Draft Protocol

OPERATIONAL EVALUATION OF THE MM5 METEOROLOGICAL MODEL OVER THE CONTINENTAL UNITED STATES: Protocol for Annual and Episodic Evaluation

Task Order 4TCG-68027015

AG-TS-90/158

Prepared for:

Mr. Pat Dolwick

Office of Air Quality Planning and Standards
U.S. Environmental Protection Agency
Research Triangle Park, NC. 27711

Prepared by:

T. W. Tesche
Dennis E. McNally
Alpine Geophysics, LLC
3479 Reeves Drive
Ft. Wright, KY 41017

and

Dr. Craig Tremback
ATMET, LLC
PO Box 19195
Boulder, Colorado 80308-9195

12 July 2002

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

Over the past half decade, emergent requirements for direct numerical simulation of urban and regional scale photochemical and secondary aerosol air quality—spawned largely by the new particulate matter (PM2.5) and regional haze regulations—have led to intensified efforts to construct high-resolution emissions, meteorological and air quality data sets. The concomitant increase in computational throughput of low-cost modern scientific workstations has ushered in a new era of regional air quality modeling. It is now possible, for example, to exercise sophisticated mesoscale prognostic meteorological models and Eulerian photochemical/aerosol models for the full annual period, simulating ozone, sulfate and nitrate deposition, and secondary organic aerosols (SOA) across the entire United States (U.S.) or over discrete subregions. Consistent with ongoing U.S. Environmental Protection Agency (EPA) programs, this study is aimed at developing high-resolution, gridded meteorological data sets that can be used to support urban and regional scale air quality modeling over the continental United States at both national and regional scales.

1.1 Overview

In this study, the team of Alpine Geophysics, LLC (AG), and Atmospheric, Meteorological and Environmental Technologies, LLC (ATMET) are performing three basic tasks: (a) assessing different options for constructing a nation-wide meteorological data base for a specified target year (1999, 2000, and/or 2001), (b), applying and evaluating the National Center for Atmospheric Research/Penn State University Mesoscale Model (MM5) meteorological model over eastern and western U.S. subdomains for 2 to 3 two-week episodes during the years of 1999 and/or 2000, and (c) exercising and testing the model over the entire U.S. for a full year at 36 km horizontal grid scale.

1.2 Model Evaluation Protocol Objectives

This project offers a rare and challenging opportunity to rigorously test the performance of the MM5 model across broad domains for long integration periods – up to a year. The objective of the evaluation is to determine whether and to what extent confidence may be placed in the model's output fields (e.g., wind, temperature, mixing ratio, diffusivity, clouds/precipitation, and radiation) that will be used as input to emissions models and the various episodic and annual photochemical-aerosol models planned for use by the five Regional Planning Organizations (RPOs). This assessment of the MM5's reliability will be addressed from phenomenological (i.e., does the model simulate key processes correctly?) and regulatory perspectives

One of the most important questions addressed in this study concerns whether the MM5 meteorological fields are adequate for their intended use in supporting a variety of air quality modeling exercises. For the reasons discussed in Chapter 7, we will not be able to answer this question definitively, yet a significant amount of information will be developed in this study that will enable us to quantify the adequacy of the MM5 modeling and to judge its suitability for use in ozone, regional haze, and acid deposition modeling studies. This protocol outlines a formal model evaluation process that we plan to implement for both episodic and annual simulations. If successful, this process will produce useful, quantitative assessments of the adequacy of the MM5 meteorological fields for a variety of regional- and national-scale air quality modeling studies.

1.3 Outline of the Evaluation Protocol

Chapter 2 sets forth our philosophy governing the MM5 model evaluation and the specific components of the approach. In Chapter 3, we present the results of a brief survey of model evaluation methods in the allied field of regional climatology to determine if there are other methods that might be added to the set of statistical measures and graphical tools to be proposed for testing the model. Chapter 4 identifies the available data base for MM5 operation and evaluation and then in Chapter 5 we identify the specific tools we expect to use in testing the model's performance and reliability at a variety of spatial and temporal scales. The specific evaluation procedures we propose for the episode and annual simulations are presented in Chapter 6. In Chapter 7 we present the multi-step evaluation process we propose to implement in order to determine the adequacy of the MM5 fields for use in air quality modeling. Our procedures for reporting and archiving the evaluation results are given in Chapter 8.

2 TECHNICAL APPROACH TO THE MM5 MODEL EVALUATION

Reflecting the constraints imposed primarily by available data resources, our proposed evaluation of the MM5 model for annual and episodic simulations will largely be “operational”. While we seek to identify and correct flawed model components, input data sets, and pre- or post-processing components should they exist, except for model results that are obviously unrealistic or non-physical, the observational data sets needed to achieve this are essentially absent. For both episodic and annual simulations, the observational data bases will consist purely of routine surface and aloft measurements performed by the National Weather Service and other state and federal agencies. The operational evaluation will focus on the ability of MM5 to estimate correctly surface and aloft wind speed, wind direction, temperature, mixing ratio, and precipitation at pertinent time and space scales.

Although limited to an operational evaluation, we propose to implement a rigorous set of statistical procedures and graphical display methods to examine the episodic and annual MM5 simulations. For each simulation we will evaluate surface and aloft wind (speed and direction), temperature, mixing ratio, and precipitation fields using available data sets. Examples of the suite of statistical performance measures (routinely calculated by MAPS) to be examined include scalar and vector mean wind speeds, standard deviations in measured and observed winds, RMSE errors (total plus systematic and unsystematic components), two model skill measures, the Index of Agreement, as well as the mean and standard deviations in modeled and observed wind speeds. Statistical measures for temperature, mixing ratio, and precipitation will include means, biases, gross errors, and the index of agreement. Complementing the numerical measures will be a variety of graphical displays produced by the MAPS and PAVE software. These displays will include state-variable times series plots, two-dimensional parameter fields, vertical profiles of predicted and observed variables, skew-T plots, and so on.

The MM5 evaluations be performed with both scientific and policy perspective in mind. For both episodic and annual simulations we will perform: (a) subregional evaluations and (b) limited time-period evaluations (e.g., monthly and seasonal). These evaluations are aimed at elucidating the model’s ability to predict key processes at smaller time scales (e.g. coastal circulation regimes) as well as defining the model’s ability to produce reliable air quality inputs at scales appropriate to PM2.5 and regional haze issues. For example, for the annual MM5 simulation, we will conduct the statistical and many of the graphical evaluations independently over domains approximately corresponding to each of the five RPO domains in the U.S (see Figure 6-3). Moreover, we will evaluate the model for individual seasons (autumn, winter, spring, and summer) in addition to the full annual cycle. For both the episodic and annual domains, we will also evaluate the model independently for daytime versus nighttime conditions. The goal of all of these additional subregional and sub-temporal evaluations is to build confidence in the use of the model for regulatory air quality decision-making and to identify potential problem areas (should they exist) in the MM5 meteorological fields.

3 REVIEW OF REGIONAL CLIMATE MODEL EVALUTIONS

3.1 Results of Review

The annual runs that we will be performing are very similar to the regional climate simulations that have been popular in the climate community over the past decade. In a regional climate simulation, a global model is run first to provide large-scale fields for the assessment of climate variations (e.g., increased CO2). A limited-area (regional) model will then be one-way nested within these global fields, where the boundaries of the regional model will be defined by interpolation from the global fields. The regional model will then be run for a timescale of months to years in an attempt to assess the more local scale effects of the large scale climate change. Both MM5 and RAMS have been used in numerous studies of this kind.

As a first step in a regional climate study, the regional model is usually evaluated to see how well it can reproduce a past time period. In this evaluation simulation, the regional model is nested into a global dataset. For example, the NCEP/NCAR Reanalysis Data is frequently used. In effect, this is virtually the same as how we will be performing the annual (and episodic runs), except that MM5 will be run in 5 day blocks, then restarted. In the regional climate simulations, it is a continuous run through the entire time period.

Since the annual MM5 run and the regional climate run configurations are so similar, we have performed a review of the recent meteorological literature to see if other researchers have discovered any new techniques that might be useful for meteorological model evaluation for air quality purposes. Following are the highlights of what we found in our review.

There are a few new techniques that are becoming popular in the meteorological community. A number of researchers have classified verification techniques in three classes (Baldwin, et.al, 2000 provide a good summary):

1) Measure-oriented results – standard techniques of mean errors, RMSE, biases, etc.
2) Distributions-oriented results – focuses on the analysis of the joint distribution of forecasts versus observations
3) Events-oriented results – determination of the how well specific events are forecast by assessing the realism of a forecast or simulation.

One of the problems of the usual techniques, especially RMS error measures, is that outlying errors are heavily weighted. This was recognized 20 years ago when mesoscale models were first being applied. The statement was frequently made that, if you want to improve your verification statistics, apply a smoother to the field. This point can be demonstrated very well from the following figures (from Baldwin, et.al., 2000):

3.2 Recommendations

We will focus on the standard statistical techniques for this project. The correct use of the statistics can provide much of the needed verifications for this project, especially if the statistics are stratified correctly. One way NOT to do it is to lump all hourly verifications together to produce a single mean error and bias score for the entire year. We propose to implement a procedure of stratification of the errors to shed light on these distributions to assess the model performance.

The errors can be stratified in numerous ways:

1) Regionally
2) Seasonal or monthly
3) Diurnally
4) Combinations of the above

The stratification of the error statistics is essential is evaluating the overall performance of the simulation. By breaking the errors down regionally, it can be determined if one region has larger errors and biases than others. The breakdown by month can point to specific period that may need to be rerun with different model options or configuration, rather than repeating the entire year. A diurnal evaluation, possibly by region and by month or season, can give valuable information regarding perhaps daytime high or nighttime low temperature biases.

Figure 3-1. Idealized Observed Precipitation Rates (from Baldwin, et al., 2000).

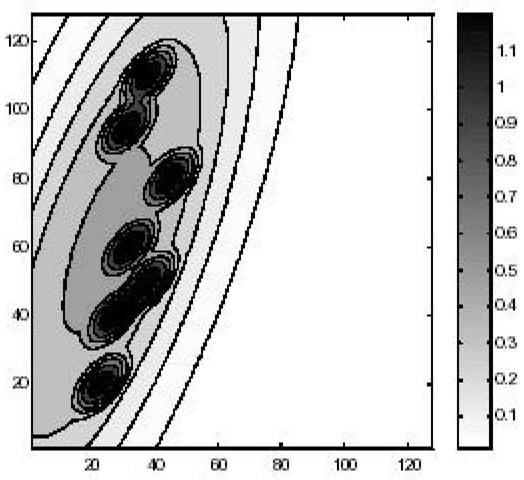


Figure 3-2. Results from *Simulation A* **(from Baldwin, et al., 2000).**

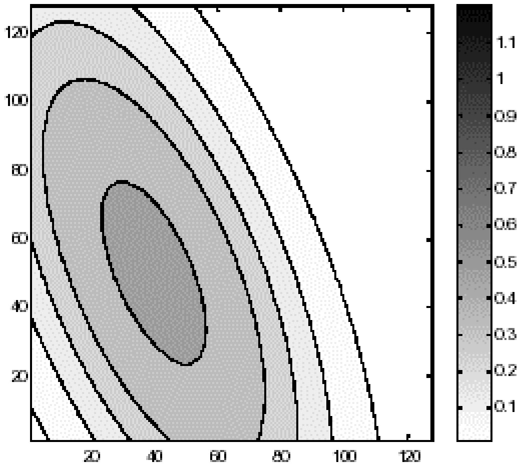


Figure 3-3. Results from *Simulation B* **(from Baldwin, et al., 2000).**

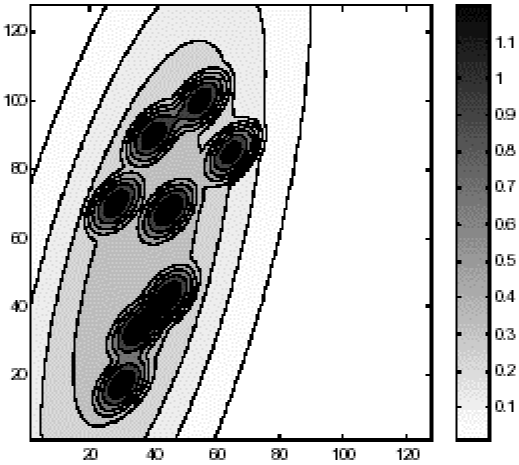


Figure 3-1 shows an idealized situation where the field is the observed precipitation rates. Figure 3-2 and Figure 3-3 show two different forecasts, with obviously different properties. By looking at the two simulations, it is clear that simulation B is a better simulation and provides more useful information to the researcher or forecaster and does give more confidence to the simulation. However, using the standard measures-oriented results, because of the slight misplacement of the field, simulation A has less error.

To look at the same two simulations from a distributions-oriented approach, Baldwin, et.al. presented the following figures.

Figure 3-4. Results from *Simulation A* **(from Baldwin, et al., 2000).**

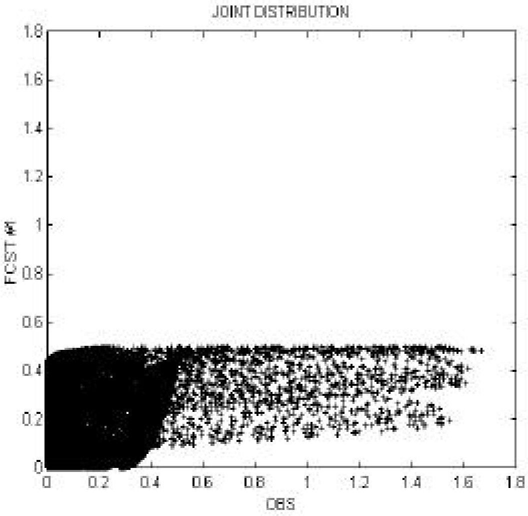
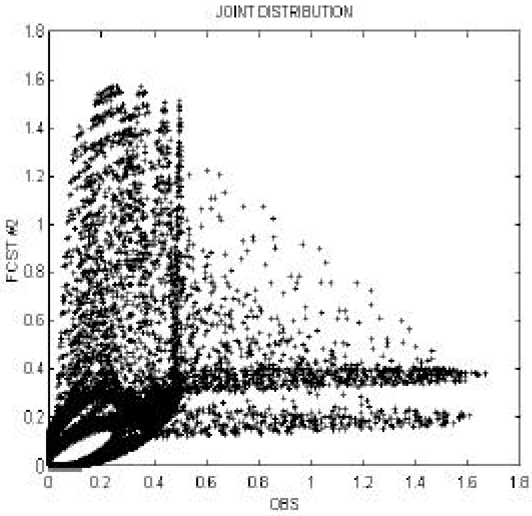


Figure 3-5. Results from *Simulation B* **(from Baldwin, et al., 2000).**



The figures show a scatterplot of the observed versus the simulated precipitation rates. Again, the differences are striking. While simulation B shows two distinct groups of points, one group where the simulated rates were less than the observed rates and one group where the simulated rates were more, simulation A shows more of a coherent group were most of the points are grouped where the simulated rates are less than the observed rates. While this can be deduced from visual examination of the original simulated fields, other quantities can be computed from the distributions-oriented techniques, such as correlation coefficients, to provide quantitative measures.

