J2.12

Real-Time Prediction of US Northeast Corridor Ozone: Results for 2001-2002 as a Benchmark for Future Forecast Systems

John N. McHenry[1],
Stuart McKeen[2], William F. Ryan[3], Nelson Seaman[4], Janusz Pudykiewicz[5],
Georg Grell[6], Ariel Stein[7], Carlie Coats[1], Jeff Vukovich[8]

[1]Baron Advanced Meteorological Systems, Research Triangle Park, North Carolina
[2]NOAA Aeronomy Laboratory, Boulder, CO
[3]The Pennsylvania State University, State College, PA
[4]NOAA/NWS Office of Science and Technology, Suitland, MD
[5]Meteorological Service Of Canada, Dorval, Quebec, Canada
[6]NOAA Forecast Systems Laboratory, Boulder, CO
[7]NOAA Air Resources Laboratory, Silver Spring, MD
[8]University of North Carolina, Chapel Hill, NC

1. INTRODUCTION

Recent advances in chemistry and meteorology models and computational efficiency have allowed the operational use of coupled chemistry-transport models (CTMs) by local air quality forecasters. The adoption of numerical model guidance by operational forecasters, the number of whom has been increasing rapidly in the past several years, depends on the reliability of numerical model guidance in critical high ozone (O_3) cases. In turn, routine use of these models by forecasters can result in a positive feedback of information to model developers to further improve model performance.

During summer 2001 and 2002, the National Oceanic and Atmospheric Administration (NOAA) conducted a pilot program designed to test existing numerical air quality prediction (NAQP) systems and their components. This pilot program was initiated in preparation for testing and deployment of an operational NAQP system at the National Centers for Environmental Prediction (NCEP) in Suitland, MD. NCEP is the NOAA office that conducts day-to-day operational numerical forecasting for the National Weather Service (NWS). The goal of the pilot was to provide information crucial to the development of an NWS forecast system capable of meeting the demands of operational forecasters.

Though not among those tested, the model chosen by NOAA/NCEP for implementation, CMAQ (Byun et al., 1999), is closely related to one of the models used in the pilot programs, the MAQSIP-RT (McHenry, et al. , 2003) model. In the pilot programs MAQSIP-RT utilized other components of the EPA Models-3 system currently administered by the Community Modeling and Analysis Systems (CMAS) Center at UNC Chapel Hill, including MM5 (Grell et al., 1994) and SMOKE (Coats, 1996; Houyoux, 2000). However, MM5 is not planned for use in the NCEP system; rather, the NCEP Eta-model (Janjic, 1994) will be used as a meteorological driver. Further, SMOKE component models have been integrated or parameterized into the NCEP system to allow optimal use of scarce computational resources at NCEP.

This abstract reports on the results of the pilot programs for MAQSIP-RT as compared to the other pilot models/components tested. Because of its good performance, the MAQSIP-RT system establishes performance benchmarks that can be used to quantify expectations for the Eta-CMAQ system currently in early testing at NCEP (Davidson, et al., 2003).

2. FORECAST MODELS TESTED

The first phase of the pilot was begun in 2001, with a second phase taking place in 2002. In the first phase, MAQSIP-RT model forecasts were produced over New England and the Northern Mid-Atlantic, running in real-time at 45km, 15km and 5km. These forecasts were run in order to demonstrate capability and assess baseline performance. Other models available within NOAA were exercised offline during Phase 1. During Phase 2, these models were also run in real-time and include the NOAA HYSPLIT-CheM model (Stein et al., 2000) and the MM5-Chem model (Grell et al., 2000).

