VERIFICATION OF THE NOAA-EPA AIR QUALITY FORECASTS FOR SUMMER 2005

Jerry Gorline*, Michael Schenk, Wilson Shaffer, and Arthur Taylor
Meteorological Development Laboratory, NWS, NOAA, Silver Spring, MD, USA

1. INTRODUCTION

In the summer of 2005, the National Oceanic and Atmospheric Administration (NOAA) in cooperation with the U.S. Environmental Protection Agency (EPA), tested the Air Quality Forecasting (AQF) capability. The AQF system links the National Centers for Environmental Prediction's (NCEP) North American weather Model (NAM) with EPA's Community Multiscale Air Quality (CMAQ) modeling system to produce gridded ground-level ozone forecast guidance. (Otte et al. 2005)

We compared the performance of two models with different configurations, namely, the developmental (5x) on a conterminous U.S. (CONUS) domain and the experimental (3x), which covered a smaller domain in the eastern U.S. The 5x developmental model was subject to change and did not run on every day of the test season. The 3x experimental model was more stable during the test period. These configurations differed also in ozone boundary conditions and in CMAQ approximations for convective mixing. Further information regarding differences in experimental and developmental test configurations is provided in McQueen et al. (2005).

We verified predicted surface ozone concentrations against observations compiled by the EPA for the different domains. We also examined the performance of the 5x model taken over the 3x domain to aid comparison of the two test configurations. Our verification metrics included categorical analyses for Fraction Correct (H), Threat Score (TS), Probability of Detection (POD), and the False Alarm Rate (FAR). We also calculated weekly, monthly, and seasonal Mean Absolute Error (MAE) and bias, where bias is forecast minus observation. Graphic products included daily spatial maps and weekly/monthly statistics displayed in the form of bar charts, scatterplots, and graphs. Specifically, we compared categorical performance of next-day maximum 8-h average ozone predictions for June – September, 2005, based on daily tests driven by the 1200 UTC NAM cycle. We compared MAE and bias of 8-h predictions for a two-week period of elevated ozone, August 1 – 15, 2005, to examine regional differences in performance.

This paper also includes a case study of 5x verification results for July 12, 2005, when Tropical Storm Dennis generated a long narrow strip of elevated surface ozone levels across the Mid-Atlantic States. This case exhibits relatively good spatial verification but also shows evidence of a reduction of surface ozone levels compared to predictions, after thunderstorm activity in the Pittsburgh, Pennsylvania area.

2. AIR QUALITY FORECAST VERIFICATION

The NWS Meteorological Development Laboratory (MDL) produced verification scores for 5x developmental tests to provide feedback for possible model configuration changes. MDL also produced verification for the 3x experimental tests to assist in the validation of 3x verification provided by NCEP. A graphic of the 2005 3x grid is given in Figure 1.

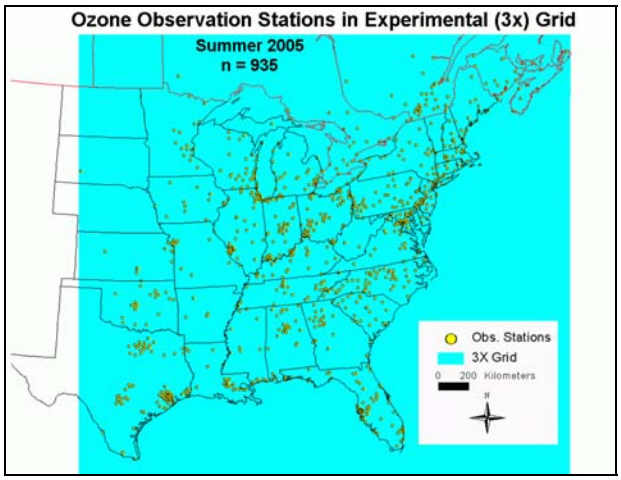

Fig. 1. 3x grid over the eastern U.S., 935 stations, 2005.

