

Predictability of Low-Level Winds by Mesoscale Meteorological Models

DARAN L. RIFE, CHRISTOPHER A. DAVIS, AND YUBAO LIU

National Center for Atmospheric Research, Boulder, Colorado*

THOMAS T. WARNER

National Center for Atmospheric Research, and Program in Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado

(Manuscript received 30 October 2003, in final form 26 April 2004)

ABSTRACT

This study describes the verification of model-based, low-level wind forecasts for the area of the Salt Lake valley and surrounding mountains during the 2002 Salt Lake City, Utah, Winter Olympics. Standard verification statistics (such as bias and mean absolute error) for wind direction and speed were compared for four models: the Eta, Rapid Update Cycle (RUC-2), and Global Forecast System of the National Centers for Environmental Prediction, and the fifth-generation Pennsylvania State University–NCAR Mesoscale Model (MM5). Even though these models had horizontal grid increments that ranged over almost two orders of magnitude, the highest-resolution MM5 with a 1.33-km grid increment exhibited a forecast performance similar to that of the other models in terms of grid-average, conventional verification metrics. This is in spite of the fact that the MM5 is the only model capable of reasonably representing the complex terrain of the Salt Lake City region that exerts a strong influence on the local circulation patterns. The purpose of this study is to investigate why the standard verification measures did not better discriminate among the models and to describe alternative measures that might better represent the ability of high-horizontal-resolution models to forecast locally forced mesogamma-scale circulations. The spatial variability of the strength of the diurnal forcing was quantified by spectrally transforming the time series of wind-component data for each observation location. The amount of spectral power in the band with approximately a diurnal period varied greatly from place to place, as did the amount of power in the bands with periods longer (superdiurnal) and shorter (subdiurnal) than the diurnal. It is reasonable that the superdiurnal power is largely in the synoptic-scale motions, and thus can be reasonably predicted by all the models. In contrast, the subdiurnal power is mainly in nondiurnally forced small-scale fluctuations that are generally unpredictable with any horizontal resolution because they are unobserved in three dimensions by the observation network.

A strong positive relationship is demonstrated between the strength of the local forcing at each observation location, as measured by the spectral power in the diurnal band of the wind component time series, and forecast skill, as reflected by an alternative verification metric, a measure of anomaly correlation. However, the mean-absolute error showed no relationship to the power in the diurnal band. Two other measures of comparison among the low-level wind forecasts, the direction climatology and the spatial variance, showed a positive correlation between forecast quality and horizontal resolution.

1. Introduction

The need for accurate transport and dispersion (T&D) forecasting techniques has become increasingly important because of the threat of the intentional release of hazardous material into the atmosphere. Particularly in areas of complex local surface forcing, and for longer transport distances, mesoscale-model-generated forecast

winds must be employed as input to the T&D models. In such situations, the products from coupled meteorological and T&D models are being used operationally by emergency managers for training and for consequence analysis. For example, during the 2002 Salt Lake City, Utah, Winter Olympics, mesoscale models were run operationally for over 3 months to provide high-resolution meteorological fields to T&D models for the Salt Lake City area and all Olympic venues.

This coupling of meteorological models with T&D models is motivation for developing better methods for objectively assessing the quality of model wind predictions. This is an especially relevant task because conventional objective measures of forecast quality sometimes seem to poorly reflect the improvement that one

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Daran L. Rife, NCAR, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: drife@ucar.edu

might intuitively expect from increased horizontal resolution. For example, Mass et al. (2002) describe the overall performance of a real-time mesoscale weather prediction system and show that there were clear improvements in the objectively measured forecast accuracy as the horizontal grid spacing was decreased from 36 to 12 km. In contrast, there were only small improvements in the objective quality as the grid spacing was decreased from 12 to 4 km. Similarly, Davis et al. (1999) showed that, in terms of conventional verification scores such as bias, mean-absolute error (MAE), and root-mean-square error (RMSE), a high-resolution (1.11-km grid increment) mesoscale model that was run operationally over the northern Utah region provided only slightly better surface temperature forecasts than did the much coarser resolution 80-km Eta Model of the National Centers for Environmental Prediction (NCEP), with the two models exhibiting 10-m wind field forecast errors of comparable magnitude. Another study of mesoscale model performance over northern Utah showed that reducing the horizontal grid spacing from 12 to 4 km produced little or no improvement in the prediction of surface temperature, relative humidity, and winds (Hart et al. 2004). A study for east-central Florida compared conventional objective verification scores from a mesoscale model, which employed a 1.25-km horizontal grid increment, to the scores from the 32-km grid-increment Eta Model (Case et al. 2002). The high-resolution model provided little objective improvement over the much coarser Eta Model.

As a contribution toward better understanding this paradox in the context of low-level winds, this study compares the forecast quality for four models that ran operationally during the Salt Lake City Olympics: 1) a specially adapted version of the fifth-generation Pennsylvania State University (PSU)–National Center for Atmospheric Research (NCAR) Mesoscale Model (MM5), 2) the NCEP Global Forecast System (GFS), 3) the NCEP Rapid Update Cycle model (RUC-2), and 4) the NCEP Eta Model. These models had horizontal resolutions that spanned almost two orders of magnitude. The coarser-resolution models are clearly more suitable for general weather prediction than for defining mesoscale wind fields for T&D calculations. Another MM5, operational during the same period, had its fine grid located over the White Sands Missile Range, and its coarse grid, with a 30-km grid increment, spanned the Salt Lake City study area. Because, except for horizontal resolution, this model was identical to the one deployed for the Olympics, more direct comparisons are possible of the effects of resolution on forecast accuracy.

It will be shown that the 10-m-AGL wind forecasts from the three NCEP models have roughly similar error in terms of standard verification measures. This is understandable given that the models' publicly available output datasets, which were used in this study, were defined on a 40-km grid for the Eta and RUC-2 models, and on a 1° (~ 111 km) grid for the GFS. The 30-km-

grid-increment MM5 had comparable error. In each case, the complex terrain of the Salt Lake City region is poorly resolved, as is its thermal forcing of the low-level winds. However, the objective performance of the MM5 that resolved the local physiographic features reasonably well with its high-resolution grid (1.33-km grid increment) was only marginally better than that of the other models. Nevertheless, the climatology of the 10-m-AGL observed wind for this area shows a strong and intuitively reasonable signal associated with the thermal forcing from the mesoscale terrain. These features should be predictable by any mesoscale model with sufficient horizontal resolution and adequate surface and boundary layer physics (assuming an accurate specification of the ground surface characteristics such as soil temperature and moisture, water surface temperatures, and snow depth).

In order to better quantify the quality of boundary layer wind predictions by mesoscale models, the following questions will be addressed:

- What is the relative accuracy of the low-level wind forecasts from the four models in terms of conventional verification statistics (bias, MAE, RMSE)? How do these statistics compare with those from "no skill" forecasts and from "perfect model" forecasts?
- Is there significant spatial variation in the strength of diurnally forced circulations, and is the high-resolution model's accuracy related to the strength of the diurnal circulations?
- What measures of forecast quality are more useful than the traditional objective metrics for evaluating the accuracy of low-level wind forecasts for T&D calculations?

This paper is organized as follows. The MM5 modeling system used in this study is described in the next section. Section 3 describes the model output data and the observational datasets. Section 4 presents the results of the conventional verification statistics, as well as an examination of the maximum and minimum forecast error that is practically realizable over the study region (details found in the appendix). A spectral decomposition of the time series of the observations into diurnal, subdiurnal, and superdiurnal bands is described in section 5. In section 6, alternative verification procedures that better discriminate among the models are presented. The paper concludes with a discussion and summary of the results.

2. Model description

The NCEP models whose products are employed in this study have been well documented in the open literature, so their specifications will not be repeated here. However, the MM5 system used in support of the Olympics has not been described elsewhere, so this section will be devoted to a summary of its characteristics.

The nonhydrostatic MM5 (Dudhia 1989, 1993; Grell

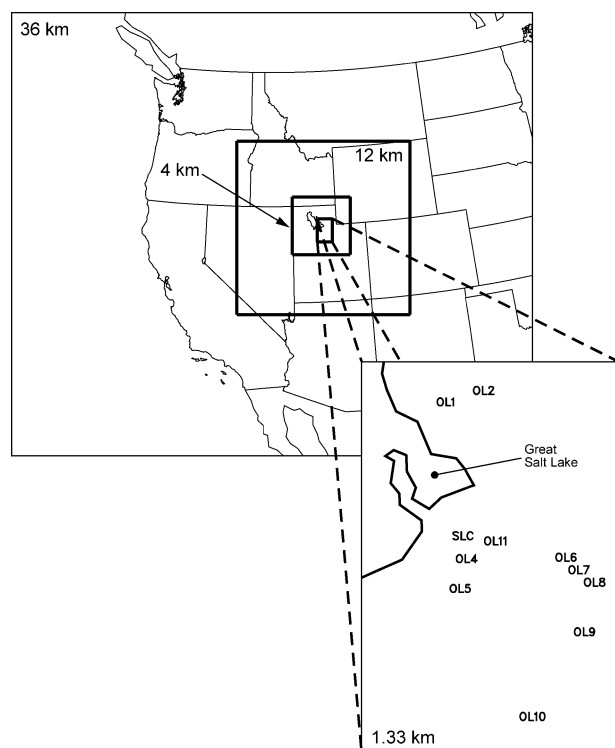


FIG. 1. Area coverage for the four MM5 computational grids. The grid increment for each grid is indicated. The expanded inner grid shows the shore of the Great Salt Lake (heavy line) and the locations of most of the Olympic-event venues (number and letter codes).

et al. 1994) is a full-physics limited-area model. It has many options for parameterization of physical processes such as moist convection and boundary layer turbulence. The version of the MM5 used in this study is part of a rapidly deployable, operational, mesogamma-scale weather analysis and forecast system that has been developed by NCAR for various U.S. Army Test and Evaluation Command facilities (Davis et al. 1999). This system is composed of two principal components. One component employs the meteorological model in four-dimensional data assimilation (FDDA) mode, wherein artificial tendency terms are used in the prognostic equations to relax the model state toward observations (Seaman et al. 1995; Stauffer and Seaman 1990; Stauffer et al. 1991). This FDDA system runs continuously, assimilating surface mesonet data, radiosonde data, satellite-derived cloud-track winds, surface-based profiler data, and Automated Commercial Aircraft Reporting System data. Because model errors may accumulate in data-sparse regions, the system is restarted from an objective analysis every 7 days. The forecast component of the system is initialized from the model-assimilated datasets at an interval of 1–3 h, with forecast durations of typically 12–36 h.

