ISSN 1047-3289 *J. Air & Waste Manage. Assoc.* **55**:1782–1796

The New England Air Quality Forecasting Pilot Program: Development of an Evaluation Protocol and Performance Benchmark

Daiwen Kang and Brian K. Eder

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

Ariel F. Stein

Fundación Centro de Estudios Ambientales del Meditrráneo, Valencia, Spain

Georg A. Grell and Steven E. Peckham

Forecast Systems Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO

John McHenry

Baron Advanced Meteorological Systems, Research Triangle Park, NC

IMPLICATIONS

Results revealed that no single metric is sufficient but rather a suite of measures is required to fully characterize a model's performance. Additionally, these measures need to be examined spatially, temporally, and over varying concentration ranges to adequately characterize a model's performance. For discrete-type evaluations, mean and normalized measures of bias and error were chosen. These revealed the following: (1) two of the three models overpredicted ozone (O_3) concentrations (mean bias ranged from +1.41 to +9.51 ppb for maximum 1 hr and from −1.16 to +8.31 ppb for maximum 8 hr), and (2) the root mean square errors produced by the models ranged from 14.63 to 21.25 ppb for maximum 1 hr and from 13.04 to 18.18 ppb for maximum 8 hr. Metrics associated with the categorical-type evaluation revealed that each model was able to achieve an accuracy >90% for the maximum 1-hr O_3 forecast, a minimum goal for the initial implementation of the new National Air Quality Forecast capability. However, this metric is heavily influenced by the very large number of correctly forecast nonexceedances. To circumvent this influence, two more stringent measures of categorical performance, the critical success index and the hit rate, were also calculated. These revealed that only a small percentage (between 6 and 36% depending on model and metric) of exceedances can be expected to be forecasted correctly. There is also a large false alarm ratio associated with each of the three models, which ranged from 64 to 87%. Evaluation results of the three prototype models have shown promise, but they have also shown that considerable work needs to be done as National Oceanic and Atmospheric Administration develops a National Air Quality Forecasting System.

ABSTRACT

The National Oceanic and Atmospheric Administration recently sponsored the New England Forecasting Pilot Program to serve as a "test bed" for chemical forecasting by providing all of the elements of a National Air Quality Forecasting System, including the development and implementation of an evaluation protocol. This Pilot Program enlisted three regional-scale air quality models, serving as prototypes, to forecast ozone (O_3) concentrations across the northeastern United States during the summer of 2002. A suite of statistical metrics was identified as part of the protocol that facilitated evaluation of both discrete forecasts (observed versus modeled concentrations) and categorical forecasts (observed versus modeled exceedances/nonexceedances) for both the maximum 1-hr (125 ppb) and 8-hr (85 ppb) forecasts produced by each of the models. Implementation of the evaluation protocol took place during a 25-day period (August 5–29), utilizing hourly O_3 concentration data obtained from over 450 monitors from the U.S. Environment Protection Agency's Air Quality System network.

INTRODUCTION

Each year, over 100 million Americans are exposed to levels of air pollution that exceed one or more health-based ambient pollutant standards. For many of them, especially those who suffer from respiratory problems, the availability of air quality forecasts, analogous to weather forecasts, could make a significant difference in how they plan their daily activities and, in turn, improve the quality of their lives. Weather forecasting, or more generally, environmental forecasting, has been one of the National Oceanic and Atmospheric Administration's (NOAA's) core missions since its inception. In response to Congressional direction (H.R. 4 Energy Policy Act of 2002 [Senate Amendment] S.517, SA 1383), which states: "The Secretary of Commerce, through the Administrator of the NOAA, shall, in order of priority as listed in section (c) establish a program to provide operational air quality forecasts and warnings for specific regions of the United States . . . " NOAA is preparing to deploy an operational National Air Quality Forecasting System. This follows NOAA's recently sponsored New England Forecasting Pilot Program that serves as a "test bed" for pollutant forecasting by providing all of the elements of a forecast system, including emission, meteorological and chemical models, and their evaluation. This pilot program enlisted three regional-scale air quality models, serving as prototypes, to forecast ozone (O_3) concentrations across the northeastern United States during the summer of 2002. The three models, which are discussed in Section 2, include: a hybrid Lagrangian model based on NOAA's Hybrid Single-Particle Lagrangian Trajectory (HYSPLIT) model,[1] an Eulerian model with coupled chemistry and meteorology developed at NOAA's Forecast System Laboratory,[2] and Multiscale Air Quality Simulation Platform (MAQSIP), another Eulerian model developed by Environmental Modeling Center of Micro-Computing Center of North Carolina (MCNC).[3]

A major component of the New England Pilot Program, and the subject of the work presented here, has been the development and implementation of an evaluation protocol, the purpose of which is three-fold. First, it is to determine which statistical metrics offer the most insight concerning model performance. Second, it is to provide feedback to the individual modelers concerning model performance, although not necessarily to determine which model performs best overall. The third and most important objective is to establish a "performance benchmark," for predictions of ground-level O_3.

This evaluation, which took place during a 25-day period (August 5–29, 2002), compared the modeling results with O_3 observations obtained from the U.S. Environmental Protection Agency's (EPA's) Air Quality System (AQS) network as discussed in Section 3. A suite of statistical metrics, of which the origins can be traced back to weather forecast verification,[4,5,6] were identified through the evaluation protocol development and are presented in Section 4. These metrics facilitated evaluation of both discrete-type forecasts (observed versus modeled O_3 concentrations) and categorical-type forecasts (observed versus modeled events [or exceedances]/nonevents [nonexceedances]) for hourly, maximum 1-hr, and maximum 8-hr O_3 forecasts produced by each of the models as discussed in Section 5. However, model intercomparisons are complex and difficult. Many factors, such as domain size, grid resolution, physical parameterizations, model complexity, development stage, and so forth, affect model performance and are handled very differently in the three prototype modeling systems. In addition, sensitivity to emissions data as well as other input fields may be as large as sensitivity to different models. Statistics should, therefore, be interpreted with care when comparing a model against other models, and small differences in performance should not receive undue emphasis.

Description of the Modeling Systems

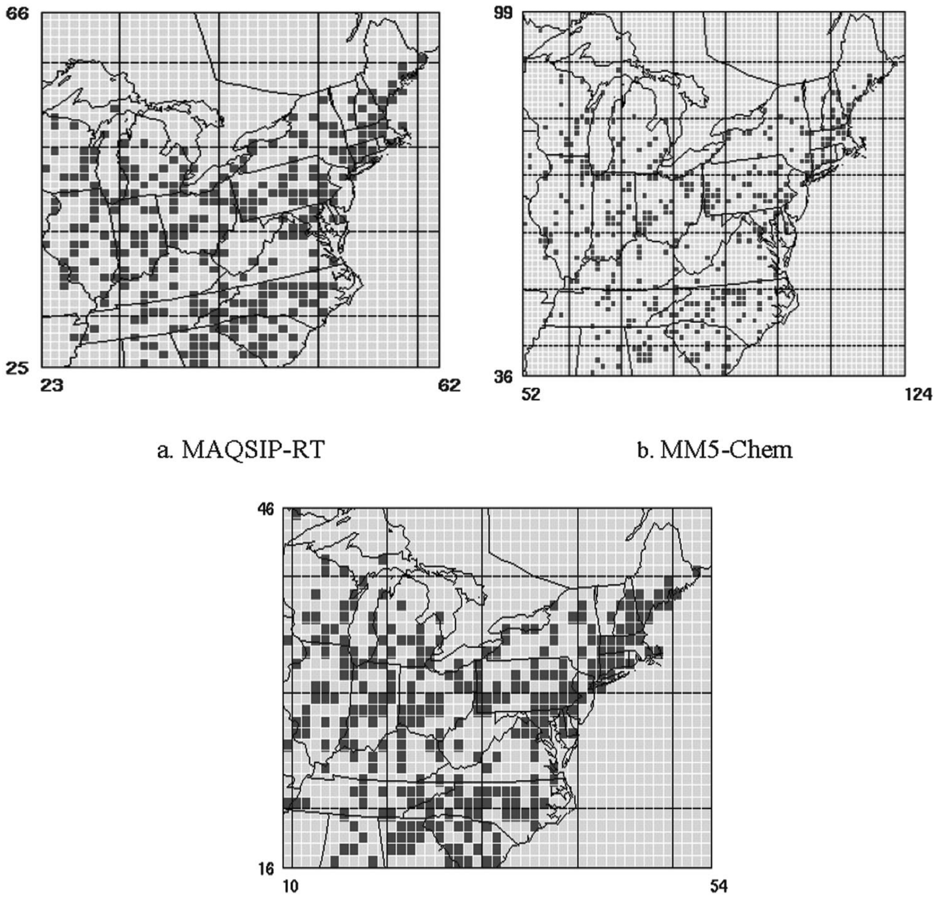
Brief summaries of the key attributes (meteorological model, chemical mechanism, emission, and horizontal and vertical resolution) of each of the three models are provided below. For spatial comparability, similar (although not identical) subdomains were extracted from the original modeling domains of each model (Figure 1). Similarly, for temporal comparability, the evaluation focused on a 25-day period (August 5–29, 2002), although longer simulation, though dissimilar, periods were available, depending on the model. For complete model descriptions, including information on transport, diffusion, and deposition schemes, refer to the citations provided in each model section.

