Evaluations of Mesoscale Models' Simulations of Near-Surface Winds, Temperature Gradients, and Mixing Depths

STEVEN R. HANNA AND RUIXIN YANG

George Mason University, Fairfax, Virginia

(Manuscript received 3 April 2000, in final form 21 September 2000)

ABSTRACT

Mesoscale meteorological models are being used to provide inputs of winds, vertical temperature and stability structure, mixing depths, and other parameters to atmospheric transport and dispersion models. An evaluation methodology is suggested and tested with simulations available from four mesoscale meteorological models (Fifth-Generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model, Regional Atmospheric Modeling System, Coupled Ocean–Atmosphere Mesoscale Prediction System, and Operational Multiscale Environmental Model with Grid Adaptivity). These models have been applied by others to time periods of several days in three areas of the United States (Northeast, Lake Michigan area, and central California) and in Iraq. The authors' analysis indicates that the typical root-mean-square error (rmse) of hourly averaged surface wind speed is found to be about 2–3 m s^{-1} for a wide range of wind speeds for the models and for the geographic regions studied. The rmse of surface wind direction is about 50° for wind speeds of about 3 or 4 m s^{-1}. It is suggested that these uncertainties in wind speeds and directions are primarily due to random turbulent processes that cannot be simulated by the models and to subgrid variations in terrain and land use, and therefore it is unlikely that the errors can be reduced much further. Model simulations of daytime mixing depths are shown to be often within 20% of observations. However, the models tend to predict weaker inversions than are observed in interfacial layers capping the mixing depth. The models also underestimate the vertical temperature gradients in the lowest 100 m during the nighttime, which implies that the simulated boundary layer stability is not as great as that observed, suggesting that the rate of vertical dispersion may be overestimated. The models would be able to simulate better the structure of shallow inversions if their vertical grid sizes were smaller.

1. Introduction and background

Because of the improved computational speed and resolution of mesoscale meteorological models, their use in providing inputs to mesoscale atmospheric dispersion models is increasing (Lyons et al. 1995; Stauffer et al. 1993; Seaman 2000). Mesoscale meteorological models can now provide inputs on horizontal grids with dimensions of 10 km or less, making them useful in complex terrain, urban, coastal, and other spatially inhomogeneous geographical regions. Recent advances in four-dimensional data assimilation (FDDA) allow the model simulations to be ''nudged'' toward the latest observations. Furthermore, the models run sufficiently fast that they can be used for real-time predictions of transport and dispersion rather than solely for retrospective analyses. In this paper, we focus on evaluation of model outputs of interest to transport and dispersion models, such as near-surface wind speed and direction, near-surface vertical temperature gradient, mixing depth, and vertical temperature gradients in the capping inversion.

Some of these same issues are addressed by the model developers. For example, Pielke and Uliasz (1998) discuss the limitations and strengths of the Regional Atmospheric Modeling System (RAMS) mesoscale meteorological model (Pielke et al. 1992) for atmospheric dispersion applications. They give examples of the magnitudes of the spatial and temporal variability in the atmosphere and describe how the variability can lead to differential advection and delayed diffusion.

A series of papers by the contributing developers of the Fifth-Generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al. 1994) addresses various aspects of model evaluation. For example, Seaman's (2000) review paper summarizes the relations between meteorological models and air quality models. Descriptions and evaluations of MM5 in various geographic domains, with an emphasis on air quality applications, are described by Seaman and Michelsen (2000) for the northeastern United States, by Seaman et al. (1995) and Tanrikulu et al. (2000) for central California, and by Shafran et al. (2000) for the Lake Michigan area. It is

Corresponding author address: Dr. Steven R. Hanna, SCS Mail Stop 5C3, 103 Science and Technology I, George Mason University, Fairfax, VA 22030-4444.
E-mail: shanna@gmu.edu

found that the results are improved by FDDA and by use of a relatively shallow (about 10 m) lowest grid layer. The model can simulate observed winds with a root-mean-square-error (rmse) of about 2 m s^{-1} and is capable of simulating specific phenomena such as low-level jets.

Cox et al. (1998) discuss a comprehensive meteorological model evaluation exercise using several models applied on a relatively coarse grid (about 80 km) to four geographic domains (Korea, Middle East, Central America, and the continental United States). The authors made use of model performance criteria prescribed by the U.S. Air Force, Air Weather Service and found that the near-surface temperature forecasts were usually within 2°C of observations, the wind speed forecasts were usually within 2.5 m s^{-1} of observations, and the wind direction forecasts were usually within 30° of observations.

The above examples of mesoscale meteorological model evaluations were all carried out by the model developers. Even the U.S. Air Force–sponsored study, reported by Cox et al. (1998), was primarily performed by the company who markets RAMS. Even though these ''self-evaluation'' exercises are usually performed in a fair fashion, there is always the ''appearance'' of a conflict of interest. Therefore it is desirable that an independent unbiased group carry out the evaluations.

Independent evaluations of mesoscale meteorological models have been reported by Tesche and McNally (1996, 1999) and Tesche et al. (1997), who developed and applied a general software system for evaluating mesoscale meteorological models used to provide inputs to regional photochemical grid models. Tesche and McNally usually make the model runs themselves, although sometimes they do make use of runs provided by the developers. Their methodology emphasizes relevant model performance measures such as the mean bias and rmse of boundary layer variables such as wind speed and temperature. These and similar model performance measures for air quality applications are described by Lyons et al. (1995), Seaman et al. (1995), Seaman and Michelsen (2000), Shafran et al. (2000), and Tanrikulu et al. (2000), among others. However, Tesche and colleagues provide an important contribution that has stimulated others to perform more thorough evaluations of boundary layer characteristics.