The model used in this application had four nested, two-way-interacting, computational grids, which are depicted in Fig. 1. The four grids had grid increments of

36, 12, 4, and 1.33 km, and mesh sizes of 70×82 , 82×82 , 82×82 , and 97×64 , respectively. All grids used 36 unevenly spaced vertical computational levels, extending from approximately 15 m to about 17 km AGL. The distribution of vertical levels provided the greatest resolution in the PBL and near the tropopause. The PBL parameterization employed was the Medium-Range Forecast (MRF) model PBL scheme, as implemented in the GFS (Hong and Pan 1996). Grids 1 and 2 utilized the Grell (1993) cumulus parameterization, with no convection parameterized on grids 3 and 4. Longwave and shortwave radiation interact with the clear atmosphere, cloud, precipitation, and the ground (Dudhia 1989). The explicit cloud microphysical scheme of Hsie et al. (1984) was used and includes improvements to allow ice-phase processes below 0°C (Dudhia 1989). The surface energy and water budgets were computed using a multilayer soil model (Dudhia 1996; Dudhia et al. 2002). The substrate soil moisture varied with time in response to predicted rain and/or snow accumulation, snowmelt, and evaporation from the ground surface. The dominant vegetation type at each model grid point was specified through the use of the U.S. Geological Survey Earth Resources Observing System 1-km dataset (Loveland et al. 1995), with climatological values of albedo, graybody emissivity, aerodynamic roughness length, and thermal inertia assigned to each category. Lateral-boundary conditions for the outer grid, grid 1, were defined using linear temporal interpolations between 3-hourly Eta Model analysis and forecast fields (with 40-km grid increment). The Hat Island, Utah, lake temperature observations were used to specify the Great Salt Lake surface temperatures in the model. Forecasts of 12-h duration were initiated every 3 h, with all the system specifications determined by the fact that calculations needed to be performed on a 32-node Linux PC cluster. The terrain representations of each of the four modeling systems over the Salt Lake City Olympics study area are presented in Fig. 2. As noted earlier, the MM5 that was centered over the White Sands Missile Range, whose coarse-grid solution was used for comparison in the Salt Lake City area, was identical to the Olympics system except for horizontal resolution and the location of the lateral boundary.

3. Observations and model-forecast data

a. Observations

The grid-4 region contains much complex terrain and a dense network of about 200 surface observation sites operated by a variety of agencies such as the National Weather Service (NWS), the Utah Department of Transportation, the Utah Department of Air Quality, and the National Resources Conservation Service. The observation data were obtained in real time, primarily through the MesoWest network (Horel et al. 2002) and the NWS Telecommunications Gateway, during the 3 February to

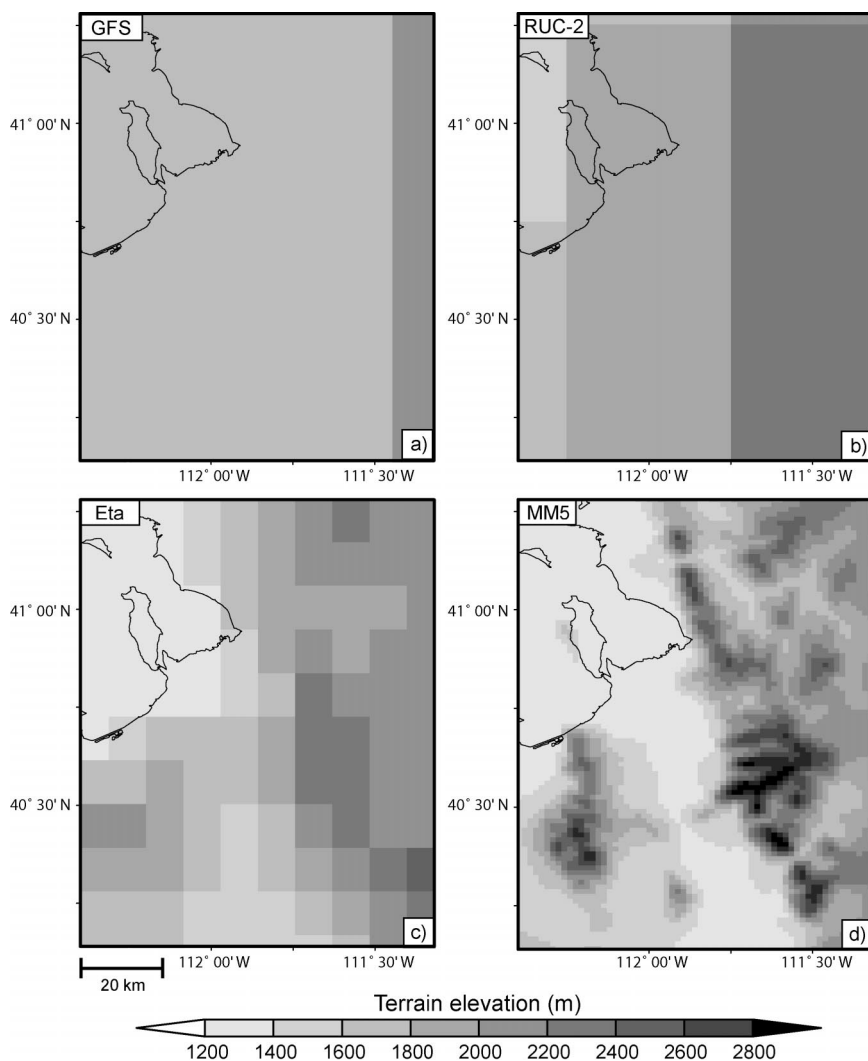


FIG. 2. Terrain representation for the (a) GFS, (b) RUC-2, (c) Eta Model, and (d) MM5 over the study area, as defined on their native grids. Elevation (m) is defined on the scale at the bottom.

30 April 2002 study period. Data were employed for verification from the 28 observation stations for which at least 80% of the possible measurements existed in the 86-day record (Fig. 3). The station siting characteristics are quite variable in terms of the local terrain and vegetation.

b. Model data

Output from the NCEP models was obtained from archives that contained the highest-resolution publicly available datasets for each model during the study period. The GFS data were available from forecasts initialized at 0000 and 1200 UTC on a 1° grid. The RUC-2 data were available from forecasts initiated every 3 h, starting at 0000 UTC each day, on a grid with a 40-km horizontal increment. The Eta Model forecasts initialized at 0000 and 1200 UTC were available with the

same horizontal and vertical resolution. The MM5 data from the four computational grids were archived on their native grids from forecasts initiated every 3 h, beginning at 0000 UTC each day. The output files from all four models included several surface layer fields, such as 10-m-AGL winds and 2-m-AGL temperature and humidity. The MM5 surface layer fields were computed by extrapolating, using similarity theory (Stull 1998), from the lowest model computation level of 15 m AGL to the 2-m-AGL observation level for temperature and humidity, and to the 10-m-AGL level for the wind. Similar methods are used by NCEP for extrapolation to the observation levels. The temporal frequency of the forecast output from the NCEP and MM5 models was 3 and 1 h, respectively. Only the 3-hourly output from the MM5 was used in the present study, corresponding to the forecast output times of the RUC-2 model. The GFS and Eta Model forecasts initialized at 0600 and 1800 UTC

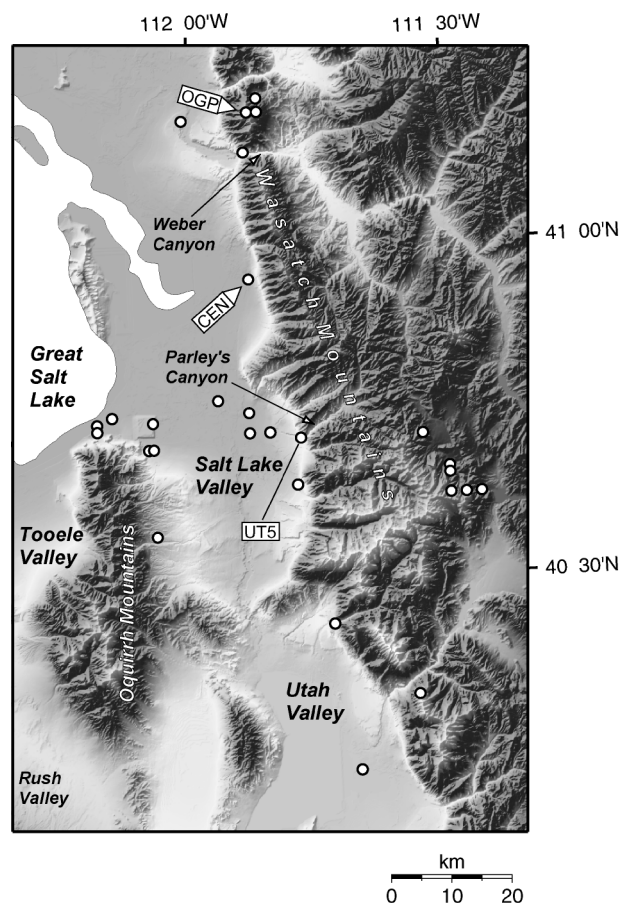


FIG. 3. The locations of the 28 surface observation stations (white circles) used for the study, within the grid-4 region. These sites were selected based on the high reliability and timeliness of their reports (all stations reported at least 80% of the time during the study period). Also displayed is the actual topography of the region. The number and letter codes identify stations referenced in the text.

were unavailable from the archives and thus were not used in this study.

4. Conventional verification statistics

The verification was performed within the MM5 grid-4 area (Fig. 3) for days on which forecasts from all four models were available. The general approach was to validate forecasts at the observation sites. This was done by bilinearly interpolating the surface layer fields from the model output files to the observation sites, using observations that reported within a 10-min window centered on each forecast valid time. In addition, the winds from the RUC-2, Eta, and MM5 models were rotated to an earth-relative reference coordinate to provide the direction from true north. No attempt was made to compensate for the difference between model and actual terrain elevations at the observation sites. As noted, only the stations that reported at least 80% of the possible times were used in the verification.

Three conventional verification scores were calculated for the wind fields: bias, MAE, and rmse. These statistics were computed as a function of forecast lead time, and then further stratified by time of day. The representativeness error associated with this study is estimated in the appendix and is used for defining the maximum MM5 skill (or minimum error) that is practically achievable over the study region, given the properties of the forecasting systems and the verifying observations. This error results from the fact that the observation defines conditions at a point, whereas the model forecast represents a grid-box average. Based on this estimate, the representativeness errors for 10-m-AGL wind speed and direction, under well-mixed PBL conditions with this MM5 model resolution in complex terrain, are 1.15 m s^{-1} and 14.6° , respectively. In addition, conventional cup and vane anemometers are generally accurate to within $\pm 0.3 \text{ m s}^{-1}$ and $\pm 3^\circ$ for wind speed and direction, respectively (W. Dabberdt 2003, personal communication). This yields a practically realizable minimum error for a wind speed and direction forecast by a perfect model of 1.45 m s^{-1} and 17.6° , respectively (assuming the errors are additive). The opposite error bounds are also estimated in the appendix. These are the thresholds beyond which the forecasts have 1) no skill, and 2) no skill beyond what could be achieved through the use of simple procedures such as diurnal persistence. The no-skill MAE scores are approximately $2\text{--}3 \text{ m s}^{-1}$ for wind speed and $80^\circ\text{--}90^\circ$ for wind direction, depending on the time of day. For the diurnal persistence forecasts, the MAE scores range from 65° to 80° for wind direction and 1.6 to 1.9 m s^{-1} for wind speed.

