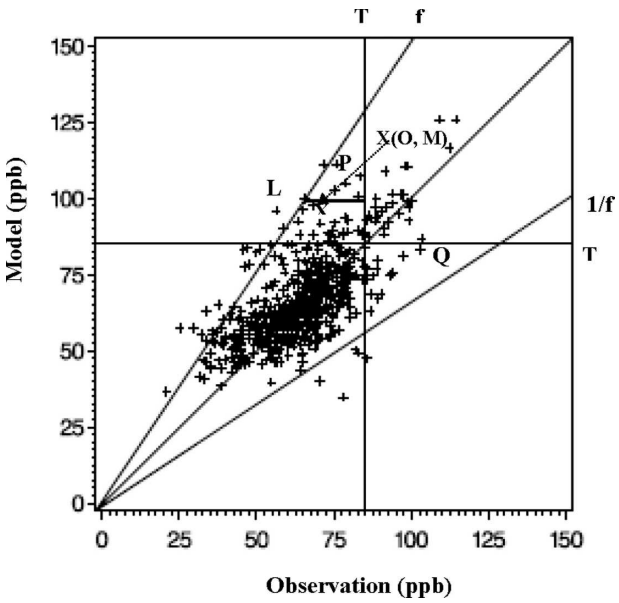

FIG. 2. Example scatterplot for the definition of WSI (see text).

The CSI indicates how well exceedances were predicted by considering false alarms and missed forecasts of exceedances:

$$\mathrm{CSI} = \left(\frac{b}{a+b+d}\right) \times 100\%. \tag{2}$$

The CSI is not influenced by correctly forecast nonexceedances, which can be disproportionately large for some species. A CSI of 50% indicates that one-half of the forecast and observed exceedances were correct.

Metric H, which is similar to the CSI and is often called the probability of detection, indicates the percentage of actual exceedances that are correctly forecast:

$$H = \left(\frac{b}{b+d}\right) \times 100\%. \tag{3}$$

3. New categorical metrics

The categorical statistics discussed in section 2 are defined by the numbers of paired data points found in the quadrants defined by threshold values T as shown in Fig. 1. Although informative, each metric can only represent the model's performance limited to some aspects. To illustrate some of the limitations of traditional categorical statistics, consider $x(O, M)$ (Fig. 2), which represents a paired data point (O is the observed value and M is the modeled value) that lies within quadrant A (forecast exceedance that did not occur) but also lies within a factor line f (fixed M/O value) inside a triangle designated as P. In comparing model forecasts with observations, f is often used to represent reasonable model performance and reflects inherent uncertainties in both the model and measured values. This individual forecast, though considered a "failure" or false alarm using traditional categorical metrics, in actuality may be considered a "success" if inherent uncertainties associated with representing the variability associated in comparing grid and point values, as well as intrinsic model process representation and measurements, are factored into the analysis. The same is true for points falling into quadrant D but within the lower factor line (triangle Q). In accord with this concept, three new metrics are proposed (based on traditional categorical metrics) to account for the uncertainties associated with both model process representation and measurements.

a. Weighted success index

The WSI gives credit for points located in the triangles P and Q that are formed by the threshold lines T and the factor line f (Fig. 2). The choice of f is empirical and is based on rules of thumb (Hanna 2006). Analysis of real-time O_3 forecasts for the past 2 yr has shown that about 80% of the data points are located within a factor of 1.5 on prediction–observation scatterplots; thus, in this study, f is set to 1.5.

If a data point $x(O, M)$ is within P ($M > T > O$ and $M < fT$), the length L of the line that passes through x and intersects both T and f can be computed as

$$L = T - (1/f)M. \tag{4}$$

Length L can then be used to define an intermediate parameter (IP) that represents WSI on the prediction side:

$$\mathrm{IP} = 1 - \frac{T-O}{L} = 1 - \frac{T-O}{T-(M/f)} = \frac{M-fO}{M-fT}. \tag{5}$$

The values of P are between 0 and 1.