3. MODEL INTERCOMPARISON RESULTS: PHASE 1

The Phase 1 MAQSIP-RT forecast results were evaluated using standard EPA episodic regulatory

* *Corresponding author address*: John N. McHenry, Chief Scientist, Baron Advanced Meteorological Systems, Email: john.mchenry@baronams.com. Phone: (919) 248-9237; Fax: (919) 248-9245

model metrics and compared against several operational forecast methods already available at that same time. These included a statistical model in routine use in Philadelphia (Ryan et al., 2000), the Canadian CHRONOS model (Pudykiewicz, et al., 1997), a persistence forecast of peak 8-hour-average ozone in New England, and the official forecasts issued by state forecasters in the New England states (McHenry et al., 2003).

To conduct the comparison, a typical high-ozone episode was chosen which characterizes challenging forecast situations in New England and the northern mid-Atlantic. The episode occurred August 1-10, 2001 and was initiated with the establishment of a surface high and associated upper-level ridge centered over central MD. Early on August 2, an area of low pressure developed southeast of Cape Hatteras (HAT). Onshore flow was enhanced as the center of high pressure moved offshore, providing a cooler, cleaner maritime air mass to the southern Mid-Atlantic while in New England, winds re-circulated as high pressure passed to the south. Between Aug. 3-5, a frontal boundary aligned zonally and became quasi-stationary along a line from Portland, ME, to Pittsburgh, PA. Between Aug. 6-10, this boundary washed out as high-pressure built back across the whole region. A brief respite occurred in northern New England on Aug. 8, with the arrival and quick departure of a back-door cold front. On Aug. 9, the upper level ridge oscillated back eastward. Boundary layer winds backed to the west-southwest, the band of highest O_3 became oriented directly along the I-95 Corridor, and peak concentrations rose. A vigorous cold front approached the region on August 10, bringing the episode to a close.

MAQSIP-RT is able, in forecast mode, to meet several key performance criteria for regulatory models that are exercised with analyzed, rather than forecast, meteorological fields (Table 1). Gross error, in percent, in the 15-27% range throughout the episode, meets the EPA performance criteria of 35%. Model bias shows a good deal of day-to-day variation but overall is –9.7% when normalized, which is within the EPA performance criteria of ± 5-15%. Mean absolute error is in the 11-21 ppbv range, and rmse, which gives a rough estimate of forecast consistency, is in the 16-26 ppbv range.

Evaluation in PHL showed that for the episode, the mean absolute error (MAE) for MAQSIP-RT was 12.1 ppbv and compared well to the statistical model (11.5-12.9 ppbv). The PHL-expert-modified public forecast was the best forecast with an MAE of 8.0 ppbv. MAQSIP-RT out-performed the raw statistical forecasts by consideration of the median absolute error (7.3 ppbv compared to 9.6-12.0 ppbv). As the difference in mean and median error suggests, day-to-day skill of both statistical and numerical models varied. The regression model was not able to resolve the northward extent of the advection of maritime air on August 2 and over-predicted in the range of 27-34 ppbv. On August 4, the regression model carried a better forecast of cloud cover and a subsequent reduction in temperature and so provided better forecasts than MAQSIP-RT although still retained an over-prediction of 10-20 ppbv. Regression model skill was better than MAQSIP-RT on August 7-8 as MAQSIP-RT under-predicted across the PHL region, but the regression model was less skillful on August 9-10 with a tendency to under-predict. Overall, the skill of MAQSIP-RT was as good or better than the raw statistical guidance. The consistent performance of MAQSIP-RT by this measure was unexpected as domain-wide peak 1-hour-average O_3 is a difficult measure by which to evaluate numerical forecast models.

TABLE 1. MAQSIP-RT day-to-day model performance measures for August 1-10, for 1-hour-average concentrations across the 15km NE domain. All measures are based on a 60 ppbv threshold.