Figure 4 summarizes the diurnal characteristics of the MAE for 10-m-AGL wind direction and speed over the grid-4 area for all four models, for the analyses and 12-h forecasts. The bounding diurnal-persistence, no-skill, and perfect-model forecast curves are also shown. For each 3-hourly forecast valid time, there are between about 1650 and 2100 pairs of observations and forecasts from a particular model. It is evident from Figs. 4a and 4b that there are fairly substantial wind-direction errors associated with the forecasts from all four models over this region, with MAEs ranging from about 50° to 80° , depending on the model and time of day. Wind speed MAEs from all models cluster around 2 m s^{-1} (Figs. 4c,d). Additionally, the overall magnitude of the error grows slightly as the forecast length increases. For example, the average MM5 wind-direction MAEs for the analyses and 12-h forecasts are 61° and 66° , respectively. The wind-direction forecasts from MM5 are generally superior at all lead times to those from the other models, although this difference in error is modest ($5^\circ\text{--}15^\circ$). Even though the objective verification scores of the three NCEP models are similar to those of MM5, statistical significance tests on the differences in scores among all the models indicate that an MAE difference of more than about 2° or 0.08 m s^{-1} is significant at the

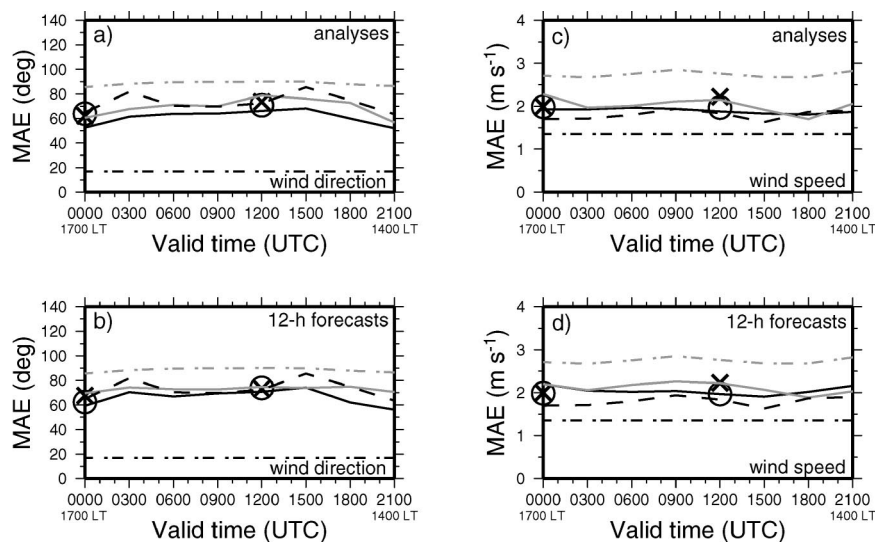


FIG. 4. Diurnal characteristics of the grid-4 average MAE for 10-m AGL wind direction (left panels) and wind speed (right panels) for the NCEP and MM5 models during the 3 Feb to 30 Apr 2002 period, calculated as a function of forecast lead time. Also displayed are the corresponding statistics from the diurnal persistence, random “no skill,” and “perfect” model forecasts.

Legend

- MM5
- Eta
- × GFS
- RUC-2
- - Diurnal persistence
- - Random “no skill” forecast
- - “Perfect” model forecast

99% confidence level. The 30-km MM5 wind-direction and wind speed statistics were similar to those for the NCEP models (not shown).

Interesting features of the wind-direction statistics are the two maxima in the MM5 error curves at 0300 and 1500 UTC. The average time of sunset was about 0130 UTC (1830 LT) in the Salt Lake City region during the study period, which marks the transition from daytime-unstable to nighttime-stable PBL conditions, with the opposite transition occurring at about 1330 UTC (0630 LT).¹ The fact that the maximum wind-direction forecast errors from MM5 tend to occur near these times may indicate that the model does not adequately represent the complex characteristics of the transition between PBL regimes. For example, Stewart et al. (2002) show that the observed near-surface wind field is highly variable over the Salt Lake City region during these transition periods.

The results of the analysis of maximum and minimum skill shown in Fig. 4 illustrate that for all times of the day, the typical wind speed forecast errors from the NCEP and MM5 models are slightly larger than the perfect-model threshold value. However, the wind

speed errors from the random no-skill forecasts are not greatly larger than the errors from the perfect forecasts, and they together form a narrow window (width of $\sim 1.5 \text{ m s}^{-1}$) within which the model skill levels fall. The wind speed forecasts from all the models are very similar in skill to that of the diurnal-persistence forecast. The wind-direction MAE curves for the no-skill and perfect-model forecasts define a range of about 70° . Here, the model MAEs are about 20° – 30° better than the no-skill forecast and 40° – 50° worse than the perfect forecast. For all lead times, the MM5 wind-direction MAE is 5° – 25° lower than that of diurnal persistence.

The RMSE for 10-m-AGL wind speed and direction shows similar relative scores of the models. Finally, the bias errors for 10-m-AGL wind direction and speed show that the models all perform similarly, with small biases (not shown).

5. Spectral decomposition of the observed 10-m wind field

Quantitatively defining the power in the diurnal component of the time series of the observed wind at each location is important because one of the major potential benefits of high-resolution mesoscale models is for capturing the diurnal forcing by the local topography and other surface contrasts. Thus, in those areas where there

¹ Sunrise and sunset at Salt Lake City occurred at 0736 and 1748 LT (1436 and 0048 UTC), respectively, on 3 February 2002 and at 0527 and 1923 LT (1227 and 0223 UTC), respectively, on 30 April 2002.

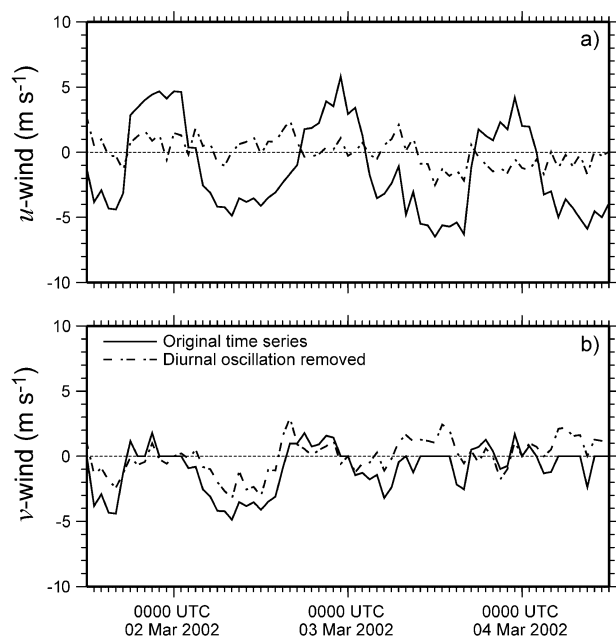


FIG. 5. Observed 10-m-AGL (a) zonal and (b) meridional wind components at Parley's Canyon (UT5) in the eastern Salt Lake Valley for three diurnal cycles. The 3-day average at each point in the time series is computed, and the result is subtracted from the original series to highlight the variability on time scales less than the diurnal.

is a large observed diurnal component, the potential benefit of mesoscale models is great. Features with longer-than-diurnal periods may be viewed as synoptic scale, and therefore are reasonably representable by all the models considered here. Motions with subdiurnal time scales include mesoscale circulations that are not forced by the diurnal heating cycle. These may result from orographic or other landscape forcing, perhaps far upstream, or from nonlinear interactions. Given the sparse nature of the radiosonde network, these mesoscale features are not represented well, or at all, in three dimensions by the observation network, and therefore are not in the model initial conditions. Unless they are locally generated through nondiurnal forcing, they are not deterministically predictable by any model, no matter how good the resolution and physics. Thus, knowing the percent of the total power of the observed wind that is in the diurnally forced component can provide insight into the potential benefit of employing high-resolution mesoscale models. To help visualize the subdiurnal and diurnal components of the spectra, Fig. 5 illustrates three diurnal cycles from location UT5 (see Fig. 3). A superdiurnal change in the wind direction is also apparent.

To perform this analysis for the study period, the time series of observed 10-m zonal and meridional wind components at each of the 28 stations for the 86-day study period were spectrally decomposed using a discrete Fourier transform, and the energy in three frequency bands was computed: the diurnal motions with periods of 22–26 h, the longer-period, superdiurnal mo-

tions, and the motions with subdiurnal periods. Weighted linear temporal interpolation was used to fill the data-void periods. Although the possibility exists for contamination of the time series spectra by interpolation, it was found that changes in the spectra from the use of more sophisticated interpolation techniques (e.g., cubic splines) were negligible. Prior to spectral decomposition, each time series was detrended.

The results of the spectral analysis show that the actual and relative amounts of spectral power in each band vary greatly with location. At the mountaintop station, OGP (Fig. 3), about 67% of the power is in the superdiurnal band for both wind components, only ~2% is in the diurnal band, and approximately 31% is in the subdiurnal band. In contrast, at locations near the mountain slopes (CEN and UT5), the relative and absolute amount of diurnal power in the zonal component is much larger; at CEN the diurnal band contains 28% of the power and at UT5 it contains 33%. However, the less predictable subdiurnal band has more energy here as well. Farther away from the mountain slope in the valley, there is relatively little energy in any of the bands. These results are consistent with Davis et al. (1999), who found that the amplitude of nonrecurring circulations operating on short time scales over the northern Utah area was at least as large as that for the systematic (diurnal) circulations.

To illustrate the spatial variability in the amount of spectral power in the diurnal band, Fig. 6 shows the diurnal spectral power in each wind component for the 28 stations, plotted against the corresponding amplitude of the average diurnal oscillation of that wind component (e.g., $\overline{u_{\max} - u_{\min}}$). This illustrates that there is a large station-to-station variation in diurnal power and that there is the intuitively expected positive correlation between diurnal power and the average magnitude of the diurnal oscillation in the low-level wind. The relationship between the quality of the forecasts and the spatial variability of energy in the three spectral bands will be investigated in the next section.