During 2005, MDL generated categorical verification metrics for a subset of the 5x developmental domain (5x3). This 5x3 subset was equivalent to the 3x domain which allowed comparisons in performance to the 3x model predictions. We concentrated our analyses on the 8 hour (8-h) averaged forecasts. The 1-h averaged forecasts contained very few observed threshold exceedances because the threshold was higher: 125 parts per billion (ppb)

*Corresponding author: Jerry L. Gorline, Meteorological Development Laboratory, NWS, 1325 East West Highway, Silver Spring, MD. 20910; fax: 301-713-9316; phone: 301-713-1768; e-mail: jerry.gorline@noaa.gov

for the 1-h threshold vs. 85 ppb for the 8-h threshold.

All 8-h predictions or observations that were equal to or greater than the threshold during a predefined 24 hour period were counted as exceedances. The 24 hour window for counting exceedances was midnight to midnight, beginning on day two of the 1200 UTC CMAQ forecast period. The 8-h exceedance window began on projection 24 and ended on projection 47. The 1200 UTC CMAQ predictions were backward averaged and the time stamp was the end time for the forecast period. For example, an 8-h prediction for 1200 UTC was valid starting 0400 UTC and ending 1200 UTC.

We populated two-by-two contingency tables as follows,

	Observed
Forecast	| a | b |
	| c | d |

where, a = forecast, observed (yes/yes)
b = forecast, no observed (yes/no)
c = no forecast, observed (no/yes)
d = no forecast, no observed (no/no).

The corresponding scores were computed as follows:

$$H = (a + d)/(a + b + c + d) \quad (1)$$
$$TS = a /(a + b + c) \quad (2)$$
$$POD = a/(a + c) \quad (3)$$
$$FAR = b/(a + b) \quad (4).$$

H is the fraction of correctly predicted cases: both above (a) and below (d) the exceedance threshold. The air quality statistics for this performance measure were dominated by the number of correctly predicted non-exceedances (d). TS is the number of correct predicted exceedances (a), divided by all predicted or observed exceedances (a + b + c). POD is the fraction of observed exceedance conditions that are correctly predicted and FAR is the fraction of exceedance predictions that are incorrect. High skill is a function of both low FAR and a high POD. A large number of false alarms are an indication of over-predicting exceedance events. For a more detailed discussion about two-by-two contingency table analyses, see Wilks (1995).

3. COMPARISON OF THE 5X MODEL ON THE 3X DOMAIN TO THE 3X MODEL

We compared the performance of the 5x3 developmental tests to the 3x experimental tests. EPA provided ozone observations for 1,290 sites within the CONUS domain; 935 sites are in the eastern U.S. (3x) domain. The statistics were calculated by month and by season for June through September, 2005. Tables 1 a-d show 8-h monthly contingency table results for the 5x model (5x), 5x model on the 3x grid (5x3), and the 3x model (3x). The contingency table results for the entire season are shown in Table 1e. For the 8-h results, the sample sizes for a, b, and c, were sufficient for reliable statistical analyses. If an observation or interpolated model data value for a station was missing, we excluded that station from our calculations. We used the same station list for the 5x3 and 3x comparisons but because of differences in missing predictions from occasional interruptions to testing cycles for the two domains, the total number of verification cases for each data set was not identical.

Table 1a. Contingency results for June, 2005.

200506	5x, 8-h	5x3, 8-h	3x, 8-h
a	171	114	59
b	978	271	371
c	391	241	162
d	25358	12887	11255
H	0.949	0.962	0.955
TS	0.111	0.182	0.100
POD	0.304	0.321	0.267
FAR	0.851	0.704	0.863

Table 1b. Contingency results for July, 2005.

200507	5x, 8-h	5x3, 8-h	3x, 8-h
a	252	136	124
b	1273	684	716
c	707	255	237
d	32591	24963	24113
H	0.943	0.964	0.962
TS	0.113	0.127	0.115
POD	0.263	0.348	0.343
FAR	0.835	0.834	0.852

Table 1c. Contingency results for August, 2005.