MAQSIP

MAQSIP-real time (RT) is a highly optimized version of MAQSIP, a comprehensive Eulerian grid model developed by MCNC-North Carolina Supercomputing Center (now Baron Advanced Meteorological Systems).[3] MAQSIP-RT uses a modified version of the Carbon Bond IV chemical mechanism.[7] Emissions used in the model were from EPA's National Emissions Trend 1996 Emission Inventory, which were then processed through Sparse Matrix Operator Kernel Emissions, a highly efficient emission processing system.[8] The meteorology is provided by MM5 (the Fifth Generation Penn State/NCAR Mesoscale Model).[9] The model domain covered the eastern United States using a 45-km horizontal grid spacing and 31 σ-coordinate vertical layers. Results from the surface layer, which is 38-m thick, are used for this evaluation.

MM5-Chem

NOAA's Forecast Systems Laboratory MM5-Chem modeling system[2] is a multiscale Eulerian air pollution prediction system based on MM5, which is coupled with the Regional Acid Deposition Model chemical mechanism.[1] In this system the chemical kinetic mechanism is embedded within the meteorological model structure. As a result, emissions, deposition, photolysis, and chemical-transport-transformation calculations are performed



a. MAQSIP-RT

b. MM5-Chem

c. HYSPLIT-CheM

Figure 1. Maps of the modeling subdomains used in the evaluation for (a) MAQSIPRT, (b) MM5-Chem, and (c) HYSPLIT-CheM. Cells with observations are denoted.

"online" as part of the MM5 simulation. Emissions used in the model were from the EPA's National Emissions Trend 1996 Emission Inventory data set. The model domain contains 27-km horizontal grid cells and 30 vertically stretched layers. Results from the surface layer (16 m) are used in this study.

HYSPLIT-CheM

NOAA's Air Resources Laboratory's HYSPLIT-Chemistry Model (CheM) is a hybrid Lagrangian-meteorological/Eulerian-chemical modeling system.[1] As with the other two models, HYSPLIT-CheM utilizes meteorological input from MM5. Like MAQSIP, emissions are from the EPA's National Emissions Trend 1996 Emission Inventory and are processed through Sparse Matrix Operator Kernel Emissions. HYSPLIT-CheM assumes that the entire pollutant mass at each emission source is uniformly distributed among a number of "particles," each of which may be thought of as a capsule containing the various chemical species. These particles are advected, dispersed, and deposited throughout the simulation domain. The Carbon Bond IV mechanism is used for chemical transformations, which are solved for the entire concentration field between each advection/dispersion time step. The model domain covers the eastern United States using 50-km horizontal grid spacing and 10 vertical layers. Results from the surface layer (75 m) are used in this study.

O_3 Data

The O_3 data used in this evaluation were obtained from the EPA's AQS (formerly the Aerometric Information Retrieval System). This database contains a multitude of hourly aerometric data, including O_3 concentrations (measured in ppb), collected by state and local agencies at thousands of locations nationwide. Depending on model domain, between 464 and 472 AQS monitors were used. The monitor locations are shown in Figure 1. For those model grid cells containing more than one monitor, the average concentration computed from all of the monitors was used.

STATISTICAL TECHNIQUES

As discussed in the introduction, one of the main objectives of this work was to determine which statistical metrics offer the most insight concerning model performance. A review of germane literature revealed an overabundance of potential metrics (many interchangeable), with varying advantages and disadvantages.[4,5,6,10] Ultimately, a suite of metrics, of which the origins can be traced back to weather forecasting, were selected and calculated that facilitate evaluation of both discrete-type O_3 forecasts and categorical-type O_3 forecasts. Additionally, skill scores, which provide a measure of the relative accuracy of the O_3 forecasts (with respect to persistence), were calculated.

Discrete Statistics

For the discrete forecast evaluation, basic summary statistics along with two standard and widely used measures of bias, the mean bias (*MB*) and the normalized mean bias (*NMB*), and error, the root mean square error (*RMSE*) and normalized mean error (*NME*), were selected and are defined below:

$$MB = \frac{1}{N}\sum_{1}^{N}(C_m - C_o) \qquad (1)$$

$$NMB = \frac{\sum_{1}^{N}(C_m - C_o)}{\sum_{1}^{N} C_o} \cdot 100\% \qquad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(C_m - C_o)^2} \qquad (3)$$

$$NME = \frac{\sum_{1}^{N}|C_m - C_o|}{\sum_{1}^{N} C_o} \cdot 100\% \qquad (4)$$

Where C_m and C_o are modeled and observed concentrations, respectively.

Categorical Statistics

For the categorical forecast evaluation, the models' accuracy (*A*), bias (*B*), hit rate (*H*), false alarm rates (*F*), false alarm ratio (*FAR*), and critical success index (*CSI*) were calculated, based on observed exceedances and nonexceedances versus forecast exceedance and nonexceedances for both the 1- and 8-hr O_3 standard. A graphical representation of the variables (*a*, *b*, *c*, and *d*) used to formulate the categorical metrics is presented in Figure 2, where *a* would represent a forecast 1-hr exceedance (>125 ppb) that did not occur; *b*, a forecast 1-hr exceedance that did occur; *c*, a forecast 1-hr nonexceedance that did not occur; and *d*, a nonforecast 1-hr exceedance that did occur.

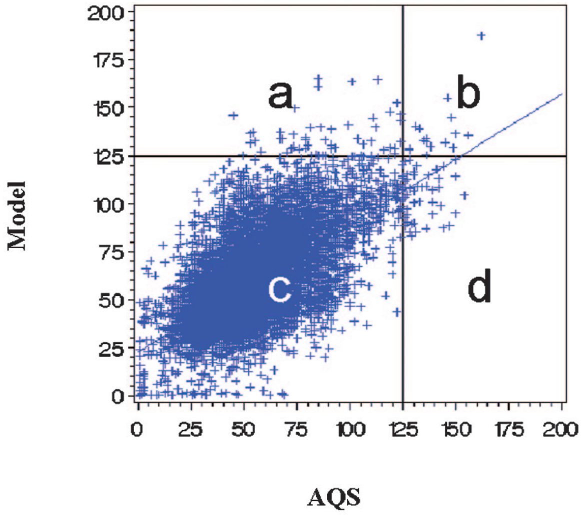


Figure 2. Example plot for categorical evaluation.

Accuracy (*A*) measures the percentage of forecasts that correctly predict an exceedance or nonexceedance and is given by:

$$A = \left(\frac{b + c}{a + b + c + d}\right) \cdot 100\% \qquad (5)$$

As will be discussed in Section 5.2, A is strongly influenced by the number of correctly forecast nonexceedances (*c*), which is invariably very large; hence, care must be taken in its interpretation. The bias (*B*) indicates, on average, if the forecasts are underpredicted (false negative) or overpredicted (false positives).

$$B = \left(\frac{a + b}{b + d}\right) \qquad (6)$$

A value of 1 would indicate no bias, values <1 indicate underprediction, and >1 indicates overprediction. The false alarm rate (*F*) is the proportion of nonexceedances that were incorrectly forecast.

$$F = \left(\frac{a}{a + c}\right) \cdot 100\% \qquad (7)$$

Similar to *A*, *F* is also strongly influenced by the number of correctly forecast nonexceedances (*c*), which is invariably very small. To avoid the influence of large numbers of nonexceedances, the false alarm ratio (*FAR*) measures the percentage of times an exceedance was forecast when none occurred.

$$FAR = \left(\frac{a}{a + b}\right) \cdot 100\% \qquad (8)$$

Smaller numbers are of course desirable, with a *FAR* = 0 indicating no false alarms, and a *FAR* of 50% indicating that half of the forecast exceedances did not actually occur. The *CSI* indicates how well both forecast exceedances and actual exceedances were predicted.

$$CSI = \left(\frac{b}{a + b + d}\right) \cdot 100\% \qquad (9)$$

Unlike the *A*, the *CSI* is not affected by a large number of correctly forecast nonexceedances. A *CSI* of 50% would indicate that half of the forecasted and actual exceedances were correct. Finally, the hit rate (*H*), which is similar to the *CSI*, indicates the percentage of actual exceedances that were forecasted. It is sometimes also called probability of detection (POD).

$$H = \left(\frac{b}{b+d}\right) \cdot 100\% \qquad (10)$$

Skill Scores

Skill scores (SS) were also calculated as part of this evaluation. This metric refers to the relative accuracy of a forecast (with respect to a reference forecast), which can be interpreted as a percentage of improvement over the reference forecast.[6] In terms of O_3 forecasts, the most convenient reference is the persistence forecast. Atmospheric variables often exhibit statistical dependence with their own past or future values. This dependence through time is usually known as persistence. Persistence forecast for O_3 is basically using today's values (observed maximum 1-hr and/or 8-hr concentrations) to make tomorrow's forecast.