6. Alternative verification procedures

This section summarizes alternative measures of forecast quality. A quantitative measure is the anomaly correlation between the observed and forecasted time series of the wind. Two qualitative measures of forecast quality that have special relevance to T&D winds are comparisons of the analyzed and forecast winds in terms of their climatology and spatial variance.

a. The anomaly correlation

The anomaly correlation (AC; Wilks 1995) is a measure of the degree of correspondence in phase and amplitude of anomalies in the observed and forecasted time series, in this case the time series of the wind components at the observation locations. In contrast to the

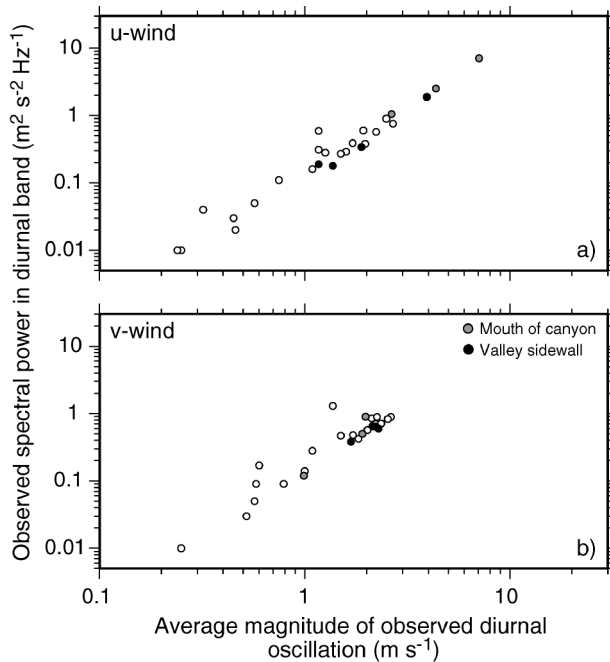


FIG. 6. Spatial variability in the amount of spectral power in the diurnal band. Displayed is the observed diurnal spectral power in each wind component for the 28 stations, plotted against the corresponding magnitude of the average observed diurnal oscillation (e.g., $u_{\max} - u_{\min}$) for the 3 Feb to 30 Apr 2002 period. Those stations located at the mouth of a canyon or very near a valley sidewall are indicated.

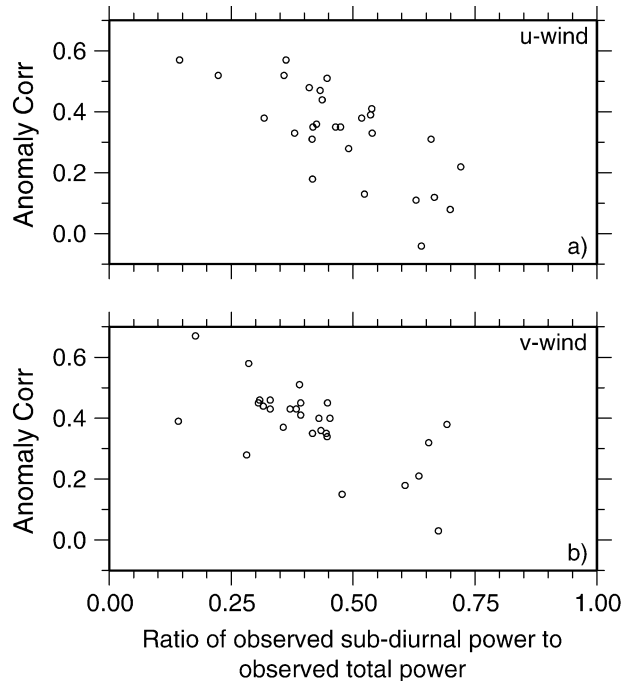


FIG. 8. Same as Fig. 7, except for subdiurnal spectral power.

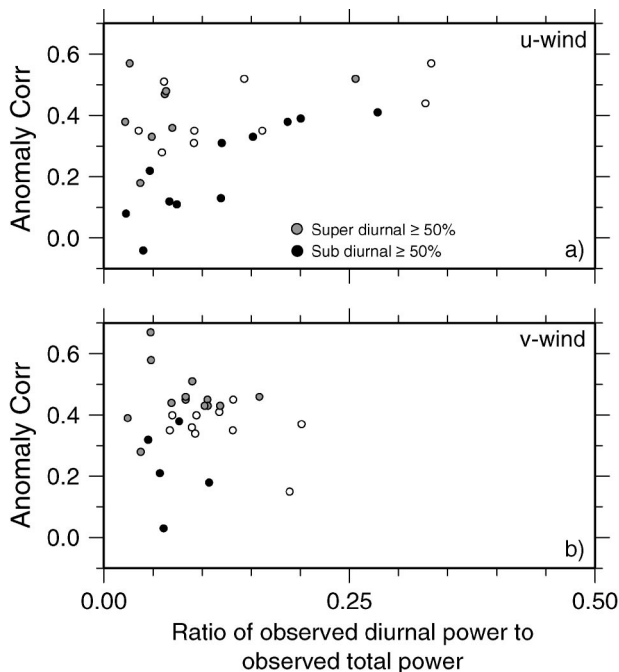


FIG. 7. Anomaly correlation score at each station for the MM5 1.33-km 12-h forecasts plotted as a function of the ratio of observed diurnal spectral power to the observed total power for the 3 Feb to 30 Apr 2002 period. Those stations with at least 50% of the observed power in the subdiurnal or superdiurnal band are indicated.

standard measures of quality like MAE, the AC is designed to reward for good forecasts of the pattern of the observed field, with less sensitivity to the correct magnitudes of the field variable. The AC will be primarily applied to the MM5 output to illustrate how the AC forecast skill is related to the strength of the diurnal power at different locations in the study area. Figure 7 shows the AC for each observation location plotted against the fraction of the total power in the diurnal band. A similar relationship exists when the AC is plotted against the actual diurnal power. For the zonal component of the wind, locations with a greater amount of diurnal power (e.g., ratio > 0.15), have ACs that are relatively large (AC > 0.4), indicating more skillful predictions.² For locations with smaller diurnal power, the AC depends on whether the subdiurnal or superdiurnal part of the spectrum dominates. At locations where the superdiurnal power is large (gray circles), the AC is typically high because synoptic-scale motions are relatively predictable. By contrast, the locations that are dominated by subdiurnal features with lower predictability have typically low AC. Even though the meridional component does not exhibit large diurnal power at any stations, the same general relationships prevail. Figure 8 illustrates that the MM5 AC is inversely related

² The anomaly correlation is generally evaluated relative to the reference value of 0.6, which represents the subjective cutoff for "useful" forecast skill (Wilks 1995). However, this threshold was developed in the context of 500-hPa-height forecasts for global-scale meteorological models and may not be appropriate for near-surface wind field predictions.

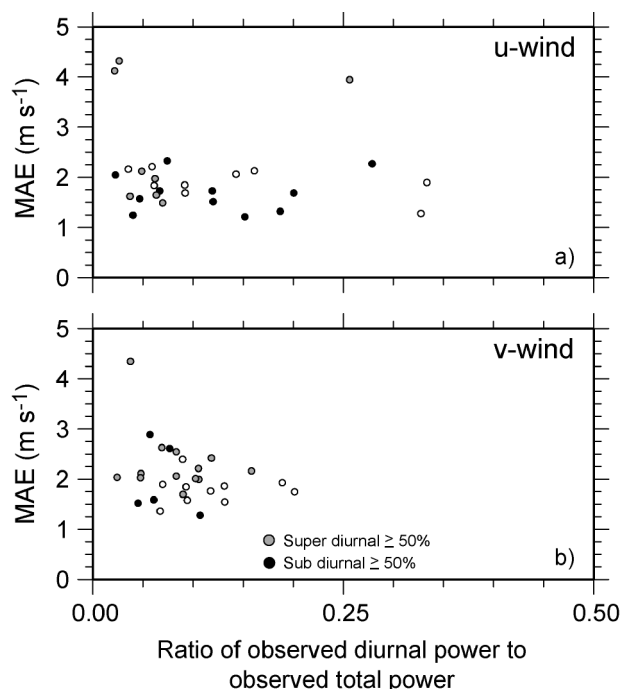


FIG. 9. Same as Fig. 7, except for MAE score.

to the degree to which the spectrum is dominated by subdiurnal power, as expected. In contrast, the MAEs for each station exhibit no relationship to the distribution of power in the three spectral bands, as demonstrated in Fig. 9 for the diurnal power.

For the RUC-2 model, the overall relationship between observed diurnal power and the AC is similar to that of the 1.33-km MM5. Additionally, the average AC score is comparable to that of the MM5. This result raises the question of why the mean AC scores for the two models are so similar. Wilks (1995) notes that the behavior of the AC is qualitatively similar to the RMSE, so the similarity of the AC score for the two models is perhaps to be expected. Another consideration is that the study period contained cloudy episodes, in which synoptic-scale forcing prevailed over the diurnally forced circulations. Evidence of this is that the largest contribution from the diurnal spectral band at any station within the grid-4 region amounted to only 33% of the total power in the zonal wind component. To further investigate this issue, the AC analysis was focused on periods in which diurnal forcing dominated the region. A set of objective criteria was developed to identify days where diurnally driven flows dominated the local circulation patterns. First, days with a large diurnal temperature cycle were identified from the 28-station composite time series (i.e., $T_{\max} - T_{\min} \geq 8.0^{\circ}\text{C}$). Also, periods from this subset were considered only if the overall trend from one day to the next was small, ensuring that no large-scale regime change was occurring. Specifically, $T_{24} - T_{00} \leq 0.25(T_{\max} - T_{\min})$.

Twenty-four days meeting the above criteria were

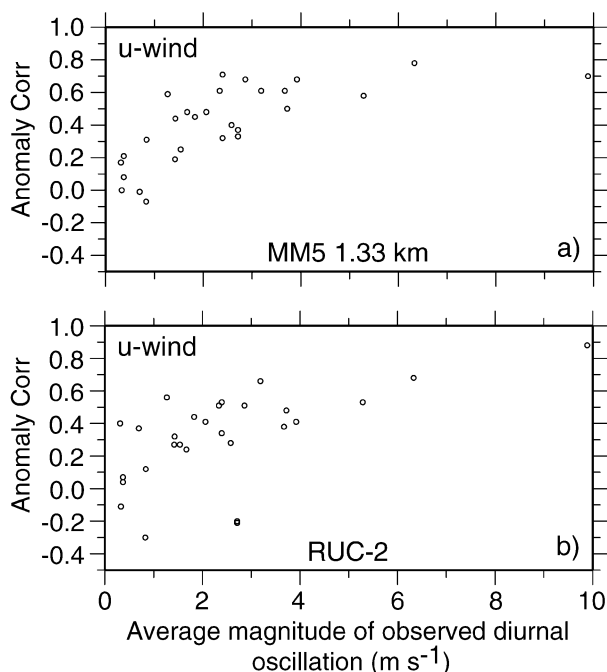


FIG. 10. Anomaly correlation score at each station for days in which diurnal forcing dominated during the 3 Feb to 30 Apr 2002 period. Displayed is the anomaly correlation plotted as a function of the magnitude of the averaged observed diurnal oscillation (e.g., $\frac{u_{\max} - u_{\min}}{u_{\max} + u_{\min}}$) for the (a) MM5 1.33-km and (b) RUC-2 12-h forecasts.