Most of the new work in the past several years, especially for the short to medium range forecasting, has been done relative to the events-based verification, especially when applied to the quantitative precipitation forecasting (QPF). Precipitation is a field that integrates the effects of all of the state variables (winds, temperature, moisture, pressure), then adds the complexities of the cloud parameterizations. QPF is probably the most difficult of all parameters to simulated and forecast, due to the sometimes stochastic nature of the precipitating elements and the difficulties inherent in the verification of the precipitation, since it may occur on a smaller scale than can be resolved with the regional model grid.

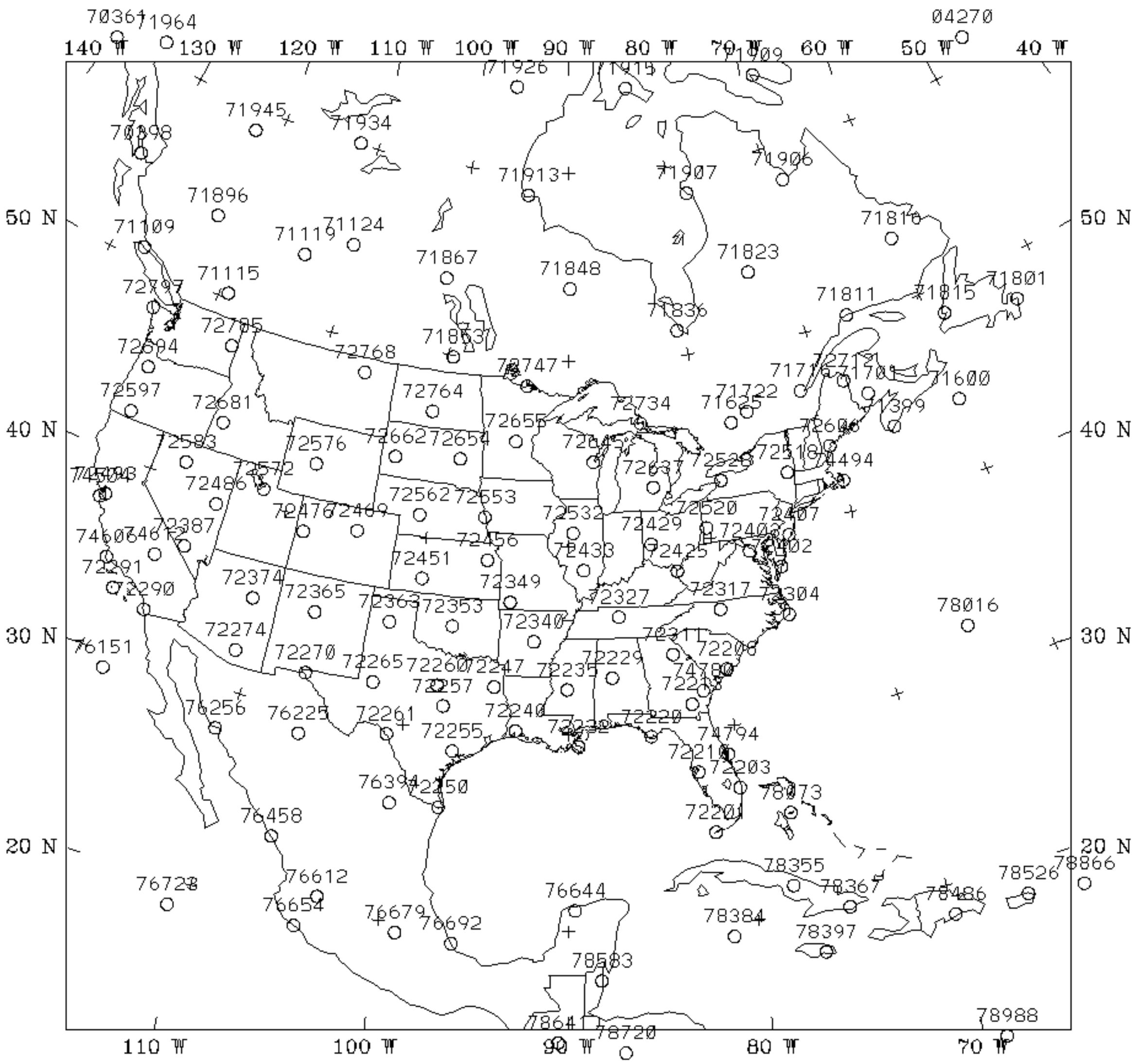
Most of the QPF verification work has focused on the class of techniques that could be categorized as pattern recognition or pattern-matching. For example, Ebert and McBride (2000) investigated the overlap of contoured rain amounts between the observed and predicted, looking at displacement, amplitude, and shape errors.

Pattern matching techniques are probably not informative for the basic meteorological state variables, and are beyond the scope of this study. However for the future, for annual and episodic simulations, the same techniques could possibly be applied to specific meteorological events, such as land and sea breezes. In the past, these were evaluated in a more qualitative sense, while the events-based techniques have the possibility of performing those types of evaluations in a quantitative sense. But even a better possibility for the future is the use of pattern-matching techniques for verification of ozone and other pollutant species, which can be spatially distributed (similar to precipitation) relative to the photochemical models. While the implementation of the event-based techniques is beyond the scope of this project, we will continue monitor the developments in the meteorological community.

4 DATA SUPPORTING THE EVALAUTION

The predominant sources of meteorological data available for the model application and evaluation include the NOAA Forecast Systems Laboratory (FSL), the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the UNISYS weather map archive, and the National Climatic Data Center (NCDC). Surface and aloft wind speed, wind direction, temperature, and moisture measurements will be obtained for these agencies and reformatted for use in comparisons with model predictions. The specific NCAR data set to be used is DS472.0, the hourly airways surface data. The primary data set available for comparing model performance aloft is the FSL RAOB archive. Precipitation will be analyzed using the NCDC TD3240 hourly precipitation data. Synoptic features (i.e. sea level pressure) will be analyzed by subjective comparison with the UNISYS synoptic weather charts. Figure 4-1 shows a reasonably complete list of the locations of the routine upper air meteorological data available for use in MM5 operation and performance testing.

Figure 4-1. Location of the NWS Upper Air Sounding Sites for Use in the MM5 Meteorological Modeling.



5 AVAILABLE STATISTICAL AND GRAPHICAL TOOLS

The MM5 operational performance evaluation will include the calculation and analysis of several statistical measures of model performance and the plotting of specific graphical displays to elucidate the basic performance of the model in simulating atmospheric variables. The specific statistical measures we propose to calculate are defined below. In addition, we also identify the suite of graphical procedures we propose to use to identify various feature of model performance. All of these procedures have been employed extensively in other prognostic model performance testing (see, for example, Doty et al., 2002; Tesche and McNally, 2001; Tesche et al., 2002, Emery et al., 2001). These evaluation procedures are incorporated into the most recent version of Alpine's public-domain Model Performance Evaluation, Analysis, and Plotting Software (MAPS) system.

5.1 Operational Evaluation of Surface Fields

5.1.1 Mean and Global Statistics

Several statistical measures will calculated as part of the MM5 evaluation. In the definitions below, the variable Ö represents a model-estimated or derived quantity, e.g., wind speed, wind direction, PBL height, mixing ratio, precipitation amount, or temperature. The subscripts e and o correspond to model-estimated and observed (i.e., measured) quantities, respectively. The subscript i refers to the i-th hour of the day.

Mean Estimation (M_e). The mean model estimate is given by:

$$M_e = \frac{1}{N} \sum_{i=1}^{N} \Phi_{ei}$$

where N is the product of the number of simulation hours and the number of ground-level monitoring locations providing hourly-averaged observational data. $\text{Ö}_{e\ i}$ represents the model-estimate at hour i.

Mean Observation (M_o). The mean observation is given by:

$$M_o = \frac{1}{N} \sum_{i=1}^{N} \Phi_{oi}$$

Here, Ö_{oi} represents the observations at hour i.

Average Wind Direction. Because wind direction has a crossover point between 0 degrees and 360 degrees, standard linear statistical methods cannot be used to calculate the mean or standard deviation. The method proposed by Yamartino (1984) performs well in estimating the wind direction standard deviation. Specifically, this quantity is calculated by:

$$s_a = \arcsin(b)\ [\,1 + 0.1547\ b^3\,]$$

where:

$$\boldsymbol{b} = \left[\, 1.0 - \left[\, (\overline{\sin\ \boldsymbol{a}})^{2} + (\overline{\cos\ \boldsymbol{a}})^{2} \,\right] \,\right]^{1/2}$$

Here, α is the measured hourly or instantaneous wind direction value.

5.1.2 Difference Statistics

Residual (d_i). For quantities that are continuous in space and time (i.e., wind speed, temperature, pressure, PBL height, species concentrations) difference statistics provide considerable insight into the model's performance, temporally and spatially. Difference statistics are based on the definition of a residual quantity. A mixing ratio residual, for example, is defined as:

$$d_i = c_e(x_i, t) - c_o(x_i, t)$$

where d_i is the i-th residual based on the difference between model-estimated (c_e) and observed (c_o) mixing ratio at location x and time i. In the definitions that follow, we shall use the letter c to denote any continuous atmospheric variable (e.g., temperature, precipitation amount, PBL height).

Standard Deviation of Residual Distribution (SDr). The standard deviation of the residual distribution is given by:

$$SD_r = \left(\frac{1}{N-1} \sum_{i=1}^{N} (d_i - \mathrm{MBE})^2 \right)^{0.5}$$

where the residual is defined as:

$$d_i = c_e(x_i, t) - c_o(x_i, t)$$

and MBE is the first moment, i.e., the mean bias error, defined shortly. This statistic describes the "dispersion" or spread of the residual distribution about the estimate of the mean. The standard deviation is calculated using all estimation-observation pairs above the cutoff level. The second moment of the residual distribution is the variance, the square of the standard deviation. Since the standard deviation has the same units of measure as the variable (e.g., meters/sec for wind), it is used here as the metric for dispersion. The standard deviation and variance measure the average "spread" of the residuals, independent of any systematic bias in the estimates. No direct information is provided concerning subregional errors or about large discrepancies occurring within portions of the diurnal cycle although in principle these, too, could be estimated.

Mean Bias Error (MBE). The mean bias error is given by:

$$MBE = \frac{1}{N} \sum_{i=1}^{N} (c_e(x_i, t) - c_o(x_i, t))$$

where N equals the number of hourly estimate-observation pairs drawn from all valid monitoring station data on the simulation day of interest.

Mean Normalized Bias Error (MNBE). The mean normalized bias error, often just called the bias, is given by:

$$MNBE = \frac{1}{N}\sum_{i=1}^{N}\frac{(c_e(x_i, t) - c_o(x_i, t))}{c_o(x_i, t)} \text{ x } 100\ \%$$

Mathematically, the bias is derived from the average signed deviation of the mixing ratio (or temperature) residuals and is calculated using all pairs of estimates and observations above the cutoff level.

Mean Absolute Gross Error (MAGE). The mean gross error is calculated in two ways, similar to the bias. The mean absolute gross error is given by:

$$MAGE = \frac{1}{N}\sum_{i=1}^{N} | c_e(x_i, t) - c_o(x_i, t) |$$

Mean Absolute Normalized Gross Error (MANGE). The mean absolute normalized gross error (or simply 'gross error') is:

$$MANGE = \frac{1}{N}\sum_{i=1}^{N}\frac{| c_e(x_i, t) - c_o(x_i, t) |}{c_o(x_i, t)} \text{ x } 100\ \%$$

The gross error quantifies the mean absolute deviation of the residuals. It indicates the average unsigned discrepancy between hourly estimates and observations and is calculated for all pairs. Gross error is a robust measure of overall model performance and provides a useful basis for comparison among model simulations across different model grids or episodes. Unless calculated for specific locations or time intervals, gross error estimates provide no direct information about sub-regional errors or about large discrepancies occurring within portions of the diurnal cycle.

Root Mean Square Error (RMSE). The root mean square error is given by:

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N} | \Phi_{ei} - \Phi_{oi} |^2 \right]^{1/2}$$

The RMSE, as with the gross error, is a good overall measure of model performance. However, since large errors are weighted heavily, large errors in a small subregion may produce large a RMSE even though the errors may be small elsewhere.

Least Square Slope and Intercept Regression Statistics. A linear least-squares regression is performed to calculate the intercept (a) and slope (b) parameters in the following equation:

$$\hat{\Phi}_{ei} = a + b\ \Phi_{oi}$$

This regression is performed for each set of hourly (or instantaneous) data to facilitate calculation of several error and skill statistics.

Systematic Root Mean Square Error (RMSE$_s$). A measure of the model's linear (or systematic) bias may be estimated from the systematic root mean square error given by:

$$RMSE_s = \left[\frac{1}{N} \sum_{i=1}^{N} | \hat{\Phi}_{ei} - \Phi_{oi} |^2 \right]^{1/2}$$

Unsystematic Root Mean Square Error (RMSE$_u$). A measure of the model's unsystematic bias is given by the unsystematic root mean square error, that is:

$$RMSE_u = \left[\frac{1}{N} \sum_{i=1}^{N} | \Phi_{ei} - \hat{\Phi}_{ei} |^2 \right]^{1/2}$$

The unsystematic difference is a measure of how much of the discrepancy between estimates and observations is due to random processes or influences outside the legitimate range of the model.

A "good" model will provide low values of the root mean square error, RMSE, explaining most of the variation in the observations. The systematic error, RMSE$_s$ should approach zero and the unsystematic error RMSE$_u$ should approach RMSE since:

$$\mathrm{RMSE}^2 = (\mathrm{RMSE}_S)^2 + (\mathrm{RMSE}_U)^2$$

It is important that RMSE, RMSE$_s$, and RMSE$_u$ are all analyzed. For example, if only RMSE is estimated (and it appears acceptable) it could consist largely of the systematic component. This bias might be removed, thereby reducing the bias transferred to the photochemical model. On the other hand, if the RMSE consists largely of the unsystematic component (RMSE$_u$), this indicates further error reduction may require model refinement and/or data acquisition. It also provides error bars that may used with the inputs in subsequent sensitivity analyses.

5.1.3 Skill Measures

We will calculate three skill measures as follows.

Index of Agreement (I). Following Willmont (1981), the index of agreement is given by:

$$I = 1 - \left[\frac{N\ (RMSE)^2}{\sum_{i=1}^{N} \left(| \Phi_{ei} - M_o | + | \Phi_{oi} - M_o | \right)^2} \right]$$

This metric condenses all the differences between model estimates and observations into one statistical quantity. It is the ratio of the cumulative difference between the model estimates and the corresponding observations to the sum of two differences: between the estimates and observed mean and the observations and the observed mean. Viewed from another perspective, the index of agreement is a measure of how well the model estimates departure from the observed mean matches, case by case, the observations' departure from the observed mean. Thus, the correspondence between estimated and observed values across the domain at a given time may be quantified in a single metric and displayed as a time series. The index of agreement has a theoretical range of 0 to 1, the latter score suggesting perfect agreement.