As stated above, it is best for the model evaluations to be carried out by an independent unbiased group who make the runs themselves and who therefore have access to the full set of model output files. The evaluations can alternatively be carried out by an independent group who obtain the output files from the group who made the original runs. This is the methodology used in the current paper. Because sets of meteorological model simulations and observations were available from the studies carried out by Tesche and McNally and others, it was very cost-effective and efficient to make use of these existing databases, which were developed as part of extensive regional ozone research studies. The assumption is made that the modelers optimized their simulations through selection of the best combinations of inputs such as FDDA methods, planetary boundary layer parameterizations, and grid resolution.

A drawback of the philosophy of making use of other persons' mesoscale meteorological model outputs is that often not all of the desired outputs are available in the basic files that can be acquired. For example, most models can output friction velocities, surface heat fluxes, and turbulent energy components. However, these parameters are not included in the sets of saved files sent to us for analysis.

The mesoscale meteorological model evaluations described in this paper involve a mix of case studies with different subgroups of models, depending on how the particular case studies were carried out in the original project. For example, the Ozone Transport Assessment Group (OTAG) study in the eastern United States and the Lake Michigan Ozone Study (LMOS) both used the MM5 and RAMS models. The central California SARMAP (defined in section 3b) study used only the MM5 model. The Iraq study used three models—MM5, the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS; Hodur 1997), and the Operational Mesoscale Environmental Model with Grid Adaptivity (OMEGA; Bacon et al. 2000).

2. Mesoscale meteorological model evaluation methodology

The mesoscale meteorological model variables used in our evaluations were limited by the amount of information in the files provided to us by the persons who ran the models. Consequently, our evaluations stressed the hourly values of the near-surface wind speed and direction, the mixed-layer averaged wind speed and direction, the mixing depth, the temperature gradient in the capping inversion during the day, and the temperature gradient in the 100-m-deep layer near the surface in the nighttime. Simulations of these variables could be compared with observations at many surface meteorological stations and by radiosondes at a few special stations.

The depth of the lowest model level was relatively shallow in all these applications. For example, for the OTAG study, the lowest RAMS and MM5 grid layers were 10 m and 9 m, respectively. For the SARMAP study, the lowest MM5 grid layer was 9 m. The center point of these lowest layers has an elevation similar to that of the level of the wind instruments. Also, it is of interest that, in the lowest 1000 m, the models typically have about 10 or 12 layers. At an elevation of about 1000 m, the models typically have a grid increment of about 100 or 200 m.

There are several alternate measures of the vertical temperature structure that could be evaluated. For example, the temperature jump in the capping inversion may be more important than the temperature gradient in determining whether a buoyant thermal can penetrate the inversion. As another example, the amount of energy required to break the surface-based nocturnal inversion layer may be more important than the temperature gradient in determining how quickly the boundary layer becomes well-mixed in the morning. However, we stress the vertical temperature gradient in both cases because it is often used in parameterizations of vertical diffusivity coefficients, which are important for air quality applications.

Air quality applications also make use of surface flux scaling parameters such as the friction velocity u_*, the temperature scale T_*, and the turbulent velocity components. Although these parameters can be output by most models, they are not generally available to us in the files saved for the current applications.

It is important to recognize that there are major differences between the observations used in this study and the model simulations. The in situ observations are taken at a point by an anemometer, a thermometer, or a radiosonde, whereas the model simulations represent a spatial mean over a volume determined by the horizontal and vertical grid spacing, with typical dimensions of, say, 10 km $\times$ 10 km horizontal dimensions by 10- to 100-m vertical dimensions. The simulated wind speed generally represents a short-term average on the hour. Observed in situ meteorological variables such as wind speed are affected by stochastic fluctuations in wind speed and spatial variability within the grid square. Also, the standard National Weather Service (NWS) wind observations represent about 3-min averages on the hour. The stochastic spatial fluctuations in observed surface wind speed are observed to be about a factor of 2 over distances of 5 or 10 km (Hanna and Chang 1992). These considerations suggest that there are differences expected between the observations and the simulations due simply to the differences in the averaging volumes and averaging times that they represent.

The independent evaluation of mesoscale meteorological models is further complicated by the models' use of FDDA, so that some of the observations used in our evaluations have already been used to nudge the solution in the FDDA process (e.g., Seaman et al. 1995). Even though the nudging coefficient is usually very small, the process still compromises the concept of independence between model results and observations. In some evaluation exercises, the modelers have made several optional runs in which part of the data is used in the FDDA process and another part of the data is used for independent model evaluation. In the evaluations in this paper, all observed data are used, including those used for the FDDA exercise.

As suggested by Hanna (1989), Tesche et al. (1997), and Shafran et al. (2000), the mean bias (the average simulated value minus the average observed value) and the rmse (square root of the individual differences between the simulated and observed values) can be considered as primary performance measures for the wind speed and direction comparisons. These measures are given as absolute values with their original units or are sometimes given as relative values for which the absolute values are divided by the average of the observed value. These performance measures are applied to surface observations of wind speed and direction. It should be noted that, if the lowest model depth greatly exceeds the height of the wind observation (usually about 10 m), either the model simulations or the observations of wind speed should be adjusted using standard boundary layer wind profile formulas so that the heights of the observed and modeled wind speeds are as similar as possible. For the evaluations in this paper, this adjustment was not made because the height of the observation was generally within a factor of 2 of the mid-height of the model's lowest level. Even for a factor-of-2 difference in heights z (say 10 vs 20 m), the wind speeds u given by the logarithmic formula [$u = (u_*/0.4) \ln(z/z_0)$] would differ by only about 20%, assuming a surface roughness length z_0 of about 0.03 m, typical of rural areas. Note that u_* is the friction velocity in the logarithmic formula for wind speed. For the analysis of the differences between observed and simulated mixing depths and vertical temperature profiles, scatter-plots are used, as well as comparisons of median values. Contingency tables are employed in which there are qualitatively determined categories of vertical temperature gradients.

In the following sections, various combinations of the above-described model performance measures are applied, depending on the available model outputs and observations for each geographic domain. The final section summarizes the results found on the individual domains.