In addition to presenting persistence over time, O_3 also displays persistence over space.[11] To distinguish the two kinds of persistence, the former is called temporal persistence and the latter spatial persistence. Considering the current density of the AQS monitoring stations, we can make a spatial persistence forecast using Location A's observed values as Location B's forecast, if A is the nearest location to B among all of the available AQS stations within the model domain.

Statistically, the persistence forecast and model forecast can be expressed as:

$$P = \mu + E_P \qquad (11)$$

$$M = \mu + E_M \qquad (12)$$

where P is the forecast value by persistence forecast, M is the value forecast 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 (10) and (11), the SS can be defined as:

$$SS = \frac{E_P - E_M}{E_P} \times 100\% \qquad (13)$$

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 *SS*). This definition of *SS* is the same as the generic form:

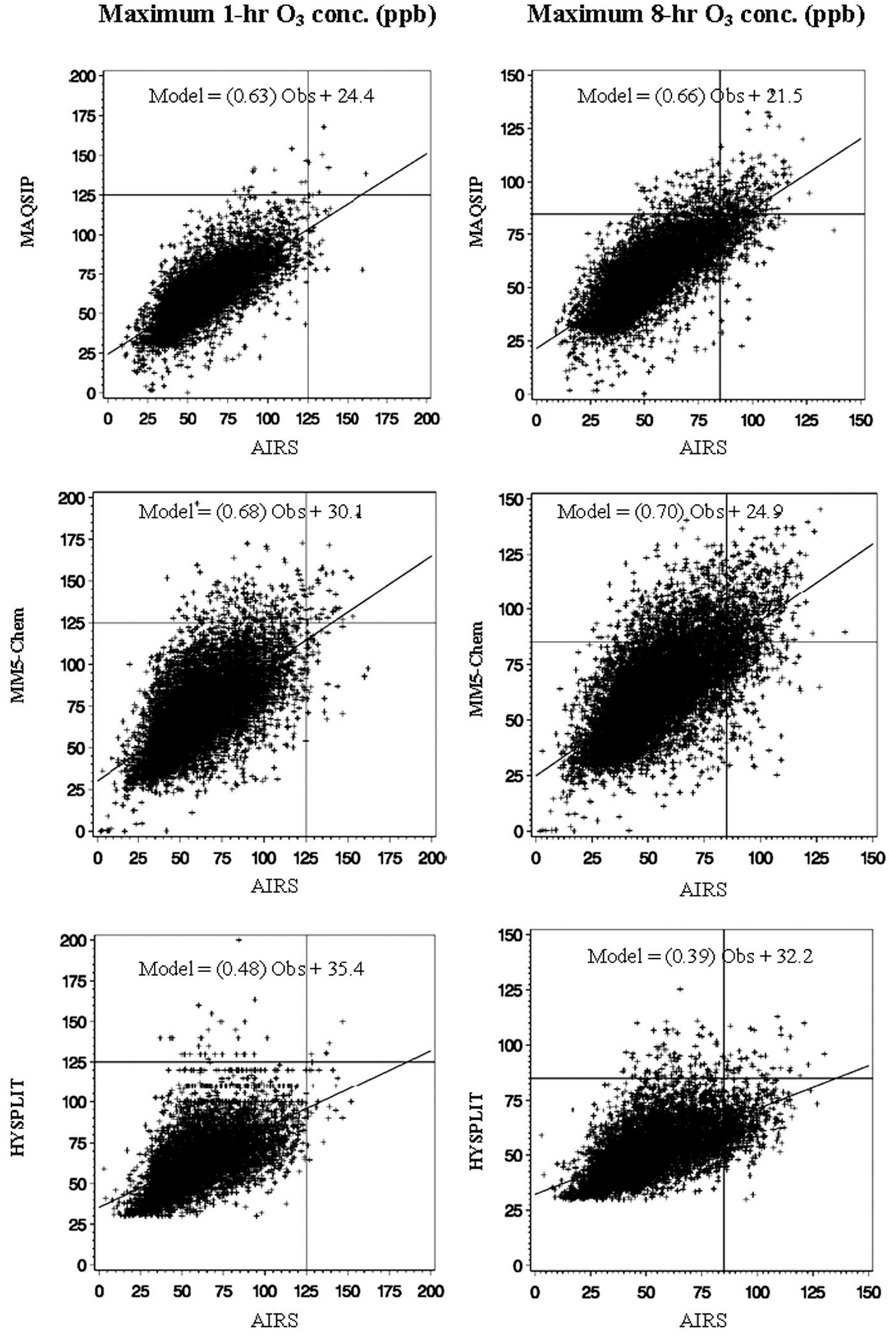


Figure 3. Scatter plots of the model versus AQS for both maximum 1-hr (left panels) and maximum 8-hr (right panels) O_3 concentrations (ppb) with exceedance thresholds, least-squares regression lines, and coefficients provided.

$$\left(SS_{\text{ref}} = \frac{E - E_{\text{ref}}}{E_{\text{perf}} - E_{\text{ref}}} \times 100\%\right) \tag{14}$$

where a perfect forecast would have a zero error ($E_{\text{perf}} = 0$)

Table 1. 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

RESULTS

Discrete Evaluations

Overall Performance and Summary Statistics. Scatter plots of the model forecasts versus AQS observations (for both the maximum 1- and 8-hr O_3 concentrations) are provided in Figure 3. In addition to illustrating the exceedance threshold areas (which were used in calculation of the categorical statistics), the plots also provide least-squares regression lines (and coefficients) associated with each evaluation. As evident from the regression lines (all have intercepts ≥20 ppb), most of the overprediction common to each model occurs at the lower concentrations. (Note: HYSPLIT does not forecast concentrations <30 ppb). All of the models underpredict the higher O_3 concentrations.

As seen in Table 1, which provides results for the discrete forecasts, the three models varied in their ability to accurately predict the 1-hr and 8-hr maximum O_3 concentrations. Both MAQSIP and MM5-Chem overpredict the maximum 1-hr and 8-hr concentrations as indicated by their positive MBs and NMBs. HYSPLIT overpredicts the maximum 1-hr concentrations but underpredicts maximum 8-hr forecast. For the 8-hr predictions, HYSPLIT provided the best performance with an MB of only −1.16 ppb (NMB −2.13%); whereas for the 1-hr prediction, MAQSIP performed slightly better (MB 1.41 ppb, NMB 2.24%) than HYSPLIT (MB 3.2 ppb, NMB 5.13%), and both were much better than MM5-CHEM (MB 9.51 ppb, NMB 15.01%). In terms of error, MAQSIP outperformed the other models for both the 1- and 8-hr maximum forecasts, producing the lowest RMSEs (14.63, 13.04 ppb) and NMEs (17.96, 18.55%), respectively. MAQSIP also provided better correlation coefficients for both the 8-hr (0.76) and 1-hr (0.74) as compared with MM5-Chem (0.68, 0.64) and HYSPLIT (0.60, 0.57). Note that for each model, the correlation coefficient associated with the 8-hr maximum was slightly greater than the 1-hr maximum.

Evaluations over Different Concentration Ranges. In addition to performing the evaluation over the entire data set, the same error and bias metrics are also calculated over different concentration ranges that correspond with EPA's Air Quality Index based on the 8-hr O_3 concentrations. As seen in Figure 4, the concentrations are grouped "good"

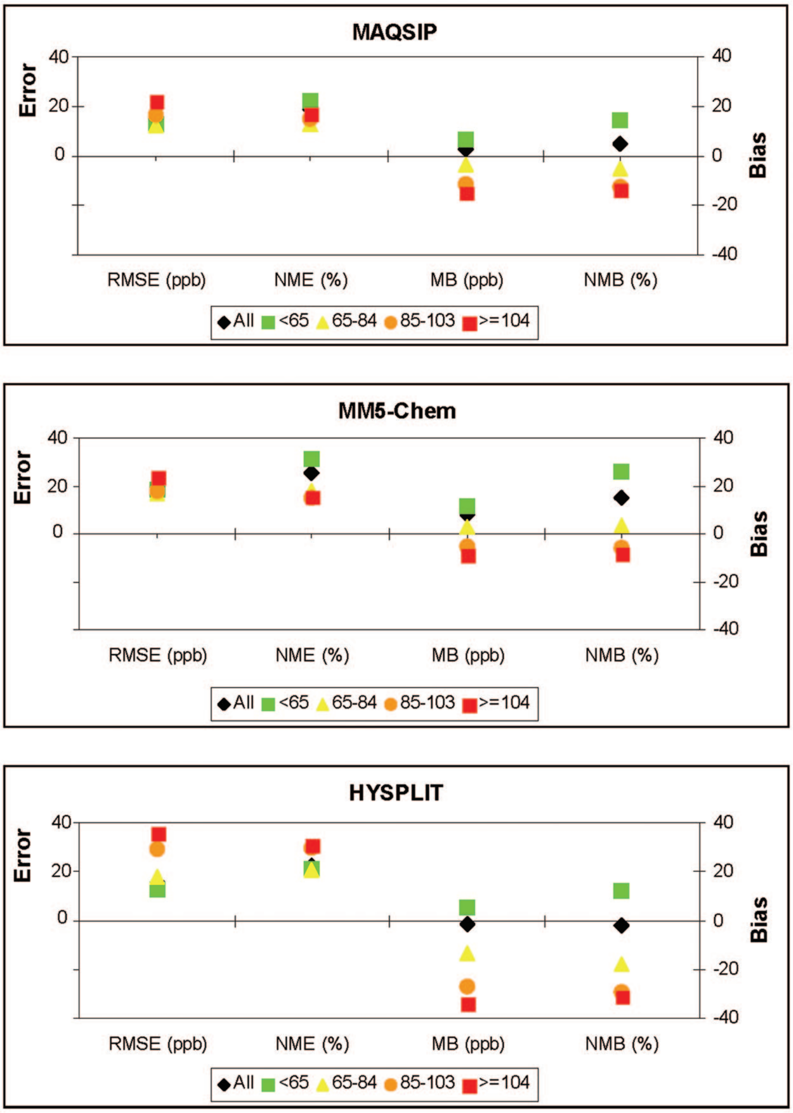


Figure 4. Errors and biases over concentration ranges corresponding to EPA's Air Quality Index for maximum 8-hr forecast.

