The Operational Mesogamma-Scale Analysis and Forecast System of the U.S. Army Test and Evaluation Command. Part II: Interrange Comparison of the Accuracy of Model Analyses and Forecasts

YUBAO LIU,* THOMAS T. WARNER,*+, ELFORD G. ASTLING,# JAMES F. BOWERS,# CHRISTOPHER A. DAVIS,* SCOTT F. HALVORSON,# DARAN L. RIFE,* RONG-SHYANG SHEU,* SCOTT P. SWERDLIN,* AND MEI XU*

*National Center for Atmospheric Research, Boulder, Colorado
+ Department of Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado
# U.S. Army Dugway Proving Ground, Dugway, Utah

(Manuscript received 27 November 2006, in final form 19 September 2007)

ABSTRACT

This study builds upon previous efforts to document the performance of the U.S. Army Test and Evaluation Command’s Four-Dimensional Weather Modeling System using conventional metrics. Winds, temperature, and specific humidity were verified for almost 15,000 forecasts at five U.S. Army test ranges using near-surface mesonet data. The primary objective was to use conventional metrics to characterize the degree to which forecast accuracy varies from range to range, within the diurnal cycle, with elapsed forecast time, and among the seasons. It was found that there are large interrange differences in forecast error, with larger errors typically associated with the ranges located near complex orography. Similarly, significant variations in accuracy were noted for different times in the diurnal cycle, but the diurnal dependency varied greatly among the ranges. Factor of 2 differences in accuracy were also found across the seasons.

1. Introduction

Very similar mesogamma-scale modeling systems have been running operationally at five U.S. Army Test and Evaluation Command (ATEC) ranges for over 5 years, providing an unusual opportunity for model verification in a variety of different climate zones. Aberdeen Test Center (ATC) is in a temperate, humid, coastal setting; the Cold Regions Test Center (CRTC) has a continental sub-Arctic climate; Dugway Proving Ground (DPG) is in a continental, midlatitude “cold” desert; and Yuma Proving Ground (YPG) and White Sands Missile Range (WSMR) are in subtropical hot deserts influenced by a monsoon. A description of this ATEC Four-Dimensional Weather (4DWX) Modeling System is found in Liu et al. (2008), the first paper in this series. Briefly, 4DWX uses the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5; Dudhia 1993; Grell et al. 1994) in a real-time four-dimensional data assimilation (RTFDDA) cycle in which the model is initialized using continuous data assimilation (observation nudging), typically producing 24–48-h forecasts every 3 h. High horizontal resolution is achieved through the use of three or four two-way-interacting, nested, computational domains.

There are multiple motivations for evaluating the accuracy of forecasts from these operational mesoscale modeling systems. First, as with any other operational model, the forecast accuracy needs to be documented in order to objectively and quantitatively assess the benefit of routine system changes, upgrades, or bug fixes. For example, there has been a general consensus in the modeling community that afternoon near-surface wind speeds are underestimated by mesoscale models in general (Zhang and Zheng 2004). The fact that this negative speed bias was observed in the verification statistics for the 4DWX system forecasts from all of the ranges confirmed that the problem was not region specific [e.g., related to inappropriate vertical momentum mixing due to erroneous roughness-length estimates; Liu et al. (2006)], which allowed 4DWX model devel-
opers to search for a more universal solution to the problem. In this case, Liu et al. (2006) identified and corrected a long-standing problem with the Medium-Range Forecast (MRF) Model boundary layer (BL) parameterization, which is used widely by the mesoscale modeling community.

One of the motivations for deploying customized, local-area, high-horizontal-resolution modeling systems such as the 4DWX system is that it is hoped that they can outperform coarser-resolution models, such as those run at the national level by the National Weather Service’s (NWS) National Centers for Environmental Prediction (NCEP) and by the Department of Defense (DoD) (Mass et al. 2002). Even though the potentially greater accuracy of the local-area models is only one of the motivations for their use, it is nevertheless important to develop metrics that demonstrate the advantages of the higher-resolution models, with one eventual goal perhaps being to define an optimal resolution.

Each ATEC test center has a somewhat different set of critical weather variables for which forecasts are required to support its mission, and these variables tend to be emphasized in the verification process. Because all ranges require forecasts of low-level winds and other variables, these variables are a verification priority at all ranges and will be the focus of this paper.