200508	5x, 8-h	5x3, 8-h	3x, 8-h
a	230	204	197
b	872	689	818
c	474	230	238
d	31006	23118	24668
H	0.959	0.962	0.959
TS	0.146	0.182	0.157
POD	0.327	0.470	0.453
FAR	0.791	0.772	0.806

Table 1d. Contingency results for Sept., 2005.

200509	5x, 8-h	5x3, 8-h	3x, 8-h
a	84	60	53
b	382	332	318
c	244	93	81
d	29507	15267	11999
H	0.979	0.973	0.968
TS	0.118	0.124	0.117
POD	0.256	0.392	0.396
FAR	0.820	0.847	0.857

Table 1e. Contingency results for 2005 season.

2005	5x, 8-h	5x3, 8-h	3x, 8-h
a	737	573	433
b	3505	2189	2223
c	1816	924	718
d	118462	89319	72035
H	0.957	0.967	0.961
TS	0.122	0.155	0.128
POD	0.289	0.383	0.376
FAR	0.826	0.793	0.837

The June results are given in Table 1a. For June, the 8-h POD for the 3x model was 17% lower than for the 5x3 POD and the FAR for the 3x was 21% higher than for the 5x3 FAR. There were data gaps which may have affected the June results. Specifically, 3x predictions were not available for June 1 – 9 and June 24 – 28, 2005, while 5x predictions were not available for June 25 – 28, 2005. Although the 5x3 and 3x samples are not identical, the 8-h POD and FAR are very similar for July – September and for the 2005 season.

Comparing 5x on the CONUS grid to 5x3 model runs, for July - September, the 8-h POD for the 5x tests were about 24% - 34% lower than for 5x3 POD. We investigated whether the lower 5x POD values were the result of greater under-predicting in California than for the rest of the CONUS domain.

We compared 5x contingency table results to results for California stations only. We examined performance for August 1 – 15, 2005, an extended period of elevated ozone. For maximum 8-h predictions, 5x POD was 0.402, but for California, the POD was only 0.088, or 78% lower than the POD for the entire CONUS. For California, there were only 14 correctly predicted threshold exceedances, with 145 observed exceedances that were not predicted. For the entire CONUS grid, there were 214 correctly predicted threshold exceedances and 318 observed exceedances that were not predicted. The much lower POD for California was a function of a much higher number of misses from under-prediction, compared to the rest of the CONUS grid.

4. GRAPHICAL 5X VS. 3X COMPARISONS

We compared performance of developmental 5x3 testing vs. experimental 3x testing using H, POD, and FAR for the summer of 2005, specifically, June 15 – August 13. The samples are not identical but we used the same 3x station list for the comparisons. Figure 2a shows the number of hits for the 8-h maximum daily exceedance predictions where there were ten or more observed exceedances. The black diamonds are the number of observed exceedances, the red circles are the 5x3 hits and the blue triangles are the 3x hits. Figure 3b shows the POD and Figure 2c shows the FAR for the 5x3 and 3x runs. Again, the red circles are the 5x3 results and the blue triangles are the 3x results.

Figure 2 a-c shows that the 5x3 results are generally similar to the 3x results. The one exception occurred for predictions valid June 30. This day contained a small sample of observed threshold exceedances. The 3x predictions provided six correct threshold exceedance hits while the 5x3 provided only one hit. Due to an interruption in testing during June 24 – 28, 2005, both tests for June 29, providing forecasts valid for June 30, 2005, were necessarily cold starts. It takes several days for the emissions budget to reach steady-state after a cold start, so the forecast statistics are much different than if the model had run the previous day. Ignoring the June 30, 2005 results, we find that the performance of the 5x3 model was similar to the 3x model, especially after July 8, 2005.