Date	Bias (ppbv)	Mean absolute error (ppbv)	Gross error (%)	rms error (ppbv)
August 1	-12.6	15.7	20.5	20.8
August 2	-9.0	12.3	15.9	16.2
August 3	0.5	11.0	15.5	14.0
August 4	12.5	17.9	26.8	20.8
August 6	-3.5	16.0	21.9	20.6
August 7	-14.4	17.4	20.2	23.1
August 8	-18.0	21.2	27.3	25.9
August 9	-7.7	20.6	19.5	20.6
August 10	2.6	15.9	17.2	15.9

When evaluated against monitor-specific forecasts in the NE U.S., MAQSIP-RT improved on expert forecasts, persistence, and the CHRONOS numerical model by a variety of traditional *discrete* (bias, MAE, rmse, IA) measures, taken over the whole set of monitors. MAE results are shown in Figure 1.

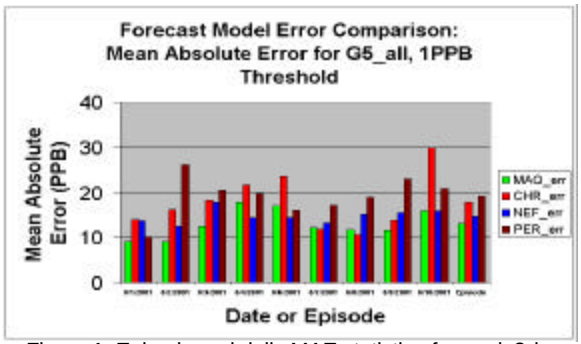

Figure 1. Episode and daily MAE statistics for peak 8-hour-average O_3 at all forecasted monitors MAQSIP-RT (MAQ) is given in green, NE forecasts (NEF) in blue, CHRONOS (CHR) in red and Persistence (PER) in brown.

Table 2. Forecast performance measures for selected sub-regions (all measures in ppbv).

Sub-region	MAQSIP	CHRONOS	NE Forecasts	Persistence
Coastal	Monitors = 20		N= 177	Mean Peak O_3 = 72.8
Bias	+1.9	+9.3	+5.8	-1.6
MAE	13.8	18.7	16.4	24.4
RMSE	17.2	23.4	20.3	29.1
Index of Agreement	0.77	0.66	0.71	0.45
Western-Rural	Monitors = 32		N=277	Mean Peak O_3 = 57.9
Bias	+4.4	+13.2	+11.6	-0.7
MAE	11.3	16.1	14.3	14.8
RMSE	14.9	20.0	17.2	17.8
Index of Agreement	0.74	0.70	0.68	0.58
I-95 Corridor Interior	Monitors = 20		N=118	Mean Peak O_3 = 77.9
Bias	+0.7	+17.7	+6.7	-12.6
MAE	15.0	20.0	13.2	21.6
RMSE	18.4	24.8	16.3	24.8
Index of Agreement	0.68	0.61	0.77	0.44

Table 3. Contingency table for threshold forecasts.

		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

Table 4. Threshold skill measures.

Measure	Symbol	Formula
Accuracy	A	$\frac{a+d}{a+b+c+d}$
Bias	B	$\frac{a+b}{a+c}$
False Alarm Ratio	F	$\frac{b}{a+b}$
Probability of Detection (POD)	H	$\frac{a}{a+c}$
Heidke Skill Score	HSS	$\frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}$
Critical Success Index	CSI	$\frac{a}{a+b+c}$
Pierce Skill Score	PSS	$\frac{a}{(a+c)}-\frac{b}{(b+d)}$

In addition (Table 2), MAQSIP-RT performed best in two key sub-regions, the western-rural monitors that "define" the regional background O_3 concentrations, and the coastal monitors that are often subject to abrupt air mass changes. The expert forecasts were slightly better in the interior of the I-95 corridor, reflecting model difficulties in resolving the effects of steep near-urban precursor gradients as well as forecaster experience in this environment.

Because air quality forecasts are issued to the public in the form of color codes, forecast skill at high O_3 thresholds are an important measure of performance. For a variety of threshold (*categorical*) forecast skill measures (Tables 3, 4), MAQSIP-RT outperformed the CHRONOS model and provided results similar to the expert New England forecasters for an 8-hour-average of 85ppbv as the threshold (Table 5). This threshold represents the cutoff between relatively good and relatively poor air quality and is used to trigger Ozone Action Day advisories in New England.