Some results of forecast verification studies for the 4DWX system have been published previously for specific variables, verification methods, and locations. For example, the 4DWX system was deployed to support the real-time hazard assessment for accidental releases of toxic industrial chemicals or terrorist incidents during the Olympics in Salt Lake City, Utah; Athens, Greece; and Turin, Italy. The near-surface and BL wind forecasts for these deployments have been objectively verified using conventional and newly developed object-based metrics (e.g., Rife et al. 2004). In addition, forecasts for ATEC ranges have been experimentally verified using the new metrics (e.g., Rife and Davis 2005). The present paper complements this previous work by focusing on additional forecast variables and by including all of the ATEC ranges. Because the newer metrics are experimental, and their strengths and weaknesses are still under study, only the conventional accuracy statistics will be included in this paper. Note that the accuracy statistics for the Electronic Proving Ground and the Redstone Technical Test Center 4DWX systems are not included here because they did not commence operation until 2006. Section 2 focuses on the verification of near-surface variables using conventional metrics, with emphasis on commonalities and differences among the ranges. The final section of the paper provides a discussion and summary.

2. Interrange comparison of verification of forecasts of near-surface variables

This section summarizes the verification of forecasts of near-surface wind, temperature, and humidity at the ATEC ranges. Because any analysis and interpretation of model accuracy should be done with knowledge of what has been learned in previous studies, the first subsection below provides a brief review of other verification studies of these variables that have been performed using the 4DWX and other mesogamma-scale modeling systems. The second subsection describes the current verification of the forecasts of these variables at the ATEC ranges.

a. Previous related verification studies

Several previous efforts have verified 4DWX modeling system analyses and forecasts of near-surface variables, especially winds. For example, Davis et al. (1999) and Rife et al. (2002, 2004) illustrate the model’s accuracy in predicting diurnal thermally forced BL winds associated with complex orography, the Great Salt Lake, and salt flats in the Great Basin Desert near DPG. The former paper suggests that nonsystematic (nonperiodic) circulations have low predictability. This point is substantiated in Rife et al. (2004), which uses the spectral decomposition of time series of observed near-surface winds to show that the model forecast accuracy is lower for locations where a larger fraction of the spectral power is on subdiurnal rather than diurnal time scales. That is, the results suggest that models can represent mesoscale circulation only if they are defined by observations in the initial conditions or are generated internally through local forcing. For cases with weak local surface forcing, predictability is low if observations are not available to represent mesoscale processes in three dimensions in the initial conditions. The Rife et al. (2004) results further show that, even when the model has skill in defining mesoscale features, small temporal and spatial offsets in the forecasts of these features (e.g., lake breezes, mountain-valley circulations, salt breezes) can cause large penalties in the model skill when traditional accuracy metrics are used. For the 40 days of model forecasts during the Salt Lake City 2002 Winter Olympic Games, the mean absolute error in the 12-h forecasts of the 10-m above ground level (AGL) wind was $60^\circ$–$75^\circ$ in the direction, and 1.75–2.00 m s$^{-1}$ in the speed.

Only a few of the many studies evaluating mesogamma-scale forecasts of near-surface variables by modeling systems other than the 4DWX system will be mentioned here. Case et al. (2002) applied a mesogamma-scale version of the Regional Atmospheric Mod-
el ing System (RAMS) for summer seasons in east-central Florida. They provide standard error statistics for near-surface variables for this area (which has fairly flat terrain) that are sometimes less than those obtained by Rife et al. (2004) using the 4DWX system in a complex-terrain area. The discussion in the next subsection addresses the effects of terrain complexity. A similarity in the Case et al. (2002) and Rife et al. (2004) studies is that the statistics from the mesogamma-scale RAMS and 4DWX models do not differ greatly from those of the much-coarser-resolution NCEP Eta Model. Similarly, Hart et al. (2004), who used the MM5 to evaluate the effect of model horizontal resolution on conventional performance statistics for near-surface variables, conclude that a factor of 3 change in horizontal resolution has no significant impact. Their simulations, which were for the same area near Salt Lake City studied by Rife et al. (2004) using the 4DWX system, produced verification statistics comparable to those reported by Rife et al. (2004). Such issues in mesoscale model verification challenge us to better understand the causes of variations with season, time of day, and location in the standard, commonly used skill metrics such as the root-mean-square error (RMSE).