3. Evaluations of MM5 and RAMS on eastern U.S. and central California domains

The MM5 (Grell et al. 1994) and RAMS (Pielke et al. 1992) meteorological models have been run by others as part of extensive studies of ozone episodes in the eastern United States. MM5 was run as part of a similar study of ozone episodes in the Central Valley of California. During most of these ozone episodes, the meteorological conditions were marked by light winds associated with widespread high pressure in the summer. The observation files and the model simulation files are available from studies carried out by Tesche et al. (1997) and Tesche and McNally (1996, 1999), who were responsible for the original independent meteorological model evaluations on these domains. However, the files available from the Tesche et al. (1997) and Tesche and McNally (1996, 1999) reports and/or from files saved by the investigators are abridged in the sense that not all surface and vertical sounding data are included, and the model output files do not include some boundary layer parameters such as heat fluxes and friction veloc-

TABLE 1. Model performance measures for surface wind speed and direction for RAMS and for MM5 for the 12-km OTAG grid in the eastern United States. Bias is defined as simulated minus observed. Rmse is root-mean-square error.

Day of 1995	194	195	196	197	198	199	200	201	202	All Days
Speed—mean observed (m s^{-1})	3.4	3.0	2.6	2.8	3.2	2.9	3.1	3.1	3.0	3.0
Speed—RAMS mean bias (m s^{-1})	−0.2	−0.2	−0.3	−0.1	0.0	−0.1	0.2	0.2	−0.1	−0.1
Speed—MM5 mean bias (m s^{-1})	0.1	−0.1	0.0	0.2	0.4	0.3	0.7	0.9	1.7	0.5
Speed—RAMS rmse (m s^{-1})	1.8	1.6	1.6	1.6	1.6	1.5	1.5	1.6	1.7	1.6
Speed—MM5 rmse (m s^{-1})	1.8	1.8	1.8	1.7	1.8	1.7	1.7	1.8	2.4	1.9
Direction—RAMS mean bias (°)	1	6	−17	−33	−23	−14	−14	−9	−2	−12
Direction—MM5 mean bias (°)	−12	−10	−11	−12	−10	−13	−14	−15	−30	−14
Direction—RAMS rmse (°)	77	52	76	79	68	76	70	64	64	76
Direction—MM5 rmse (°)	45	48	60	49	46	48	45	50	69	51

ities. The files that are available for our study could be considered to be ‘‘basic’’ files consisting of routine surface observations and radiosonde soundings, as well as standard model outputs for comparison with these routine observations.

a. Eastern United States

The MM5 and RAMS meteorological model simulations were reported by Tesche and McNally (1996) for a nine-day period in the eastern United States in July of 1995 when regional ozone concentrations were high (and hence wind speeds were low, because the region was covered by a large high pressure system). The MM5 model is described by Grell et al. (1994) and the RAMS model is described by Pielke et al. (1992). The so-called Blackadar PBL scheme was used in these MM5 runs, and the lowest vertical grid depth was about 10 m for both models. This study is part of the OTAG project. Both MM5 and RAMS were ‘‘triple nested,’’ with a 108-km-grid outer domain covering the entire United States, a 36-km-grid medium domain covering the eastern 2/3 of the United States, and a 12-km-grid inner domain covering the east–central United States. Tesche and McNally (1996) provide more details concerning the models and the observations, as well as the results of their own statistical evaluations. Our statistical tests are similar to theirs but have been carried out independently.

The NWS observations of wind speed and direction are made at a height of 10 m, which is comparable to the height of the lowest model layer in MM5 and RAMS. The gridcell heights above ground level in the lowest 200 m for MM5 are 9, 34, 74, 133, and 211 m and for RAMS are 10, 33, 67, 101, 135, 174, and 211 m. At an elevation of about 1000 m, the vertical cell thicknesses are about 185 m for MM5 and about 220 m for RAMS. Table 1 contains an example of model performance statistics for RAMS and MM5 hourly averaged wind speed and direction simulations for the 12-km-grid inner OTAG domain for each of the nine days and for all days combined. For this nine-day period, dominated by high pressure and light winds (averaging about 3 m s^{-1}), the day-to-day results are fairly consistent: the mean bias usually has an absolute value less than about 0.5 m s^{-1} for wind speed and about 15° for wind direction, and the rmse is usually about 1.5–2.0 m s^{-1} for wind speed and 50°–80° for wind direction. Dividing by the mean wind speed of about 3 m s^{-1}, the relative mean bias in wind speed is calculated to be about 15% and the relative rmse is about 50%–70%. It is seen that, on average over ‘‘all days,’’ MM5 tends to slightly overestimate the wind speed (the mean bias is 0.7 m s^{-1} for MM5), but RAMS has little mean bias. Over all days, MM5 has a slightly lower rmse than RAMS does in its wind direction simulations (58° for MM5 vs 76° for RAMS). However, in such light wind situations, wind directions are known to be variable and unreliable.

The question arises as to whether the FDDA methods are affecting the model performance measures. This effect is unknown, although it is expected that the influence of FDDA on the model performance measures is not as strong as expected because of the small magnitude of the FDDA nudging coefficient.

Confidence limits for the performance measures in Table 1 were calculated using Hanna’s (1989) bootstrap resampling software, showing that these differences in performance measures between models are significant at the 95% confidence level. The reason for the significance of the results is the large number of data points, given that 216 hours of data were available from about 360 observing sites. Confidence intervals are inversely proportional to the square root of the number of data points. Even if the degrees of freedom were reduced by a factor of 2 because of the influence of FDDA, the differences would still be significant.

b. Central California

A comprehensive research program known as the SARMAP project took place in the San Joaquin River area in central California (Seaman et al. 1995; Tesche et al. 1997; Tanrikulu et al. 2000). The objectives of the SARMAP study were similar to those of the OTAG study described above. The SARMAP study is based on the San Joaquin Valley Air Quality Study and Atmospheric Utility Signatures (Ranzieri and Thuillier 1991) field experiment, which provides the first two letters (SA) of the SARMAP acronym. The last four letters of

TABLE 2. Model performance measures for surface wind speed and direction for MM5 for the 4-km grid in the central California SARMAP domain. Bias is defined as simulated minus observed.