(<65 ppb), "moderate" (65–84 ppb), "unhealthy for sensitive groups" (85–103 ppb), and "unhealthy" (≥104 ppb). For reference, the error and bias metrics for the entire dataset are also presented. The advantage of performing such an evaluation is best illustrated when examining the bias associated with HYSPLIT (Figure 4, bottom panel). Overall, HYSPLIT's MB (−1.16 ppb) and NMB (−2.13%) are very small (smallest of the three models), yet segregated biases reveal a different pattern. The small overall bias results from an overestimation of low concentrations (of which there are a great number) and a very large underestimation of the high concentrations (of which there are a small number). This pattern, whereas most evident with HYSPLIT, is not unique to HYSPLIT, as all three of the models underpredict the highest concentrations while overpredicting the lowest concentrations.

The dependence of model error on concentration range is generally not as strong as that of model bias on concentration range as denoted by the tighter grouping depicted by the RMSEs and NMEs. The one exception being that HYSPLIT RMSEs are considerably larger for higher concentrations (>85 ppb) when compared with lower concentrations.

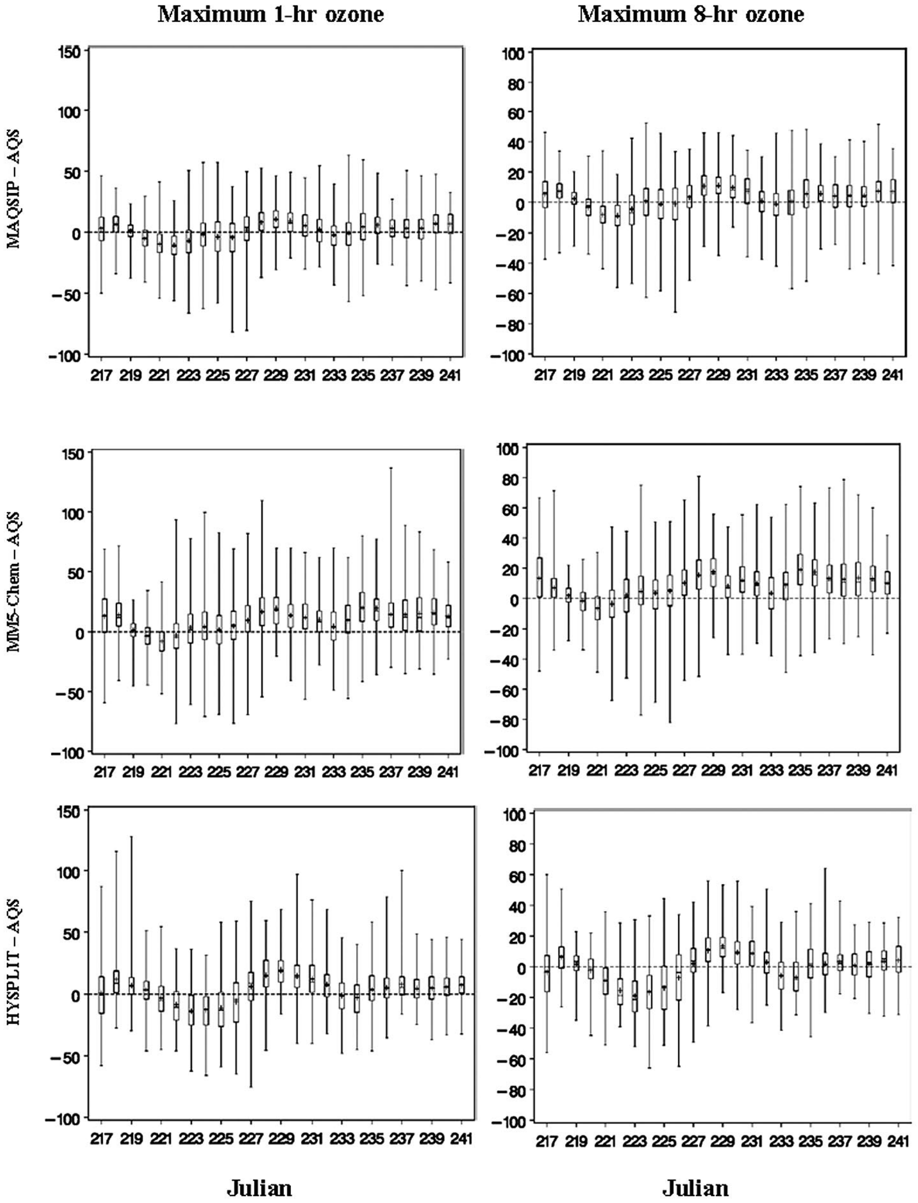


Figure 5. Boxplots of the variation across time of the (Model–AQS) for both 1- and 8-hr maximum O_3 concentrations (ppb).

Evaluations over Time

Evaluation of model performance over time is shown in Figure 5, where boxplots (denoting 75th, 50th, 25th percentiles, maximum, minimum, and mean) of simple bias (Model-AQS) are provided for each of the 25 days. Of the three models, MAQSIP generally exhibits the smallest bias variability across time and HYSPLIT the largest. It is interesting to note that the timing of the fluctuations of the bias above and below the zero bias line is generally "in-phase." This may be attributable to the fact that all three of the models used the same meteorological model (MM5) and that errors attributable to the meteorology may be perturbating through the forecasts; however, additional study is needed.

Examination of the diurnal performances of each model forecast (Figure 6) revealed subtle yet intriguing differences. Boxplots of the diurnal bias (model–observed) reveal that both MAQSIP and MM5-Chem persistently overpredict throughout the diurnal cycle with the smallest positive bias (best performance) occurring between 12 and 21 GMT (mainly daylight hours) for MAQSIP and 2 and 12 GMT (mainly nighttime hours) for MM5-Chem. HYSPLIT produces the largest bias differences over the diurnal cycle with overpredictions during the night and especially during the early morning hours. During the period from 15 to 22 GMT, the biases become smaller and even negative for the 16–20 GMT period.

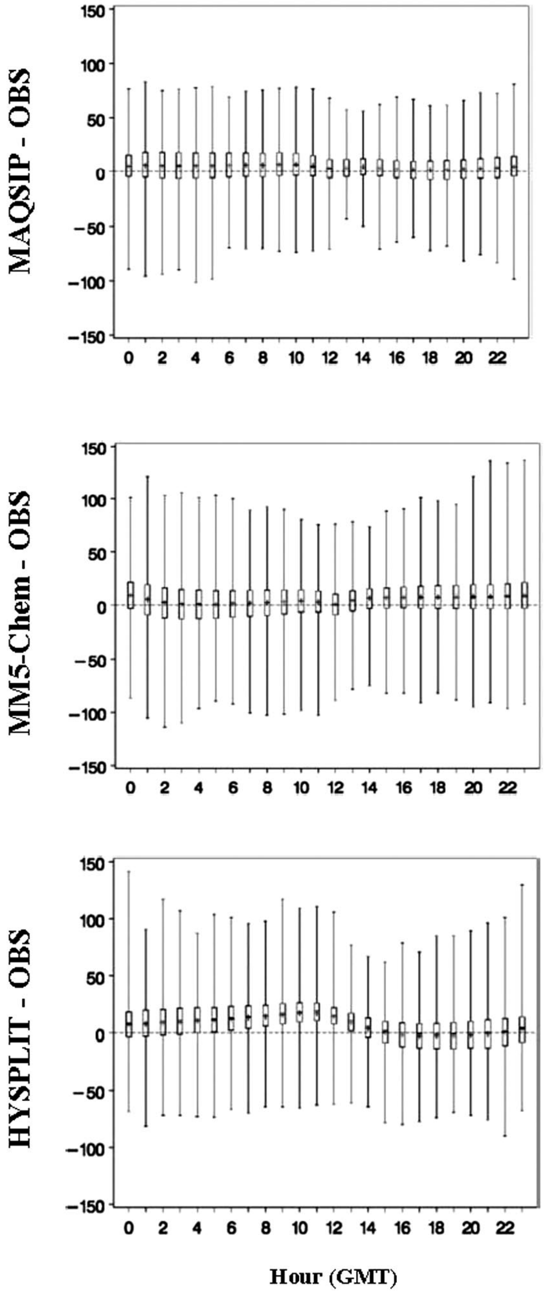


Figure 6. Boxplots of the diurnal variation (Model–AQS).