RMS Skill Error ($\mathbf{Skill_e}$). The root mean square error skill ratio is defined as:

$$Skill_E = \frac{RMSE_u}{SD_o}$$

Variance Skill Ratio ($\mathbf{Skill_{var}}$). The variance ratio skill is given by:

$$Skill_{Var} = \frac{SD_e}{SD_o}$$

5.2 Graphical Evaluation Tools

Over the years, a rich variety of graphical analysis and display methods have been developed to evaluate the performance of mesoscale meteorological models. Besides the statistical measures described in the preceding section, there are a number of procedures for graphically representing model results and observations that allow for direct comparison between them. Therefore, in selecting the graphical evaluation tools for portraying the MM5 episodic and annual simulation results, we will draw from among several approaches. In many instances, the differences in how modeled and measured quantities are treated in certain of these graphical techniques are more a matter of preference than correctness. Each graphical technique requires some assumptions that influence the outcome. However, by using a variety of graphical approaches, it is possible to examine the MM5's performance from different viewpoints and thus gain a clearer understanding of the results.

In application of the graphical techniques described below, we will focus the 36 km results for the annual simulation and the 12 km results for the episodic simulations. The parameters to be emphasized include but are not necessarily limited to bias, relative error, root mean square error, and index of agreement. These measures will be plotted in various ways for temperature, wind speed, wind direction, water vapor mixing ratio and precipitation. The graphical tools will be used to examine model performance both at the surface and aloft.

5.2.1 Graphical Displays Emphasizing Residual and Skill Measures

The graphical displays we propose to use for examining various residual, error and skill measures are included in the current set of displays produced by the MAPS software. These include, but are not limited to:

> The temporal correlation (time series) between point estimates and observations;

> The spatial distribution (gridded fields) of estimated quantities;

> The correlation among hourly pairs of estimates, observations, residuals, and distributions;

> The variation in spatial mean, bias and error estimates as functions of time and space;

> The degree of mismatch between volume-averaged model estimates and point measurements; and

> Log p/Skew-T plots of wind, temperature and mixing ratio.

These plotting methods are exemplified in the many recent MM5 and RAMS model evaluation studies (see, for example, Doty et al., 2002; Tesche and McNally, 2001; Emery et al., 2001; Tesche, et al., 2002).

5.2.2 Graphical Displays Emphasizing Spatial Fields

We seek a complimentary means of displaying and inter-comparing modeled and measured fields and plan to consider in greater detail the use of spatial interpolation methods. Specifically, we wish to examine the spatial variation in the differences between measurements and predictions across the modeling domain and not just at the monitoring stations where measurements are available. Two general approaches are available. The first is to simply interpolate the measurements to a grid mesh and then compare the gridded “observations” with measurements. This method, described in detail by Doty et al., (2002), allows one to compare predicted and measured ground-level two-dimensional fields of, say, bias and root mean square error (RMSE) statistics for temperature, water vapor mixing ratio, wind direction, wind speed, and precipitation. On the 36 and 12 km grids, the observed meteorological measurements would be interpolated, producing a gridded array of observed variables on the same grid as the modeled fields. These measured and modeled fields can be plotted and compared separately, or subtracted to produce residual fields of differences. Figure 5-1 is an example of how these residual fields might be displayed. In principal, the method could be used to compare the two-dimensional spatial fields of bias and RMSE for surface temperature, water vapor mixing ratio, wind direction, wind speed, and precipitation.

There are a number of critical assumptions inherent in the method outlined by Doty et al., (2002) which we believe may not be valid across the wide geographical domains and full annual cycle that we must address in the annual modeling. Spatial interpolation of measurements in regions of elevated terrain, sharp land-water contrasts, or in areas where sparse measurements are available all introduce potentially severe limitations on this first method. At present, we do not believe it is technically valid for the proposed episodic and annual MM5 evaluations.

The second method, and the one we propose to investigate further in order to arrive at a final recommendation, involves calculation of residuals (differences between prediction and observation) at the measurement sites and then interpolation of the residuals in space to produce the desired two-dimensional displays such as shown in Figure 5-1. At the moment, we are considering alternative methods of spatial interpolation, including the method of Doty et al., (2002) used in the SAMI RAMS meteorological modeling. This method used a Barnes-type analysis scheme to produce gridded variables. However, the Barnes (1973) interpolation method is not an accurate indication of actual conditions in high terrain areas because the observations are almost entirely made from locations outside the mountainous areas. For example, if the model is performing correctly in situations where the temperatures are decreasing with height, then there will be a cool "bias" over high terrain areas. Accordingly, we plan to investigate other methods for spatially interpolating the prediction-observation residuals across the domain(s) of interest. We invite any suggestions the EPA may have on suitable methods for spatially depicting the differences between modeled and observed meteorological parameters and in particular specific methods for conducting the interpolation. Once we have developed a final recommendation, it will be incorporated into this protocol as a update revision.

5.3 Operational Evaluation of Aloft Fields

We will perform an evaluation of the MM5 model performance in simulating the upper level horizontal winds and temperatures for the episodic and annual average simulations. Due to data limitations, this evaluation will necessarily be limited to comparisons of means, bias and errors in daily averaged wind speed, wind direction, and temperatures, vertically integrated from the surface to an appropriate level in the atmosphere (e.g., 400mb). This analysis of vertically-averaged mean statistics is intended to provide a coarse indication of the MM5's performance in reproducing the vertical wind and thermodynamic structures of the various episodes to be modeled. In particular, we will present tabular summaries of the predicted and observed, vertically-integrated horizontal winds and temperatures for all modeling days based on measurements made on the various NWS upper air reporting sites over the U.S. shown in Figure 4-1. From experience, we expect fairly good agreement in these comparisons due to the fact that aloft temperature and wind observations from the NWS radiosondes will be employed in the FDDA nudging scheme. However, the MM5 weighting coefficients used in the nudging will be fairly small so that the aloft fields will not be under a heavy constraint to match the observations locally. These statistical comparisons are intended to shed light on the degree of confidence one may place in modeled aloft wind and temperature patterns. However, this test will be insufficient by itself to 'validate' the reasonableness of the model predictions aloft.

Further insight into the aloft MM5 model performance aloft will be provided through the development of skew-T plots of the modeled and observed wind and thermodynamic profiles. Using standard analysis software obtained from NCAR and incorporated into MAPS, these plots will be developed for every available rawinsonde sounding during the modeling episodes. Figure 5-2 presents an example of a skew-T plot for the Jacksonville, FL site on 22 April 1999 at 1400 LST. The solid red and blue lines correspond to the observed and predicted winds, respectively. The thinner red and blue lines denote the mixing ratio observations and predictions.

5.4 Evaluation of Diagnosed Planetary Boundary Layer Height Fields

Historically, the mixed layer height (i.e., mixing height) was examined routinely as part of photochemical model evaluations since earlier models such as the UAM-IV were fundamentally tied to the 'mixing height' concept. Current generation photochemical models like CAMx are not formulated in this manner; instead, the characterization of vertical mixing is related to the distribution of turbulent kinetic energy via the meteorological model's simulation of the coupled mass, momentum and energy equations. Other models (e.g. CMAQ) that may be applied using the MM5 data bases developed in this study do require specification of mixing height. A more useful concept in this analysis is the planetary boundary layer (pbl) height which may be defined as that part of the troposphere that is directly influenced by the presence of the earth's surface and responds to surface forcings with a time scale of about an hour or less (Stull, 1988). During the daytime, this PBL definition corresponds fairly well with the so-called mixed layer height determined from the inflection in the potential temperature, moisture, and trace gas concentration profiles derived from vertical sounding devices (e.g. radiosondes). We propose to compare the maximum MM5 modeled and observed afternoon mixing heights for each modeling day over an appropriate domain or subdomain(s). While this comparison is an important one because of its implications for air quality modeling, it is a difficult one to make unambiguously because of the nature of mixing heights.

While the mixing height can be determined directly from the MM5 output, mixing height is not a variable that is directly obtained from traditional rawinsonde measurements. Estimates of mixing heights or, alternatively PBL heights, are typically determined manually or with objective algorithms that examine the vertical profiles of temperature, moisture, light scattering, or trace gas (e.g., ozone) concentrations. In most cases the mixing height is defined as that height where there is a significant change in one or more of these properties. A simple objective procedure for estimating mixing height was developed many years ago by Holzworth using near-surface temperature measurements and the morning upper air temperature soundings.

Given the full range of meteorological and physical conditions to be considered in the annual modeling, manual determination of mixing height from the various soundings locations shown in Figure 4-1 is clearly impossible. However, a carefully defined objective procedure may have some merit in providing an efficient, yet fairly reliable method for estimating mixing heights to compare with MM5 predictions. The method we are currently exploring in this regard is similar to the original Holzworth method, but uses the 0000 UTC sounding and the midday surface temperatures to estimate the height at which a dry adiabat intersects the upper air stable layer. By restricting this analysis to afternoon, non-coastal and non-mountainous regions, we hope to be able to remove the influences of local thermally-induced circulations that would confound the simple method developed by Holzworth. We plan to investigate this and possibly other methods for objectively determining daily maximum afternoon mixing heights over mid-continental locations within the domain(s) of interest. We invite any suggestions the EPA may have on suitable methods for estimating the mixing heights using routine surface and aloft sounding data, keeping in mind the fundamental technical difficulties in prescribing on method that covers all conditions within the annual cycle. If we are successful in developing a credible method for estimating daily maximum mixing heights at the rawinsonde locations, we will present the method as a final recommendation, to be incorporated into this protocol as a update revision. If it is possible to estimate mixing heights reliably, we will calculate mixing height residuals (predicted minus observed) and contour these across the domain(s) of interest to produce graphical displays of modeled PBL height fields such as those shown in Figure 5-3, an example over the Lower Lake Michigan Region.

5.5 Evaluation of Precipitation Fields

Evaluation of the precipitation fields represents another area where additional investigation (and hopefully interaction with EPA) is needed before arriving at a final methodology. Several methods are being explored for potential use in evaluating the MM5 precipitation forecasts. Each of these have been used recently in major prognostic model evaluation studies in support of air quality programs; none of the methods is fully satisfactory. Below we review these methods briefly.

One method involves the calculation of standard rainfall statistics and the use of contingency table categories as shown in Table 5-3. The categories of correct "no rain" forecasts, false alarms, misses, and hits are denoted by Z, F, M, and H, respectively. The following statistical measures are defined:

$$PBIAS = \frac{F + H}{M + H} \tag{5-1}$$

$$POD = \frac{H}{M + H} \tag{5-2}$$

$$FAR = \frac{F}{F + H} \tag{5-3}$$

$$ANR = \frac{Z}{Z + F} \tag{5-4}$$

$$HK = POD + ANR - 1 \tag{5-5}$$

Each of these statistics can be calculated for various precipitation thresholds, e.g., 0.2, 2, 5, 10, 15, 25, 35, 50, and 75 mm with respect to the observed 6-hour amounts at a given NWS station.

The precipitation bias (PBIAS) as defined by Equation (5-1) is the total number of model forecasts of precipitation divided by the total number of observed precipitation events. The precipitation probability of detection (POD) given by Equation (5-2) is the total number of correct model forecasts of precipitation divided by the total number of observed precipitation events. The precipitation false alarm ratio (FAR) defined by Equation (5-3) is the total number of times of model predictions of precipitation when there was none observed divided by the total number of model forecasts of precipitation. The accuracy for non-rain events (ANR) given by

Equation (5-4) is the total number of correct model forecasts of no precipitation divided by the total number of observed no precipitation events. Finally, the so-called Hanssen-Kuipers score (HK) described by Equation (5-5) is a composite of the POD and ANR and has a range of ±1. These statistics can computed on event, daily, monthly, seasonal, and annual bases for each monitoring station recording quantitative precipitation. Table 5-4 gives an example of how a contingency table might look; it is based on RAMS precipitation forecasts for the 12 km SAMI domain.

A second method would entail comparing gridded MM5 precipitation fields on an event total, monthly, seasonal, or annual basis with gridded fields of precipitation. The principal challenge here is to select an appropriate spatial interpolation algorithm for use in interpolating between and extrapolating from rain gauge measurement sites. Several methods are available in the literature and we propose to examine leading candidates to see if they are reasonable for this application. Should a suitable method be identified and if it can be implemented within the budget constraints of the MM5 evaluation, this approach yields graphical comparisons similar to those depicted in Figure 5-4.

A third approach involves various graphical methods for depicting the time and space variability of precipitation over an event, a day, a week, month, season or year. For example, Figure 5-5 gives a time series plots of the daily precipitation totals derived from the measured and RAMS predicted values averaged across all reporting stations in the 12 km and 24 km SAMI domains over the southeastern U.S. The biggest challenge in the use of this type of display is in selecting the appropriate averaging time to use.

As noted above, it is especially challenging to devising a sound scheme for comparing point rainfall measurements with spatially distributed (i.e., gridded) model predictions. Another method that could be adapted to the annual and episodic simulations is depicted in Figure 5-6. In this comparison, the top panel presents the gridded daily precipitation totals (in mm) across the domain for a particular day (here, 7 August 1993.) In the bottom panel, the daily total measurements (also in mm) are depicted utilizing the same color coding scheme. Comparing the predicted and observed rainfall totals on this day gives a qualitative understanding of the adequacy of the modeled precipitation rates for the period analyzed. In addition, scatter plots of predicted and observed total precipitation might be helpful (see, for example Figure 5-7).

Another methods under consideration involve calculation of cumulative predicted and observed distributions of rainfall as a function of time throughout the month, season, or annual cycle.

We invite any suggestions the EPA may have on suitable methods for evaluating the MM5 precipitation predictions using the surface NWS rain gauge measurements. We will continue to explore the methods outlined here and others that may come to our attention. When we have developed a suitable method(s) for comparing modeled versus observed precipitation for the annual and episodic simulations, we will present them as a final recommendation, to be incorporated into this protocol as a update revision.

Table 5-1. Statistical Measures and Graphical Displays to be Used in the MM5 Operational Evaluation of Surface Fields.