In a similar way, for points in Q ($O > T > M$ and $O < fM$)—that is, an observed but not forecast exceedance—IP is defined as

$$\mathrm{IP} = \frac{O-fM}{O-fT}. \tag{6}$$

For any other conditions, IP = 0.0. Then the WSI is defined as

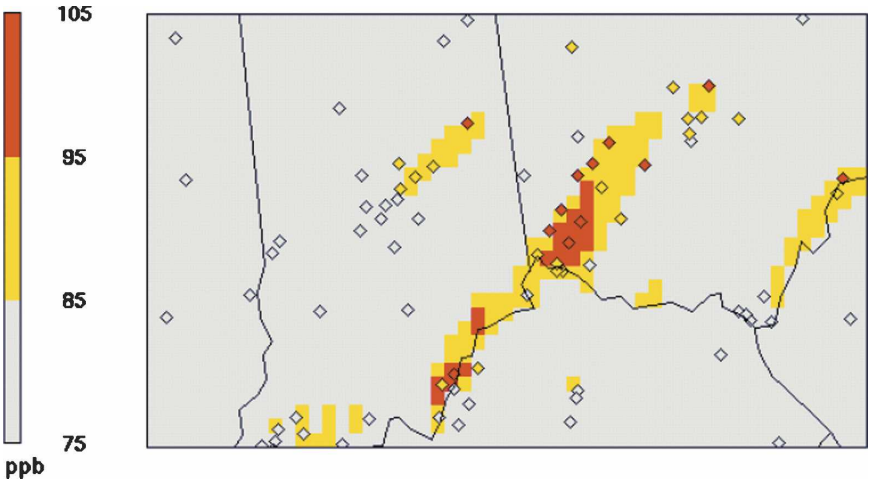

FIG. 3. Sample predicted (background) and observed (diamond overlay) maximum 8-h O_3 mixing ratios (ppb).

$$\mathrm{WSI} = \frac{b + \sum_{1}^{n} \mathrm{IP}}{a + b + d} \times 100\%, \tag{7}$$

where n is the number of data points. Values of WSI range from 0% (worst possible forecast) to 100% (perfect forecast). As seen from the definitions, both WSI and CSI have the same denominator, but the numerator in the WSI definition contains the intermediate term that credits the points within the triangles P and Q (Fig. 1). For a perfect forecast or for a no-event (neither observed nor forecast exceedances exist) forecast, WSI and CSI are the same.

b. Area hit rate

The H defined by Eq. (3) indicates the percentage of observed exceedances that were forecast, where the forecast exceedances are only from the grid cell in which the monitor is located. In some cases, the monitor may be located just at the edge or corner of the model grid cell, which may not best represent the conditions of the observation site. The air quality model forecast will also reflect spatial and temporal errors in the simulation of meteorological features (frontal systems, precipitation, cloud cover, etc.), especially with increasingly finer model resolutions. On the other hand, air quality forecasts are typically issued for regions such as metropolitan areas. Wherever an exceedance is forecast to occur within the area, a warning of the exceedance will be issued for the whole area. As illustrated in Fig. 3, it is often the case that observed exceedances (red or orange diamonds) are only one or two grid cells away from the model forecast exceedances (dark or light background). From this practical consideration, a new aH metric is developed that reflects both the spatial uncertainties of the model forecast and practical considerations typically employed by a local forecaster.