Table 5. Skill results for forecast methods. "Blend" refers to a 50-50 weighted average of both numerical forecasts.

	MAQSIP	CHR	NEF	PER	Blend
H	0.49	0.28	0.45	0.19	0.49
F	0.13	0.26	0.22	0.14	0.19
B	1.06	1.47	1.61	0.98	1.48
A	0.80	0.68	0.77	0.76	0.80
CSI	0.34	0.20	0.39	0.10	0.41
PSS	0.37	0.12	0.38	0.05	0.41
HSS	0.38	0.13	0.42	0.05	0.46

4. MODEL INTERCOMPARISON RESULTS: PHASE 2

Phase 2 was planned to coincide with the first of two northeast US field programs, the New England (http://airmap.unh.edu/about/NEAQS.cfm) AQ Study, NEAQS-2002. This project utilized the NOAA research vessel Ronald H. Brown and involved more than 20 partner institutions. In addition to the heavily instrumented ship, a G1 Gulfstream research aircraft operated by the U.S. Department of Energy's (DOE) Pacific Northwest National Laboratory (PNNL) also collected data with instruments developed at both PNNL and DOE's Brookhaven National Laboratory.

Phase 2 data was evaluated against monitor observations in the NE US for the period Aug 5-29, 2002. As in the phase 1 evaluation, both discrete and categorical analyses were performed on the operational model results. Further, skill scores were determined using the persistence forecast as a baseline measure of skill. Statistically, the persistence forecast and model forecast can be expressed as:

$$P = \mu + E_P \quad (1)$$
$$M = \mu + E_M \quad (2)$$

where P is the forecasted value by persistence forecast, M is the value forecasted by a model, μ is the true value, and E_P and E_M are the errors associated with persistence forecast and model forecast, respectively. If the model forecast outperforms the persistence forecast, then E_M must be smaller than E_P. Based on (1) and (2), the skill score *SS* can be defined as:

$$SS = \frac{E_P - E_M}{E_P} \times 100\% \qquad (3)$$

where E_P and E_M can be any valid error metrics such as RMSE and NME (in this study RMSE is used to calculate the skill score). In fact, this definition of SS is exactly the same as the generic form ($SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\%$) defined by Wilks (1995) and considering that a perfect forecast would have a zero error ($E_{perf} = 0$).

Table 6 presents the discrete evaluation results comparing the three models, where *mb* is the mean bias in ppb, *nmb* is the normalized mean bias, *nme* is the normalized mean error, and R is the correlation coefficient. Table 7 presents the categorical results; Table 8 the skill score results.

Table 6: Phase 2 (NEAQS) Model Intercomparison: discrete evaluation results.

	MAQSIP-RT		MM5-Chem		HYSPLIT-CHeM	
	Max 1-hr	Max 8-hr	Max 1-hr	Max 8-hr	Max 1-hr	Max 8-hr
mb (ppb)	1.41	2.75	9.51	8.31	3.2	-1.16
nmb (%)	2.24	5.02	15.01	15.1	5.13	-2.13
nme (%)	17.96	18.55	25.81	25.38	23.42	22.46
rmse (ppb)	14.63	13.04	21.25	18.18	19.05	15.84
R	0.74	0.76	0.64	0.68	0.57	0.60

Table 7: Phase 2 (NEAQS) Model Intercomparison: categorical evaluation results.