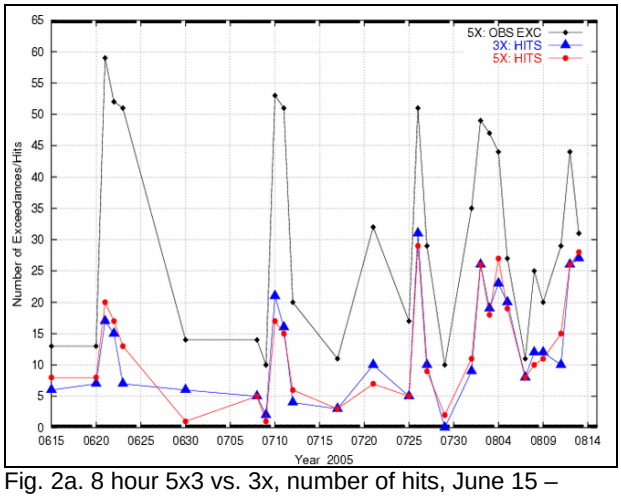

Fig. 2a. 8 hour 5x3 vs. 3x, number of hits, June 15 – August 13, 2005, 935 stations.

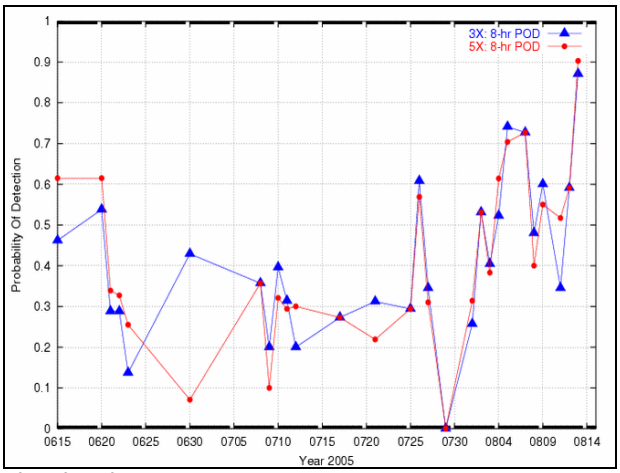

Fig. 2b. 8 hour 5x3 vs. 3x, POD, June 15 – August 13, 2005, 935 stations.

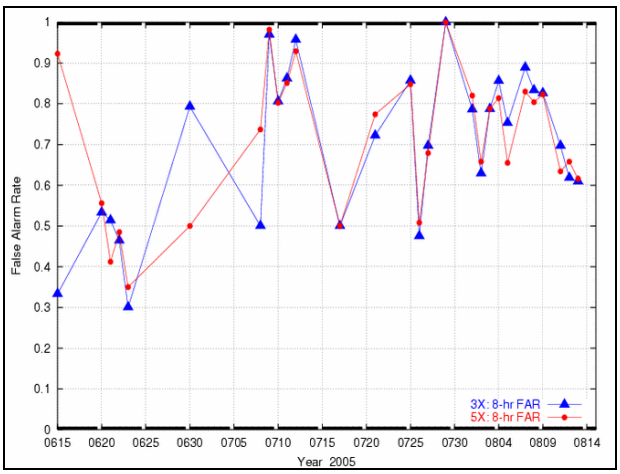

Fig. 2c. 8 hour 5x3 vs. 3x, FAR, June 15 – August 13, 2005, 935 stations.

Figure 3a compares the mean bias in predicted 8-h average concentrations for the different test configurations. The mean observed peak in 8-h ozone occurs at projection 12 (0000 UTC), day one, and projection 36 (0000 UTC), day two. Similarly, Figure 3b compares the MAE. Figures 3a and 3b show that the 3x model demonstrated a consistently higher bias and MAE than the 5x3 model.

