DEVELOPMENT OF NEW CATEGORICAL METRICS FOR AIR QUALITY MODEL EVALUATION AND THEIR APPLICATION TO ETA-CMAQ FORECASTS

Daiwen Kang*[$], Rohit Mathur[+], Kenneth Schere[+], Shaocai Yu[$], and Brian Eder[+]

Atmospheric Sciences Modeling Division, Air Resources Laboratory,
National Oceanic and Atmospheric Administration, RTP, NC, USA
[+]On assignment to National Exposure Research Laboratory,
U.S. Environmental Protection Agency, Research Triangle Park, NC, USA
[$] On assignment from Science and Technology Corporation,
10 Basil Sawyer Drive, Hampton, VA, USA

1. INTRODUCTION

One important attribute of an air quality forecast system is how the system predicts exceedance and non-exceedance events which are evaluated using categorical metrics. Current categorical metrics used in model evaluations (Kang et al., 2005) 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 direct observation-forecast pairs. However, the observations and the model forecast represent different spatial and temporal scales. For example, observations are point measurements at fixed locations, while model forecasts represent volume average concentrations. 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. To avoid the “clear-cut” effect, in this paper three new categorical metrics, Weighted Success Index (WSI), area Hit rate (aH), and area False Alarm Ratio (aFAR), are developed. In 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 – 1:1 line on scatter plot). The aH is defined as a Hit if an observed exceedance is matched within a fixed area (adjacent grid cells) surrounding the observation location (central cell). The concept of aH resembles the manner in which forecasts are usually issued. In reality, a warning would be issued for a region of interest (such as a metropolitan area) if an exceedance is forecast to occur anywhere within the region. Similarly, aFAR is defined to reflect false alarm ratios of the forecast system based on the same concept as in the aH definition. In this paper, these metrics are demonstrated using the Eta-CMAQ (Community Multiscale Air Quality) forecasts during the period of June and July 2005.

2. THE ETA-CMAQ FORECAST SYSTEM

The Eta-CMAQ Air Quality Forecast (AQF) system is based on the National Centers for Environmental Prediction’s (NCEP’s) Eta model (Black 1994; Rogers et al., 1996) and EPA’s CMAQ Modeling System (Byun and Ching 1999). A brief summary of the linkage between the Eta and the CMAQ models, relevant to this study, is presented below. Additional details can be found in Otte et al. (2005). 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 U.S. 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 concentrations are forecast over the eastern U.S. using a 12-km horizontal grid spacing on a Lambert Conformal map projection. There are 22 layers in the vertical domain, which are set on a sigma coordinate extending from the surface to 100 hPa. The chemical fields for CMAQ are initialized using the previous forecast cycle. The primary Eta-CMAQ model forecast for next-day surface-layer O_3 is based on the current day’s 12 UTC cycle, and the products are issued daily no later than 1330 LST. The target forecast period is local midnight through local midnight (04 UTC to

[*] *Corresponding author:* Daiwen Kang, US EPA, Mail Drop E243-03, NERL/AMD, RTP, NC 27711; e-mail: kang.daiwen@epa.gov

O_3 UTC for the eastern U.S.). Hourly, near real-time, O_3 (ppb) data obtained from EPA's AIRNow program are used in this study (http://www.epa.gov/airnow).

3. EXISTING CATEGORICAL METRICS

For the categorical forecast evaluation, the model's Accuracy (A), Bias (B), Hit Rate (H), False Alarm Ratio (FAR), and Critical Success Index (CSI) are typically examined for both the 1- and 8-hour O_3 standard. A graphical representation of the formulation of the categorical metrics (for the maximum 8-hr O_3) is presented in Figure 1, where **a** represents the number of forecast 8-hr exceedances (O_3 concentrations >= 85 ppb) that were not observed, **b** represents the number of correctly forecast 8-hr exceedances, **c** represents the number of correctly forecast 8-hr non-exceedances, and **d** represents the observed 8-hr exceedances that were not forecast.
Accuracy (A) measures the percentage of forecasts that correctly predict an exceedance or non-exceedance and is given by:

$$A = \left(\frac{b+c}{a+b+c+d} \right) \times 100\% \qquad (1)$$

In air quality forecast evaluation, A can be strongly influenced by the number of correctly forecast non-exceedances (c), which is invariably very large; hence care must be taken in interpretating.