Day of 1991	215	216	217	218	All Days
Speed—mean observed (m s^{-1})	3.4	3.1	2.9	2.8	3.1
Speed—mean bias (m s^{-1})	1.2	1.4	1.5	2.3	1.5
Speed—relative bias (%)	36.0	43.0	49.0	79.0	51.0
Speed—rmse (m s^{-1})	2.5	2.4	2.3	2.7	2.5
Direction—mean observed (°)	266	279	296	271	278
Direction—mean bias (°)	3	−2	−11	2	−2
Direction—MM5 rmse (°)	61	61	65	76	66

the acronym stand for Regional Meteorological and Air Pollution (RMAP). The MM5 model was run by Seaman et al. (1995) on a 4-km grid for the 3–6 August 1990 SARMAP ozone episode, although the current evaluations used a later model run. The vertical grid structure was similar for SARMAP and OTAG. A variety of PBL schemes and FDDA assumptions were tested in the SARMAP study. The final MM5 runs used in our evaluations used the so-called turbulent kinetic energy Gayno–Seaman scheme for the PBL and used a detailed FDDA method. A dataset containing the MM5 model simulations and the observations was obtained from McNally and Tesche (1998, personal communication), who also extensively evaluated the model simulations (McNally and Tesche 1998). The MM5 model evaluations described below are based on observations of surface winds at NWS stations plus several additional surface stations set up specifically for the SARMAP study.

Table 2 contains our comparisons of observations and MM5 simulations for surface wind speeds and directions for the four days of the episode and for the averages over the four days, based on the information acquired from McNally and Tesche (1998, personal communication). The mean observed wind speed is 3.1 m s^{-1} and the mean MM5 simulated wind speed is 4.6 m s^{-1}, for a mean bias of 1.5 m s^{-1}. The rmse in surface wind speeds is about 2.5 m s^{-1}, which is approximately the same as the standard deviations of the observed or the simulated wind speeds. There is little day-to-day difference in the performance measures over the four days of the ozone episode. Because FDDA was used, the simulations are not entirely independent of the observations. The effect of the FDDA is difficult to assess a priori, because the nudging coefficient is relatively small. Certainly, if statistical confidence limits were calculated, the degrees of freedom should be reduced to reflect the lack of independence.

The surface wind direction comparisons in Table 2 suggest that the four-day mean simulated and observed wind directions are within 2°. It should be noted that these means are calculated using scalar wind directions rather than vector wind directions. The rmse or scatter in simulated versus observed wind directions averages about 66°. Similar rmse for wind direction were found for the northeast U.S. OTAG episode reported in Table 1 and discussed in section 3a. Again, the same comment applies that the observed wind directions tend to be uncertain and unreliable as wind speeds approach zero.

4. Evaluations of simulations of boundary layer vertical temperature gradients in the Lake Michigan Ozone Study

a. Background of available data

Tesche and McNally (1999) report on the results of an evaluation of the MM5 and RAMS mesoscale meteorological models for two intensive field study periods (26–28 June 1991 and 16–18 July 1991) during LMOS. The meteorological conditions during the LMOS ozone episodes were similar to those during the OTAG and SARMAP ozone episodes discussed previously. The MM5 model was run over a nested 108/36/12/4-km-grid domain centered over Lake Michigan, and the RAMS model was run using a 13.5/4.5-km nesting scheme centered over a similar domain. Vertical grid spacing was about 10 m at the surface, about 50 m at a height of 100 m, and about 200 m at a height of 1000 m. During the intensive field study, vertical profiles were obtained by radiosondes at three sites at 2 times the usual frequency (i.e., every 6 h) and from a few additional stations. The emphasis of the Tesche and McNally (1999) report was on comparison of surface observations with simulations, yielding mean bias and rmse values for wind speed and direction that are similar to those reported in Tables 1 and 2 for the OTAG and SARMAP domains.

A few examples of observed and simulated MM5 and RAMS vertical profiles were included in the Tesche and McNally (1999) report and were plotted on so-called skew T–log$p$ diagrams. Because we were interested in developing and testing methods for evaluating models using vertical temperature profiles, we obtained a full set of the LMOS skew T–log$p$ diagrams from nine radiosonde stations. The analysis reported here concerns only those radiosonde stations [Green Bay (GRB), Wisconsin, and Peoria (PIA), Illinois] for which both MM5 and RAMS model simulations were available concurrently. These vertical profiles from GRB and PIA were available at 0000, 0600, 1200, and 1800 local time.

Examples of the observed and simulated vertical profiles of temperature are given in Fig. 1. The left part of

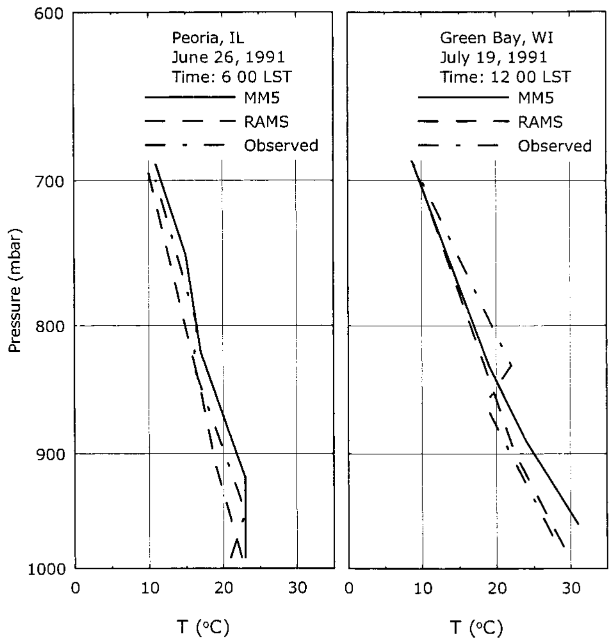

FIG. 1. Examples of vertical profiles of observed and MM5 and RAMS model-simulated temperature from the LMOS domain. The left figure shows a typical nighttime profile (taken on 26 Jun 1991 in Peoria, IL) and the right figure shows a typical daytime profile (taken on 19 Jul 1991 in Green Bay, WI). Note that a pressure difference of about 100 hPa is equivalent to a height difference of about 1000 m.