b. Results of the verification of forecasts of near-surface variables at all ATEC ranges

This section uses conventional metrics to illustrate how the accuracy of wind, temperature, and specific humidity forecasts varies from range to range, during the diurnal and seasonal cycles, with forecast lead time, and as a function of the forecast cycle (eight per day). When possible, the above dependencies are ascribed to properties of the model, the verification metrics, and local climate features. The forecasts for all ranges are compared with range mesonet data, NWS routine aviation weather report (METAR) data, and non-NWS mesonet data for the area of model grid 3. All of the range mesonet data had the same sample rate and averaging period. Figure 1 illustrates the five grid 3 areas, including the orography and the locations of observations used in the verification. Table 1 lists the number of range mesonet observations and METARs that were used in the verification. Both the DPG and CRTC operational systems have a grid 4 with a 1.11-km grid increment embedded within this grid 3 area. For model verification at DPG and CRTC, each observation was compared with forecasts from the highest-resolution grid available at that location.

Table 2 summarizes the specific factors at each range that might contribute in various ways to differences in predictive accuracy. All of the ranges except ATC have nearby complex orography and experience the associated processes such as diurnal upslope-downslope circulations, katabatic flows, and channeling. Both ATC and DPG are close to sea or lake coastlines and can be affected by related thermal circulations. The ATC is on the west shore of the Chesapeake Bay. The nearby water temperature is specified using measurements from three buoys in the bay and the nearby coastal Atlantic Ocean. The water surface temperature of the Great Salt Lake is based on analyses from the NCEP North American Model (NAM), with adjustments in temperature for the nearshore water. The water temperature is updated daily, but the diurnal variation is not simulated. For small lakes that are not resolved by the NAM, the water temperature is defined by the average of the 24-h NAM surface air temperature. Overall, the landscape properties are sufficiently variable in the horizontal at all ranges to potentially generate local circulations [e.g., the salt-flat breezes documented for DPG by Rife et al. (2002)].

Besides the geographic differences, the ranges are affected by different prevailing synoptic environments. Both WSMR and YPG experience significant impacts from the North American monsoon in the summer. Snowfall and snow cover can affect the range-scale circulations at ATC, DPG, and CRTC in the winter, especially when snow–no-snow boundaries are present in the fine meshes. The continuous cycling process of the modeling system allows the model physics to simulate and track the snow cover evolution. Another possible source of variations in forecast accuracy between ranges is the use of the fourth, higher-resolution, grid at two of the ranges.

In addition to the range-specific factors listed in Table 2, general factors that can cause range-to-range predictability differences include the following.

- The accuracy with which the model physics represents the different prevailing processes; the accuracy of forecasts of physical processes that prevail in different climates and geographic areas depends on different components of the model physics. For example, the veracity of the land surface physics package determines the accuracy with which summer monsoon circulations are forecast at WSMR and YPG. The accuracy with which the model predicts coastal cyclogenesis influences the winter-season forecast skill at ATC.
- The degree to which the model solution contains small space- and time-scale features, which we know verify more poorly using conventional metrics. As was mentioned earlier, small time and space offsets of physical features in fields with a lot of structure verify more poorly than do smooth forecasts. Thus, conven-
FIG. 1. Grid 3 areas within which the verification of forecasts was performed for the different ATEC ranges. Shown are the terrain elevation (colors) and the locations of the range mesonet (stars) observations, as well as METARs and other surface observations (triangles), used for verification. Red lines show range and geographic boundaries, and a white inset box defines the grid 4 area for CRTC and DPG.
tional skill statistics are typically going to be poorer in geographic areas where there are mesoscale structures from local forcing like orography.

- The type, density, and distribution of the available meteorological data that are assimilated to define forecast initial conditions and used in verification. Clearly, in general, the more data that are available to initialize a forecast, the better the forecast skill.

- The accuracy with which (invariant and variable) land surface properties are defined in the forecast initial conditions. Land surface properties in some geographic are less well defined than in others, for example because human-induced changes have been more extensive since the last survey. Thus, the skill with which surface-forced features are forecast will be regionally variable.