Statistical Measure	Graphical Display
***Surface Winds* (ms^{-1})**	
Vector mean observed wind speed	Vector mean modeled and observed wind speeds as a function of time
Vector mean predicted wind speed	Scalar mean modeled and observed wind speeds as a function of time
Scalar mean observed wind speed	Modeled and observed mean wind directions as a function of time
Scalar mean predicted wind speed	Modeled and observed standard deviations in wind speed as a function of time
Mean observed wind direction	RMSE, RMSE$_s$, and RMSE$_u$ errors as a function of time
Mean predicted wind direction	Index of Agreement as a function of time
Standard deviation of observed wind speeds	Surface wind vector plots of modeled and observed winds every 3-hrs
Standard deviation of predicted wind speeds	
Standard deviation of observed wind directions	
Standard deviation of predicted wind directions	
Total RMSE error in wind speeds	
Systematic RMSE error in wind speeds	
Unsystematic RMSE error in wind speeds	
Index of Agreement (I) in wind speeds	
SKILL$_E$ skill scores for surface wind speeds	
SKILLvar skill scores for surface wind speeds	
***Surface Temperatures* (^{0}C)**	
Maximum region-wide observed surface temperature	Normalized bias in surface temperature estimates as a function of time
Maximum region-wide predicted surface temperature	Normalized error in surface temperature estimated as a function of time
Normalized bias in hourly surface temperature	Scatter plot of hourly observed and modeled surface temperatures
Mean bias in hourly surface temperature	Scatter Plot of daily maximum observed and modeled surface temperatures
Normalized gross error in hourly surface temperature	Standard deviation of modeled and observed surface temperatures as a function of time
Mean gross error in hourly surface temperature	Spatial mean of hourly modeled and observed surface temperatures as a function of time
Average accuracy of daily maximum temperature estimates over all stations	Isopleths of hourly ground level temperatures every 3-hr
Variance in hourly temperature estimates	Time series of modeled and observed hourly temperatures as selected stations
Surface Mixing Ratio (gm/Kg)	
Maximum region-wide observed mixing ratio	Normalized bias in surface mixing ratio estimates as a function of time
Maximum region-wide predicted mixing ratio	Normalized error in surface mixing ratio estimates as a function of time
Normalized bias in hourly mixing ratio	Scatter Plot of hourly observed and modeled surface mixing ratios
Mean bias in hourly mixing ratio	Scatter Plot of daily maximum observed and modeled surface mixing ratios
Normalized gross error in hourly mixing ratio	Standard deviation of modeled and observed surface mixing ratios as a function of time
Mean gross error in hourly mixing ratio	Spatial mean of hourly modeled and observed surface mixing ratios as a function of time
Average accuracy of daily maximum mixing ratio	Isopleths of hourly ground level mixing ratios every 3-hr
Variance in hourly mixing ratio estimates	Time series of modeled and observed hourly mixing ratios at selected stations

Table 5-2. Statistical Measures and Graphical Displays to be Used in the MM5 Operational Evaluation of Aloft Fields.

Statistical Measure	Graphical Display
Aloft Winds (ms^{-1})	
Vertically averaged mean observed wind speed aloft for each sounding	Vertical profiles of modeled and observed horizontal winds at each sounding location
Vector averaged mean predicted wind speed aloft for each sounding	
Vertically averaged mean observed wind direction aloft for each sounding	
Vertically averaged mean predicted wind direction aloft for each sounding	
Aloft Temperatures (^{0}C)	
Vertically averaged mean temperature observations aloft for each sounding	Vertical profiles of modeled and observed temperatures at each sounding location
Vertically averaged mean temperature predictions aloft for each sounding	
Aloft Mixing Ratio (gm/Kg)	
Vertically averaged mean mixing ratio observations aloft for each sounding	Vertical profiles of modeled and observed mixing ratios at each sounding location
Vertically averaged mean mixing ratio predictions aloft for each sounding	

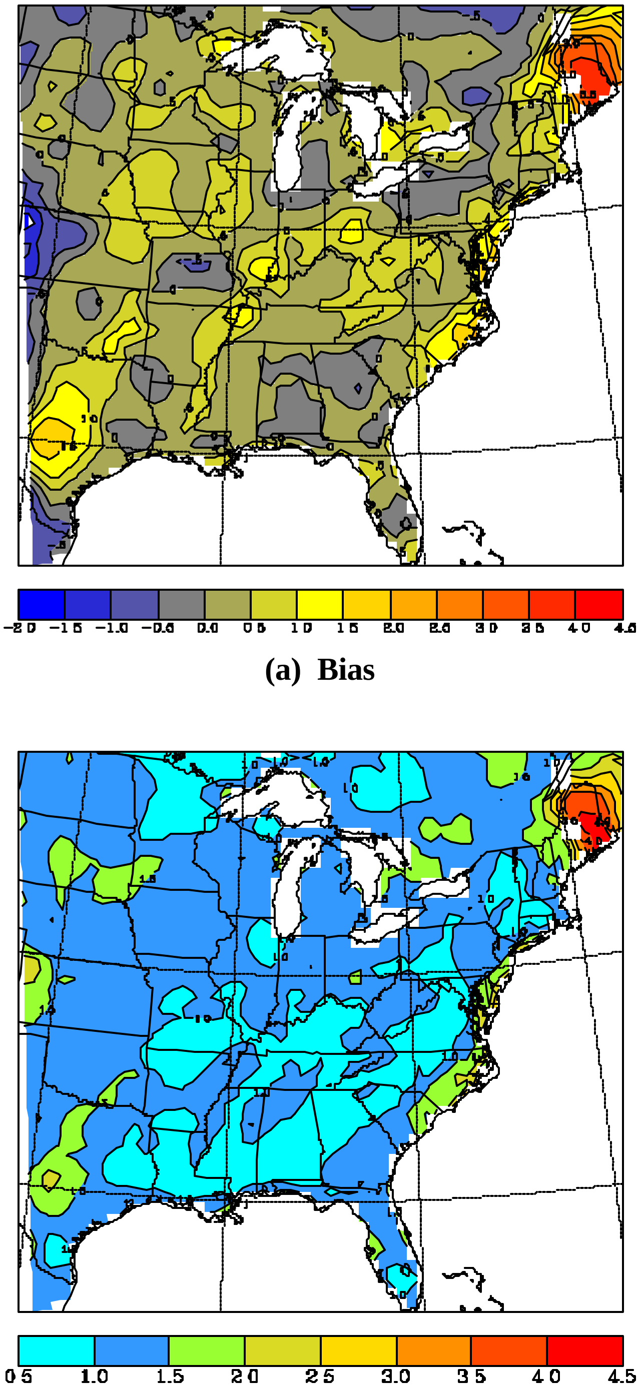
Table 5-3. Contingency Categories For Comparing 6-Hr MM5 Precipitation Estimates with 6-Hr NWS Observations.

Event	Predicted	
Observed	No Rain	Rain
No Rain	Z	F
Rain	M	H

Table 5-4. Rainfall Statistics for Various Thresholds for 6-Hr RAMS Model Precipitation Compared With 6-Hr NWS Observed Values for July 1995 episode. (Source: Doty et al., 2002).

STATISTIC	THRESHOLD (mm)								
	0.2	**2**	**5**	**10**	**15**	**25**	**35**	**50**	**75**
BIAS	1.9802	0.2018	0.0247	0	0	0	0	0	-99
POD	0.2129	0.0263	0	0	0	0	0	0	-99
FAR	0.8925	0.8696	1	-99	-99	-99	-99	-99	-99
ANR	0.9303	0.9962	0.9996	1	1	1	1	1	1
HK	0.1432	0.0225	-0.0004	0	0	0	0	0	-99
OBS EVENTS	202	114	81	48	32	19	7	3	0

Figure 5-1. Spatial Difference Fields of Bias and Root Mean Square Error in Modeled Minus Observed Surface Wind Fields (10 m) for July 1995 Over the 48 Km SAMI Grid Domain (Source: Doty et al., 2002).



(a) Bias

(b) RMSE

Figure 5-2. Comparison of Observed (red) and Predicted MM5 (blue) Upper Air Wind, Temperature, and Mixing Ratios at Jacksonville, FL on 22 April 1999 at 1500 LST (Source: Tesche et al., 2002).

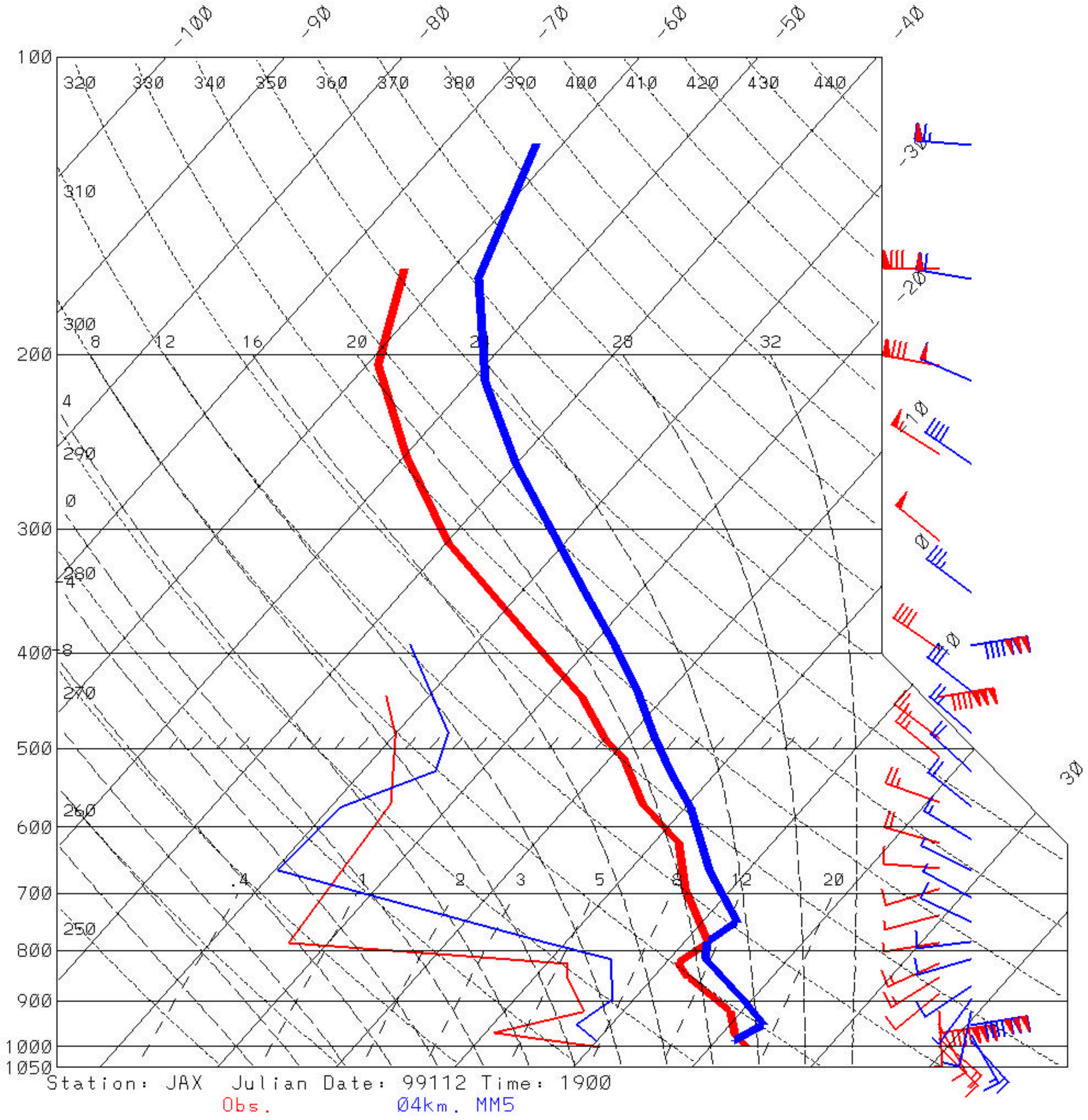
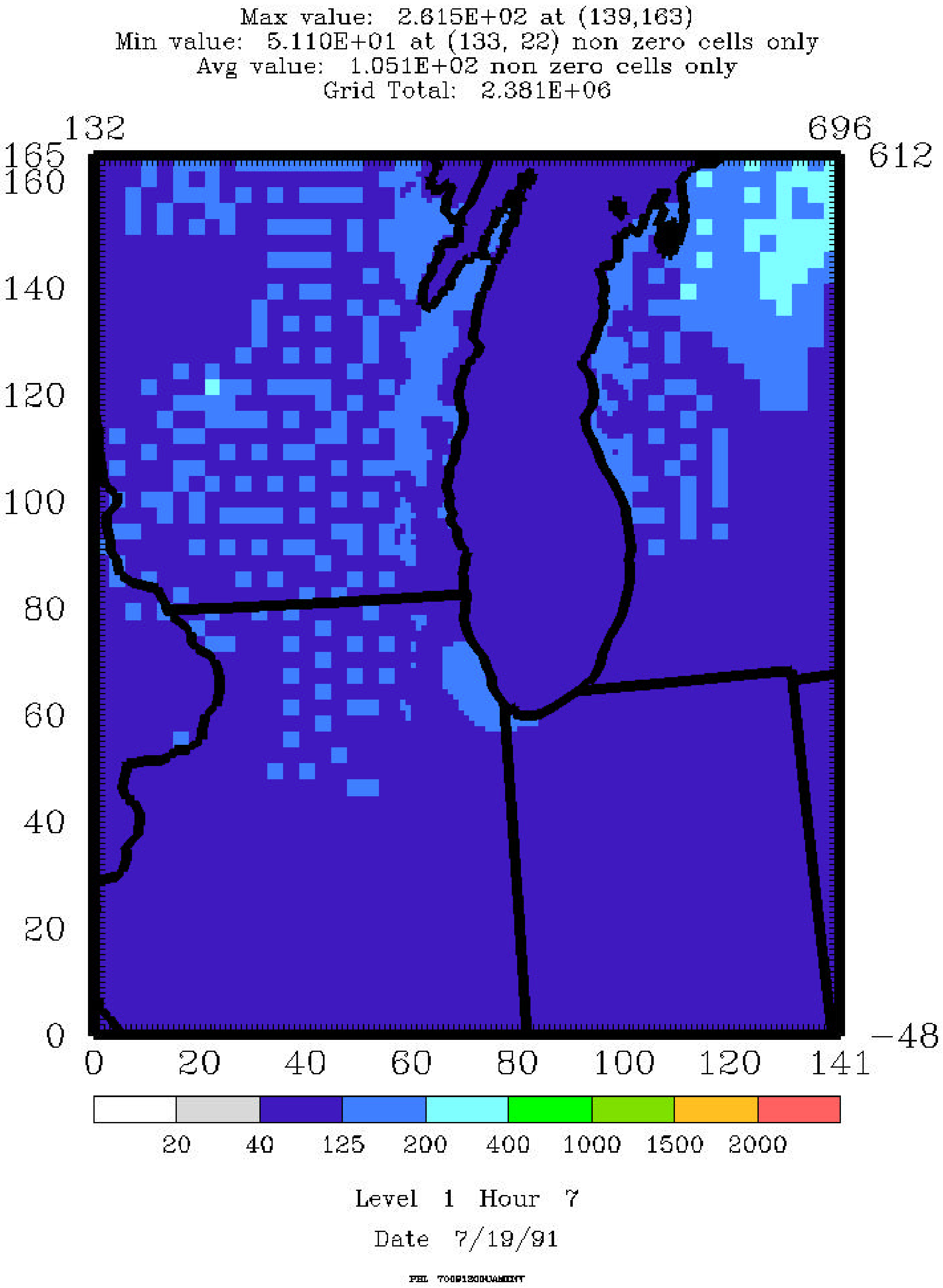
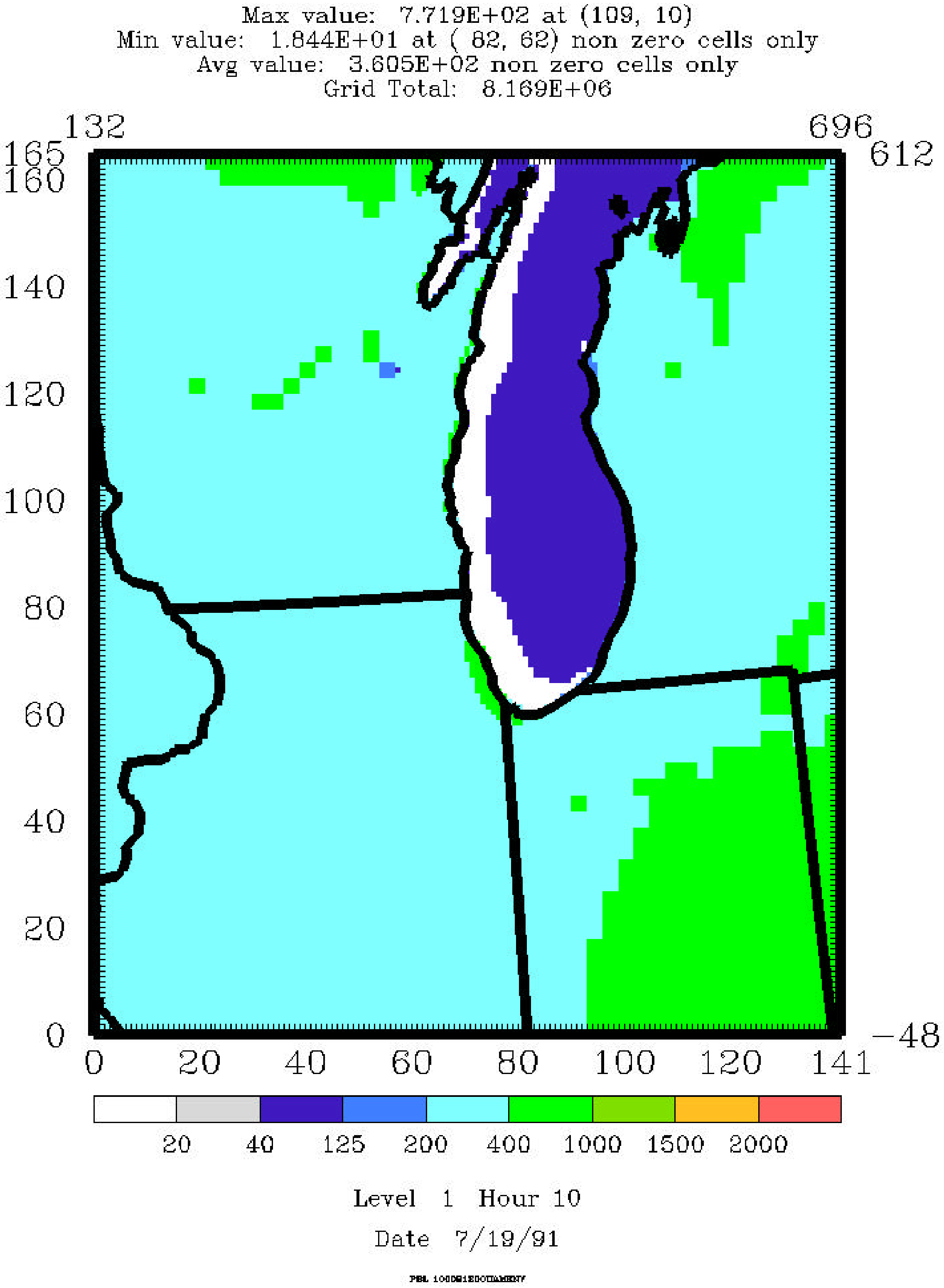