The Bias (B) indicates, on average, if the forecasts are underpredicted (false negative) or overpredicted (false positives).

$$B = \left(\frac{a+b}{b+d} \right) \qquad (2)$$

A value of 1.0 indicates no bias (i.e., a perfect forecast), values < 1.0 indicate underprediction, and values >1.0 indicate overprediction.

The False Alarm Ratio (FAR) measures the percentage of times an exceedance was forecast and did not occur.

$$FAR = \left(\frac{a}{a+b} \right) \times 100\% \qquad (3)$$

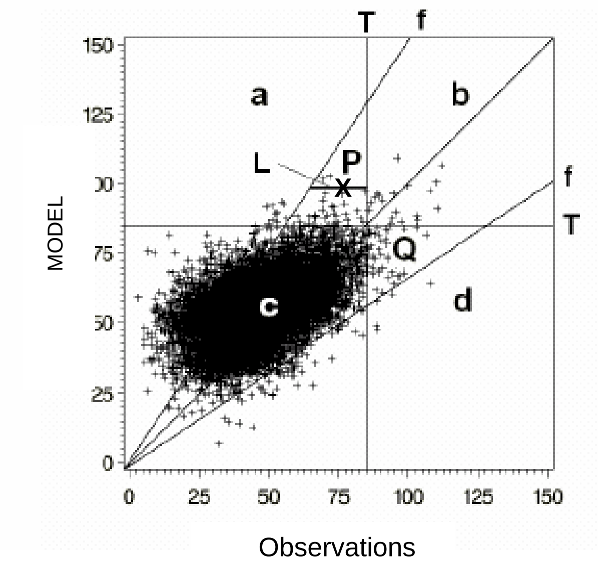

Fig. 1. Example scatter plot for the categorical evaluation and definition of WSI (see text)

Smaller values of FAR are desirable, with a FAR = 0 indicating no false alarms and a FAR of 50% indicating that half of the forecast exceedances were not observed.

The CSI indicates how well both forecast exceedances and actual exceedances were predicted.

$$CSI = \left(\frac{b}{a+b+d} \right) \times 100\% \qquad (4)$$

Unlike accuracy, the CSI is not influenced by correctly forecast non-exceedances. A CSI of 50% indicates that half of the forecast and observed exceedances were correct. Finally, the Hit Rate (H), which is similar to the CSI, indicates the percentage of actual exceedances that were forecast. It is also called Probability Of Detection (POD).

$$H = \left(\frac{b}{b+d} \right) \times 100\% \qquad (5)$$

4. NEW CATEGORICAL METRICS

The categorical statistics discussed in section 3 are defined by the numbers of paired data points found in the quadrants defined by threshold values (T) as shown in Figure 1. While informative, these metrics are not infallible in that they do not always represent the model's performance accurately. To illustrate some of the limitations of existing categorical statistics, consider x(M,O) (Figure 1) representing a paired data point (M is the modeled value, O the observed) that lies within area *a* (forecast exceedance that did not occur) but also lies within a factor line *f* inside a triangle designated as *P*. This individual forecast, though considered a “failure” or false alarm from a categorical standpoint, in actuality may be considered a “success”, if inherent uncertainties (1) in both the model and measured values as well as (2) those associated with representing the variability associated in comparing grid and point values, are factored into the analysis. The same is true for points falling into area *d*, but within the lower factor line (triangle *Q*).

4.1 Weighted Success Index (WSI)

A new metric is proposed, called the weighted success index (WSI), that gives some credit for points located in the triangles *P* and *Q*, while penalizing points in area *a* and *d* but outside the triangles. The value of the factors (*f*) used to determine successful model performance was set to 1.5 in this example.