	MAQSIP-RT		MM5-Chem		HYSPLIT-CHeM	
	Max 1-hr	Max 8-hr	Max 1-hr	Max 8-hr	Max 1-hr	Max 8-hr
A (%)	99.16	85.82	96.96	76.17	98.98	89.53
B	0.58	0.74	2.34	1.43	1.36	0.30
CSI (%)	9.68	18.10	9.81	17.60	8.33	5.79
POD (%)	13.95	26.72	29.81	36.38	18.18	7.12
FAR (%)	76.0	64.04	87.24	74.58	86.67	76.27

Table 8: Phase 2 (NEAQS) Model Intercomparison : skill scores for RMSE

Conc. (ppb)	TSS (%)			SSS (%)		
	MAQSIP	MM5-Chem	HYSPLIT	MAQSIP	MM5-Chem	HYSPLIT
All	9.57	-21.98	-15.85	9.75	-31.42	-15.52
<40	2.37	-34.86	-24.82	10.50	-26.29	-3.13
40-79	13.11	-39.58	-13.31	9.62	-50.85	-18.16
80-119	6.53	-1.51	-23.99	9.72	-3.65	-19.81
>=120	-32.39	-22.25	-59.74	7.88	12.67	1.57

5. PERFORMANCE EXPECTATIONS FOR ETA-CMAQ

Currently, NOAA/NCEP has *established a single metric*—a categorical accuracy of 90%--as a target for acceptable performance for Eta/CMAQ (Davidson, et al., 2003). Since accuracy itself is very strongly influenced by box "d" in the contingency table (which does not include any of the high ozone episode days pertinent to planning ozone action days/health alerts), it is a relatively weak measure of performance. However, the CSI & POD scores are more relevant measures—while much more challenging to achieve at a high level of skill. Further, as shown here, additional evaluation and performance metrics—using the broad range of approaches identified above—should also be added. Since MAQSIP-RT is currently the most skillful model of all those tested in the 2001-02 pilot, it establishes a reference for eventual community acceptance of the emerging Eta/CMAQ system.

REFERENCES

Coats, C.J. Jr., 1996: High-performance algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system. *Proc. Ninth AMS Joint Conference on Applications of Air Pollution Meteorology with A&WMA*, Amer. Meteor. Soc., Atlanta, GA, 584-588.

Davidson, Paula, 2003: National air quality forecast capability: first steps toward implementation. *Program Overview, National Air Quality Forecast Capability,* NWS, Office of Sci/Technology, http://www.nws.noaa.gov/ost/air_quality

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note, NCAR/TN-398+STR, 122 pp.

Grell, G. A., S. Emeis, W. R. Stockwell, T. Schoenemeyer, R. Forkel, J. Michalakes, R. Knoche, W. Seidl, 2000: Application of a multiscale, coupled MM5/chemistry model to the complex terrain of the VOTALP valley campaign. *Atmospheric Environment* **34** 1435-1453.

Houyoux, M.R., J.M. Vukovich, C.J. Coats, Jr., N.W. Wheeler, and P.S. Kasibhatla, 2000: Emission inventory development and processing for the seasonal model for regional air quality (SMRAQ) project. *J.. Geophys. Res.*, *Atmospheres*, **105**(D7), 9079-9090.

McHenry, J.N., W.F. Ryan, N.L. Seaman, C.J. Coats Jr., J. Pudykeiwicz, S. Arunachalam, and J.M. Vukovich, 2003: A real-time eulerian photochemical model forecast system: overview and initial ozone forecast performance in the NE US corridor. Accepted for publication, *Bull. Amer. Meteor. Soc.*

Pudykiewicz, J., A. Kallaur, and P. K. Smolarkiewicz, 1997: Semi-Lagrangian modelling of tropospheric ozone. *Tellus*, **49B**, 231-248.

Ryan, W. F., C. A. Piety, and E. D. Luebehusen, 2000: Air quality forecasts in the mid-Atlantic region: Current practice and benchmark skill, *Wea. Forecasting*, **15**, 46-60.

Stein, A.F., Lamb D., and Roland R. Draxler, 2000. Incorporation of detailed chemistry in to a three-dimensional Lagrangian-Eulerian hybrid model: Application to regional tropospheric ozone. Atmospheric Environment, 34, 4361-4372.