- The complexity of the topography and other surface properties, which causes representativeness errors of point observations relative to model grid-box-average values. Verification involves comparison of point observations with model values that correspond to grid-box averages. Where topography causes sign-

ificant spatial complexity in the atmosphere, point observations are less likely to correspond to larger-scale averages represented by the model, even for “perfect” model forecasts. See Rife et al. (2004) for further discussion of this issue.

The accuracy statistics discussed in this section are based on ~15,000 forecasts performed during the period 1 January–31 December 2003, a period that was chosen because the model configurations were relatively invariant. Both bias and RMSE were calculated by interpolating from the model grid to the observation locations using bilinear interpolation from the four neighboring grid points. Winds are measured at 10 m above ground level (AGL), and temperature and specific humidity are measured at 2 m AGL. Because the lowest model computational level is at about 15 m AGL, it was necessary to use the Monin–Obukhov similarity theory to extrapolate the wind, temperature, and humidity predictions to the height of the observations for verification purposes. The formulation of the Monin–Obukhov theory used in the diagnostics is an integral part of the MM5 surface layer physics module (Grell et al. 1994). Winds with speeds of less than 0.5 m s\(^{-1}\) were not used in the calculation of the direction error. Much of the interpretation of the following results will be deferred until the summary and discussion in section 3.

1) Wind speed and direction

Figure 2a shows the wind direction RMSE for forecast hours 10–12, as a function of coordinated universal time (UTC), for each of the five ranges. All 3-hourly

---

**Table 1. Number of range mesonet and METAR observations used in the verification for each range (within grid 3).**

<table>
<thead>
<tr>
<th>Range</th>
<th>No. of range obs</th>
<th>No. of METAR obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aberdeen Test Center</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Cold Regions Test Center</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Dugway Proving Ground</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>White Sands Missile Range</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>Yuma Proving Ground</td>
<td>13</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 2. Properties of the weather and modeling systems at the ATEC ranges that could affect objective verification statistics.**

<table>
<thead>
<tr>
<th>Range</th>
<th>Physiographic setting</th>
<th>Grid increments (km)</th>
<th>Important mesoscale physical processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aberdeen Test Center</td>
<td>Slight orographic variation; complex coastline of Chesapeake Bay and Atlantic Ocean; dense vegetation</td>
<td>30/10/3.33</td>
<td>Sea/bay breeze; moist convection</td>
</tr>
<tr>
<td>Cold Regions Test Center</td>
<td>North slope of the Alaska mountain range; taiga vegetation</td>
<td>30/10/3.33/1.11</td>
<td>Katabatic/drainage flow during most times of day and seasons</td>
</tr>
<tr>
<td>Dugway Proving Ground</td>
<td>Flat intermountain desert valley with salt flats; surrounding nearby complex orographic features; sparse vegetation</td>
<td>30/10/3.33/1.11</td>
<td>Upslope/drainage flow from mountains, penetrating into valley; salt breeze; lake breezes from Great Salt Lake and Utah Lake</td>
</tr>
<tr>
<td>White Sands Missile Range</td>
<td>Flat intermountain desert valley with high-albedo gypsum sands and low-albedo lava flows; surrounding nearby complex orographic features; sparse vegetation</td>
<td>30/10/3.33</td>
<td>Monsoon convection; upslope/drainage flow from mountains, penetrating into valley; thermal effect of the white sands</td>
</tr>
<tr>
<td>Yuma Proving Ground</td>
<td>Flat desert, with nearby complex orography; very sparse vegetation</td>
<td>30/10/3.33</td>
<td>Monsoon convection; upslope/drainage flow from mountains, penetrating into valley</td>
</tr>
</tbody>
</table>
forecast cycles are represented here. For example, for each range, the data plotted for 1800, 1900, and 2000 UTC are from hours 10–12 of the 365 forecasts initialized at 0800 UTC. This type of plot isolates accuracy differences as a function of time during the diurnal cycle for a given forecast lead period. Some ranges, such as ATC, show very little variation in RMSE as a function of the diurnal cycle. In contrast, the YPG
RMSE varies by $\sim 20^\circ$ (~25%), with the highest values at night and in the early daylight hours. One of the most significant differences among the ranges is the much lower wind direction RMSE for ATC. The RMSE for ATC is less than 50° for all times of the day, whereas for all of the other ranges the diurnal average is 75° or more. A possible explanation is that ATC is the only range that is not strongly influenced by thermally forced circulations associated with nearby complex orography (see Table 2). Thus, forecasts for ATC suffer much less from the aforementioned skill penalty associated with the position and timing errors in forecasts of mesoscale features of orographic origin.