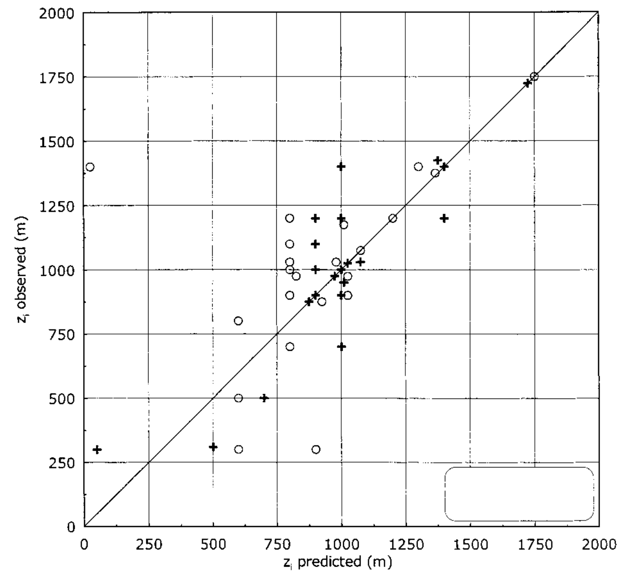

FIG. 2. Observed vs MM5 (pluses) and RAMS (circles) simulations of daytime mixing depths for the Peoria and Green Bay sites on the LMOS domain.

the figure shows a typical nighttime example of observed and simulated vertical profiles of temperature from Peoria at 0600 local time on 26 June 1991. The observed surface-based inversion is stronger (in terms of dT/dz) than the simulated inversions. As will be seen later, the result is often repeated in the nighttime vertical temperature profiles.

The right part of Fig. 1 shows a typical daytime example of temperature profiles, from Green Bay at noon local time on 19 July 1991, with mixed layers up to heights of about 1000 m (assuming a pressure difference of 100 hPa is equivalent to a height difference of about 1000 m). However, the observed profile has a stronger capping inversion than the simulated profiles. The strong inhibiting effect of the observed capping inversion is also indicated by a rapid decrease in observed water vapor mixing ratio above the mixing depth (not shown on Fig. 1 but shown on the original skew T–log$p$ diagrams). The simulated profiles do not show a strong capping inversion, and the variation of water vapor mixing ratio is smoother. These differences would also have strong effects on vertical mixing of pollutants. The probable reason for the inability of the models to simulate a 100-m-deep capping inversion is that the models' vertical grid increment is only about 200 m at that height, thus not allowing the shallow stable layer to be adequately resolved.

The analysis was broken down into nighttime profiles (0000 and 0600 LST) and daytime profiles (1200 and 1800 LST). The following key measures were estimated from each of the 43 available profiles: lower-level vertical temperature gradient dT/dz [°C (100 m)$^{-1}$], z_i (mixing depth) if lower layer is adiabatic, temperature gradient in intermediate layer above lower layer, and capping inversion (if any) and dT/dz in that layer. These measures were selected because of their relevance to air quality models. The magnitudes of dT/dz were calculated by best-fitting a straight line by eye to the data.

b. Analysis of daytime (1200 and 1800 LST) mixing depth z_i

There were 20 profiles observed during the daytime (1200 and 1800 LST) at the two stations (GRB and PIA) with well-mixed conditions and a well-defined mixing depth. These observed mixing depths ranged from 300 to 1750 m, with a median of 1000 m. The MM5 and RAMS model simulations of daytime mixing depths were estimated by us by eye from the vertical temperature and dewpoint profiles. In general, as seen in the scatterplot in Fig. 2, the MM5 and RAMS model simulations of mixing depths agreed fairly well with these observed mixing depths, with medians of 1000 and 850 m, respectively. About 60% of the simulated mixing depths are within ±20% of the observations. However, for low observed mixing heights (say, 300 m), the simulations are seen to be in error by a factor of 2–4. Furthermore, each of the models had one instance in which they simulated an inversion at the time a well-mixed layer was observed.

TABLE 3. Occurrence of three categories of temperature gradients in capping inversions during the daytime for the Peoria and Green Bay sites on the LMOS domain. Numbers are given for observations and for simulations by the MM5 and the RAMS mesoscale meteorological models.

	Observed	MM5	RAMS
1) nearly dry adiabatic to wet adiabatic [$-1°C\ (100\ m)^{-1} < dT/dz <$ wet adiabatic]	4	15	16
2) wet adiabatic to weakly stable [wet adiabatic $< dT/dz < 1°C\ (100\ m)^{-1}$]	7	7	6
3) weak to strong inversion [$1°C\ (100\ m)^{-1} < dT/dz$]	11	0	0

c. Analysis of daytime (1200 and 1800 LST) capping inversion

The magnitude of the vertical temperature gradient in the daytime capping inversion is important because it is a measure of the rate of vertical diffusion or pollutant fluxes through the capping inversion. In section 4a, the qualitative result was mentioned that the model-simulated capping inversions were not as strong as the observed capping inversions, and the model-simulated water vapor mixing ratios were more smoothly varying across the capping inversion than were the observed mixing ratios. The current section quantifies these results for the vertical temperature gradients and presents comparisons in contingency tables.