Area hit is defined as

$$\mathrm{aH} = \left(\frac{\mathrm{Ab}}{\mathrm{Ab} + \mathrm{Ad}}\right) \times 100\%, \tag{8}$$

where Ab is the number of exceedances that are both observed and forecast but the forecast is any exceedance that occurs in the designated area centered at the monitor location. Parameter Ad is the number of observed exceedances that are not forecast within the designated area centered at the monitor location. In this illustration, the area includes the grid cell in which the monitor resides (i.e., the center cell) and the adjacent cells (Fig. 4). The value of aH depends on the size of the selected area. If the area only covers the center cell, then aH collapses to H. In general, the larger the size of the chosen area is, the larger the aH values will be. However, the concept of area hit rate also intrinsically incorporates the uncertainties associated with grid resolution. For a grid resolution of Δ, features occurring on spatial scales of less than 2Δ cannot be resolved by the model. Thus, as an initial demonstration of the application of the metrics, we examine them on model–observation pairs on scales of 2Δ and 3Δ. For the Eta Model–Community Multiscale Air Quality (CMAQ) forecast system with a 12-km horizontal grid spacing, the area includes either one or two cells on each side of the central cell; in this way the area covers 9 and 25 grid cells and forms a square of 36 km $\times$ 36 km or 60 km $\times$

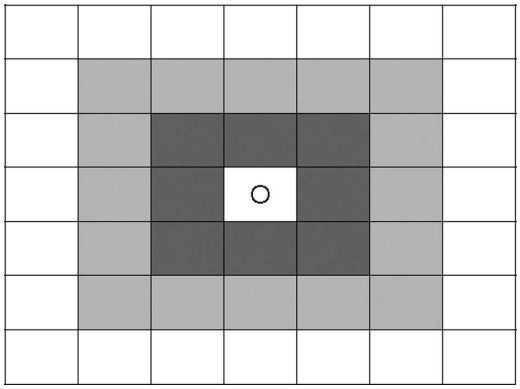

FIG. 4. Illustration of the "area" categorical metrics: the grid cell in which an observation site is located is marked with a letter "O," the dark-gray cells are the adjacent cells one cell away from the O cell, and light-gray cells are the adjacent cells two cells away from the O cell.

60 km (Fig. 4), respectively. However, if the observation site is located at the edge or the corner of the modeling domain, then only the adjacent cells that reside within the domain are counted.

c. Area false-alarm ratio

The aFAR is defined using the same spatial concept as that of aH. In mathematical terms, aFAR is defined as

$$\text{aFAR} = \left(\frac{\text{Aa}}{\text{Aa} + \text{Ab}}\right) \times 100\%, \tag{9}$$

where Aa is the number of forecast area exceedances that were not observed and Ab is the number of forecast exceedances that were observed. When an area contains multiple observation sites, all of the observation sites within the selected area will be considered in aFAR; if no exceedances are observed at any sites located within this area, false alarms are recorded (all of the sites contribute to Aa). If an exceedance is recorded at any of the observation sites, then no false alarm is counted, even for the sites that did not observe exceedances (i.e., all sites contribute to Ab).When the area only includes the cell that contains the monitor, aFAR becomes FAR.

4. Case study

The new categorical metrics (WSI, aH, and aFAR) are compared with their traditional counterparts (CSI, *H*, and FAR) using the Eta Model–CMAQ real-time O_3 forecasts for the period from 13 June to 31 July 2005. The national air quality forecasting (AQF) system is based on the National Centers for Environmental Prediction (NCEP) Eta Model (Black 1994; Rogers et al. 1996) and the U.S. Environmental Protection Agency (EPA) CMAQ modeling system (Byun and Ching 1999; Byun and Schere 2006). The Eta Model provides the meteorological fields for input to CMAQ. The processing of the emission data for various pollutant sources has been adapted from the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system (Houyoux et al. 2000) using input from the EPA national emission inventory. The Carbon Bond chemical mechanism (version 4.2) is used to represent the photochemical reactions. Detailed information on transport and cloud processes in the CMAQ is described in Byun and Ching (1999). For this application, O_3 mixing ratios are forecast for the eastern United States using a 12-km horizontal grid. There are 22 sigma layers extending from the surface to 100 hPa. The chemical fields for CMAQ are initialized using the previous forecast cycle. The primary Eta Model–CMAQ model forecast for next-day surface layer O_3 is based on the current day's 1200 UTC cycle. The target forecast period is from local midnight through local midnight (the hours beginning 0400–0300 UTC for the eastern United States). Hourly, near-real-time O_3 observations obtained from the EPA's "AIRNow" program (at the time of writing, information was available online at http://airnow.gov) are used in this study. Additional details can be found in Otte et al. (2005).