A second factor could cause greater error in complex terrain. Even mesogamma-scale models have representativeness errors in which the grid-box average wind is not representative of the conditions at the location of the observations. This will be greatest where there are finescale orographic effects. Also, convection in semiarid areas is focused over higher orography, and the timing, location, and structure of this convection and its effects on local winds and other variables are difficult to predict. Figure 3 illustrates the wind direction RMSE for the 10–12-h lead time forecasts for the outer grid (30-km grid increment) of the WSMR system. It can be seen that, even with this coarse-grid simulation, there clearly is a spatial relationship between the wind direction RMSE and the complexity of the orography. These results indicate that mesoscale model forecast verification over complex terrain remains challenging for a variety of reasons.

A third factor affecting the predictability at ATC and not the other ranges is the existence of a sea breeze from the Atlantic Ocean and the Chesapeake Bay. On the one hand, the existence of this diurnally forced phenomenon has the potential of increasing the overall predictability. However, if the timing and strength of the sea breeze are not forecast correctly, the accuracy metrics used here could strongly penalize the forecast. In Fig. 2b, the RMSE is plotted for each range as a function of elapsed forecast time out to 15 h. It shows that the RMSE typically increases by about $10^\circ$–$15^\circ$ during the first three to four forecast hours, and then levels off. This very short period of error growth is consistent with that estimated by Davis et al. (1999).

The initial error for ATC is less than half that for DPG.

Figures 2c and 2d show how the RMSE varies for each of the eight daily forecasts, for WSMR and ATC, respectively. All cycles at WSMR show a pronounced error minimum (amplitude of about 10°) during the local afternoon hours, regardless of when this falls during the forecast. This can also be seen in the WSMR curve in Fig. 2a, which applies for the 10–12-h forecast period. Note that the error actually decreases after initialization for the forecasts initialized at 1700 UTC (1000 LT). This RMSE minimum is less pronounced for ATC, but still exists. At other ranges, such as YPG (not shown), the amplitude of the variation is over 30°, with the minimum at 1700 LT.

Figures 2e and 2f show the seasonal variations in wind direction RMSE for the 10–12-h forecasts at ATC and WSMR. For both ranges, the summer forecasts have higher average RMSEs, indicating possible contributions from convective winds that are not forecasted, or stronger thermally forced circulations whose incorrect timing and position are severely penalized by the accuracy metric. At ATC, the winter season RMSE is lowest, but at WSMR there is no single season with clearly the lowest RMSE.

For brevity, wind speed statistics are not shown graphically. For this quantity, there is less variation in the RMSE among the ranges than is seen with wind direction. For example, both ATC and YPG have similar 10–12-h forecast RMSEs of 2.1–2.2 m s$^{-1}$. The largest 10–12-h forecast wind speed RMSE is for WSMR, and varies from 2.4 to 2.8 m s$^{-1}$, depending on time of day. It is interesting that the ATC wind direction RMSEs are lowest in winter relative to other seasons (Fig. 2f), but the speed RMSEs are highest in winter. The latter is associated with the strong synoptic weather forcing during winter. Wind speed and direction biases also are not illustrated. For the DPG and WSMR 10–12-h forecasts, the wind direction bias typically does not exceed $\pm 10^\circ$ for all seasons and times of day, but for ATC and YPG the bias at some times exceeds $15^\circ$–$25^\circ$.

2) Temperature

Figure 4a, which is analogous to Fig. 2a, shows the temperature RMSE for forecast hours 10–12 as a function of UTC. As with wind direction, the temperature RMSE is lower at ATC than at the other ranges. At times, the RMSE varies among ranges by almost a factor of 2, and for individual ranges the error varies with time of day by 15%–30%. All ranges except CRTG have a temperature RMSE minimum in the morning. As with wind direction, the smaller temperature RMSE for ATC could be related to smoother orography. When a plot similar to that in Fig. 3 is prepared for temperature (not shown), there is a clear tendency for larger RMSEs to exist in areas of high orography.