Evaluations over Space

To investigate the performance of the forecast models over space, correlation coefficients (R), MBs, and RMSEs for both maximum 1-hr and maximum 8-hr forecast were plotted across each model domain (Figures 7-12). Relative frequency distributions (histograms) of the different parameters are also provided. All three of the models generally have better correlation coefficients (Figures 7 and 8) in the northeastern part of the model domains and smaller correlation coefficients across the Appalachian mountain ranges (possibly related to the high elevation) stretching from Georgia to western Virginia and West Virginia. Among the three models, MAQSIP presents the best correlation with observation (the first quantiles of correlation coefficients are 0.63 for maximum 1-hr and 0.68 for maximum 8-hr forecast, meaning that >75% monitoring locations have correlation coefficients >0.63 and 0.68 for maximum 1-hr and 8-hr forecast, respectively), HYSPLIT the least (first quantiles are 0.38 for maximum 1-hr and 0.43 for maximum 8-hr), and MM5-Chem in between (first quantiles are 0.48 for maximum 1-hr and 0.56 for maximum 8-hr).

The distribution of MB for the maximum 1-hr forecast in Figure 9 indicates that there are more overpredictions than underpredictions for all three of the models, but there are more underpredictions by MAQSIP than MM5-Chem and HYSPLIT. However, for the maximum 8-hr forecast (Figure 10), HYSPLIT displays balanced underpredictions and overpredictions across the domain, whereas MAQSIP shifts to overprediction, and MM5-Chem is similar to its maximum 1-hr forecast. Each of the models tends to overpredict across the western edge of their domains, especially MM5-Chem. This may be related to the boundary conditions established for the models.

The RMSEs for each of the models are provided in Figures 11 and 12. RMSEs associated with HYSPLIT CheM are generally higher in eastern sections, stretching from New England down into North Carolina (especially for the maximum 1-hr forecast). A different RMSE pattern is seen with MM5-Chem, where the largest errors are generally found around metropolitan areas. MAQSIP's errors are more evenly distributed over space, with a slight tendency for lower RMSEs to be found in northern sections of the domain.

Categorical Evaluations

As shown in Table 2, the accuracy (A) for each model prediction, which indicates the percentage of forecasts that correctly predict an exceedance or nonexceedance, is >90% for maximum 1-hr forecast and 76–90% for maximum 8-hr forecast. The accuracy of each model's 1-hr exceedance/nonexceedance prediction is considerably better than its 8-hr prediction, and in fact approaches 100% (perfection); however, care must be taken in interpretation of this metric, as it is greatly influenced by the overwhelming number of correctly forecast nonexceedances (area c in Figure 2). To circumvent this inflation (which is common when evaluating the prediction of rare events like O_3 exceedances), the CSI is often a better metric of model performance. The CSI provides a measure of how well the exceedances were predicted, without regard to the large occurrence of correctly predicted nonexceedances. For our evaluation, the CSIs for the 8-hr exceedance ranged from 18.1% for MAQSIP and 17.6 for MM5-Chem to 5.8% for HYSPLIT. This indicates that MAQSIP and MM5-Chem were approximately three times

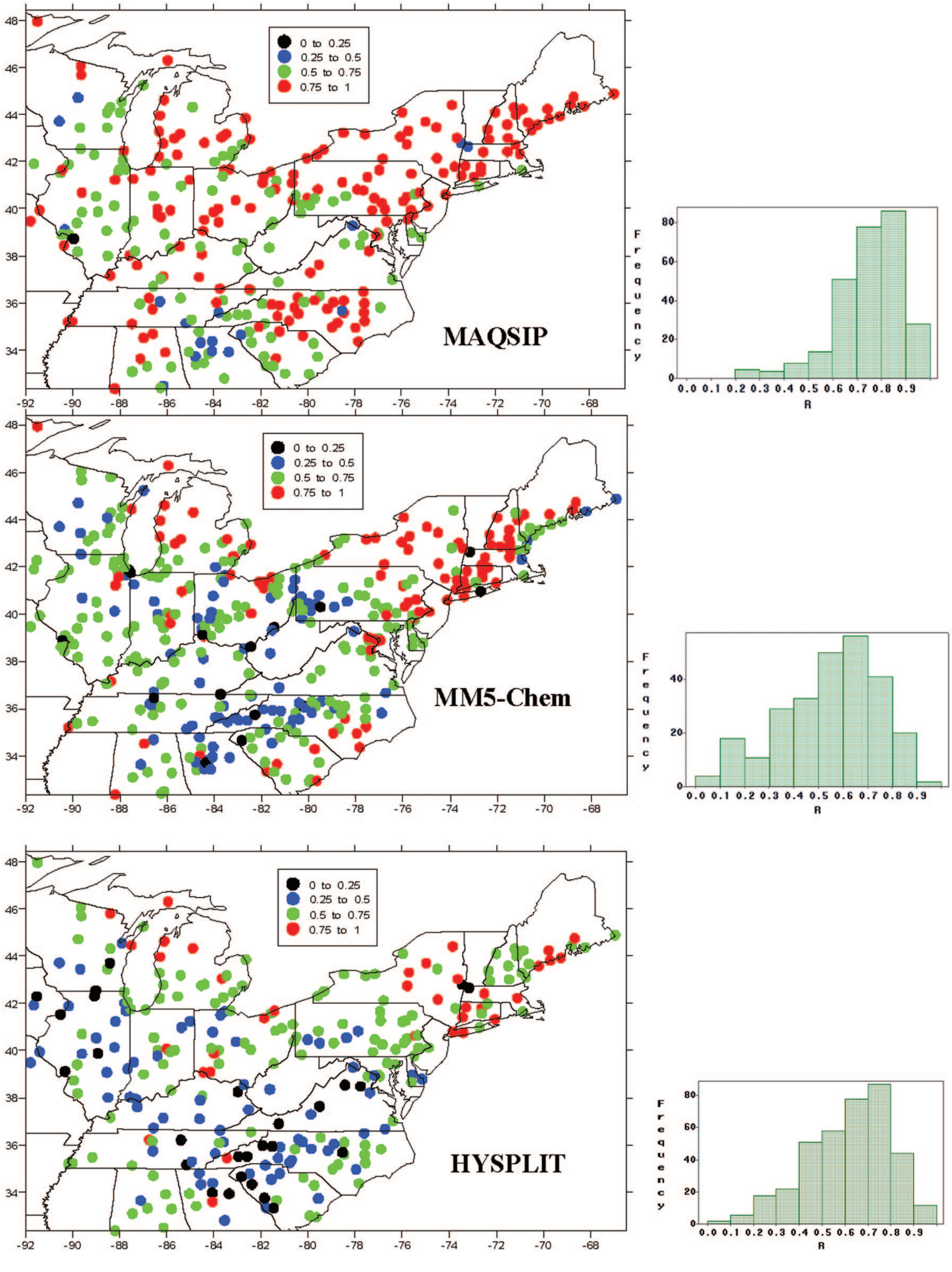


Figure 7. Maximum 1-hr correlation coefficient over space.

better than HYSPLIT at accurately predicting 8-hr exceedances. The ability of the models to predict 1-hr exceedances was more similar, though considerably lower, ranging from 9.7% for MAQSIP to 8.3% for HYSPLIT.

The hit rate (H) metric is similar to the CSI, in that it measures the number of times a model predicted an exceedance when one actually occurred. In our evaluation, MM5-Chem had the largest Hs (36.4% for 8-hr, 29.8% for 1-hr), followed by MAQSIP (26.7%, 14%) and HYSPLIT (7.1%, 18.2%). Note that only HYSPLIT has a smaller H for the 8-hr forecast than its 1-hr forecast.