The forecast domain covers the eastern United States, which includes more than 850 AIRNow monitoring sites. During this forecast period, 1083 exceedances (maximum 8-h O_3 mixing ratios $\geq$ 85 ppb) were observed, which resulted in a CSI of 19.2% and a WSI of 59.8% when the factor *f* (Fig. 2) is set at 1.5 (80% of the observation–forecast pairs are within a factor of 1.5). This apparent increase in skill with WSI (when compared with CSI) indicates that there are many observation–forecast pairs in *P* and *Q* (see Fig. 2) that the CSI metric categorizes as failures. When the proximity of these data points to the acceptable factor line is taken into consideration by the weighting associated with the WSI, a more representative measure of the model's performance is obtained.

As seen from Fig. 5, about 40% of the exceedances during this period are forecast if aH is calculated like *H* (i.e., using an area of 1 $\times$ 1 grid cell). If the area is expanded to 3 $\times$ 3 grid cells, then the aH increases to about 70%. About 30% of the observed exceedances are forecast in a grid cell that is adjacent to the cell in which the monitor is located. When the area is expanded from 3 $\times$ 3 to 5 $\times$ 5 grid cells, the aH is increased by another 10%. This result indicates that the majority of exceedances are captured by the forecast system within the 3 $\times$ 3 grid cell area. The values of

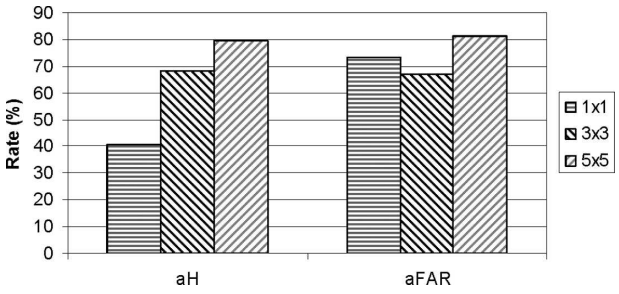

FIG. 5. The aH and aFAR for all observations calculated by direct match (1 × 1), 9-gridcell area (3 × 3), and 25-gridcell area (5 × 5) during the period from 13 Jun to 31 Jul 2005.

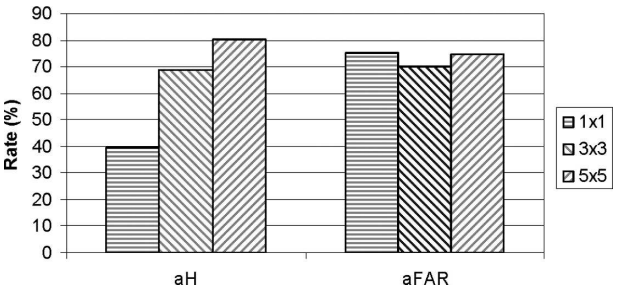

FIG. 6. As in Fig. 5, but for only the urban and suburban regions.

aFAR are smallest when it is calculated over a 3 × 3 grid cell area but increase when calculated over the 5 × 5 grid cell area. These results are consistent with analyses by Rao et al. (1997) that demonstrate that ambient O_3 mixing ratios and its exceedances have a spatial scale of 15 km. The 3 × 3 grid cell area roughly covers the spatial scale of 18 km around a monitoring station.