There were 22 observed daytime temperature profiles, for which the simulated and observed capping inversion temperature gradients were arbitrarily divided into the three categories listed in Table 3, in which the occurrence is entered for observed and for MM5 and RAMS simulated gradients. Table 3 reveals that most of the observed vertical temperature profiles are marked by weak to strong capping inversions. However, none of the simulated profiles are marked by weak to strong capping inversions. Instead, the simulated profiles are capped by temperature gradients that are closer to dry or wet adiabatic. As mentioned above, because the model grid increments are about 200 m at an elevation of 1000 m, it is not expected that the model could adequately simulate the vertical temperature structure in a capping inversion with typical depth of 100 or 200 m.

d. Analysis of nighttime (0000 and 0600 LST) vertical temperature gradients in lowest 100 m

The vertical temperature gradient in the lowest part of the nighttime boundary layer is important to transport and dispersion processes because it is used to parameterize the vertical diffusivity coefficient and the Richardson number, which are measures of the amount of vertical mixing. In the overview in section 4a, the qualitative result was mentioned that the model-simulated near-surface vertical temperature gradients were less than the observed vertical temperature gradients at night. It is postulated that the difference is due to the relatively coarse model vertical resolution.

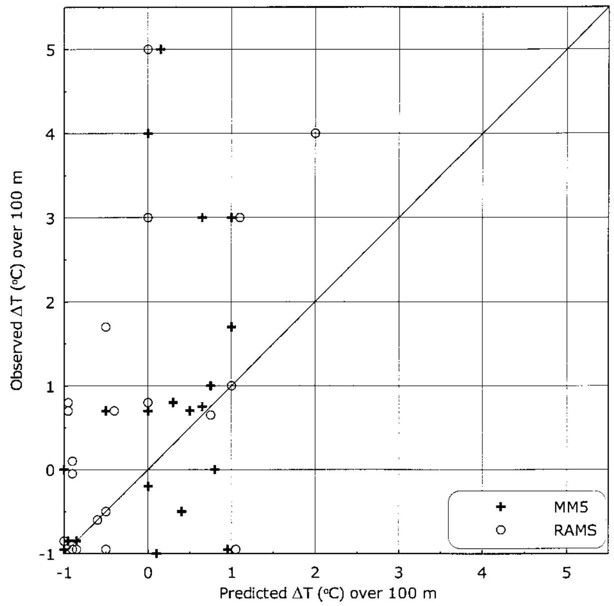

FIG. 3. Observed vs MM5 and RAMS simulations of nighttime temperature gradients in the lowest 100 m for the Peoria and Green Bay sites on the LMOS domain.

Figure 3 contains a scatterplot of the temperature difference in the lowest 100 m for the 20 observed vertical temperature profiles plotted against the temperature difference for the simulated vertical temperature profiles. A layer depth of 100 m is arbitrarily chosen for these comparisons because it represents the depth over which pollutants released near the surface are typically mixed at night. The gradients were estimated by fitting a straight line by eye to the plotted profiles. The bias of the simulations toward less stability can clearly be seen. In particular, there are 10 points on the figure where relatively strong vertical temperature gradients between 1° and 5°C $(100\ m)^{-1}$ are observed, while there is only one point on the figure [at 2°C $(100\ m)^{-1}$] where such strong stable vertical temperature gradients are simulated. The models currently use grid levels of 10, 33, 67, and 101 m (for RAMS) and 9, 34, 74, and 133 m (for MM5). Grid increments of 5 m or less would be needed to resolve the observed temperature gradients.

5. Use of meteorological data from Iraq for 10–13 March 1991 to evaluate OMEGA, COAMPS, and MM5

The U.S. Department of Defense (DOD) has been applying combinations of several meteorological models and transport and dispersion models to assess the dosages of chemical agents that may have been released from Khamisiyah, Iraq, during the time period from 10 to 13 March 1991. OMEGA (Bacon et al. 2000), COAMPS (Hodur 1997), and MM5 (Grell et al. 1994) had been applied to the scenario and had been separately evaluated using limited surface meteorological data and data from a few rawinsonde stations in the area. The current study uses the existing model output files and observation files, which were provided by Dr. R. Babarsky of the Office of the Special Assistant for Gulf War Illnesses (OSAGWI).

The comprehensive DOD study is described in an unpublished OSAGWI report. However, the individual modelers have separately published the results for their own models (Bacon et al. 2000; Westphal et al. 1999; Warner and Sheu 2000). The simulations by the three models could make use of only limited observations in Iraq because of the sparse network and because not all of the observations were publicly released. Bacon et al. (2000) describe the OMEGA model applications and include some model evaluation statistics. For example, for 18 surface sites, they report that the median value of the mean bias for the wind speed is about -0.5 m s^{-1} (i.e., an underestimation) and the median wind speed rmse is about 4.6 m s^{-1}. Westphal et al. (1999) present many figures that allow qualitative comparisons to be made between the COAMPS model simulations and the observations but include no quantitative results. The MM5 model applications by Warner and Sheu (2000) test four combinations of options for inputs of large-scale analysis, for boundary layer parameterization, and for desert roughness length. MM5 is run with three ''nested'' horizontal grid sizes (30, 10, and 3.3 km). The so-called Medium Range Forecast model PBL parameterization scheme (Hong and Pan 1996) was used for MM5 in the applications used in the OSAGWI report and in our evaluations. Warner and Sheu (2000) conclude that the surface wind speed and wind direction rmse values are fairly consistent for the four input options, averaging about 2.3 m s^{-1} and 60°, respectively. The wind speed rmse values are somewhat larger than those found on the OTAG, SARMAP, and LMOS domains, probably due to the reduced availability of input data such as surface winds and soil moisture.

a. Comparisons with Hafar Al'Batin rawinsonde observations

The emphasis of our analysis is on the lowest 1000 m of the atmosphere, because the alleged release of chemical agent took place at the ground surface. The observed vertical profiles at the rawinsonde station in Hafar Al'Batin, Saudi Arabia, had adequate vertical resolution (several points in the lowest 1000 m) for use in our model evaluation. Note that the 0000 UTC (0300 LST) soundings are interpreted as ''night'' and the 1200 UTC (1500 LST) soundings are interpreted as ''day.''