Measures of bias (B), which for a categorical forecast indicates if forecast exceedances (1-hr and 8-hr), are underpredicted (B <1) or overpredicted (B >1) and vary

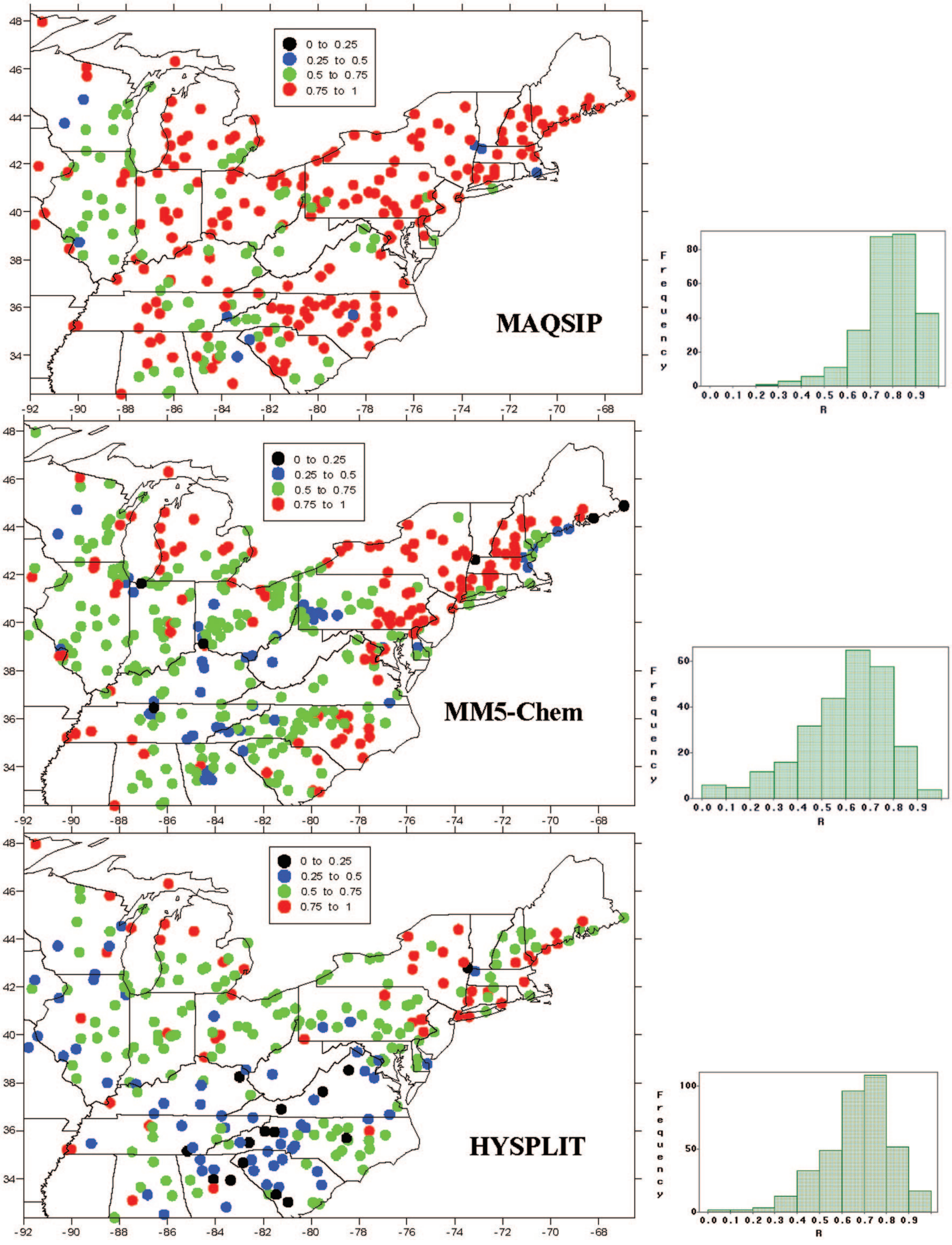


Figure 8. Maximum 8-hr correlation coefficient over space.

across models and even within models (i.e., HYSPLIT). MAQSIP underpredicts both 1-hr and 8-hr exceedances (B: 0.58 and 0.74, respectively), whereas MM5-Chem overpredicts both, especially the 1-hr (2.34 and 1.43, respectively). HYSPLIT greatly underpredicts the 8-hr (0.30), yet overpredicts the 1-hr (1.36).

A fifth categorical metric, the false alarm rate (F) is the proportion of nonexceedances that were incorrectly forecast. As expected, the Fs are low (ranging from 0.14% to 2.06%) for all three of the models, except that MM5-Chem presents an F value of 8.13% for the maximum 8-hr forecast. Again, to circumvent the influence of large

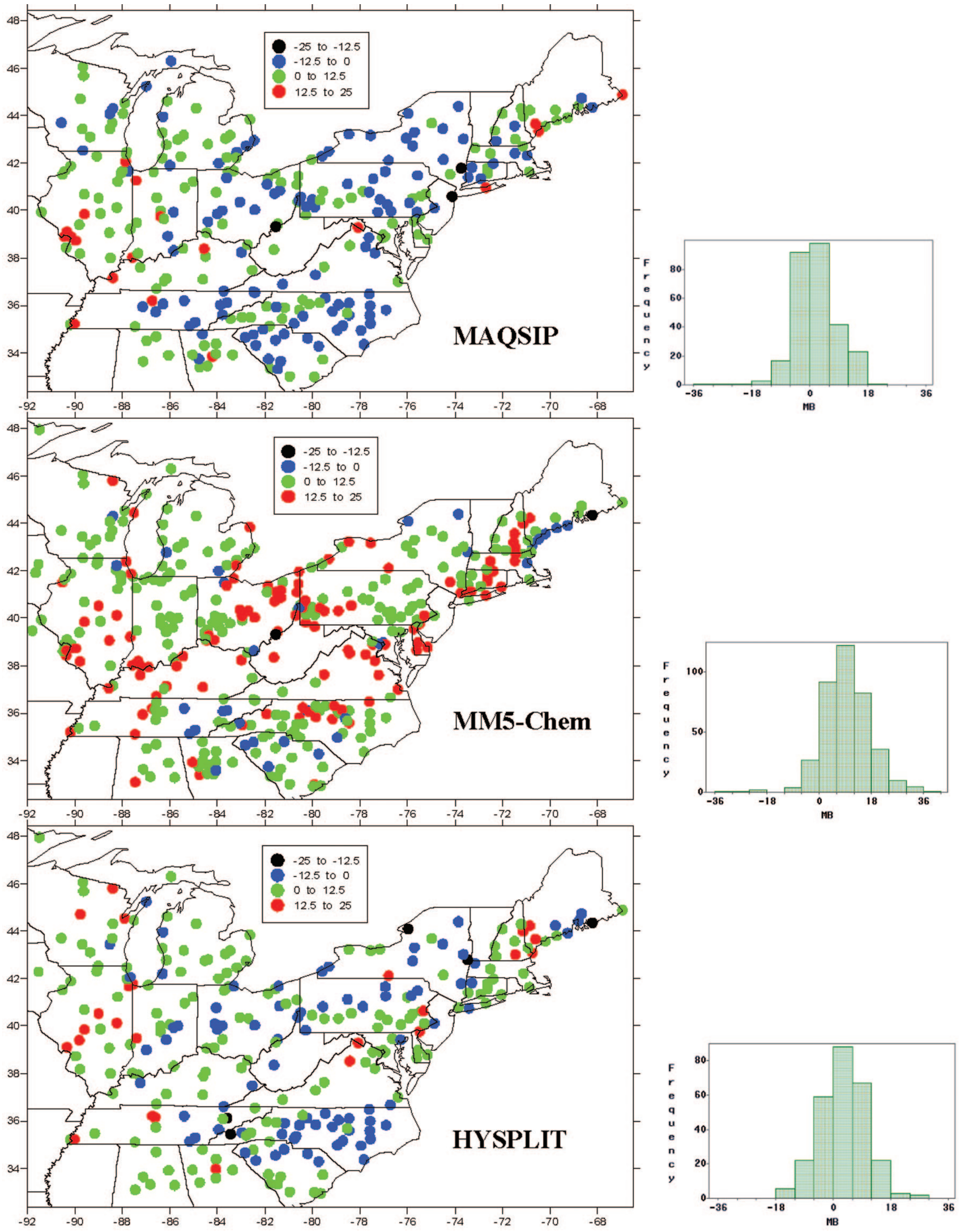


Figure 9. Maximum 1-hr MB distribution over space.

number of nonexceedances involved in the metric of F, the last categorical metric, the FAR, indicates the number of times that the model predicted an exceedance that did not occur. The FARs are high for the 1-hr forecast (ranging from 76% for MAQSIP to 87.2% for MM5-Chem) and slightly lower for the 8-hr forecast (from 64% for MAQSIP to 76.3% for HYSPLIT).

SS Results

Temporal and spatial SS for each model and for both the maximum 1-hr and maximum 8-hr O_3 forecast are found in Table 3. Of the three models, only MAQSIP outperformed forecasts based solely on persistence. Against temporal persistence, MAQSIP performed 9.57% and 10.93% better for the maximum 1-hr and 8-hr forecasts, respectively. It performed