Figure 5-3. MM5 Planetary Boundary Layer Heights over the Urban-Scale (4 km) Domain on 19 July 1991 (Source: Tesche and McNally, 2001).



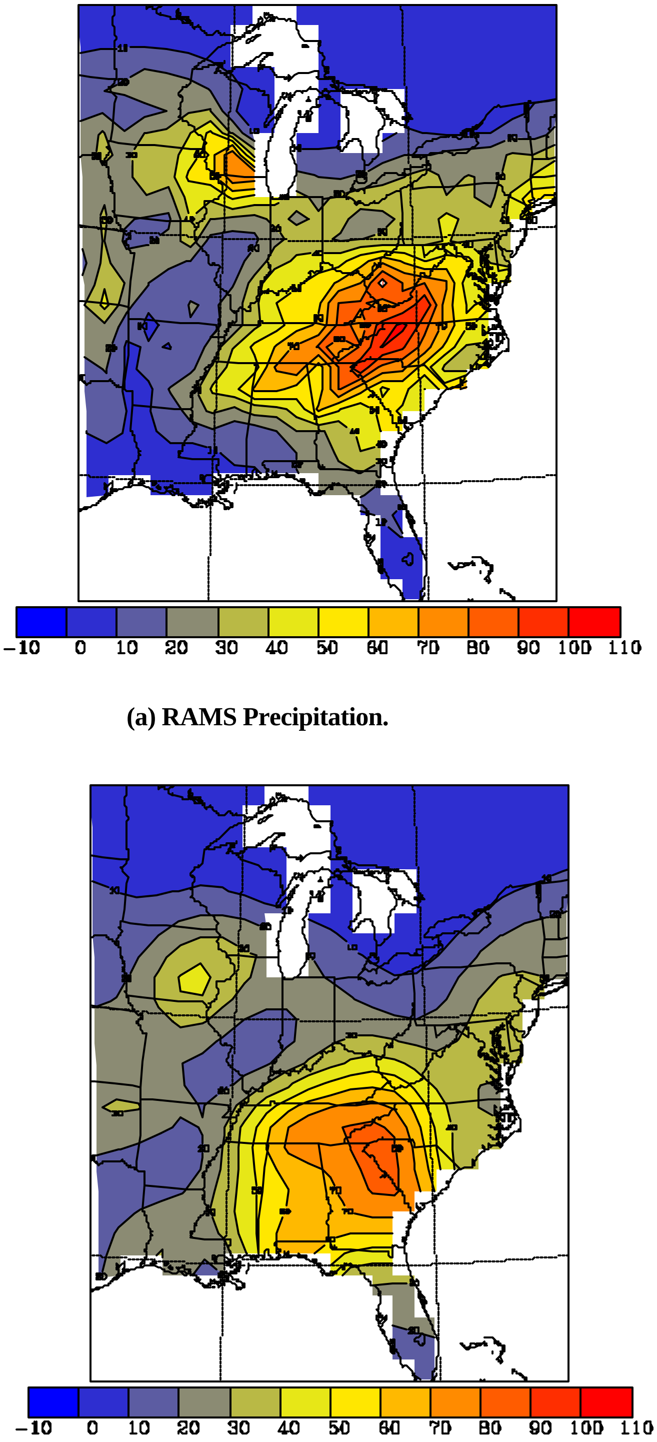
(a) 0700 CST

Figure 5-3. Concluded.



(b) 1000 CST

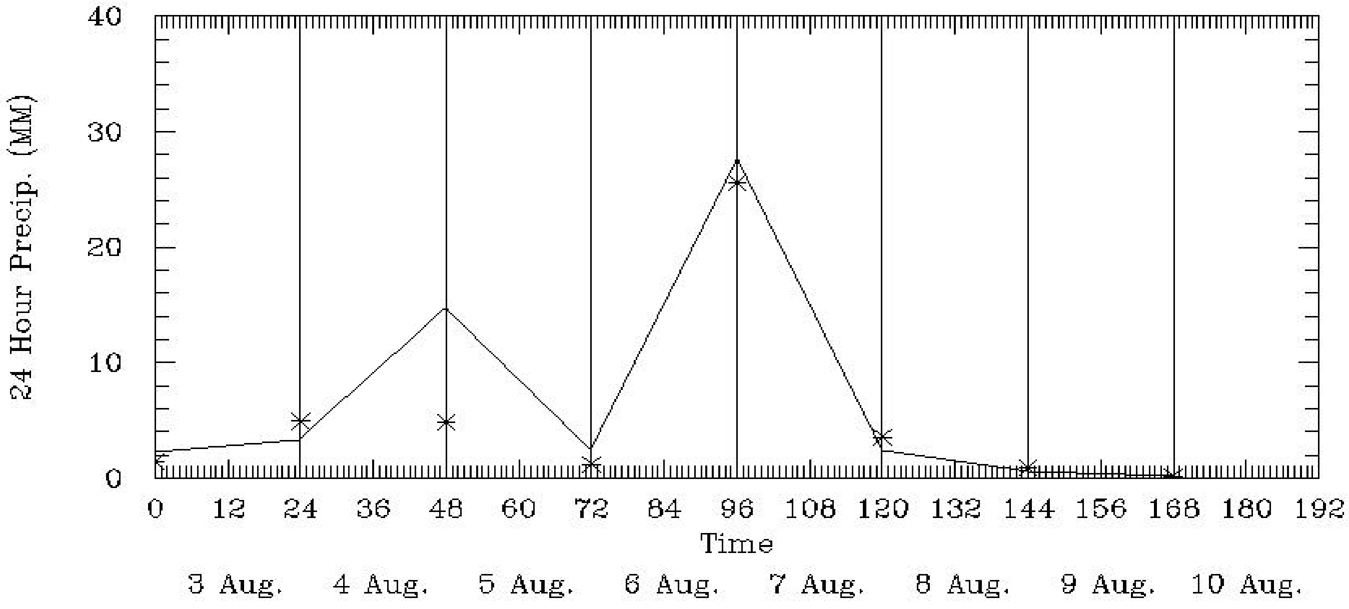
Figure 5-4. Comparison of Modeled and Spatially Analyzed Precipitation Fields (mm) Over the Southeastern U.S. for March 1993 Over a 96 km SAMI Domain. (Source: Doty et al., 2002).



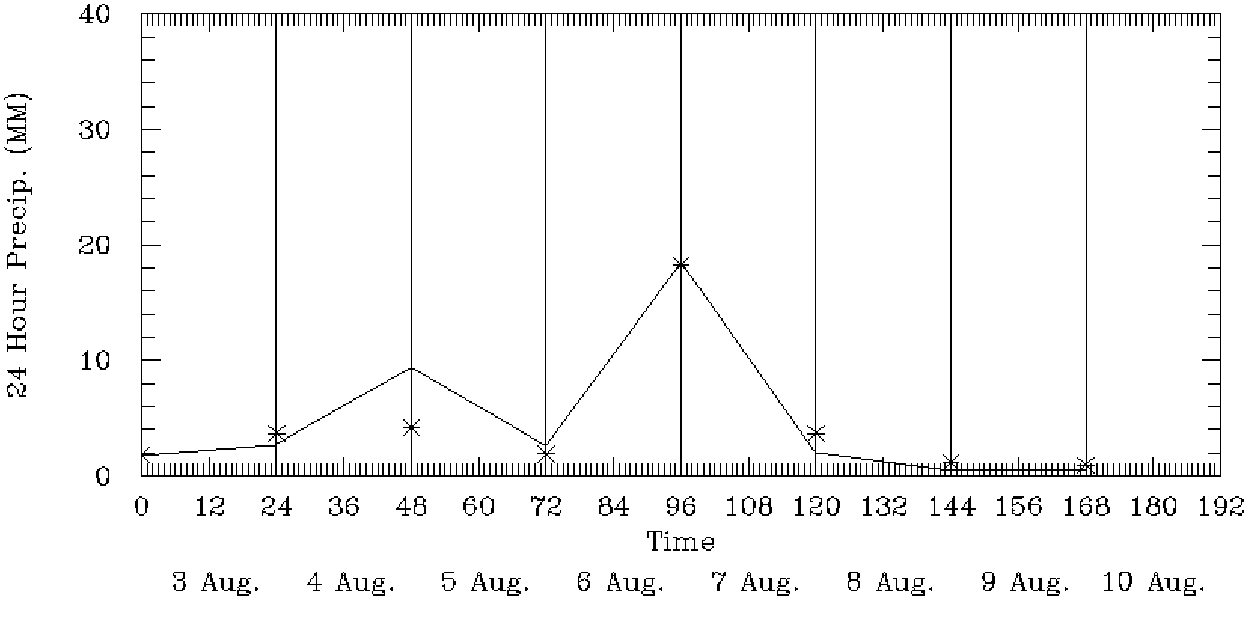
(a) RAMS Precipitation.

(b) Spatially Interpolated Precipitation.

Figure 5-5. Spatial Mean Daily Precipitation (mm) for the 3-12 August 1993 SAMI Episode. (Source: Tesche and McNally, 2002).