identified during the study period. Because the 24 diurnally dominated days represented distinct, short time series, the wind fields during these periods could not be spectrally decomposed. Thus, an alternative approach was used to estimate the strength of the diurnal forcing of the wind field at each station: the amplitude of the average diurnal oscillation. In the previous section, it was shown that there is a strong correlation between diurnal spectral power and the amplitude of the diurnal oscillation in the low-level wind (Fig. 6), demonstrating the efficacy of estimating the diurnal power in this manner. Figure 10 presents the results of the AC analysis for days in which diurnal forcing dominated. In terms of the zonal wind component, the mean AC for the MM5 is about 0.41 compared to 0.32 for the RUC-2, which represents a nearly 30% improvement to the grid-averaged skill. For both models, there is only a weak relationship between the AC and average diurnal oscillation for the meridional component (not shown). This may be related to the fact that the mountain ranges in the Salt Lake City region have a mainly north-south orientation, which imparts a larger diurnal signal to the zonal component at many stations.

b. Wind climatologies

A measure of forecast quality is how well the observed near-surface, wind field climatology is reproduced by the model. In the context of the T&D of haz-

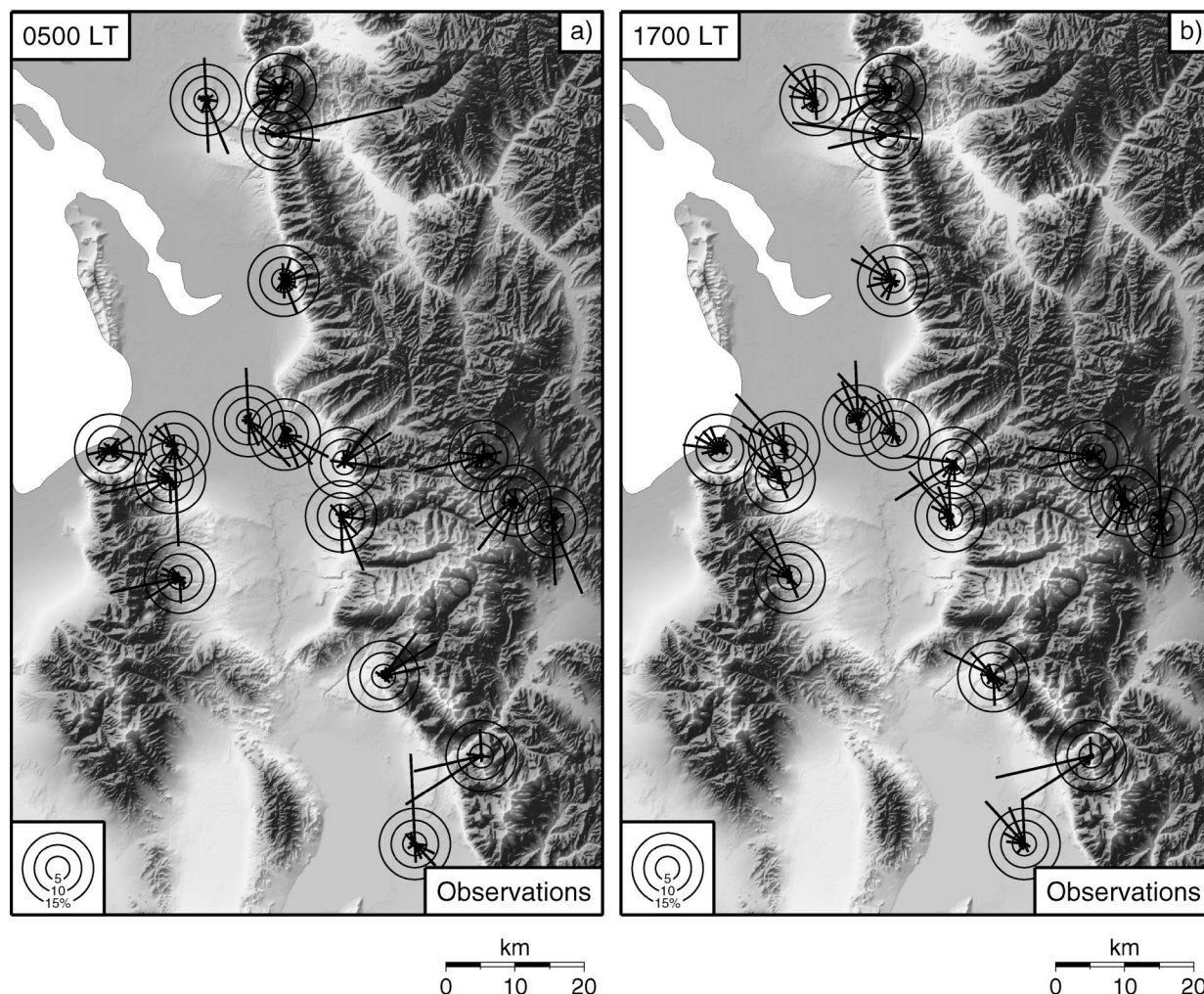


FIG. 11. The 10-m wind-direction climatology at (a) 0500 LT (1200 UTC), and (b) 1700 LT (0000 UTC) for each observation station over the grid-4 region. The climatology is based on stations with 68 or more reports (80% of the total possible number) during the 3 Feb to 30 Apr 2002 period. The percent occurrence of each 20° direction increment is indicated by the circles (see inset). Note that some reports are omitted from the figure to enhance legibility.

ardous material, this metric is important when a model is used for defining the source regions that represent the greatest potential threat to a receptor location. That is, for a particular month and time of day, what is the upwind direction? For that purpose, the relevant forecast quality metric is not related to whether the model can predict source–receptor relationships on a case-by-case basis, but rather whether the *statistics* of the wind directions can be represented.

For assessment of the ability of the models to define the 10-m wind field climatology, observed and model climatologies for the study period were constructed for the 28 observation locations in the grid-4 region. The model climatologies were computed for different forecast lengths. The observed 10-m-AGL wind-direction climatologies are displayed in Fig. 11 for two times of the day: late afternoon and late night. The complex topography of this region produces a variety of thermally

driven wind systems, including valley winds, slope winds, and the lake–land breeze (Stewart et al. 2002). Each of these circulations is apparent in the observed early morning [0500 LT (1200 UTC)] wind-direction climatology for the Salt Lake City region (Fig. 11a). At this time, the flow is dominated by downslope winds along the flanks of the Oquirrh and Wasatch Mountains, and downvalley winds on the Salt Lake valley floor that are possibly weakly reinforced by the offshore flow induced by the Great Salt Lake to the northwest. In addition, several major canyons issue into the Wasatch Front valleys from the east, and two of these—Parley’s Canyon and Weber Canyon—exhibit a pronounced outflow at this time. The strongly channeled easterlies emerging from Weber Canyon may be steered along the axis of the river valley, as they turn anticyclonically into the plain below (Fig. 11a). During the late afternoon at 1700 LT (0000 UTC; Fig. 11b), the winds illustrate

upslope flows, and upvalley flows that are in phase with the Great Salt Lake breeze.

The corresponding climatologies based on the 12-h forecasts from the four models at the late-night time (0500 LT) are presented in Figs. 12a–e.³ For the GFS (Fig. 12a), as expected, most of the stations have nearly the same climatology, as defined by the coarse-resolution analyses and forecasts that do not resolve the Salt Lake Valley, the neighboring mountains, and the Great Salt Lake. A shift in the dominant GFS wind direction for this time of day occurs during the study period, possibly because the slope flows reverse too early in the season, as the time of sunrise changes. The RUC-2 climatology for this time (Fig. 12b) is dominated by southeasterly flow everywhere, an unrealistic situation. For example, the observed downslope flow along the mountains on both sides of the valley is not reproduced by the model. The Eta Model wind distribution (Fig. 12c) shows more variability at some stations than does the RUC-2, and less at others, but there is still no clear downslope pattern. It is worth noting that although the Eta Model is run with 12-km grid spacing, the output data are interpolated to a grid having a horizontal spacing of 40 km. This interpolation to a coarser-resolution grid probably has a deleterious effect on the Eta Model wind-direction climatologies. The MM5 climatology from the 30-km grid-increment version (Fig. 12d) is similar to that of the GFS. The 1.33-km MM5 climatology (Fig. 12e) appears to most closely resemble the observations. The stations on the east side of the valley, near the mountains, have a dominant downslope easterly flow. On the west side of the valley, the forecast direction is much more variable than observed, but it is generally from the higher elevations to the west. Note that the 1.33-km MM5 forecast is the only one that correctly represented the southwesterlies at the mountaintop locations in the southeast corner and north-central part of the grid. The model climatologies for the other forecast lengths and times of day show similar results.

Whereas the wind-rose climatologies described above show the frequency distribution of observed and forecast wind directions at each location, resultant wind vectors display the 86-day mean speed and direction and therefore offer a somewhat different and complementary view of the time-averaged model performance. The resultant 10-m wind vectors for the observations and for the 12-h forecasts from the four models are shown in Fig. 13 for 0500 LT only, but the results are qualitatively similar for other times. At the valley stations near the mountain slopes, and at the mountaintop stations, the 1.33-km MM5 average vector compares

more favorably to the observed vector than do the vectors from the other models. However, at other locations, the 1.33-km MM5 provides little or no improvement.

c. Spatial variance of winds

Another technique for objectively evaluating the models is to examine the degree to which their forecasts correctly replicate the observed horizontal spatial variance of the low-level wind field over the study area. This property of the model solution is especially relevant for T&D applications because the near-surface T&D process is highly related to the variance in the wind field. Horizontal variance is defined here as the departure of the observed (or forecast) value of a variable from the average over the grid-4 area. A model solution that represents a rich array of circulation features (deterministically or not) will exhibit a high degree of horizontal variability, whereas ones with a smooth representation of the same atmosphere will not. Again, even though a model variance that corresponds well with the observed variance that does not mean that there is feature-for-feature correspondence. But having a realistic amount of statistical variability should be a desirable attribute of a model solution, for example, for plume dispersion.

Comparison of the observed instantaneous horizontal variance of 10-m-AGL wind direction over the grid-4 region with that from the model forecast is presented in Fig. 14 and illustrates this aspect of model performance.⁴ Although it is not possible for the NCEP models to exhibit much spatial variability over this limited geographic area because of the coarse resolution of their computational and output grids, it is nevertheless revealing to compare them to the MM5 in this respect to quantify the added variance that should be obtainable through a higher-resolution model. As anticipated, the 1.33-km MM5 variance in the 10-m wind field over the region is clearly superior to those produced by the NCEP models.

7. Discussion and summary

This paper illustrates shortcomings associated with the use of conventional verification statistics for the assessment of quality in near-surface wind predictions, it shows some inherent limits to current mesoscale predictability, and it suggests some alternative ways of viewing the forecast quality for the low-level winds.

Mesoscale circulations in the vicinity of complex orography, surface-type contrasts, and coastlines, at least partly result from the signatures imparted to the

³ Slight differences appear in the model wind field climatologies at some closely neighboring stations, which may seem counterintuitive for the NCEP models that had relatively coarse-resolution output grids. Such differences arise as a consequence of interpolation from the model grid to the observation sites and the fact that there are occasional station-dependent gaps in the observational record.