If a data point x(O, M) is within Triangle *P*, the length of the line that passes through x and intercepts with both the threshold line and factor line (L) can be computed as:

$$L = T - \frac{1}{f}M \qquad (6)$$

L can then be used to define the Weighted Success index of Model forecast (WSM) :

$$WSM = 1 - \frac{T-O}{L} = 1 - \frac{T-O}{T-\dfrac{M}{f}} = \frac{M-fO}{M-fT} \qquad (7)$$

The values of WSM are between 0 and 1 for points within the factor lines. For points outside the factor lines, the values are negative and their magnitude is dependent on the factor value (but limited to -1 for symmetry and to prevent outliers from dominating the weighting).

Similarly for a point in the Triangle *Q*, that is observed exceedance but not forecast, an Weighted Success index of Observations (WSO) is:

$$WSO = \frac{O - fM}{O - fT} \qquad (8)$$

WSM and WSO are then used to calculate the WSI:

$$WSI = \frac{b + \sum WSM + \sum WSO}{a+b+d} \times 100\% \qquad (9)$$

Values of WSI range from -100% (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 two more items which credit the points within the quadrants *a* and *b* (Fig. 1). CSI can only be non negative numbers, but WSI can be any number between -100% and 100%. When the factor lines are set to 1, WSI reduces to CSI. For a perfect or no event (neither observed nor forecast exceedances exist) forecasts, WSI and CSI are the same.

4.2 Area Hit (aH)

The Hit Rate (H) 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 forecast will also reflect spatial and temporal errors in simulation of meteorological features (frontal system, precipitation, cloud cover, etc.), especially with increasingly finer model resolutions. On the other hand, air quality forecast are typically issued for relatively large areas such as a metropolitan area. Wherever an exceedance is forecast within the area, a warning of the exceedance will be issued for the whole area. As Figure 2 shows, some observed exceedances (red or orange diamonds) are only one or two grid cells away from the forecast exceedances (red or orange background).

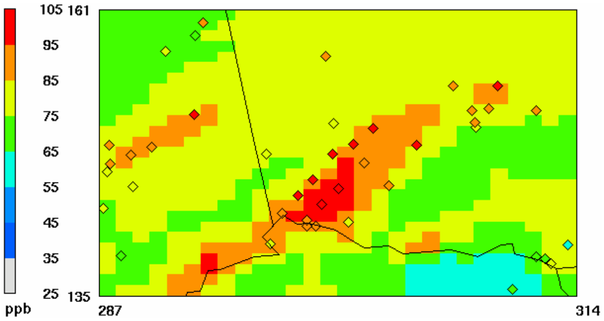

Fig. 2. Model predicted (background) and observed (diamonds) maximum 8-hr O_3 concentrations (ppb). The observations are overlaid over the model predictions.

From this practical consideration, a new metric, Area Hit (aH), is developed, which reflects both the spatial uncertainties of the model forecast and the application of the model forecast done by a local forecaster.

Area Hit (aH) is defined as:

$$aH = \left(\frac{Ab}{Ab + Ad} \right) \times 100\% \qquad (10)$$

where *aH* is the area hit, *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. *Ad* is the number of observed exceedances that are not forecast within the designated area centered at the monitor location. The area including the grid cell (center cell) in which the monitor resides and the adjacent cells are used. 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 area is chosen, the larger aH values will be. However, in order to effectively evaluate performance of a forecast system, the size of the area cannot be too large. For the Eta-CMAQ forecast system with a 12-km horizontal resolution, 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 36x36 km or 60 x 60 km (Fig. 3), respectively. However, if the observation site is located at the edge or the corner of the modeling domain, then only the adjacent cells which reside within the domain are counted.

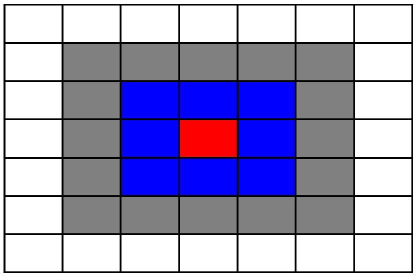

Fig. 3. Illustration of the “area” categorical metrics. The central grid cell where a monitor is located is marked red. Blue cells are the adjacent cells one cell away from the central cell, while grey cells are the adjacent cells two cells away from the central cell.