Other reasons why temperature RMSEs could be higher in regions with complex orography include the following. 1) The true terrain elevation is different from the elevation in the model because of the model resolution; consequently, the mountain stations experience
an ambient large-scale temperature climate that is different than that at their model elevations. 2) Radiation parameterizations tend to be tuned for lower elevations where there are more verification stations. Also, the parameterizations do not adequately represent the effect on insolation of the shorter column of more pristine atmosphere in mountains. 3) Terrain shadowing is not represented in the model, but affects the actual surface energy balance in mountains at low sun angles. 4) The smoother-than-real orography in the model should make thermally forced circulations weaker, meaning that pooling of cool air in valley floors at night is slower to occur, and subsidence warming in the valleys during the day is less intense. 5) Over high-elevation deserts, the amplitude of the diurnal cycle is typically large, possibly leading to larger errors.

As is the case with wind direction, Fig. 4b shows that the relative accuracy of the forecasts is generally related to the initial error, with the greatest increase in error during the 0–5-h period, followed by much slower...
increases from 6 to 9 h. At the end of the 15-h forecast, the RMSE is $\sim 2.4^\circ$C for ATC, while for WSMR it is $\sim 3.8^\circ$C. Figures 4c and 4d show the variation in temperature error during the forecast period for two example ranges, with the statistics shown for each cycle. For WSMR, there is a clear error minimum in the late morning for the forecasts that span that time of day, and there are prominent error maxima at 0700 and 1900 LT. The plots for ATC show two definite minima, one at about 0900 LT and one at about 1900 LT. Time-of-day variations in temperature RMSE of similar amplitude are seen for all ranges except CRTC. Forecast-error decreases during certain periods perhaps imply that better substrate temperature forecasts during that

---

**Fig. 4.** As in Fig. 2 but here the plots are of temperature RMSE, the observations are at 2 m AGL, and (e) applies to CRTC.
time of day reduce the error in the atmospheric forecast that accumulated earlier.\(^1\) Figures 4e and 4f show seasonal differences in the accuracy of the 10–12-h forecasts, for CRTC and ATC. For ATC, the errors in the nighttime forecasts only differ by \(\sim 5\%\) among the seasons. The mean daytime error is about the same for all seasons, but the error in winter has a somewhat different diurnal pattern. In contrast, CRTC shows an error in winter that is almost twice that in summer. Temperature biases for all ranges are typically \(\pm 0^\circ–2^\circ\)C, with CRTC having a 2°C warm bias in the winter, irrespective of the time of day (the diurnal variation is very small in winter). The summer CRTC error shows a maximum during the nighttime twilight hours, which may be related to the higher errors noted by Rife et al. (2004) during the twice-daily BL transition periods.

3) SPECIFIC HUMIDITY

There appears to be little relationship between the ambient water vapor content of the atmosphere at a range and the specific humidity RMSE. For example, Fig. 5a shows that the ATC RMSE is one of the lowest, even though the range’s annual average specific humidity is about 2–5 times larger than those of the other ranges that are in arid and sub-Arctic climates. In contrast to wind direction and temperature RMSEs, there does not seem to be a relationship between specific humidity RMSE and complex orography (based on a map of the type in Fig. 3; not shown). The daily average RMSE varies by about a factor of 2 among the ranges, with WSMR and YPG (in subtropical deserts and with a monsoon season) having the highest values. The times during the diurnal cycle of maximum and minimum error vary among the ranges, with the amplitude being \(\sim 20\%–40\%\) of the mean. At ATC, for example, the amplitude in the diurnal RMSE variation is \(\sim 45\%\) of the mean, with the maximum near 1200 LT (Figs. 5a and 5d). The RMSE approximately doubles during the forecast period, with most of the increase during the first 4–5 h (Fig. 5b). An extreme example of error growth is seen for ATC, where forecasts initialized in the early morning, before the time of maximum errors, show RMSE growths of a factor of 3 in 6–8 h (Fig. 5d).

It is curious that the specific humidity error minimum at WSMR (Figs. 5a and 5c) occurs at about the same time as the minimum in the temperature error there (Figs. 4a and 4c). For most ranges, summer RMSEs tend to be the highest, and those in the winter the lowest (Figs. 5e and 5f). The 10–12-h forecasts for all ranges but ATC show a wet bias, with a WSMR bias of 1–2 g kg\(^{-1}\) and the rest closer to 0.5 g kg\(^{-1}\) for all times of the day. For ATC, there is a very slight dry bias at night, which grows larger, to 1 g kg\(^{-1}\), during the day.