Quantitative comparisons of wind speeds averaged over the lowest 1000-m layer for the seven soundings were made but were influenced by the relatively high (10–15 m s^{-1}) observed winds on 13 March 1991. Those observed high wind speeds on 13 March tended to be underestimated by about 20% or 30% by all three models. On average over the entire four-day period, the OMEGA, COAMPS, and MM5 models underestimate the mean wind speeds by about 22%, 27%, and 14%, respectively, according to our calculations of these performance measures.

The estimates of layer-averaged wind direction by all three models are within about 15° of each other and the observations, on average. However, the MM5 model simulations of wind directions average about 16° less (i.e., in a counterclockwise direction) than the COAMPS model simulations of wind directions, which implies that the simulated cloud transport directions would differ by the same amount.

When attention is focused on the surface winds at Hafar Al'Batin, it is found that, on average, the OMEGA and COAMPS models underestimate the surface wind speed by almost a factor of 2. However, the MM5 model simulations of surface wind speeds are fairly good, on average. Average absolute errors for the seven time periods are about 3–4 m s^{-1} for the OMEGA and COAMPS models (due to their large mean biases) and are about 2 m s^{-1} for the MM5 model.

Daytime (1200 UTC or 1500 LST) mixing depths and nighttime (0000 UTC or 0300 LST) inversion strengths were compared for the OMEGA simulations and the radiosonde observations at Hafar Al'Batin. The COAMPS and MM5 simulations of these variables could not be assessed because detailed temperature soundings were not available from OSAGWI. The cell centroid heights (above ground level) for OMEGA in the lowest 1000 m were 14, 45, 80, 121, 168, 223, 285, 356, 439, 534, 642, 768, 912, and 1077 m. During the day, the observed mixing depths are seen to be higher (by a factor of about 2) than the OMEGA-simulated mixing depths. Therefore, simulations of pollutant plumes released in the mixed layer near the ground would be constrained to too shallow of a layer if the simulated temperature profiles were used. During the night, there is roughly a 500-m average error in specification of inversion depth, which implies that vertical dispersion estimates will also have uncertainties. When the vertical temperature gradients, in degrees Celsius per 100 meters, of the four observed and simulated nighttime inversions are compared, the average observed and OMEGA model-simulated temperature gradients are 2° and 1.3°C per 100 meters, respectively. This tendency for the simulated vertical temperature gradients to be less than the observed vertical temperature gradients was also seen in section 4 for the MM5 and RAMS models applied to the Lake Michigan domain and is probably caused by the relatively coarse vertical grid resolution. Even though the OMEGA model has a few more grid levels than the RAMS or MM5 models in the lowest 1000 m, there is still insufficient resolution to simulate vertical phenomena with scales of 10 m near the surface or 200 m aloft.

b. Evaluations using observations of wind speed and direction from surface stations in the Iraq area and nearby for 10–15 March 1991

The number of surface stations acquired from OSAGWI in the Iraq area with ''good'' wind data on any hour ranged from 3 to 18. The mean observed surface wind speed was about 9 m s^{-1}. The mean OMEGA, COAMPS, and MM5 model-simulated surface wind speeds are about 1.4 m s^{-1} lower, about 1.5 m s^{-1} lower, and about 1.7 m s^{-1} higher, respectively, than the mean observed surface wind speeds. The surface wind speed rmse is about 5 or 6 m s^{-1} for all models. The mean OMEGA, COAMPS, and MM5 model surface wind direction biases are $-17°$, $-13°$, and 5°, respectively. A negative bias means the simulated wind vector is to the left (i.e., counterclockwise) of the observed wind vector. The surface wind direction rmse is about 60° for all models. The wind speed rmse values for the surface stations in the Iraq area are about 2 times as large as those calculated for other geographic domains in previous sections. A possible reason for the differences may be due to the fact that the models had to be run on this domain with minimal information on boundary conditions and minimal data assimilation. Furthermore, the level of quality control on the observations may be less in the Khamisiyah area, as suggested by the relatively large fraction of missing data.

6. Conclusions

Simulations by four mesoscale meteorological models (MM5, RAMS, COAMPS, and OMEGA) have been compared with boundary layer observations on four geographic domains, with emphasis on parameters that are of interest to transport and dispersion models. The parameters include surface wind speed and direction, boundary layer–averaged wind speed and direction, daytime mixing depth and capping inversion strength, and nighttime low-level vertical temperature gradient. The results are fairly consistent from site to site and from model to model.

For summertime light-wind periods, with mean wind speeds of 3 m s^{-1} observed by a well-calibrated and well-sited set of wind monitors, the model mean biases over many monitors and hours are usually 1 m s^{-1} or less for wind speed and 10° or less for wind direction. However, the rmses are consistently about 2 m s^{-1} in wind speed and 60° in wind direction. Because of the 60° rmse in wind direction, it is possible that the model-simulated transport direction of a smoke plume may have large errors during the first hour or two of travel. This is no surprise given that wind directions are known to be variable and unreliable when winds are light.

For geographic regions such as Iraq, where the terrain is complex and where there are limited data for input to the mesoscale meteorological models, biases in mean wind speed can approach a factor of 2 and the rmse in wind speed simulations can be as high as 6 m s^{-1}.

It is expected that most of the remaining uncertainties in model simulations of near-surface wind speeds and directions will be difficult to reduce even as models are improved, because the uncertainties are due to random stochastic or turbulent fluctuations. Subgrid variation in terrain and land use also must contribute to this variability.