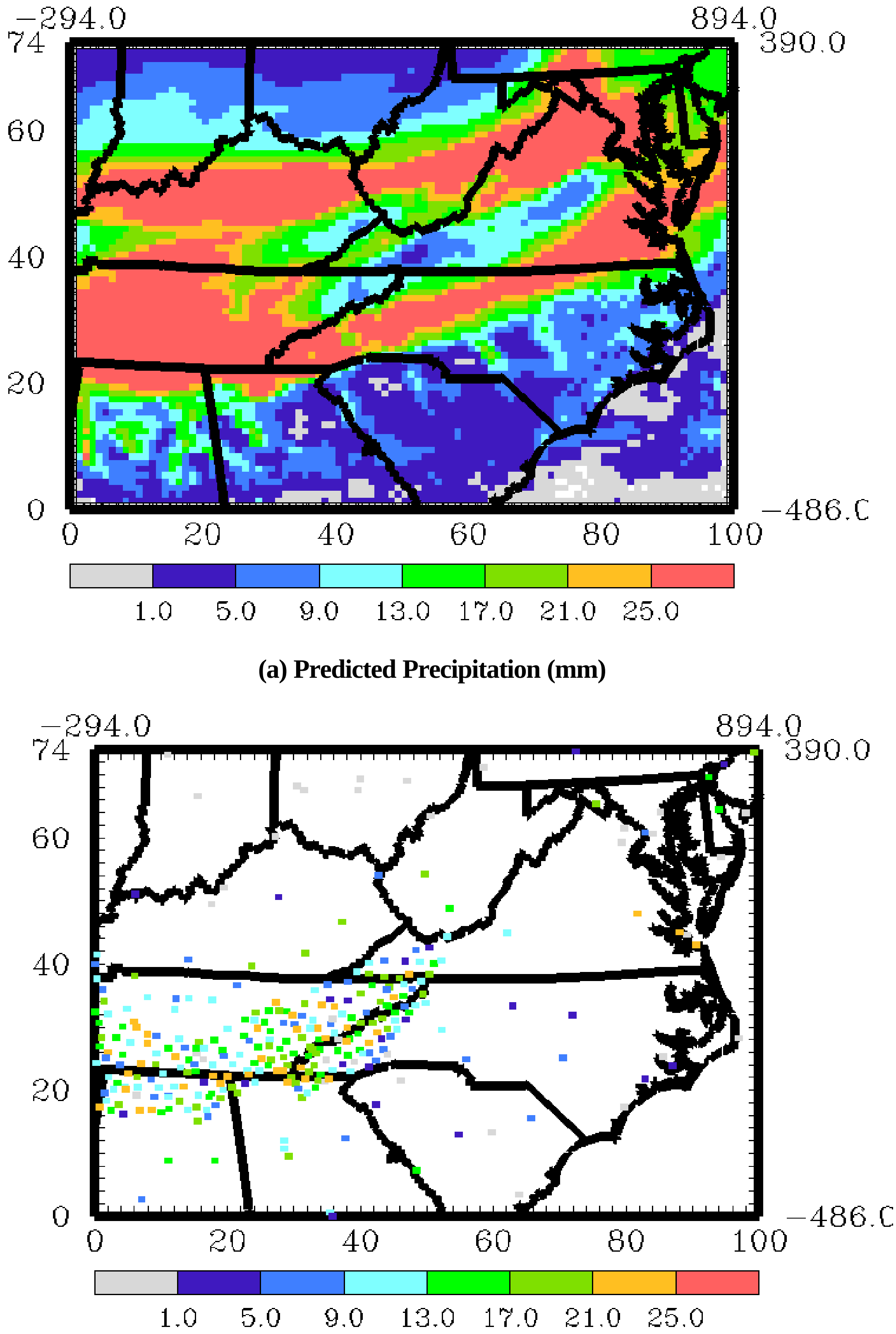


(a) 12 Km Grid



(b) 24 Km Grid

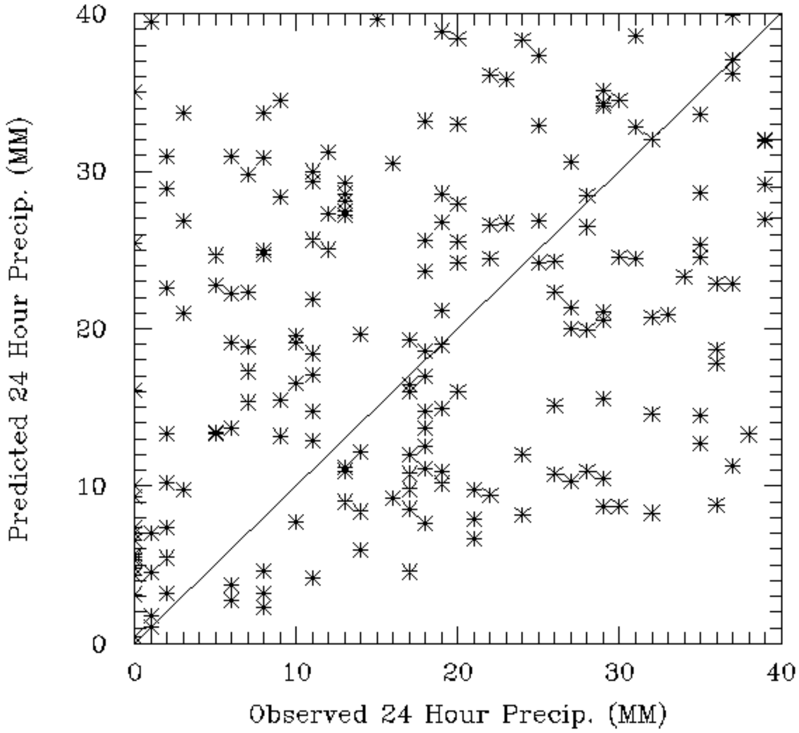
Figure 5-6. Daily Precipitation Fields for 7 August 1993. (Source: Tesche and McNally, 2002).



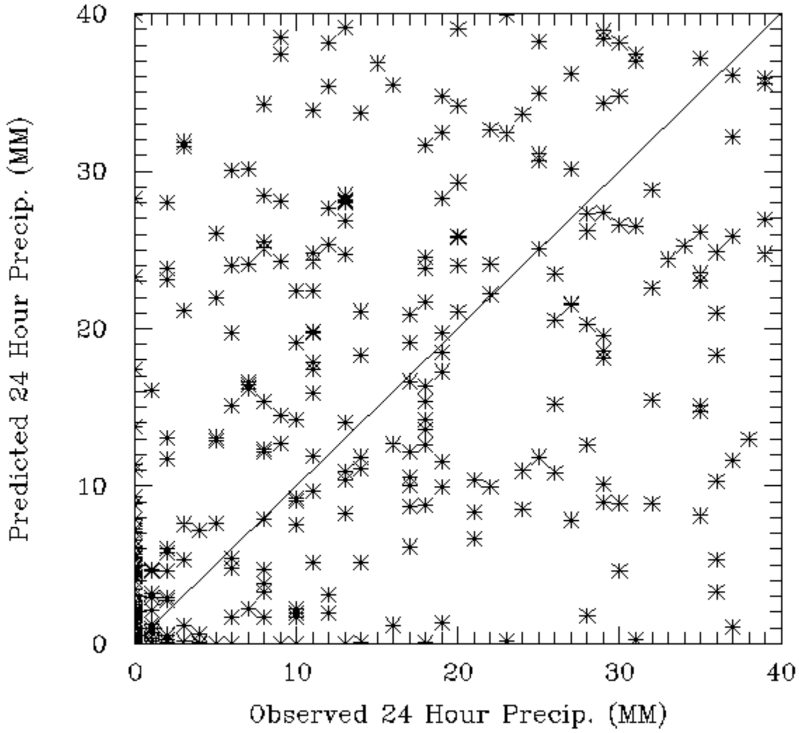
(a) Predicted Precipitation (mm)

(b) Measured Precipitation (mm)

Figure 5-7. Scatter Plot of Daily Maximum Rainfall (mm) for 7 August 1993. (Source: Tesche and McNally, 2002).



(a) 12 Km Grid



(b) 24 Km Grid

6 APPLICATION OF THE EVALUATION TOOLS

As noted in Section 2, the MM5 operational evaluation of the annual and episodic simulations will be supported by the calculation of a comprehensive set of statistical procedures and graphical displays over the episodic and annual average modeling domains (Figures 6-1 and 6-2). These methods, and pertinent plotting examples, were presented in the previous chapter. For each simulation we will evaluate surface and aloft wind (speed and direction), temperature, mixing ratio, and precipitation fields using available data sets. Primary metrics of interest include biases, gross errors, RMSE errors (total plus systematic and unsystematic components), model skill measures, the index of agreement, as well as the mean and standard deviations in modeled and observed fields. Complementing these numerical measures will be a variety of graphical displays produced by the MAPS and PAVE software. Evaluation methods specific to the annual and episodic simulations are outlined below.

6.1 Annual Evaluation Methodology

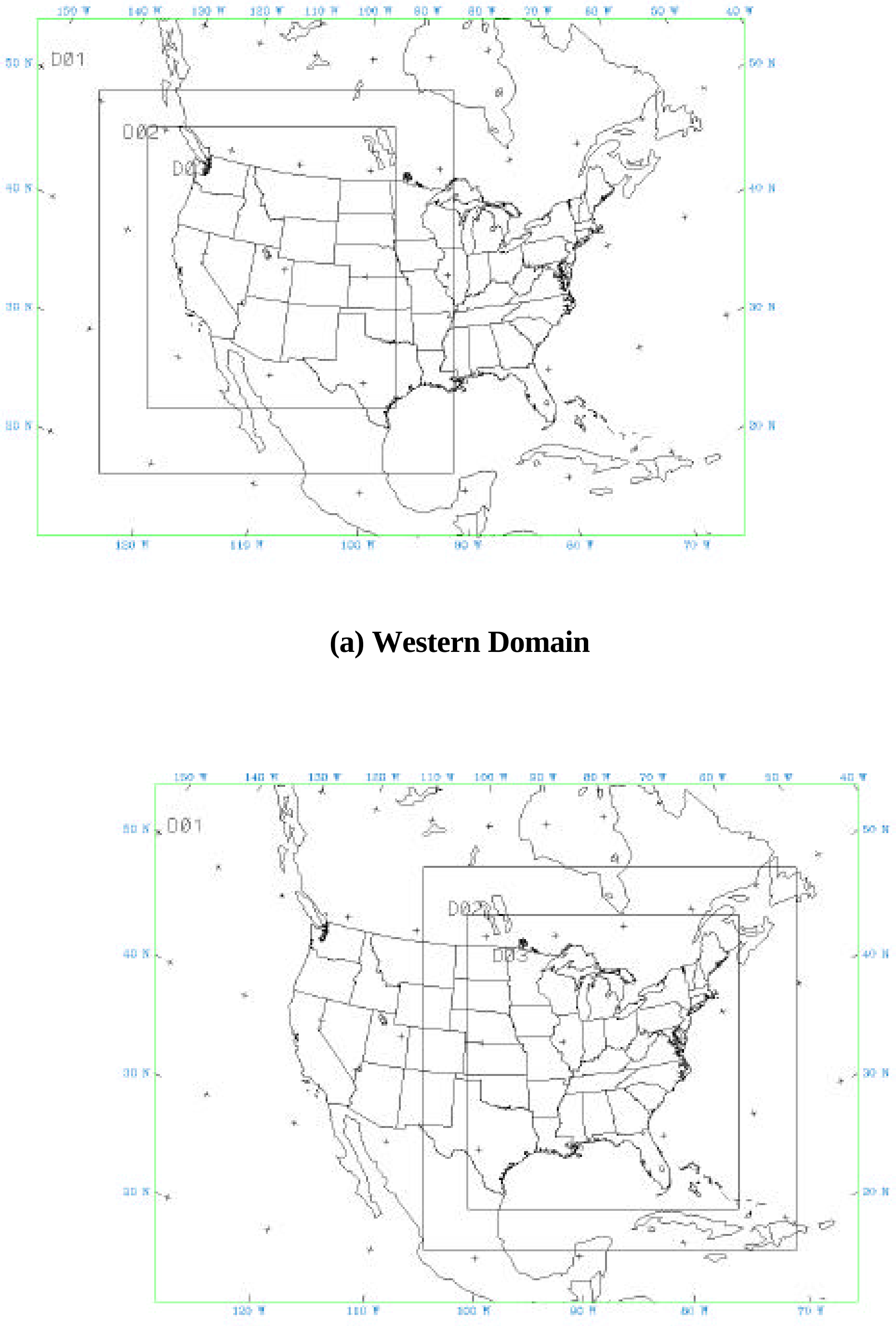
The annual MM5 evaluation will be performed with both scientific and policy perspective in mind. For the annual simulation, we will examine the model's performance across the full 36 km domain for four (4) timeframes: daily, monthly, seasonal (autumn, winter, spring, and summer), and annual averaging timeframes. We will evaluate the model over the five subregions shown in Figure 6-3. These approximately correspond to the five RPO domains. The modeling results will be evaluated for daytime versus nighttime conditions as well. The goal of these additional subregional and sub-temporal evaluations is to build confidence in the use of the model for regulatory air quality decision-making and to identify potential problem areas (should they exist) in the MM5 meteorological fields. These subregional evaluations will be aimed at elucidating the model's ability to predict key processes at the smaller time and scales (e.g. coastal circulation regimes) associated with specific RPO regions.

The full suite of statistical measures and graphical displays will be developed for each subregion and archived on CD. One purpose in archiving these subregional and subtemporal results is to provide future analysts with the basic information needed to support independent assessments of whether the MM5 fields produced in this study are actually technical credible and appropriate for use in their specific applications as opposed to the use of other modeling results or routine measurement data bases.

6.2 Episodic Evaluation Methodology

The episodic MM5 evaluations will be performed over 12 km domains but will obviously focus on narrower time scales: episodic, daily, and hourly. Though not shown presently, we will define specific subregional analysis regions within the episodic domains (analogous to those shown in Figure 6-3) to support subregional investigations and to highlight particular meteorological processes of concern to air quality modelers (e.g., land-sea breeze circulations; mountain-valley wind regimes). These subregional and sub-temporal evaluations will be focused on identifying the ability of the MM5 model to yield reliable transport fields in support of episodic air quality studies. The full suite of statistical measures and graphical displays will be developed for episodic domain, subregion, and averaging time and archived on CD.

Figure 6-1. Approximate Location of the 12 km Episodic Modeling Domains.



(a) Western Domain

(b) Eastern Domain

Figure 6-2. Approximate Location of the 36 km Annual MM5 Modeling Domain.

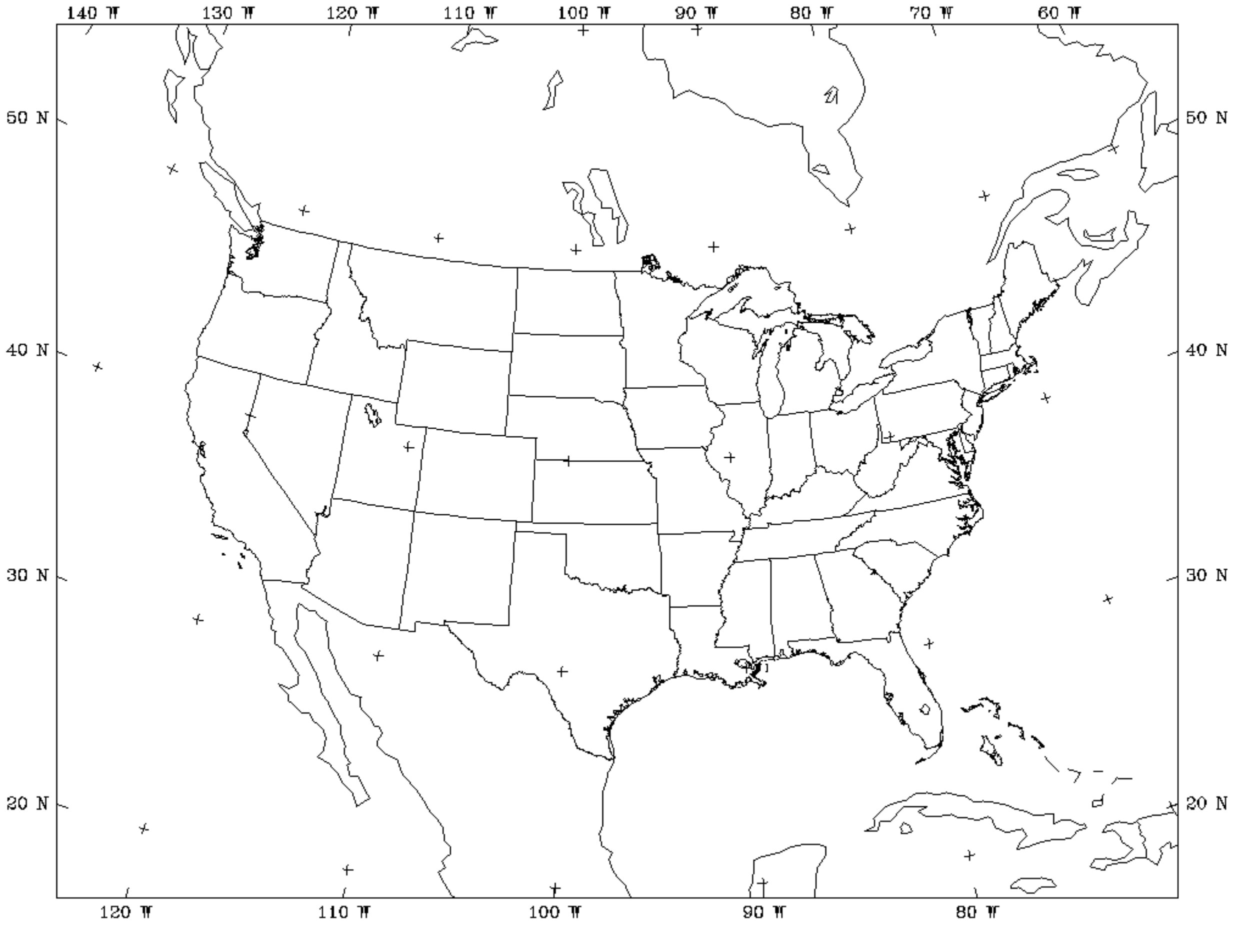
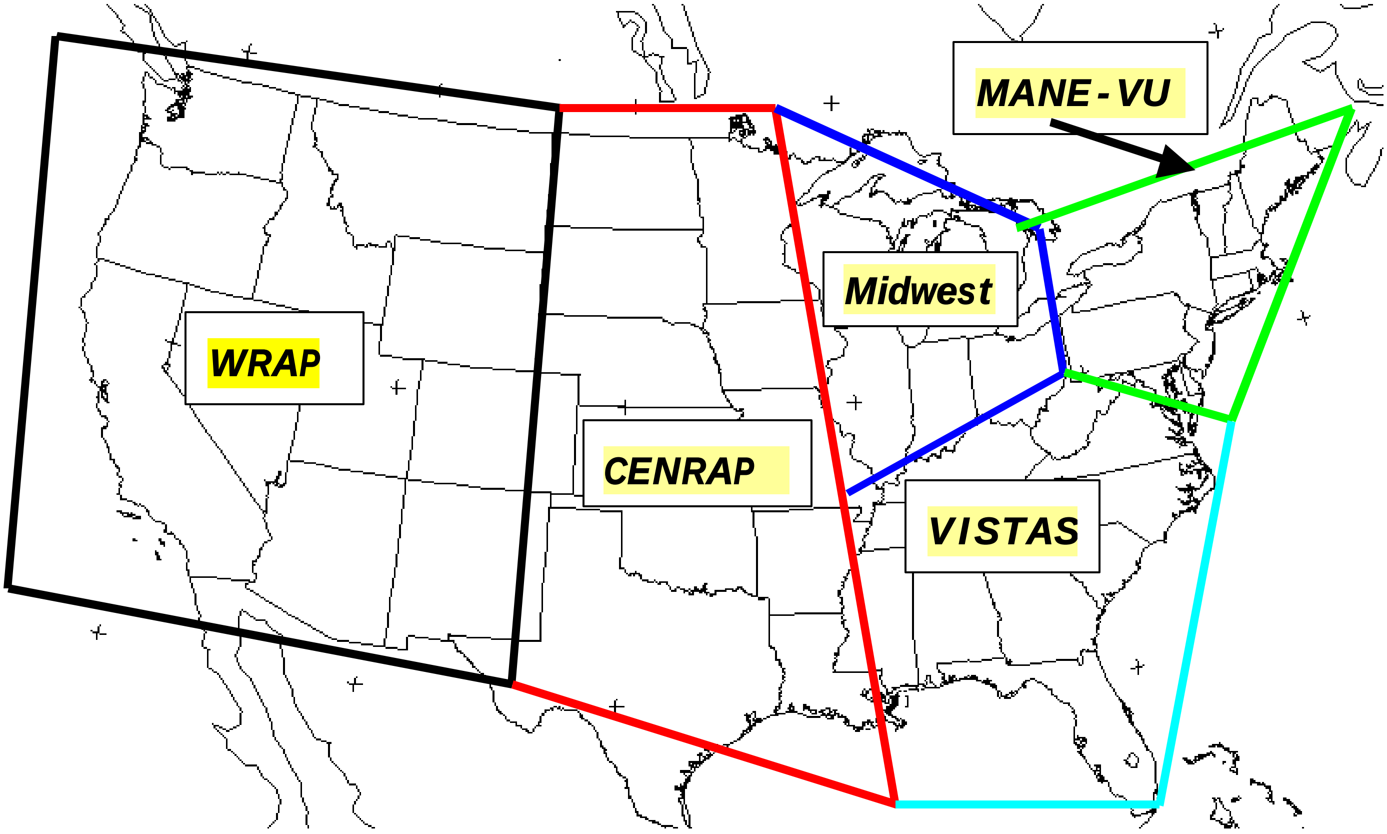