3. Summary and discussion

This study documents the accuracy of ATEC 4DWX system forecasts using conventional measures. It expands upon previous studies by comparing the accuracy of the forecasts of near-surface conditions among all of the ranges.

The interrange comparison of wind, temperature, and specific humidity forecast accuracy is based on almost 15 000 model forecasts: five ranges, eight forecasts per day, over a 1-yr period. This comparison provides insight into the degree to which accuracy varies 1) from range to range, 2) within the diurnal cycle, 3) with elapsed forecast time, and 4) among the seasons. Overall, there were large range-to-range differences in the season-aggregated RMSE. For example, for wind direction, the interrange RMSE differences for 10–12-h forecasts were much larger than those associated with applying different models, having over an order of magnitude difference in horizontal resolution, to the same area (Case et al. 2002; Rife et al. 2004). Compared with the other ranges, the ATC wind direction RMSE is anomalously low. It is hypothesized that the \(\sim 50\%\) higher RMSE at the other ranges is due to the existence of complex orography. The outer grid for the WSMR model spanned both complex and relatively flat orography, and wind direction RMSEs were about twice as large in mountainous areas. Comparable terrain-related accuracy differences were found by Warner and Sheu (2000) using a similar modeling system over the Middle East. It is interesting that Case et al. (2002) report near-surface wind direction RMSEs from a mesosigma-scale version of RAMS applied over Florida that are similar to the ATC errors reported here for similarly flat orography. Factor-of-2 differences among ranges in 10–12-h forecast temperature and specific humidity RMSEs also prevail, with values of both variables for WSMR being 2 times those for ATC, for example. This result again indicates the challenges associated with model verification in regions of complex terrain.

The season-aggregated statistics also showed considerable diurnal dependence in model performance for some ranges. For example, the 10–12-h forecasts for
YPG had RMSEs that varied by 30% of the mean RMSE for wind direction, with larger errors at night and smaller values during the day. For temperature, the YPG error varied by almost 50% of the mean RMSE, with a distinct maximum in the evening hours. Diurnal variations in specific humidity RMSE were smaller (10%–20% of the mean). When the RMSEs for individual forecast cycles were plotted, all cycles followed the same diurnally dependent error pattern, regardless of when during the forecast cycle the time of the error minima and maxima occurred. In some cases when the forecasts were initialized immediately before a diurnal minimum, the forecast error decreased after initialization.

Fig. 5. As in Fig. 2 but here the plots are of specific humidity RMSE and the observations are at 2 m AGL.
During the 15-h forecast period studied, RMSEs increased from their initial value by 50%–100%, with most of the increase occurring during the first 3–6 h of the forecast. There were considerable range-to-range and variable-to-variable differences in the amount of error growth, but there was little difference in the time interval after initialization when it occurred. For example, the YPG and ATC temperature RMSE growth occurred mostly during the first 6 h of the forecasts, but the amount of the YPG growth was twice that of ATC.

Ranges showed considerable differences in their seasonal variations in the error. For example, the ATC temperature RMSEs for 10–12-h forecasts were virtually the same for all seasons, especially at night, but the CRTC RMSEs were twice as large in the winter as in the summer. Given that CRTC is located at ~64°N, the winter months are dominated by shallow katabatic flows from the Alaska Range, and extremely intense surface-based temperature inversions (e.g., 45°C in the lowest 10 m). Neither the model’s vertical resolution in the BL, nor its land surface physics, is adequate to represent these phenomena well, so it is understandable that the error in near-surface temperature is larger in the cold season. What is perhaps surprising is that the error is not, in fact, larger than it is. Other variables showed factor-of-2 differences in RMSE for different seasons.

The bias errors in these forecasts can, of course, be removed to some degree through postprocessing the model output, but the nonsystematic errors will remain to challenge operational forecasters. However, the availability to the forecasters of the above time of day and seasonal dependence of the forecast accuracy should be of value in decisions about the degree of confidence to be placed in forecasts. In addition, the error statistics can help focus efforts to improve model performance.

Acknowledgments. This work was funded by the U.S. Army Test and Evaluation Command through an interagency agreement with the National Science Foundation. Carol Park provided editorial assistance.

References