⁴ As noted in section 3, the GFS and Eta Model output grids were available from forecasts initialized at 0000 and 1200 UTC. Therefore, the variance statistics for these two models have a ~75% smaller sample size than the corresponding statistics derived from the 8-per-day RUC-2 and MM5 forecasts.

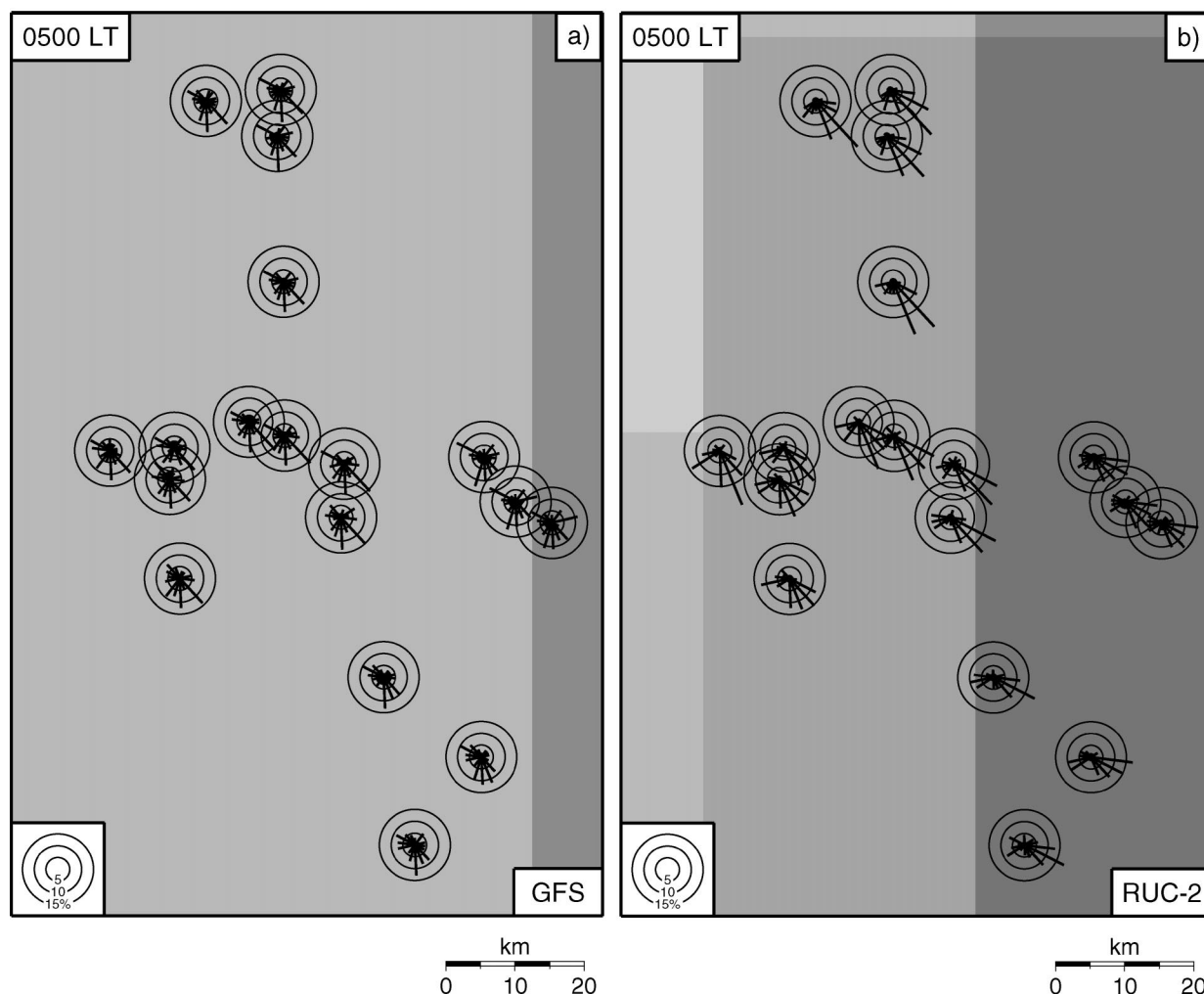


FIG. 12. As in Fig. 11, except for model-forecast climatologies at 0500 LT (1200 UTC) for the (a) GFS, (b) RUC-2, (c) Eta Model, (d) MM5 30-km model, and (e) MM5 1.33-km model. The climatologies are based on the 12-h forecasts. The model terrain is shaded as in Fig. 2.

atmosphere by differential diurnal thermal forcing. Thus, it is arguable that mesoscale model forecast quality in a particular area should depend at least partially on the strength of such diurnal forcing and the degree to which the model can represent the forcing. Nevertheless, conventional verification metrics (e.g., MAE) for the low-level-wind predictions from the high-resolution MM5 showed little relationship to surface features and no correlation with the fraction of the spectral power in the diurnally forced motions (Fig. 9). As can be anticipated from this evidence, there was little sensitivity of the MAE to the degree to which the different models resolved the local forcing (Fig. 4).

Given the situation that the conventional statistics are apparently a fairly blunt tool for assessing forecast quality, an ancillary (and unanswered) question is why there seems to be a greater impact of resolution on the wind-direction MAE than on the wind speed MAE (Fig. 4). Clearly, the zonal and meridional wind components can be greatly affected by small-scale, local, thermally driven

circulations, and thus the veracity with which we represent these quantities depends on the resolution. And, one would intuitively expect that resolving these small-scale circulations would have a similar impact on both the wind direction and speed. In terms of either wind speed or direction, the conventional measures of forecast quality do not seem to discriminate among the models as much as would be expected based on the importance of the prevailing high-resolution local forcing.

The lack of discrimination among the models by the MAE is partly explained by the modest difference between the upper and lower bounds of model skill. First, diurnal persistence, one choice for no-skill forecast, is sometimes a good predictor of flow in complex terrain, especially because it fully accounts for local effects in a way that models cannot. Second, the perfect-model forecasts exhibited fairly high errors because of the large inhomogeneities introduced to the wind field by the terrain that is not resolved by the MM5. Thus, it is un-

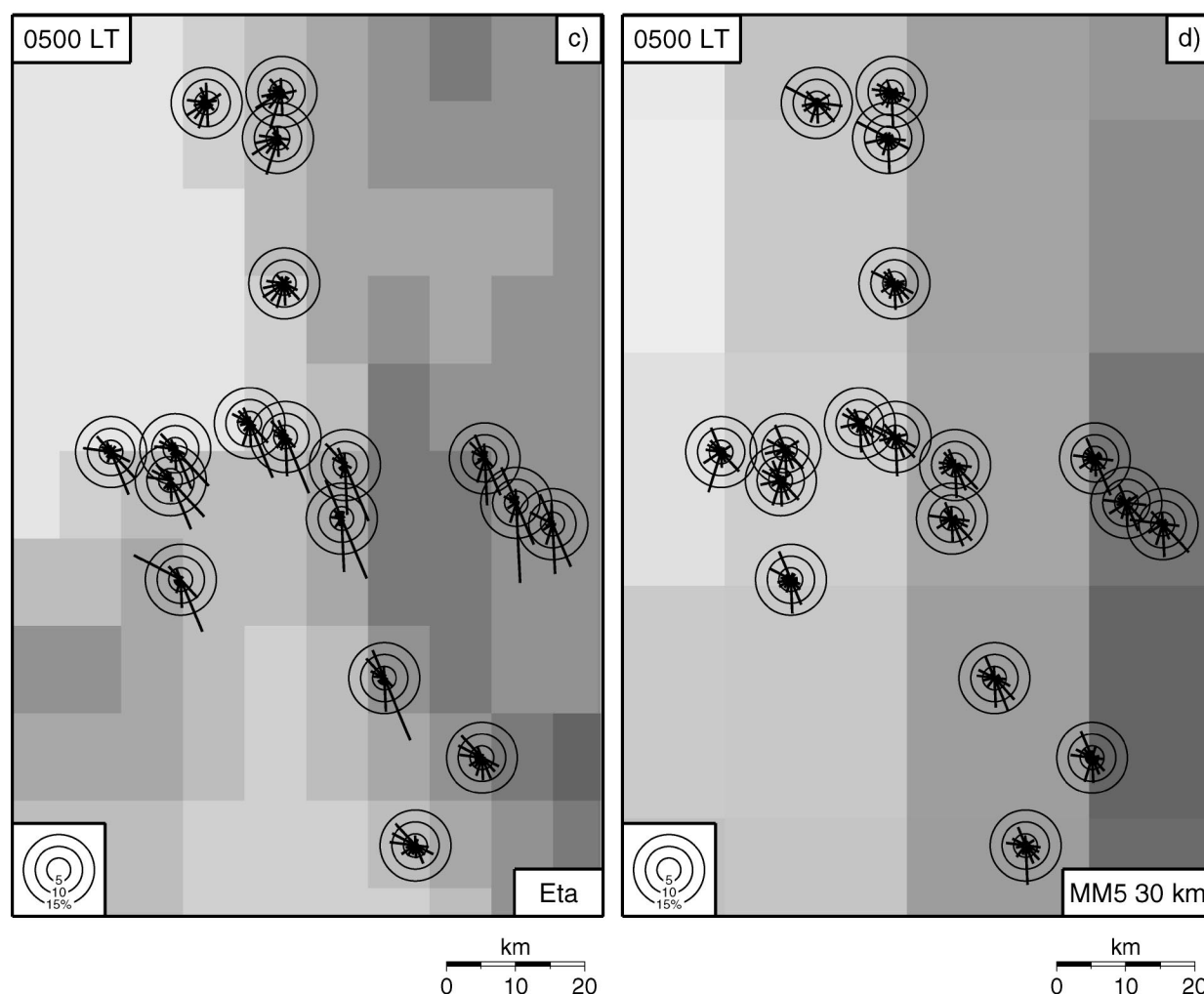


FIG. 12. (Continued)

derstandable why the upper and lower forecast bounds in Fig. 4 define such a narrow error envelope within which the model skill levels fall.

The decomposition of the observation time series into the three spectral components supports the idea that there exists significant place-to-place variation in the importance of the mesoscale diurnal forcing that can only be represented by high resolution (Figs. 6–7). There are some locations where increasing the resolution will have little if any positive impact because the diurnal power in the spectrum is very small compared with the power in the superdiurnal band (synoptic scale) and in the subdiurnal band (poorly observed and therefore largely unpredictable small scales). If representing local thermal forcing is the motivation for considering the use of a high-resolution mesoscale model, spectrally decomposing the local observations in this way could provide useful information in determining the resolution sensitivity of the forecast quality. Also, for existing mesoscale models, it would be informative to use time

series of observations to map the percent of energy in the diurnal band to assess 1) the expected variations in model forecast quality associated with location and season and 2) the potential benefit of improving the land surface physics, which drives the diurnal forcing.