Figure 6-3. Identification of Subdomains for the Annual MM5 Performance Evaluations.



\

7 A PROCESS FOR JUDGING THE ADEQUACY OF METEOROLGICAL MODEL SIMULATIONS FOR USE IN REGIONAL AIR QUALTY MODELNG

One of the most important questions to be addressed in this study concerns whether the episodic and annual MM5 meteorological fields are adequate for their intended use in supporting 8-hr ozone, secondary fine particulate, acid deposition or other regional air quality modeling studies. For the reasons discussed below, we will not be able to answer this question definitively, yet a significant amount of information will be develop that, we believe, should be of use to modelers and decision-makers. Through implementation of the methods and procedures set forth in this protocol, we believe that a framework will be established that will help users of the MM5 modeling results to determine the suitability of the results for their specific air quality model application.

7.1 Process for Judging the Suitability of MM5 Fields for Air Quality Applications

Once the full episodic and annual MM5 evaluations are completed in this project, there will still not be a definitive answer as to whether the fields are adequate as input to air quality models. This will result, in part, because there are currently no commonly accepted performance criteria for prognostic meteorological models that, if passed, would allow one to declare the MM5 fields appropriate for use[1]. For complex atmospheric modeling problems like the ones being addressed at regional scale by 8-hr ozone, secondary aerosol, haze, and acid deposition studies, it is quite doubtful that a single quantitative set of performance criteria would ever be sufficient. The question of meteorological field adequacy depends, at a minimum, upon the specific host air quality model and the nature of the modeling episode being used. Meteorological fields that might be adequate for use in the CMAQ model for an eastern U. S. episode, for example, may be quite deficient in an episode over the Gulf Coast states region since the specific needs of the air quality model and the particular chemical and physical processes that must be simulated may be quite different. Thus, quantitative statistical and graphical performance criteria, though helpful, are inherently insufficient in aiding modelers and decision-makers in deciding whether meteorological fields are adequate for air quality modeling. Other considerations must be brought to bear. Below, we present a process whereby the adequacy of the MM5 fields for use in a particular air quality study might be assessed. This process builds upon the more general evaluation process outlined by Roth, Tesche and Reynolds (1998) and implemented in connection with the RAMS/URM model applications for the Southern Appalachian Mountains Initiative (SAMI) (Doty et al., 2002) and with the MM5/CAMx applications in the 8-hr Peninsular Florida Ozone Study (PFOS) (Tesche et al., 2002).

7.2 Framing the Questions to Be Addressed

Usually air quality simulations are quite sensitive to meteorological fields. Where this sensitivity is anticipated, it is important to make an effort to develop as accurate a representation

[1] A proposed set of performance criteria has been suggested by Emery et al., (2001) drawing on the work of Tesche et al., (2001) and these will be used in the present study to gauge MM5 performance. However, outside the state of Texas, these criteria have not yet received wide-spread endorsement.

of meteorological variables as possible. Special features of the flow fields, such as eddies, nocturnal jets, drainage flows, land-sea or land-bay breezes, and vertical circulations should be adequately characterized through the meteorological modeling. In circumstances where there are significant transitions in the meteorological variables over short distances, such as along shorelines or in areas of hilly terrain, the need for finer spatial resolution that is typically specified must be considered. If inadequate attention and care are accorded meteorological modeling, there is a significant risk of developing an inaccurate representation that will be propagated into the emissions and air quality models.

Several questions should be addressed for the specific application. Examples of these questions are as follows:

Appropriateness of Model Selection:

> **Modeling Requirements**: Was a carefully written characterization made of the most important physical and chemical processes relevant to successful air quality modeling of each episode (e.g., a "conceptual model" of each simulation period)?

> **Model Selection:** Did the model selection process ensure that a suitable modeling system was chosen, properly weighing the need for mature, well-tested, publicly-available model(s) against the constraints of the specific modeling problem, characteristics of the episodes to be simulated, and the limitations of schedule and resources?

> **Model Formulation Review:** Was a rigorous evaluation and inter-comparison made between the scientific formulation of the proposed meteorological modeling system (source codes plus pre- and post-processors) versus alternative contemporary prognostic models via an open, thorough scientific review process?

> **Code Verification**: Was the fidelity of the computer coding of the proposed model confirmed with respect to its scientific formulation, governing equations, and numerical solution procedures?

Identification of Air Quality Model Needs:

> **Air Quality Input Needs:** Were the meteorological input needs of the host air quality model and supporting emissions models (e.g., biogenic, motor vehicle, area source processors) clearly identified including specification of the requisite averaging times and nested grid scales for the specific modeling episodes?

> **Air Quality Model Sensitivities:** Was the air quality model's sensitivity to key meteorological inputs established through careful consideration (including air quality model sensitivity/uncertainty simulations) of the relevant modeling episodes over the specific domain of interest? Was the effect of uncertainty in those meteorological inputs to which the air quality model is demonstrated to be most sensitive adequately defined through appropriate numerical experiments or from previous relevant studies?

Note: Identification of air quality model needs is a crucial step in the meteorological model evaluation process, yet it is most often performed superficially if at all. Pragmatic constraints of time and resources necessitate that efforts be directed at achieving the best possible meteorological performance for those variables that matter most to the overall accuracy and reliability of the air quality model. With one important exception, there is little practical benefit to be gained in devoting considerable time to improving the accuracy of a particular meteorological variable if the air quality model – in the specific application at hand -- is insensitive to that variable. The exception, obviously, is that since the meteorological variables are all inter-related though the coupled mass, momentum and energy equations, one cannot simply neglect a particular variable's importance altogether since it is tied in direct and subtle ways to the other variables. One cannot accept obviously flawed performance for one meteorological variable without exploring the impact that this has on the reliability of the other dynamical variables to which the air quality model may be especially sensitivity. Thus, particular attention should be given to those meteorological variables that have the largest uncertainty and to which the air quality model is most sensitive. This challenge can be particularly formidable when dealing with photochemical/aerosol models whose concentration and/or deposition estimates depend on several meteorological variables (mixing, transport, thermodynamic properties, precipitation) simultaneously.

Availability of Supporting Data Bases:

> **Adequate Data Available:** Were sufficient data available to test, at the ground and aloft and over all nested grid scales of importance, the model's dynamic, thermodynamic, and precipitation-related fields?

> **All Data Used**: Was the full richness of the available data base actually utilized in the input data file development, in FDDA, and in the evaluation of model performance?

Note: One of the main considerations underlying selection of modeling episodes for regulatory decision-making is the availability of special data collection programs to supplement the surface and aloft data routinely available from state and federal agencies. While attempts are made to select modeling episodes that coincide with intensive field measurement programs, in these situations it is common that the full set of supplemental measurements are not used thoroughly in the model input development and performance testing phases. At times, the availability of 'high-resolution' databases is touted in support of a particular episode selection choice yet when the modeling is actually performed and evaluated, only a fraction of the special studies data are actually used. This is most notably the case with air quality and meteorological data collected by aloft sampling platforms. Unless the high-resolution data are actually used to enhance the modeling and performance testing, their value is severely limited. Equally troublesome, selection of other candidate modeling days (supported by only routine information) may be overlooked which might otherwise be preferable modeling periods if a concerted effort to utilize special studies data is not made. Finally, as desirable as having supplemental meteorological measurements might be, unless the sampling was performed in the correct regions and includes the variables of primary importance to the air quality model, their potential to add meaningfully to the rigor of the modeling exercise will be limited. Thus, when judging the value of supplemental measurement programs, it is necessary to look beyond just their mere existence (relative to non-intensively monitored days); one must establish that these intensive data set indeed contribute to improved model performance and increased reliability. This necessitates a feedback loop to the air quality modeling exercise to ensure that the times, locations, and parameters associated with the supplemental measurements truly add to the overall quality and rigor of the study.

Results of Operational, Diagnostic, and Scientific Evaluations:

> **Full Model's Predictive Performance:** Was a full suite of statistical measures, graphical procedures, and phenomenological explorations performed with each of the models state variables and diagnosed quantities for each pertinent grid nest to portray model performance against available observations and against model estimates from other relevant prognostic simulation exercises?

> **Performance of Individual Modules:** Was there an adequate evaluation of the predictive performance of individual process modules and preprocessor modules (e.g., advection scheme, sub-grid scale processes, closure schemes, planetary boundary layer parameterization, FDDA methodology)?

> **Diagnostic Testing:** Were sufficient and meaningful diagnostic, sensitivity, and uncertainty analyses performed to assure conformance of the meteorological modeling system with known or expected behavior in the real world?

> **Mapping Methods:** Were parallel evaluations made of: (a) the output from the prognostic model and (b) the output from the 'mapping' routines that interpolate the prognostic model output onto the host air quality model's grid structure? Were any important differences between the two reconciled?

> **Quality Assurance:** Was a credible quality assurance (QA) activity implemented covering both the prognostic modeling activity as well as the mapping programs that generate air quality-ready meteorological inputs? Was the full set of hourly, three-dimensional fields examined for reasonableness even though observational data for comparison were lacking or in short supply?

Note: Such an intensive performance evaluation process is rarely, if ever, carried out due to time, resource and data base limitations. Nevertheless, it is useful to identify the *ideal* evaluation framework so that the results of the *actual* evaluation can be judged in the proper perspective. This also allows decision-makers to establish realistic expectations regarding the level of accuracy and reliability associated with the meteorological and air quality modeling process.

Comparison with Other Relevant Studies:

> **Comparisons with Other Studies:** Were the model evaluation results (statistical, graphical, and phenomenological) compared with other similar applications of the same and alternative prognostic models to identify areas of commonality and instances of differences between modeling platforms?

Note: Reflecting limited data sets for performance testing and reliable criteria for judging a model's performance, meteorological model evaluations in recent years have emphasized comparisons with other RAMS and MM5 simulations over various modeling domains and episode types as a means of broadening the scope of the evaluation. While this insight into the model's performance – when gauged against other similar applications – is useful, caution must attend such comparisons which at times are at best anecdotal. Often the reporting of previous evaluations entails grossly composited performance statistics (episode averages or averages across episodes, for example), data bases and modeling efforts of widely varying and often unreported quality, different mathematical definitions of statistical quantities, and so on. Thus, these comparisons with other studies, while occasionally providing useful perspective, are by no means sufficient for declaring a meteorological model's performance to be reliable and acceptable in a particular application. Moreover, meteorological model evaluation benchmarks developed on the basis of such historical evaluation studies must also be applied thoughtfully with these limitations in mind.

Peer Review of Specific Modeling Exercise(s):

> **Scope of Peer Review**: Was an adequate, properly-funded, independent, in-depth peer review of the model set-up, application, and performance evaluation efforts conducted?

> **Findings of Peer Review**: Was the effort judged acceptable by the peer-review?

Note: Prognostic modeling requires considerable attention to detail, careful identification of options, and complete involvement in the work. Even with this commitment, critical aspects of a modeling exercise may be treated inadequately or overlooked, most often as the result of schedule or resource constraints. Consequently, an examination of the meteorological modeling effort conducted at arm's length by individuals with appropriate expertise and who have no personal involvement in the work can be essential to avoiding inadvertent oversights and problems. Such a peer review of the effort provides another check on the work as a whole. If concerns are raised about the reliability of the modeling, yet meteorological modeling results are to be used in applying air quality models despite these concerns (e.g., due to project schedule demands), the peer review can assist in suggesting to decision-makers the weight to be given the overall air quality results the planning and management context.