Comparisons were also made with vertical temperature profiles in the boundary layer. The model simulations of daytime mixing depth were found to have little bias, although the simulated vertical temperature gradients in the capping inversions were generally smaller than the observed gradients. The model simulations of the magnitudes of nighttime low-level vertical temperature gradients were generally less than the magnitudes of the observed vertical temperature gradients, which implies that simulated vertical diffusivities and turbulent fluxes would be larger than observed. The primary reason for the bias in simulated vertical temperature gradients is expected to be that the vertical resolution in the model grid system is not fine enough to capture the observed vertical structure.

Acknowledgments. This research was sponsored by the Defense Threat Reduction Agency, with Major Thomas Smith and Mr. Ron Meris as Technical Contract Monitors. The assistance of Dr. Thomas Tesche and Mr. Dennis McNally of Alpine Geophysics in providing and interpreting data files is appreciated. Dr. Richard Babarsky of the Office of the Special Assistant for Gulf War Illnesses supplied the observations and model results for the Iraq evaluations. Dr. Pasquale Franzese of George Mason University assisted with the preparation of the graphics.

REFERENCES

Bacon, D. P., and Coauthors, 2000: A dynamically adapting weather and dispersion model: The Operational Multiscale Environmental Model with Grid Adaptivity (OMEGA). *Mon. Wea. Rev.,* **128,** 2044–2076.

Cox, R., B. L. Bauer, and T. Smith, 1998: A mesoscale model intercomparison. *Bull. Amer. Meteor. Soc.,* **79,** 265–283.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1994: A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5). NCAR/TN-398+1A, National Center for Atmospheric Research, Boulder, CO, 107 pp.

Hanna, S. R., 1989: Confidence limits for air quality model evaluations, as estimated by bootstrap and jackknife resampling methods. *Atmos. Environ.,* **23,** 1385–1398.

——, and J. C. Chang, 1992: Representativeness of wind measurements on a mesoscale grid with station separations of 312 m to 10 km. *Bound.-Layer Meteor.,* **60,** 309–324.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339.

Lyons, W. A., C. J. Tremback, and R. A. Pielke, 1995: Applications of RAMS to provide input to photochemical grid models for LMOS. *J. Appl. Meteor.,* **34,** 1762–1784.

McNally, D. E., and T. W. Tesche, 1998: Final evaluation of the MM5 model for the 3–6 August 1990 SARMAP episode over central California. Preprints, *10th Joint Conf. on the Applications of Air Pollution Meteorology with the A&WMA,* Phoenix, AZ, Amer. Meteor. Soc., 340–344.

Pielke, R. A., and M. Uliasz, 1998: Use of meteorological models as input to regional and mesoscale air quality models—limits and strengths. *Atmos. Environ.,* **32,** 1455–1466.

——, and Coauthors, 1992: A comprehensive meteorological modeling system—RAMS. *Meteor. Atmos. Phys.,* **49,** 69–91.

Ranzieri, A. J., and R. Thuillier, 1991: San Joaquin Valley Air Quality Study (SJVAQS) and Atmospheric Utility Signatures, Predictions and Experiments (AESPEX): A collaborative model program. Preprints, *84th Annual Meeting and Exhibition of AWMA,* Vancouver, BC, Canada, Air and Waste Management Association, 1–27.

Seaman, N. L., 2000: Meteorological modeling for air quality assessments. *Atmos. Environ.,* **34,** 2231–2259.

——, and S. A. Michelson, 2000: Mesoscale meteorological structure of a high-ozone episode during the 1995 NARSTO-Northeast study. *J. Appl. Meteor.,* **39,** 384–398.

——, D. R. Stauffer, and A. M. Lario, 1995: A multiscale four-dimensional data assimilation system applied in the San Joaquin Valley during SARMAP. Part I: Modeling design and basic performance characteristics. *J. Appl. Meteor.,* **34,** 1739–1761.

Shafran, P. C., N. L. Seaman, and G. A. Gayno, 2000: Evaluation of numerical predictions of boundary layer structure during the Lake Michigan Ozone Study (LMOS). *J. Appl. Meteor.,* **39,** 412–426.

Stauffer, D. R., N. L. Seaman, T. T. Warner, and A. M. Lario, 1993: Application of an atmospheric simulation model to diagnose air pollution transport in the Grand Canyon region of Arizona. *Chem. Eng. Commun.,* **121,** 9–26.

Tanrikulu, S., D. R. Stauffer, N. L. Seaman, and A. J. Ranzieri, 2000: A field-coherence technique for meteorological field-program design for air quality studies. Part II: Evaluation in the San Joaquin Valley. *J. Appl. Meteor.,* **39,** 317–334.

Tesche, T. W., and D. E. McNally, 1996: Evaluation of the MM5 model for three 1995 regional ozone episodes over the northeast U.S. Alpine Geophysics Rep. AG-90/TS80, 258 pp. [Available from Alpine Geophysics, P.O. Box 17344, Covington, KY 41017.]

——, and ——, 1999: Comparative evaluation of the MM5 and RAMS3C prognostic meteorological models over the midwestern U.S. for two 1991 LMOS intensive measurement episodes. Alpine Geophysics Rep. AG-90/TS-125, 197 pp. [Available from Coordinating Research Council, Inc., 219 Perimeter Center Parkway, Suite 400, Atlanta, GA 30346.]

——, ——, N. L. Seaman, and D. R. Stauffer, 1997: Final evaluation of the MM5 model for the 3–6 August 1990 SARMAP episode over central California. Alpine Geophysics Rep. AG-90/TS62, 303 pp. [Available from Alpine Geophysics, P. O. Box 17344, Covington, KY 41017.]

Warner, T. T., and R.-S. Sheu, 2000: Multiscale local forcing of the Arabian desert daytime boundary layer, and implications for the dispersion of surface-released contaminants. *J. Appl. Meteor.,* **39,** 686–707.

Westphal, D. L., and Coauthors, 1999: Meteorological reanalysis for the study of Gulf War illnesses: Khamisiyah case study. *Wea. Forecasting,* **14,** 215–241.