The importance of the presumably small-scale, subdiurnal motions at some locations (Fig. 8), and the argument that these features are almost completely unrepresented by the model because they are not diurnally forced or defined in the initial conditions, speak to the importance of improving mesoscale measurement systems in parallel with improving model resolution. The existence of this relatively unpredictable, and sometimes large, fraction of the wind's energy at every point must also contribute to the degree of insensitivity of the forecast quality to model resolution.

In contrast to the conventional measures of forecast quality, the AC of the wind components showed a strong positive relationship to the fraction of the spectral power that was in the diurnal component, and a strong negative

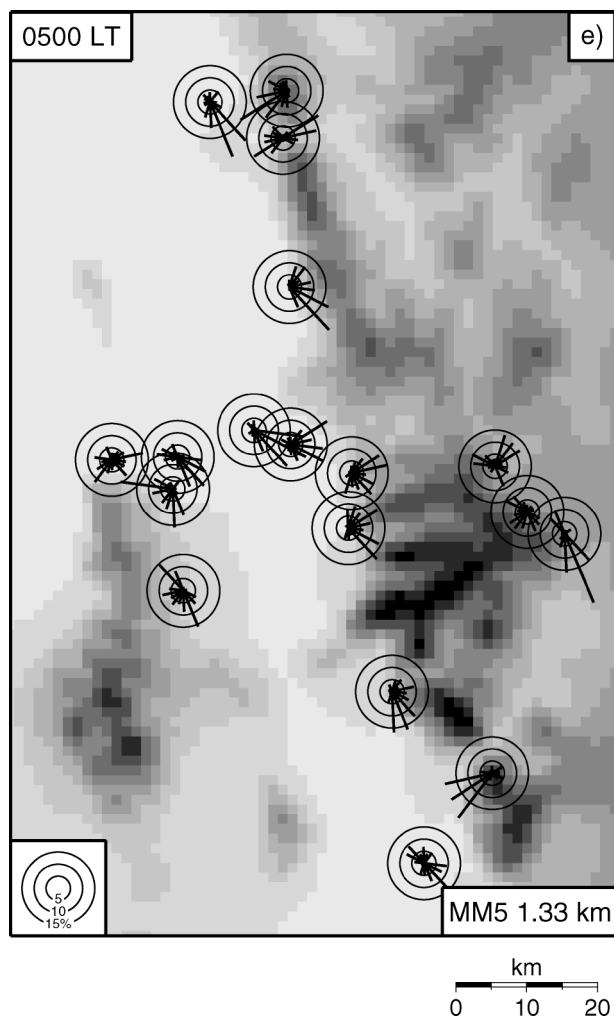


FIG. 12. (Continued)

correlation with the fraction of the power in the subdiurnal frequency band. Thus, in contrast to conventional measures, the AC illustrates that the highest-resolution model has more skill where the local diurnal forcing is greatest. The AC also showed the advantage of the MM5 over the coarser-resolution RUC-2 during periods in which diurnal forcing dominates (Fig. 10), with a nearly 30% improvement to the grid-averaged skill for the zonal wind component.

We now return to the question posed in the introduction, namely, why previous studies and the present study have had difficulty discerning deterministic skill in forecasts with fine (<10 km) grid spacing compared with forecasts using coarser grid spacing. The expectation is that in regions where terrain variation on the mesogamma and mesobeta scales is significant, diurnal heating should drive local circulations whose predictability scales with an increasingly accurate representation of terrain, and diurnal heating and cooling. However, diurnally driven and terrain-driven flows make up less of the overall spectrum of motion at many stations

than we anticipated. Observed time series were influenced by subdiurnal motions (of undetermined origin) with surprisingly large power. Owing to the low predictability of such motions (few hours or less; Davis et al. 1999), errors at these scales are effectively saturated in all models examined.

A second point pertains to the multiscale nature of terrain-induced flow. While models such as the RUC-2 or 30-km MM5 have a seemingly vastly inferior representation of terrain on scales of tens of kilometers or less (i.e., “local” terrain), the total diurnal response is determined by a composite of forcing on many scales. The coarser-grid models appear to adequately capture the larger-scale flow response to terrain. As one moves to finer grid spacing, the smaller-scale terrain features that emerge have associated motions with correspondingly smaller scales. Thus, only very near the finer terrain features are there diurnal motions forced by terrain unresolvable in the coarser models. Because characteristic terrain slopes on these scales are large, the amplitude of the diurnal fluctuations is also large. This means that subtle timing or amplitude differences between forecast and observed time evolution contribute to large errors in traditional verification statistics. Even the AC metric suffers from this problem to some extent. It is clear that object-oriented verification statistics are needed so that realistically predicted structures that suffer from relatively small spatial and temporal errors can be given adequate credit.

Thus, the resolution of our posed paradox is recognition of a combined set of factors: the importance of high-frequency (subdiurnal) motions, localization of diurnal motions near finer-scale terrain, and inadequate verification metrics. We are currently developing and testing object-based verification metrics and will report on results of their application to resolution-dependence studies of numerical forecasts of diurnally forced and terrain-induced flows.

Acknowledgments. This research was funded by the U.S. Army Test and Evaluation Command through an Interagency Agreement with the National Science Foundation. The authors gratefully acknowledge Barb Brown, Bob Sharman, and Rod Frehlich (NCAR) for enriching our understanding of verification and spectral-analysis techniques. Barb Brown is also thanked for internally reviewing the manuscript. Janice Coen (NCAR) graciously provided the Clark–Hall model output, and Tressa Fowler (NCAR) provided guidance in constructing the random no-skill forecasts.

APPENDIX

Calculation of the Maximum and Minimum Forecast Skill for Winds

It is generally well recognized that the errors produced by the model and the errors in the observations

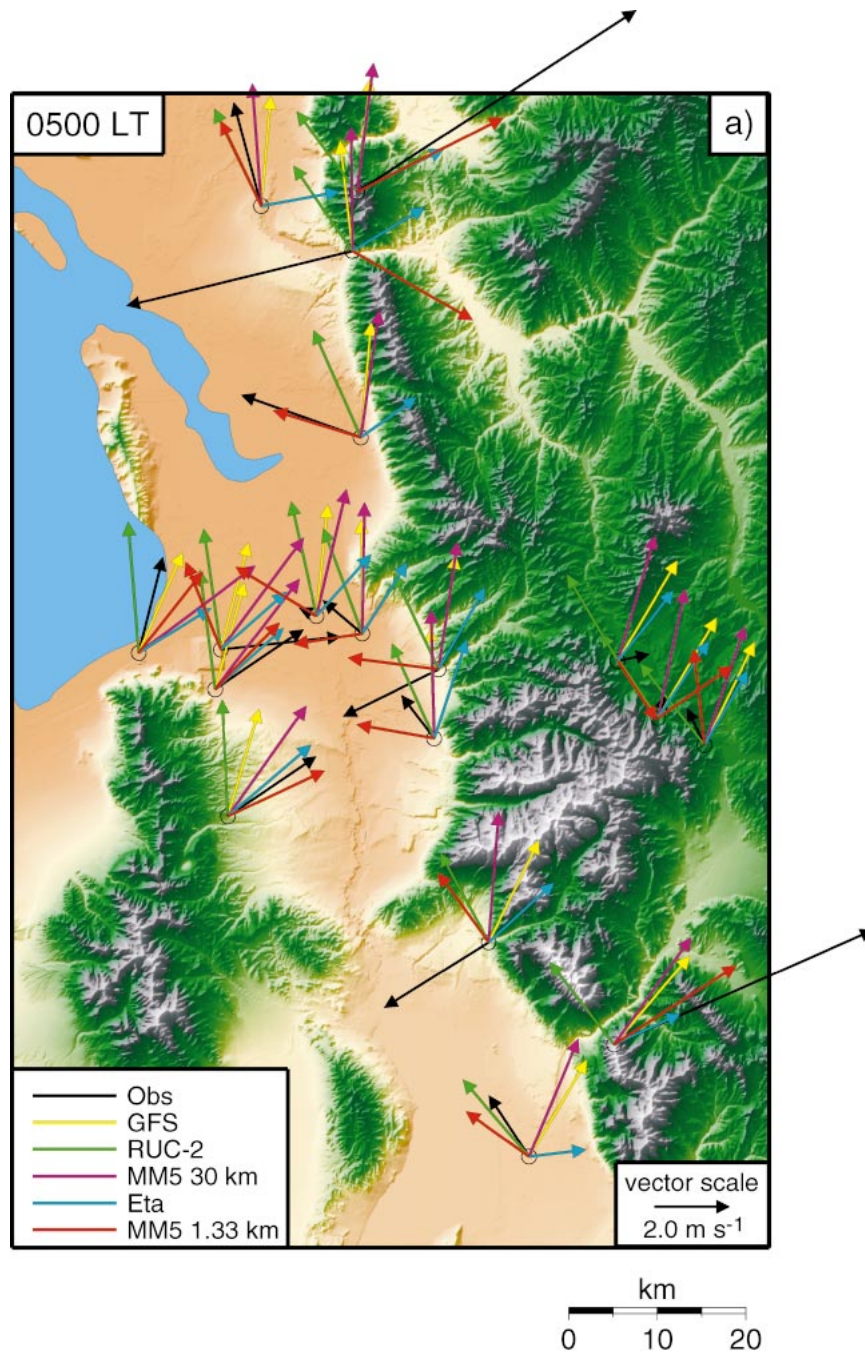


FIG. 13. The 10-m-AGL resultant wind vectors (see vector scale) at 0500 LT (1200 UTC) for each observation location within the grid-4 region. The resultant winds are based on stations with 68 or more observations and corresponding model 12-h forecasts (80% of the total possible number) during the 3 Feb to 30 Apr 2002 period.

both contribute to the total error reflected in the verification scores. The observations used for verification contain error associated with the accuracy of the instruments and calibration error. In addition, there is another somewhat less well documented source of error that impacts the conventional verification scores, which will always exist regardless of how much the model and

instrument errors are reduced. This is the representativeness error.