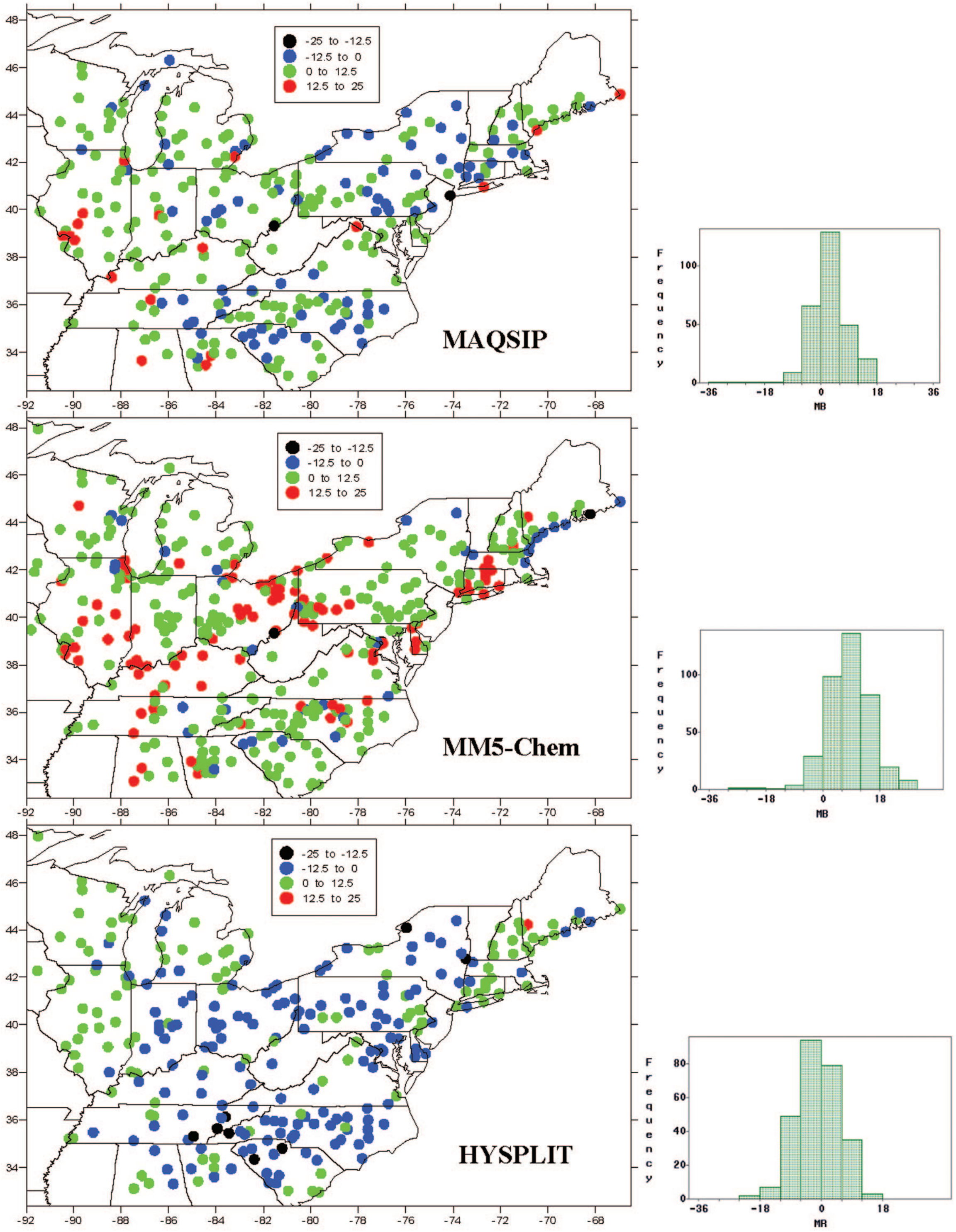


Figure 10. Maximum 8-hr MB distribution over space.

almost as well against spatial persistence, with spatial SS of 9.75% and 6.05% for maximum 1-hr and 8-hr forecast, respectively.

Neither MM5-Chem nor HYSPLIT-CheM performed better than persistence, as both produced negative SS for temporal and spatial persistence. Against temporal persistence, HYSPLIT-CheM posted a SS of −15.5% and −7.9% for maximum 1-hr and 8-hr forecasts, respectively. Against spatial persistence the SS values were −15.5% and −12.4%. MM5-Chem produced even larger negative SS (−22% against temporal persistence for both maximum 1-hr and 8-hr; −32.4% and −32.8% against spatial persistence for max. 1-hr and 8-hr, respectively.)

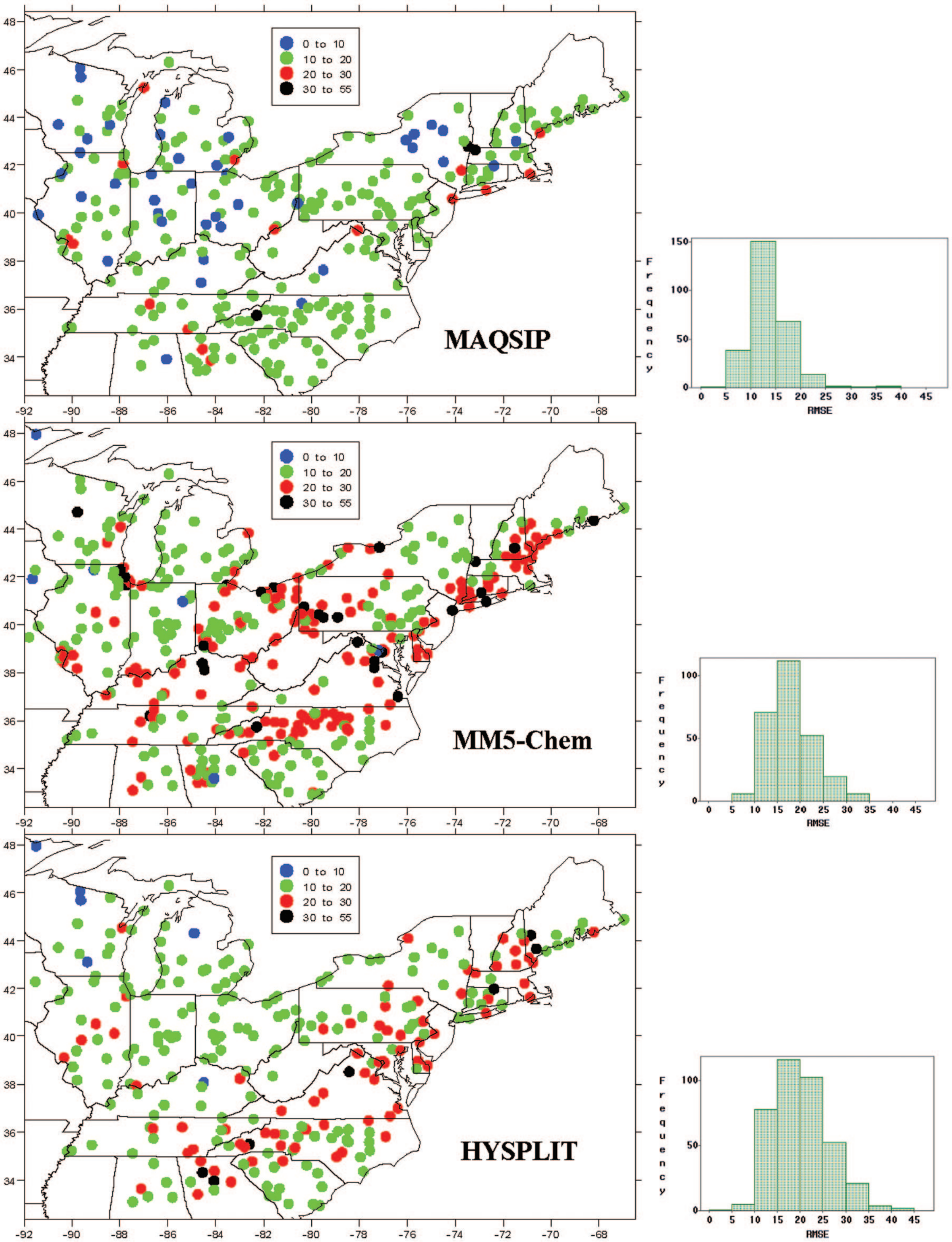


Figure 11. Maximum 1-hr RMSE distribution over space.

SUMMARY

The purpose of this research has been to develop and implement an operational evaluation protocol that will do the following: (1) determine which statistical metrics offer the most insight concerning model performance, (2) provide feedback to the individual modelers concerning their model's performance, and (3) establish a "performance benchmark" from which realistic expectations can be derived concerning the future predictions of ground-level O_3.

This research has revealed that no single evaluative measure is sufficient, but rather a suite of measures examined spatially, temporally, and over varying concentration ranges is required to fully characterize a model's performance. For discrete type evaluations, mean and normalized measures of bias and error provided insight into the models'