In addition to evaluating the air quality forecasts over the entire model domain, the forecast system is examined over urban and suburban areas where most of the human activities take place and on which air quality forecasts are primarily focused. Of the 1083 exceedances observed during the study period, 637 were observed in urban or suburban sites. Both the CSI (18.0%) and WSI (59.2%) in urban and suburban regions are slightly lower than those (19.2% and 59.8%) calculated over all sites in the modeling domain. This result indicates that the success rate of detecting exceedance events is slightly lower over urban and suburban regions than over rural locations. This, in part, is related to the inability of the model resolution to represent the spatial inhomogeneity in nitrogen oxide levels in urban areas; spatial analysis of ozone data by Rao et al. (1997) suggests a resolution of about 5 km to capture such spatial variability in ozone exceedance in urban areas. Similar to Fig. 5, Fig. 6 shows the aH and aFAR values for the urban and suburban sites calculated over different area sizes. As seen in Fig. 6, there are no statistically significant differences between the aH values in urban and suburban areas and those with all of the measurement sites (Fig. 5) when calculated over the same area size. The aFAR value in urban and suburban areas is 6.5% lower than that over the entire domain when calculated over the 5 × 5 grid cells, and the aFAR values over the 1 × 1 and 3 × 3 grid cells are about 2%–3% larger in urban and suburban areas than those over the entire domain.

As discussed earlier, air quality forecasts are typically issued for a functioning region, for example, a city, a metropolitan area, or an industrial region. The practical extension of the proposed new categorical metrics aH and aFAR is expected to provide a more representative indication of model forecast performance when compared with the traditional categorical evaluation metrics.

5. Summary

Three categorical metrics, WSI, aH, and aFAR, are developed to evaluate model performance in forecasting exceedance events. These new metrics supplement the traditional categorical metrics (CSI, H, and FAR) and provide a more relaxed but practical way to evaluate model performance when compared with the existing counterparts. The new metrics not only evaluate the exceedances and nonexceedences as their existing counterparts do for the clear-cut match, but also evaluate the "effort" or "potential" of the forecast system.

The application of the aH and aFAR metrics also reveals useful spatial performance information of forecast systems when the evaluating area expands from the 1 × 1 grid cell to the 5 × 5 grid cells. The case study shows that approximately 40% hit occurred in the direct match (1 × 1 grid cell), about another 30% hit occurred in the immediately adjacent cells (3 × 3 grid cells), and only about 10% hit would be gained by expanding to 5 × 5 grid cells.

In practice, if the information about the coverage of local forecasts is available, aH and aFAR can be calculated for the actual area over which local forecasts are issued (such as a metropolitan area). When the area included in the statistic is extended from the adjacent grid cells to local forecast regions, then the "area" or "a" in aH and aFAR is not arbitrary but rather represents actual forecast areas. Future research will include the application of these metrics to actual forecast areas.

Acknowledgments. The authors are grateful to Tanya Otte and Jonathan Pleim for their helpful and insightful

comments on initial drafts of this manuscript. The research presented here was performed under the Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under Agreement DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

REFERENCES

Black, T., 1994: The new NMC Mesoscale Eta Model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Byun, D. W., and J. K. S. Ching, Eds., 1999: Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system. U.S. Environmental Protection Agency Science Doc. EPA-600/R-99/030, 727 pp.

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

Hanna, S. R., 2006: A review of uncertainty and sensitivity analysis of atmospheric transport and dispersion models. Preprints, *28th NATO/CCMS Int. Technical Meeting on Air Pollution Modeling and Its Application,* Leipzig, Germany, NATO/CCMS, 225–237.

Houyoux, M. R., J. M. Vukovich, C. J. Coats Jr., N. M. Wheeler, and P. S. Kasibhatla, 2000: Emission inventory development and processing for the Seasonal Model for Regional Air Quality (SMRAQ) project. *J. Geophys. Res.,* **105,** 9079–9090.

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

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

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

Rogers, E., T. L. Black, D. G. Deaven, G. J. DiMego, Q. Zhao, M. Baldwin, N. W. Junker, and Y. Lin, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting.,* **11,** 391–413.