Representativeness errors arise from the fact that there is a fundamental mismatch between the spatial and temporal scales represented by the models and the observations. Conventional ground-based instruments make time-averaged measurements at a point, whereas the

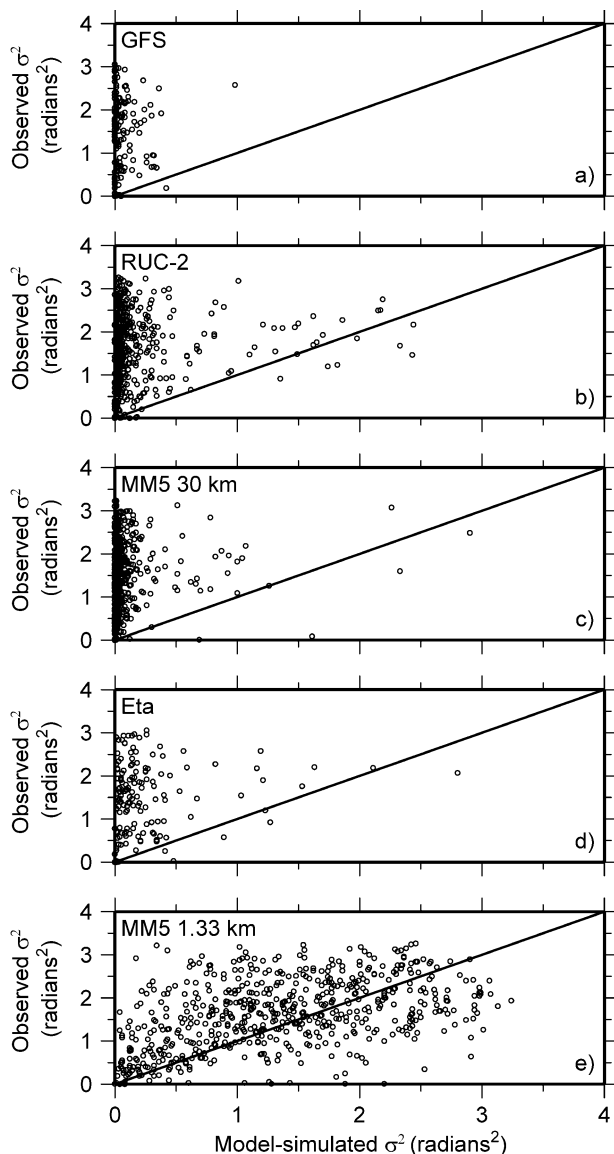


FIG. 14. Comparison between the observed spatial variance of 10-m-AGL wind direction (σ^2) over the grid-4 region and the corresponding variances from the (a) GFS, (b) RUC-2, (c) MM5 30-km model, (d) Eta Model, and (e) MM5 1.33-km model 12-h forecasts during the 3 Feb to 30 Apr 2002 study period. Each point corresponds to a single observation time.

model-predicted quantities represent spatial averages over each model grid-box volume. Representativeness error can be appreciated through the following idealized example: Suppose there exists a perfectly known near-surface wind field over a 1 km^2 area. The field is sampled at the center of this area to create a “perfect” point observation of the wind. Next, the 1 km^2 spatial average is computed, which represents the corresponding grid-box-mean value of the wind predicted by a perfect model. Despite the fact that model and observation both exactly characterize the wind field in their own way, the difference between the two will obviously not be zero

because the latter is a spatial average of the wind, while the former represents a discrete point value. This difference is termed the representativeness error, and its magnitude is dependent on a number of factors including the prevailing weather regime, the amplitude of mesoscale structures, and the geographic extent of the sampling area (or size of the model grid box).

Ideally, one would quantify the representativeness error using high quality observations from a very dense surface mesonet, where the instruments are spaced tens of meters apart over an area equal to that spanned by a single model grid box. A tractable alternative approach is to estimate the magnitude of the representativeness error using an extremely high-resolution model. The model described by Clark (1977), Clark and Farley (1984), and Clark and Hall (1991) has been used for a number of studies of finescale atmospheric phenomena such as clear-air turbulence (Clark et al. 2000), cloud microphysical processes (Farley et al. 1992; Bruintjes et al. 1994), and the interaction of forest-fire dynamics with ambient small-scale airflows (Clark et al. 1996). This model is commonly referred to as the Clark–Hall model, so that term will be used here. Clark–Hall model output was obtained for a real-data simulation over the Pinewood Springs, Colorado, area for the afternoon hours of 17 July 2002. The model employed multiply nested interactive grids, and the highest-resolution grid had an increment of approximately 50 m and encompassed a nearly 36 km^2 area. The geographic region over which the simulations were performed is characteristic of the Salt Lake City, Utah, area, in that it has a semiarid climate and contains much complex terrain and varied vegetation and substrates. The meteorological conditions during the simulation consisted of weak synoptic-scale forcing and little or no cloud cover, thus allowing thermally driven flows to dominate the local circulation patterns.

The procedure for estimating the representativeness error is as follows. First, the spatially averaged wind speed and direction are computed from the Clark–Hall model output within a stencil having dimensions of an MM5 1.33-km grid box. There are about 676 Clark model grid points for each MM5 grid box. Next, the point values of the speed and direction are determined at the stencil center. The stencil, initially located at the southwest corner of the Clark–Hall model domain, is then laterally repositioned in the domain by a distance equal to its width, and the spatial average and center-point values for speed and direction are again calculated. This process is repeated until the entire Clark–Hall model domain has been sampled in a nonoverlapping fashion. The mean difference between the grid-box-average and point values of wind speed and direction from each unique sample (36 individual paired values) is computed to produce an estimate of the representativeness error. This estimate is conservative because the Clark–Hall model with a 50-m grid increment underestimates the true amount of spatial variability that would exist in the

near-surface wind field under similar environmental conditions.

The opposite error bound is also estimated: This is the threshold beyond which the forecasts have 1) no skill, and 2) no skill beyond what could be achieved through the use of simple procedures such as persistence. In the latter case, a “diurnal persistence” forecast was computed by using the previous day’s 3-hourly observations as the forecast values for the current day. Note the fact that forecasts based on diurnal persistence will be difficult to improve upon by a model for locations and periods in which the weather variability is dominated by local diurnal forcing. For a true no-skill value, beyond seasonal climatology, the bootstrap technique of Efron and Tibshirani (1993) was used. Here, the available data throughout the entire study period are repeatedly and randomly resampled (with replacement) to yield multiple synthetic samples of the same size as the original set of observations. These samples serve as forecasts. In this study, 5000 random forecasts of size $n_{\text{obs}} = 16\,550$ were created from the entire 86-day collection of wind speed and direction observations. The forecasts are thereby constrained by the climatological distribution of the observations over the study period. Note that randomly sampling the entire body of observations has the effect of removing the diurnal signal from the dataset. Each of the 5000 random forecasts is compared with the observations at each 3-hourly verification time, and the average verification score at each time is then used to define the maximum error (or no skill) value.^{A1} This also represents a conservative estimate, since the observations themselves sometimes significantly undersample the true amount of wind field variability because of imperfect station siting and instrument exposure characteristics.

REFERENCES

- Bruintjes, R. T., T. L. Clark, and W. D. Hall, 1994: Interactions between topographic airflow and cloud/precipitation development during the passage of a winter storm in Arizona. *J. Atmos. Sci.*, **51**, 48–67.
- Case, J. L., J. Manobianco, A. V. Dianic, M. M. Wheeler, D. E. Harms, and C. R. Parks, 2002: Verification of high-resolution RAMS forecasts over east-central Florida during the 1999 and 2000 summer months. *Wea. Forecasting*, **17**, 1133–1151.
- Clark, T. L., 1977: A small scale dynamic model using a terrain following coordinate transformation. *J. Comput. Phys.*, **24**, 186–215.
- , and R. D. Farley, 1984: Severe downslope windstorm calculations in two and three dimensions using anelastic interactive grid nesting: A possible mechanism for gustiness. *J. Atmos. Sci.*, **41**, 329–350.
- , and W. D. Hall, 1991: Multi-domain simulations of the time dependent Navier Stokes equations: Benchmark error analysis of some nesting procedures. *J. Comput. Phys.*, **92**, 456–481.
- , M. A. Jenkins, J. Coen, and D. Packham, 1996: A coupled atmosphere–fire model: Convective feedback on fire line dynamics. *J. Appl. Meteor.*, **35**, 875–901.
- , W. D. Hall, R. M. Kerr, D. Middleton, L. Radke, F. M. Ralph, P. J. Neiman, and D. Levinson, 2000: Origins of aircraft-damaging clear-air turbulence during the 9 December 1992 Colorado downslope windstorm: Numerical simulations and comparison with observations. *J. Atmos. Sci.*, **57**, 1105–1131.
- Davis, C., T. Warner, E. Astling, and J. Bowers, 1999: Development and application of an operational, relocatable, mesogamma-scale weather analysis and forecasting system. *Tellus*, **51A**, 710–727.
- Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107.
- , 1993: A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and the simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121**, 1493–1513.
- , 1996: A multi-layer soil temperature model for MM5. Preprints, *Sixth PSU/NCAR Mesoscale Model Users’ Workshop*, Boulder, CO, NCAR, 49–50. [Available from D. L. Rife, NCAR, P.O. Box 3000, Boulder, CO 80307.]
- , D. Gill, K. Manning, A. Bourgeois, W. Wang, and C. Bruyere, cited 2002: PSU/NCAR Mesoscale Modeling System tutorial class notes and users’ guide: MM5 Modeling System Version 3. NCAR Tech. Memo. [Available online at <http://www.mmm.ucar.edu/mm5/doc.html>.]
- Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.
- Farley, R. D., S. Wang, and H. D. Orville, 1992: A comparison of 3D model results with observations for an isolated CCOPE thunderstorm. *Meteor. Atmos. Phys.*, **49**, 187–207.
- Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.*, **121**, 1493–1513.
- , J. Dudhia, and D. R. Stauffer, 1994: A description of the 5th-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN 398+STR, 138 pp.
- Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic Winter Games. *Wea. Forecasting*, **19**, 200–218.
- 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.
- Horel, J. D., and Coauthors, 2002: MesoWest: Cooperative mesonets in the western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–226.
- Hsie, E.-Y., R. A. Anthes, and D. Keyser, 1984: Numerical simulation of frontogenesis in a moist atmosphere. *J. Atmos. Sci.*, **41**, 2581–2594.
- Loveland, T. R., J. W. Merchant, J. F. Brown, D. O. Ohlen, B. C. Reed, P. Olson, and J. Hutchinson, 1995: Seasonal land-cover regions of the United States. *Ann. Assoc. Amer. Geogr.*, **85**, 339–355.
- Mass, C. F., D. Owens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.
- Seaman, N. L., D. R. Stauffer, and A. L. Lario-Gibbs, 1995: A multi-scale 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.
- Stauffer, D. R., and N. L. Seaman, 1990: Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic-scale data. *Mon. Wea. Rev.*, **118**, 1250–1277.
- , —, and F. S. Binkowski, 1991: Use of four-dimensional data assimilation in a limited-area mesoscale model. Part II: Effects of data assimilation within the planetary boundary layer. *Mon. Wea. Rev.*, **119**, 734–754.
- Stewart, J. Q., C. D. Whiteman, W. J. Steenburgh, and X. Bian, 2002: A climatological study of thermally driven wind systems of the U.S. Intermountain West. *Bull. Amer. Meteor. Soc.*, **83**, 699–708.
- Stull, R. B., 1988: *An Introduction to Boundary Layer Meteorology*. Kluwer Academic, 666 pp.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

^{A1} It is recognized that the random forecasts are not completely devoid of skill, since they are constructed with respect to the climatological distribution of the observations.