Figure 4 shows a spatial map of predicted ranges of maximum values of 8-h ozone (shades of blue/white) vs. observed ranges, shown as green, yellow, and red points. Predicted maximum values that exceed the 85 ppb threshold are shown in dark blue and the corresponding observations are shown as red points. In the East, predicted and observed exceedances correspond well for August 13, 2005, but there were more unpredicted observed exceedances in California. In many cases, areas of predicted 8-h exceedances were near stations which reported observed exceedances, but the interpolated predictions overlaying these stations did not exceed the threshold. The spatial maps aid in identifying these near misses. Such near misses are important for diagnostic evaluations of the testing, and can help indicate elevated ozone on a regional county-wide basis. Early in the summer, the spatial maps also indicated large areas of elevated ozone in the high

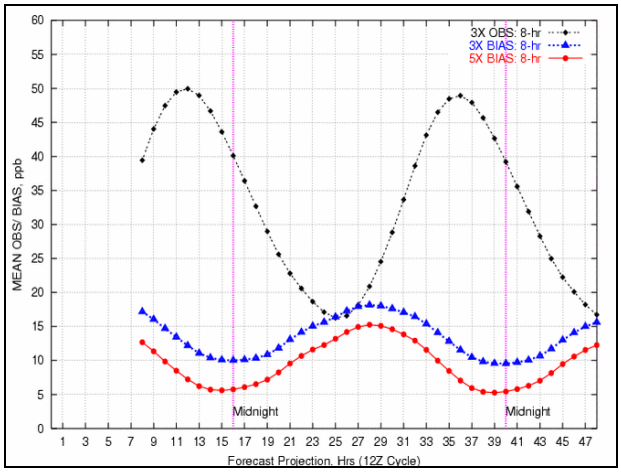

Fig. 3a. Bias, 8 hour 5x3 vs. 3x, August 1 - 15, 2005, 935 stations, 3x mean observations are in black.

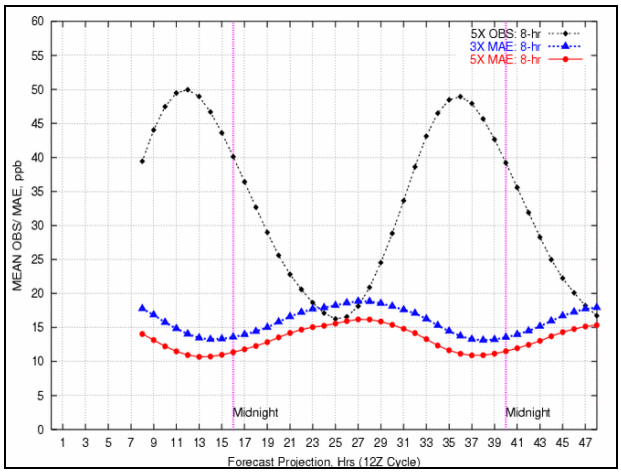

Fig. 3b. MAE, 8 hour 5x3 vs. 3x, August 1 - 15, 2005, 935 stations, 5x mean observations are in black.

terrain in the West, away from most observation stations. Although these values did not verify well at the few mountain stations, the contingency statistics were not much affected because of the dominance of stations located in the Eastern U.S. Based on evidence from spatial maps such as Figures 4 and 5, corrections to the developmental testing were made to address this issue. (McQueen et al. 2005) This information was critical in helping to diagnose performance issues for elevated terrain.

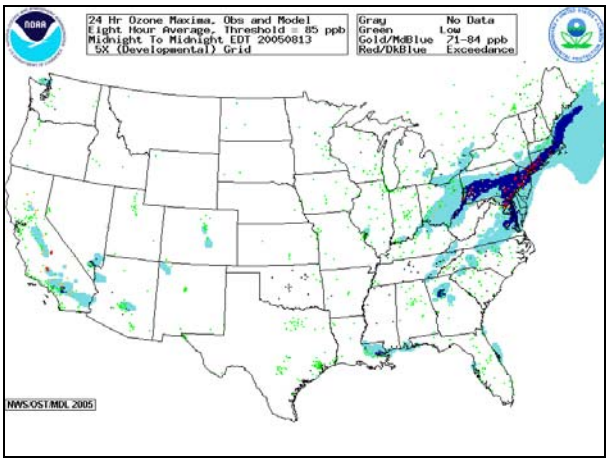

Fig. 4. Spatial map of 8-h ozone predictions with observations, 5x on CONUS grid, August 13, 2005.