Often, when a professional paper is written describing the modeling study, it undergoes "peer review" by the journal. Such efforts do not constitute the review suggested here. Journal peer review usually entails a reading of the paper, thoughtful reflection, and written commentary, perhaps a 4- to 12-hour effort. Moreover, reporting in the professional literature is necessarily condensed, and much of the detail that should be scrutinized is omitted. This is especially true for complex atmospheric modeling projects. Peer review for pre-print volumes (e.g., American Meteorological Society or Air and Waste Management Association conferences) is even less rigorous, often consisting of a cursory reading of the paper by the Session Chairperson. Peer review, as used here, refers to detailed examination and evaluation of the work conducted by experts in the field. Such experts are generally, but not limited to, those with considerable direct experience in the development, evaluation, and application of the same or very similar meteorological models. This in depth review entails the independent scientists (a) thoroughly examining the conceptual model(s) and modeling protocols prepared for the study, (b) obtaining and examining the details of the model input and output files, and (c) in many cases even running the pre- and post-processor codes and the main simulation programs to corroborate reproducibility of results and to explore inevitable technical issues that arise in such comprehensive reviews. In essence, peer review refers to immersing oneself in the materials provided. Such an effort can take *several weeks* to carry out properly.

Overall Assessment:

> **Overall Reasonableness**: Has an adequate effort been made to evaluate the quality of representation of meteorological fields generated using the meteorological model, as revealed by the full suite of statistical, graphical, phenomenological, diagnostic, sensitivity, and uncertainty investigations? What were the strengths and limitations of the actual model performance evaluation?

> **Fulfillment of Air Quality Model Needs**: How well are the fields represented, particularly in areas and under conditions for which the air quality model is likely to be sensitive?

> **Appropriate Model**: Was a sound and suitable meteorological modeling system adopted?

> **Adequate Data Base**: Was the supporting database adequate to meet input and evaluation needs?

> **Adequate Application Procedures:** Was Four Dimensional Data Assimilation (FDDA) a part of the overall modeling approach and were sufficient data available to support the activity adequately? Is the FDDA application modifying the meteorological fields base on corrections due to stochastic (or other) processes which the meteorological model is not formulated to handle, or is the FDDA correcting errors that are caused by the model formulation, structure, or configuration?

> **Quality Assurance**: Were error-checking procedures instituted, followed, and the results reported?

> **Performance Evaluation:** Were suitable procedures specified and adopted for evaluating the quality (e.g., accuracy, precision, and uncertainty) of model estimates?

Were the evaluation procedures sufficiently stressful to identify the existence of hidden, internal compensating errors in the model and/or its inputs?

> **Judging the Overall Process**: Were the criteria (i.e., benchmarks) used to judge performance appropriate for the specific air quality model application, rigorously applied, and properly communicated?

7.3 Comparison of MM5 Performance Against Newly Proposed Meteorological Model Performance Benchmarks

As discussed previously, there are no currently accepted performance criteria for prognostic meteorological models. In addition, there is valid concern that establishment of such criteria, unless accompanied with a careful evaluation process such as the one outline in this section might lead to the misuse of such goals as is occasionally the case with the accuracy, bias, and error statistics recommended for judging photochemical dispersion models. In spite of this concern, there remains nonetheless the need for some benchmarks against which to compare new prognostic model simulations.

In two recent studies (Tesche et al., 2001b; Emery et al., 2001), an attempt has been made to formulate a set mesoscale model evaluation benchmarks based on the most recent MM5/RAMS performance evaluation literature. The purpose of these benchmarks is not to assign a passing or failing grade to a particular meteorological model application, but rather to put its results into a useful context. These benchmarks may be helpful to decision-makers in understanding how poor or good their results are relative to the range of other model applications in other areas of the U.S. Certainly an important concern with the EPA guidance statistics for acceptable photochemical performance is that they are relied upon much too heavily to establish an acceptable model simulation of a given area and episode. Often lost in routine statistical ozone model evaluations is the need to critically evaluate all aspects of the model via the diagnostic and process-oriented approaches. The same must be stressed for the meteorological performance evaluation. Thus, the appropriateness and adequacy of the following benchmarks should be carefully considered based upon the results of the specific meteorological model application being examined.

Based upon the above considerations, the benchmarks suggested from the studies of Emery et al, (2001) and Tesche et al., (2001) are as follows:

Parameter	***Measure***	***Benchmark***
Wind Speed	**RMSE:**	**≤ 2 m/s**
	Bias:	**≤ ±0.5 m/s**
	IOA:	**³ 0.6**
Wind Direction	**Gross Error:**	**≤ 30 deg**
	Bias:	**≤ ±10 deg**
Temperature	**Gross Error:**	**≤ 2 K**
	Bias:	**≤ ± 0.5 K**
	IOA	**³ 0.8**
Humidity	**Gross Error:**	**≤ 2 g/kg**
	Bias:	**≤ ±1 g/kg**
	IOA:	**³ 0.6**

For each simulation, we propose to compare the MM5 results with these benchmarks. This comparison is not aimed at determining pass/fail status of a simulation day or group of days; rather, the intent is to provide an additional means of comparing the MM5 results in a quantitative manner with previous work. In the final analysis, it will be the responsibility of the users of the MM5 results to determine whether the results developed in this study are reliable and sufficiently accurate for their intended usage. To this end, a significant body of model evaluation information will be archived to support these application-specific examinations.

7.4 Assessing a Specific Model Application Need

To assist model users and decision-makers in judging the suitability of the MM5 results for their intended applications, we will initiate a process whereby this judgment can be developed in a logical manner. Table 7-1 presents a template containing the two-dozen questions that we suggest should be addressed by potential users of the MM5 meteorological fields. In performing this application-specific analysis, the user will wish to consider the various statistical, phenomenological, and graphical results presented in the final report and companion CD's which contain full details of the evaluations. However, mere statistics and graphical alone will not produce a reliable assessment of whether the MM5's use is justified in a particular application. The air quality modeling group must conduct their own analyses of their particular situation, examining (as noted above) such things as the specific chemical and physical process that must be treated well, the appropriate time and space scales of the modeling, and so on. By performing this in-depth appraisal of the air quality modeling, the modeler can then ensure that a sound match exists between the episodic or annual MM5 simulations produced in this study and his/her particular meteorological data input needs. For examples of how this process was carried out in two recent regional applications with the MM5 and RAMS, model, the reader is referred to recent studies by Doty et al. (2002) and Tesche et al. (2002)

Table 7-1. Questions to Be Addressed in Assessing the Adequacy of MM5 Meteorological Fields in a Specific Air Quality Modeling Application.

No.	Question	Assessment
	Appropriateness of Model Selection	
1	**Was a careful written characterization made of the most important physical and chemical processes relevant to successful air quality modeling of each episode (e.g., a "conceptual model" of each simulation period)?**	
2	**Did the model selection process ensure that a suitable modeling system was chosen, properly weighing the need for mature, well-tested, publicly-available model(s) against the constraints of the specific modeling problem, characteristics of the episodes to be simulated, and the limitations of schedule and resources?**	
3	**Was a rigorous evaluation and inter-comparison made between the scientific formulation of the proposed meteorological modeling system (source codes plus pre- and post-processors) versus alternative contemporary prognostic models via an open, thorough scientific review process?**	
4	**Has the fidelity of the proposed model's computer code's scientific formulation, governing equations, and numerical solution procedures been confirmed?**	
	Identification of Air Quality Model Needs	
5	**Were the meteorological input needs of the host air quality model and supporting emissions models (e.g., biogenic, motor vehicle, area source processors) clearly identified including specification of the requisite average times and nested grid scales for the specific modeling episodes?**	
6	**Was the air quality model's sensitivity to key meteorological inputs established through careful consideration (including air quality model sensitivity/uncertainty simulations) of the relevant modeling episodes over the specific domain of interest? Was the effect of uncertainty in those meteorological inputs to which the air quality model is demonstrated to be most sensitive adequately defined through appropriate numerical experiments or from previous relevant studies?**	
	Availability of Supporting Data Bases	
7	**Were sufficient data available to test, at the ground and aloft and over all nested grid scales of importance, the model's dynamic, thermodynamic, and precipitation-related fields?**	

8	**Was the full richness of the available data base actually utilized in the input data file development, in FDDA, and in the evaluation of model performance?**	
	Results of Operational, Diagnostic, and Scientific Evaluations	
9	**Was a full suite of statistical measures, graphical procedures, and phenomenological explorations performed with each of the models state variables and diagnosed quantities for each pertinent grid nest to portray model performance against available observations and predictions from other relevant prognostic modeling exercises?**	
10	**Was there an adequate evaluation of the predictive performance of individual process modules and preprocessor modules (e.g., advection scheme, sub-grid scale processes, closure schemes, planetary boundary layer parameterization, FDDA methodology)?**	
11	**Were sufficient and meaningful diagnostic, sensitivity, and uncertainty analyses performed to assure conformance of the meteorological modeling system with known or expected behavior in the real world?**	
12	**Were parallel evaluations made of (a) the output from the prognostic model and (b) the output from the 'mapping' routines that interpolate the prognostic model output onto the host air quality model's grid structure? Were sources of differences between the two reconciled?**	
13	**Was a credible quality assurance activity implemented covering both the prognostic modeling activity as well as the mapping programs that generate air quality-ready meteorological inputs? Was the full set of hourly, three-dimensional fields examined for reasonableness even though observational data for comparison are lacking or in short supply?**	
	Comparison with Other Relevant Studies	
14	**Were the model evaluation results (statistical, graphical, and phenomenological) compared with other similar applications of the same and alternative prognostic models to identify areas of commonality and instances of differences between modeling platforms?**	
	Peer Review of Specific Modeling Exercise(s)	
15	**Was an adequate, independent, in-depth peer review of the model set-up, application, and performance evaluation efforts conducted?**	
16	**Was the effort judged acceptable by the peer-review?**	
	Overall Assessment	
17	**Has an adequate effort been made to evaluate the quality of representation of meteorological fields generated using the meteorological model, as revealed by the full suite of statistical,**	

	graphical, phenomenological, diagnostic, sensitivity, and uncertainty investigations? What were the strengths and limitations of the actual model performance evaluation?	
18	How well are the fields represented, particularly in areas and under conditions for which the air quality model is likely to be sensitive?	
19	Was a sound and suitable meteorological modeling system adopted?	
20	Was the supporting database adequate to meet input and evaluation needs?	
21	Was Four Dimensional Data Assimilation (FDDA) a part of the overall modeling approach and were sufficient data available to support the activity adequately?	
22	Were error-checking procedures instituted, followed, and the results reported?	
23	Were suitable procedures specified and adopted for evaluating the quality (e.g., accuracy, precision, and uncertainty) of model estimates?	
24	Were the criteria used to judge performance appropriate for the specific air quality model application, rigorously applied, and properly communicated?	

8 REPORTING PROCEDURES

The reporting of the episodic and annual average modeling will consist of three components: (a) an episodic MM5 evaluation report, (b) an annual MM5 evaluation report, and (c) a set of CD's containing the full set of processed evaluation measures and graphical displays for the episodic and annual MM5 base case simulations. Regarding the two reports, we contemplate that these volumes would be approximately 75 to 100 pages in length, including color figures, tables and appropriate appendixes. Consistent with project resources, the focus of the narrative discussions and interpretations in the reports will be on those overall features of the MM5 simulations and performance testing results that are most pertinent to understanding the strengths and limitations of the modeling.

The companion CDs will contain the full set of evaluation results. This will include all temporal and spatial scales of evaluation pursued in this study. A brief written document, included on the CD, will guide the user though the CD and provide clear examples of the directory structure and proper interpretation of the graphical displays. A copy of this evaluation protocol will be provided on the CD to document the mathematical definitions of the various numerical measures used. If useful, some limited-scope animations of pertinent meteorological fields may be included.

REFERENCES

Baldwin, M.E., S. Lakshmivarahan, J.S. Kain, 2001, “Verification of Mesoscale Features in NWP Models”, proceedings of the Ninth Conference on Mesoscale Processes, American Meteorological Society, 30 July - 2 August, Fort Lauderdale, FL.

Barnes, S. L., 1973. Mesoscale objective analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, National Severe Storms Laboratory, Norman, OK 73069, 60 pp, [NTIS COM-73-10781].

Doty, K., T. W. Tesche, D. E. McNally, B. Timin, and S. F. Mueller, 2002. “Meteorological Modeling for the Southern Appalachian Mountains Initiative (SAMI), Final Report to the Southern Appalachian Mountains Initiative, prepared by the University of Alabama, Huntsville, Alpine Geophysics, LLC, U.S. EPA/OAQPS, and the Tennessee Valley Authority.

Ebert, E.E. and J. L. McBride, 2000. “Verification of Precipitation in Weather Systems: Determination of Systematic Errors”, *J. Hydrol*., Vol. 239, pp. 179-202.

Emery, C., E. Tai, and G. Yarwood, 2001. “Enhanced Meteorological Modeling and Performance Evaluation for Two Texas Ozone Episodes”, report to the Texas Natural Resources Conservation Commission, prepared by ENVIRON, International Corp, Novato, CA.

Mass, C. F., et al., 2002. “Does Increasing Horizontal Resolution Produce More Skillful Forecasts?”, *Bulletin of the American Meteorological Society***,** March, pp. 407-430.

Nielsen-Gammon, J. W., 2002a, “Evaluation and Comparison of Preliminary Meteorological Modeling for the August 2000 Houston-Galveston Ozone Episode”, report to the Texas Natural Resources Conservation Commission, prepared by the Department of Atmospheric Sciences, Texas A&M University, College Station, TX.

Nielsen-Gammon, J. W., 2002b, “Meteorological Modeling for the August 2000 Houston-Galveston Ozone Episode: PBL Characteristics, Nudging Procedure, and Performance Evaluation”, report to the Texas Natural Resources Conservation Commission, prepared by the Department of Atmospheric Sciences, Texas A&M University, College Station, TX.

Olerud, D., K. Alapaty, and N. Wheeler, 2000. “Meteorological Modeling of 1996 for the United States with MM5”, EPA Task Order No. CAA689805, prepared by MCNC Environmental, Research Triangle Park, NC.

Stull, R. B., 1988. An Introduction to Boundary Layer Meteorology, Kluwer Academic Publishers, Boston, MS.

Tesche, T. W., and D. E. McNally, 2001. “Evaluation of CAMx and Models-3/CMAQ over the Lower Lake Michigan Region with Inputs from the RAMS3c and MM5 Models”, prepared for the Coordinating Research Council, prepared by Alpine Geophysics, LLC, Ft. Wright, KY.

Tesche, T. W., and D. E. McNally, 2002. “Evaluation of the RAMS Meteorological Fields for Seven SAMI Modeling Episodes”, prepared for the Tennessee Valley Authority and the Southern Appalachian Mountains Initiative, prepared by Alpine Geophysics, LLC, Ft. Wright, KY.

Tesche, T. W., D. E. McNally, C. A. Emery, and E. Tai, 2001b. “Evaluation of the MM5 Model Over the Midwestern U.S. for Three 8-Hr Oxidant Episodes”, prepared for the Kansas City Ozone Technical Work Group, prepared by Alpine Geophysics, LLC, Ft. Wright, KY and ENVIRON International Corp., Novato, CA.

Tesche, T. W., D. E. McNally, C. F. Loomis, and J. G. Wilkinson, 2002. “Regional Photochemical Modeling over Central Florida for Nine 8-hr Ozone Episodes”, Final Report prepared for the Florida Department of Environmental Protection, prepared by Alpine Geophysics, LLC, Ft. Wright, KY.