4.3 Area False Alarm Ratio (aFAR)

Likewise, area False Alarm Ratio (aFAR) can be defined using the spatial concept as aH. Mathematically, aFAR is defined as:

$$aFAR = \left(\frac{Aa}{Aa + Ab} \right) \times 100\% \qquad (11)$$

where *Aa* is the number of forecast area exceedances that were not observed and *Ab* is the number of forecast exceedances that were observed. For example, an exceedance is forecast within a 3x3 or 5x5 grid-cell area in which multiple observation sites reside; if any of the observation sites observed an exceedance, then no false alarm is counted even for the sites that didn’t observe exceedances (all sites contribute to *Ab*). In other words, for any forecast exceedance, all the observation sites within the selected area will be checked; if no exceedances were observed at any sites located within this area, false alarms are recorded (all the sites contribute to *Aa*). When the area only covers the center grid cell (1x1 grid-cell area), *aFAR* becomes *FAR*.

5. CASE STUDY

The new categorical metrics (WSI, aH, and aFAR) are compared to their counterparts (CSI, H, and FAR) using the Eta-CMAQ real-time O_3 forecast for the period from June 13 to July 31, 2005. The forecast domain covers eastern US. More than 850 AIRNOW monitoring sites are located within this domain. During this forecast period, 1083 exceedances (maximum 8-hr O_3 concentrations >= 85 ppb) were observed which resulted in a CSI of 19.2% and WSI of 54.5% when the factor is set at 1.5. This marked increase in skill with WSI (compared to CSI) indicates that there are many data points slightly outside the desirable *b* quadrant **(**see Figure 1) that the strict CSI metric categorizes as failures. When the proximity of these data points 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. 4, about 40% of the exceedances during this period are forecasted as direct Hit (H) (i.e., using an area of 1x1 grid cell). When the Hit Rate is calculated based on the 3x3 grid cells, the area Hit Rate (aH) increases to about 70%. About 30% observed exceedances are forecast in the adjacent grid cell to the cell where the monitor is located. When the area is expanded from 3x3 grid cells to 5x5 grid cells, the aH is increased by other 10%. This result indicates that the majority of exceedances are captured by the forecast system within the 3x3 grid cell area. The values of aFAR, are smallest when it is calculated over 3x3 grid cell area, while it is the largest when calculated over the 5x5 grid cell area.

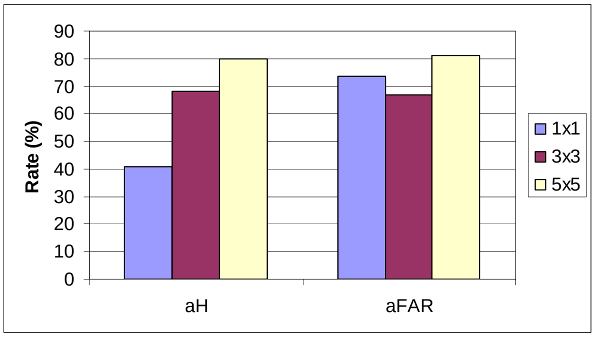

Fig. 4. Hit Rate (aH) and the area False Alarm Ratio (aFAR) calculated by direct match (1x1), 9 grid cell area (3x3), and 25 grid cell area (5x5) during the period from 6/13-7/31.

In addition to evaluating the air quality forecasts performance over the entire model domain, the forecast system is examined over urban and suburban areas where most of the human activities take place. Of the 1083 exceedances observed during the research period, 637 exceedances were observed in urban or suburban sites. Both the CSI (18.0%) and WSI (53.4%) in urban and suburban regions are slightly lower than those (19.2% and 54.5%) calculated over all the sites in the entire modeling domain. This indicates that the success rate to detect exceedance events is slightly lower over urban and suburban regions than that over rural locations. Similar to Figure 4, Figure 5 shows the aH and aFAR values for the urban and suburban sites calculated over different area size. As seen in Figure 5, there are no significant differences between the aH values in urban and suburban areas and those with all the measurement sites (Fig. 4) 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 5x5 grid cells, while the aFAR values over the 1x1 and 3x3 grid cells are about 2 to 3% larger in urban and suburban areas than those over the entire domain. In general, the Eta-CMAQ forecast system does not show a significant bias towards urban or non-urban areas in this forecast domain.