5. CASE STUDY, JULY 12, 2005, SURFACE OZONE REDUCTION ASSOCIATED WITH THUNDERSTORMS

Tropical Storm Dennis provided an interesting case study for testing predictions of strong gradients in surface ozone, and of the local impacts of thunderstorms on ozone levels during days with elevated ozone concentrations. Figure 5 shows the 5x model 8-h daily maximum ozone forecast valid July 12, 2005. Tropical Storm Dennis had moved inland from the Gulf of Mexico and was weakening over the Ohio Valley. The 5x predictions show a long narrow band of exceedances stretching from southeastern Virginia, through southwest Pennsylvania, northern Ohio, southern Michigan, and finally wrapping around the northern and western border of Illinois. This band of predicted exceedances coincided with a feeder band associated with Dennis which had tapped into some Atlantic moisture. The

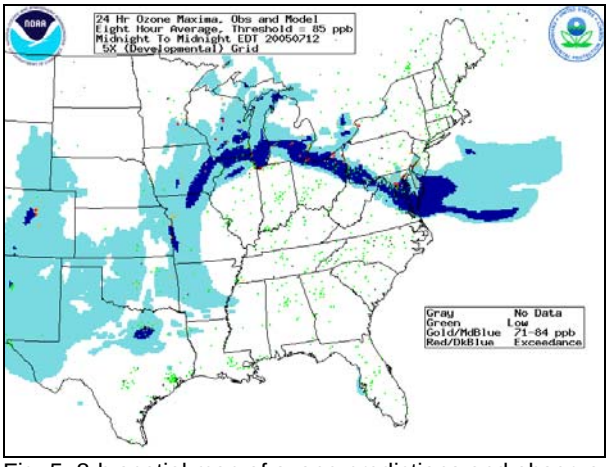

Fig. 5. 8-h spatial map of ozone predictions and observations, 5x model, July 12, 2005.

observations generally correlated well with the predicted band, with values above 85 ppb observed in a narrow band through Maryland, Delaware, northern Ohio and southern Michigan, with the exception of Pittsburgh, Pennsylvania.

We examined the observations recorded at four stations located in the narrow band of predicted exceedances for July 12, 2005. Table 2 lists observed 8-h average values for Pittsburgh, Pennsylvania, Lewes, Delaware, Eastlake, Ohio, and Lansing, Michigan. Observations in Table 2 labeled "day 1" were observations from July 11; "day 2" represents July 12.

The predicted 8-h daily maximum values were verified as hits, exceeding the 85 ppb threshold for both days over Delaware, Ohio, and Michigan, and for day 1 at the Pittsburgh station. For Day 2, the predicted value for the Pittsburgh station again exceeded the threshold, but observed values remained well below the threshold of 85 ppb. Reduced surface ozone levels in Pittsburgh, Pennsylvania, for July 12, compared to the previous day were associated with thunderstorms in the area during the midnight-to-midnight verification period. Pittsburgh, Pennsylvania, reported a thunderstorm in the area at 2300 UTC; see also the surface observations for the Mid-Atlantic States for July 12, 2005, 2300 UTC in Figure 6. The 5x prediction for Pittsburgh, Pennsylvania, would not have included the thunderstorms. Given the elevated ozone recorded in nearby areas without thunderstorms, the Pittsburgh observations are likely to have more closely matched the predicted values, had thundershowers not occurred in the area during the verification period. The Texas Air Quality Study II similarly reported reduction of high surface ozone concentrations when thunderstorms developed in the monitored area. (Texas Air Quality Study II, 2005)

Table 2. 8-h observations for four stations in band of predicted exceedances, July 12, 2005.