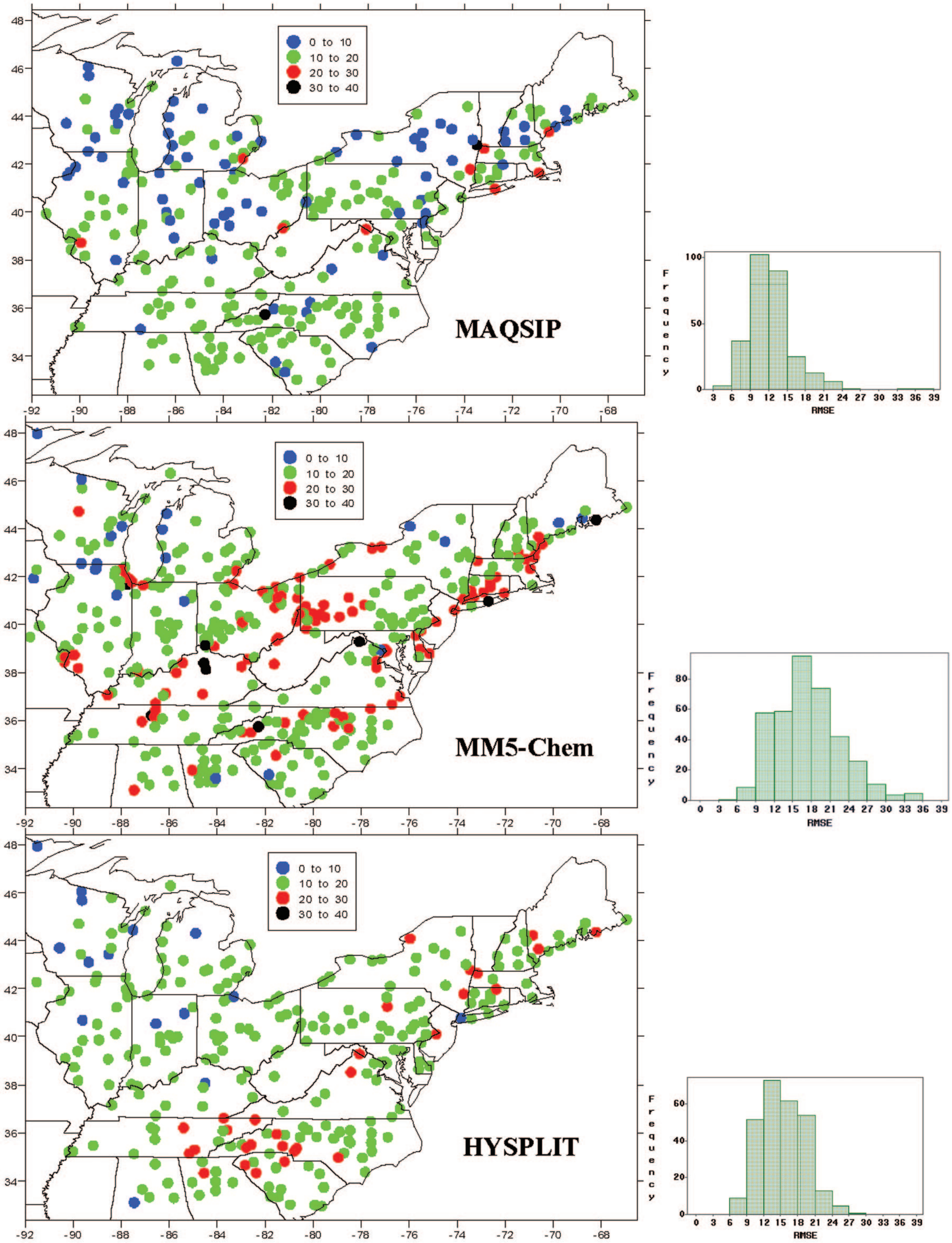


Figure 12. Maximum 8-hr RMSE distribution over space.

overall performances, as well as their performances over space and time. This type of evaluation revealed that overall, each model generally overpredicted O_3 concentrations as model biases were positive (the one exception was HYSPLIT-CheM's maximum 8-hr forecast). Measures of error were also revealing in that each model performed slightly better at forecasting the maximum 8-hr O_3 when compared with the maximum 1-hr, but that in each case, the error was substantial. The levels of bias and error varied both spatially and temporally.

Metrics associated with the categorical-type forecast evaluation were also revealing. For the maximum 1-hr forecast, each model was able to achieve an accuracy (A) level of >90%, a min for the initial implementation of the National Air Quality Forecasting Capability.[12] The false alarm rate (F) values are low for all three of the models.

Table 2. 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
H (%)	13.95	26.72	29.81	36.38	18.18	7.12
F (%)	0.14	2.06	1.42	8.13	0.45	1.30
FAR (%)	76.0	64.04	87.24	74.58	86.67	76.27

As discussed earlier, however, these metrics are heavily influenced by the number of correctly forecast nonexceedances and, therefore, are comparatively weak measures of model performance. Three additional, although more stringent, measures of categorical performance are the CSI, H, and FAR. Examination of these measures, which again circumvent the influence of correctly forecast nonexceedances, reveals that even with these "state-of-the-science" models, only a small percentage (between 6 and 36% depending on model and metric) of exceedances can be expected to be forecast correctly. There is also a large FAR associated with each of the three models, which range from 64 to 87%. Finally, examination of skill scores revealed that presently, only one of the three models, MAQSIP-RT, is better than persistence (either spatial or temporal) at forecasting maximum 1-hr and 8-hr O_3 exceedances.

The New England Forecasting Pilot Program was sponsored by NOAA to serve as a test bed for a future National Air Quality Forecasting System by providing all of the elements of a forecast "system," including the evaluation presented here. This evaluation has allowed establishment of a performance benchmark, from which realistic expectations can now be derived concerning the potential level of performance of air quality forecasting models. Although the results of the three prototype models have shown promise, they have also shown that much work lies ahead as NOAA develops a National Air Quality Forecasting System.

Table 3. SS for maximum 1-hr and 8-hr O_3 forecast.

	Temporal SS (%)			Spatial SS (%)		
	MAQSIP	MM5-Chem	HYSPLIT	MAQSIP	MM5-Chem	HYSPLIT
Max1-h	9.6	−22.0	−15.8	9.8	−31.4	−15.5
Max 8-h	10.9	−22.0	−7.9	6.0	−32.1	−12.4

DISCLAIMERS

This document has been reviewed and approved by the EPA and the NOAA for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES

1. Stein, A.F.; Lamb, D.; Draxler, R.R. Incorporation of Detailed Chemistry into a Three-Dimensional Lagrangian-Eulerian Hybrid Model: Application to Regional Tropospheric Ozone; *Atmos. Environ.* **2000**, *34,* 4361-4372.
2. Grell, G.A.; Emeis, S.; Stockwell, W.R.; Schoenemeyer, T.; Forkel, R.; Michalakes, J.; Knoche, R.; Seidl, W. Application of a Multiscale, Coupled MM5/Chemistry Model to the Complex Terrain of the VOTALP Valley Campaign; *Atmos. Environ.* **2000**, *34,* 1435-1453.
3. McHenry, J.N., Ryan, W.F.; Seaman, N.L.; Coats, C.J.; Pudykeiwicz, J.; Arunachalum, S.; Vukovich, J.M. A Real-Time Eulerian Photochemical Model Forecast System: Overview and Initial Ozone Forecast Performance in the Northeast U.S. Corridor; *Bull. AMS,* **2003** *85,* 525–548.
4. Murphy, A.H. Forecast Verification: Its Complexity and Dimensionality; *Mon. Wea. Rev.* **1991**, *119,* 1590-1601.
5. Murphy, A.H., Winkler, R.L. A General Framework for Forecast Verification; *Mon. Wea. Rev.* **1987**, *115,* 1130-1338.
6. Wilks, D.S. *Statistical Methods in the Atmosphere Sciences*. Academic Press: San Diego, CA, 1995.
7. Gery, M.W.; Whitten, G.Z.; Killus, J.P.; Dodge, M.C. A Photochemical Kinetics Mechanism for Urban and Regional Scale Computer Modeling; *J. Geophys. Res.,* **1989**, *94,* 12,925-12,956.
8. Coats, C.J. High Performance Algorithms in the Sparse Matrix Operator Kernel Emissions (SMOKE) Modeling System. In Proceedings of the *Ninth AMS Joint Conference on Applications of Air Pollution Meteorology with A&WMA*. Atlanta, GA, 1996; American Meteorological Society: Washington, DC, 1996: 548-588.
9. Grell, G.A., Dudhia, J.; Stauffer, D.R. *A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5);* NCAR Tech. Note, NCAR/TN-398+STR, National Center for Atmospheric Research: Boulder, CO, 1994
10. Jolliffe, I.T.; Stephenson, D.B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons Ltd.: New York, 2003.
11. Rao, S.T., Zurbenko, I.G.; Neagu, R.; Porter, P.S.; Ku, J.Y.; Henry, R.F. Space and Time Scales in Ambient Ozone Data; *Bull. Amer. Met. Soc.* **1997**, *78,* 2153-2166.
12. Davidson, P.M., Seaman, N.; Schere, K.; Wayland, R.A.; Hayes, J.L.; Carey, K.F. National Air Quality Forecasting Capability: First Steps Toward Implementation. Presented at the Sixth Conference on Atmospheric Chemistry: Air Quality in Mega Cities, Seattle, WA, January 2004; Poster J2 10.

About the Authors

Drs. Daiwen Kang, Brian Eder, and Ariel Stein are members of National Oceanic and Atmospheric Administration's Air Resources Laboratory. Brian Eder is currently on assignment to the National Exposure Research Laboratory, U.S. Environmental Protection Agency. Dr. Georg Grell and Steven E. Peckham are members of National Oceanic and Atmospheric Administration's Forecast Systems Laboratory. Dr. John McHenry is a member of Baron Advanced Meteorological Systems. Dr. Daiwen Kang is currently on assignment from University Corp. for Atmospheric Research, Boulder, CO 80301. Address correspondence to: Dr. Brian K. Eder, Atmospheric Sciences Modeling Division, Research Triangle Park, NC 27711; phone: +1-919-541-3994; fax: +1-919-541-1379; e-mail: eder@hpcc.epa.gov.