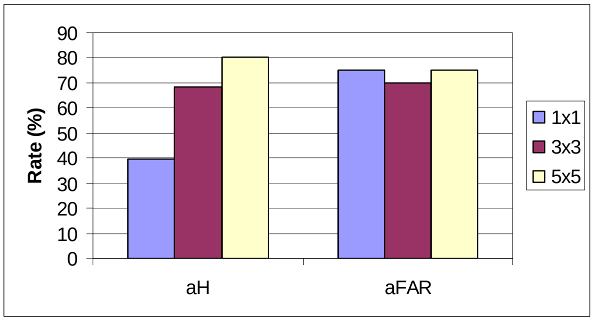

Fig. 5. Area Hit Rate (aH) and the area False Alarm Ratio (aFAR) for only the urban and suburban regions calculated by direct match (1x1), 9 grid cell area (3x3), and 25 grid cell area (5x5) during the period from 6/13-7/31.

As discussed earlier, a practical forecast is usually issued for a functioning region, e.g. a city, a metropolitan area, or an industrial region. The practical aH and aFAR values are expected to be more promising than these traditional categorical evaluation metrics. The Eta-CMAQ forecast system has demonstrated strong capability in forecasting the exceedance events, though the aFAR are still high.

6. SUMMARY

Three categorical metrics, WSI, aH, and aFAR, are developed to evaluate model performance in forecasting exceedance events. These metrics are in addition to their existing categorical metrics and provide a more relaxed but practical way to evaluate model performance compared with the existing counterparts. The new metrics not only evaluate the realization of the exceedances and non-exceedences as their existing counterparts do for the “clear-cut” match, but also evaluate the “effort” or “potential” of the forecast system to try to reach the realization.

The case study demonstrates that the Eta-CMAQ forecast system has very promising potential to forecast O_3 exceedance events during the study period and no significant difference is observed in the forecast of these exceedance events between urban and suburban and rural areas.

The way of aH and aFAR defined in this paper not only provides a guidance towards forecast model evaluation, but it also reveals very useful spatial performance information of forecast systems when the evaluating area expands from the 1x1 grid cell to 5x5 grid cells. The case study shows that ~40% hit occurred in the direct match (1x1 grid cell), about another 30% hit took place in the immediate adjacent cells (3x3 grid cells), and only ~10% hit would gain by expanding to 5x5 grid cells. In practice, if the information about the coverage of local forecasts is available, aH and aFAR can be calculated from the actual area in which local forecasts are covered (such as a metropolitan area).

7. ACKNOWLEDGEMENTS

The authors are grateful to Tanya Otte and Jonathan Pleim for their critical reviews on the paper. 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 number 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.

8. REFERENCE

Black, T. 1994: The new NMC meso-scale 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 Multi-scale Air Quality (CMAQ) modeling system, EPA/600/R-99/030, Office of Research and Development, U.S. Environmental Protection Agency.

Chang, J., et al., 1990: The regional acid deposition model and engineering model, in *National Acid Precipitation Assessment Program, Acidic Deposition, State of Science and Technology*, vol. 1, NAPAP SOS/T Rep. 4, Natl. Acid Precip. Assess. Program, Washington, D.C.

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, K., 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. In press, *JAWMA*.

Otte, T.L., G. Pouliot, J.E. Pleim, J.O. Young, K.L. Schere, D.C. Wong, P.C.S. Lee, M. Tsidulko, J.T. McQueen, P. Davison, R. Mathur, H-Y. Chuang, G. DiMego, and N.L. Seaman, 2005: Linking the Eta model with the Community Multiscale Air Quality (CMAQ) modeling system to build a national air quality forecasting system, *Wea. Forecasting*, **6**, 367-384.