Time	2200	2300	2400	0100	0200	0300	0400
PA, day1	83	94	98	98	93	83	69
PA, day2	54	58	58	55	47	40	33
DE, day1	81	86	89	91	92	91	89
DE, day2	102	107	110	111	111	109	105
OH, day1	92	97	98	95	90	84	76
OH, day2	93	95	95	93	88	82	74
MI, day1	81	87	91	94	95	93	90
MI, day2	78	83	85	85	84	82	78

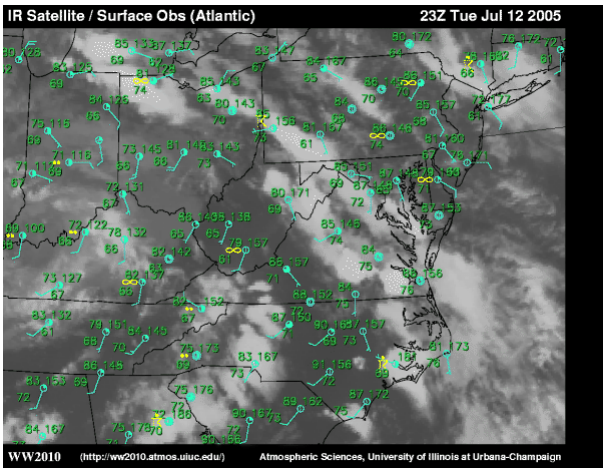

Fig. 6. Surface observations for Tuesday, July 12, 2005, 2300 UTC.

6. CONCLUSIONS

The 5x developmental testing, when verified over the experimental 3x domain, showed a lower bias and MAE in the hourly statistics compared to the 3x experimental tests, while categorical verification was similar. The lower MAE in the 5x model on the 3x grid would suggest better performance compared to the 3x model. Even though the MAE was consistently lower in the 5x statistics, this is dominated in statistical averaging by relatively low ozone concentrations that are beneath the thresholds for exceedance used in categorical verification. Two-by-two contingency table results were similar to the 3x results. The lower bias in the 5x3 statistics is also a result of more under-prediction. The increased under-forecasting is reflected in decreased POD for the 5x, which mitigates advantages from the lower MAE.

We examined 5x statistics for California only during August 1 – 15, 2005, and found that the mean bias in the 8-h forecasts dropped to near zero ppb between the hours of 0400 - 0600 UTC, a time when many 8-h exceedances were being reported by stations in California. By comparison, the mean bias in the 8-h forecasts for the 5x3 was around five ppb. The lower bias in California during a critical time in the verification period supports the lower POD for the 5x runs on the CONUS grid, compared to the 5x runs on the 3x grid.

Finally, simulations of strong ozone gradients that developed on the leading edge of the system associated with Tropical Storm Dennis were verified, except where localized (not-predicted) thunderstorm activity was observed. Such cases reveal a wealth of diagnostic information when verified/analyzed within a regional context in addition to domain-wide statistics.

7. ACKNOWLEDGMENTS

This paper was prepared in cooperation with development activities carried out under the auspices of NWS's National Air Quality Forecast Capability. We would like to thank NCEP for providing the surface ozone concentration forecasts and the EPA for providing the associated verifying observations.

8. REFERENCES

McQueen, J.T., P. C. Lee, M. Tsidulko, G. DiMego, T. Otte, J. Pleim, G. Pouliout, J. Young, D. Kang, P. M. Davidson, and N. Seaman; 2005: Update to and Recent Performance of the NAM-CMAQ Air Quality Forecast Model at NCEP operations, *17th Conference on Numerical Weather Prediction. Amer. Meteor. Soc.,Washington, D.C.,* 12A.2.

Otte, T. L., G. E. Pouliot, J. E. Pleim, J. O. Young, K. L. Schere, D. C. Wong, P. C. S. Lee, M. Tsidulko, J. T. McQueen, P. Davidson, R. Mathur, R. H. 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*: **20**, 367–384.

TCEQ, cited 2005: Texas Air Quality Study II. [Available online at http://www.tceq.state.tx.us /implementation/air/airmod/texaqs-files /TexAQS_II.html].

Wilks, D. S., 1995: *Statistical Methods in the Atmosphic Sciences: An Introduction.* Academic Press, 238 – 241.