DOES INCREASING HORIZONTAL RESOLUTION PRODUCE MORE SKILLFUL FORECASTS?
The Results of Two Years of Real-Time Numerical Weather Prediction over the Pacific Northwest

by Clifford F. Mass, David Ovens, Ken Westrick, and Brian A. Colle

Decreasing grid spacing in mesoscale models to less than 10–15 km generally improves the realism of the results but does not necessarily significantly improve the objectively scored accuracy of the forecasts.

A major question confronting the meteorological community regards the benefits of increasing horizontal resolution in short-range (1–3 day) numerical weather prediction models. As computer power has increased, operational model resolution has followed—from a horizontal grid spacing of 381 km in the National Meteorological Center barotropic model in the late 1950s to 12 km in the current version of the National Centers for Environmental Prediction (NCEP) Eta Model. With the recent acquisition of a massively parallel computer system, NCEP plans to increase the horizontal resolution of the Eta Model into the single digits over the next few years, while its proposed successor, the Weather Research and Forecasting model, is being designed for resolutions of 1–10 km. In some locations, the U.S. Navy's Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS) mesoscale model (Hodur 1997) is being applied with horizontal grid spacings below 10 km. In the private sector, several companies are experimenting with high-resolution forecast systems; for example, the MyCast prediction service provided by Digital Cyclone, Inc., makes use of the Penn State/NCAR Mesoscale Model version 5 (MM5) applied at 6-km grid spacing to provide forecasts for major urban areas of the continental United States. Finally, as discussed in Mass and Kuo (1998), dozens of groups are experimenting with high-resolution numerical weather prediction, with a number of efforts applying grid spacing of less than 15 km.
Although there is a clear trend toward higher-resolution numerical weather prediction, the value of very high-resolution modeling (here defined as 10-km grid spacing or better) compared to other forecasting approaches is still an open question. In fact, a spirited debate has developed over the use of increasing computational resources for either continuing the trend to higher resolution or in running a collection of ensemble forecasts at lesser resolution to produce probabilistic predictions. For example, considering both predictability theory and the limitations of current observing systems, Brooks et al. (1992) and Brooks and Doswell (1993) suggest that increasing horizontal (and vertical) resolution will produce little if any improvement in convective forecasts. Instead they proposed a Monte Carlo ensemble approach in which a limited number of mesoscale model simulations drive a collection of cloud-scale simulations. In contrast, some investigators (e.g., Koch 1985; Droegemeier 1990, 1997) have suggested that deterministic short-term forecasts may be quite useful for a major subset of convective events. This paper attempts to contribute to the discussion regarding the benefits of resolution by first reviewing the findings of previous studies and then presenting the results of a multiyear verification of mesoscale forecasts in the Pacific Northwest. After examining a case study in which resolution was varied, an analysis of the resolution issue is found in the closing section.

**PREVIOUS STUDIES.** During the past two decades, a number of studies have examined the effects of horizontal resolution on forecast accuracy, with some recent papers investigating the impact of grid spacing of less than 10 km. In most cases, increasing resolution produces better-defined and more realistic structures (evaluated subjectively). Few studies have demonstrated that forecast accuracy, measured objectively over an extended period of time, increases as grid spacing decreases below approximately 10–15 km. The following section reviews some of this literature on horizontal resolution and its effect on prediction skill.

**Convective systems.** Nonhydrostatic models employing grid spacings of 4 km or less have shown substantial success in duplicating the observed structural evolution of a range of mesoscale convective systems (e.g., Weisman et al. 1988; Skamarock et al. 1994; Droegemeier et al. 1994). For example, Weisman et al. (1997), using a nonhydrostatic Advanced Regional Prediction System (ARPS) model have shown that decreasing horizontal grid spacing to less than 5 km can produce realistic convective evolutions (e.g., Adlerman and Droegemeier 2001, manuscript submitted to Mon. Wea. Rev.). For example, simulated storms did not produce observed cyclic mesocyclogenesis until grid spacing dropped below 1.5 km. Bernadet et al. (2000), simulating four high plains convective events from approximately 80- to 2-km grid spacing using the Colorado State Regional Atmospheric Modeling System (RAMS) model, found that 2 km was necessary to capture convection explicitly. Nielson-Gammon and Strack (2000) examined the effects of horizontal resolution (36-4-km grid spacing) on maximum precipitation during three extreme rainfall events over Texas. They found that without convective parameterization, a grid spacing of 6 km or smaller was necessary to consistently achieve observed rainfall rates. Gallus (1999) examined the variation in precipitation forecast skill of the Eta Model for three heavy summertime precipitation events as the grid spacing was decreased from 79 to 12 km. He found that when convective parameterization schemes were active and produced the bulk of the precipitation, increasing resolution had little benefit. In a simulation of Hurricane Danny, Kuo et al. (2001) found substantial improvement in the radius of maximum wind and the eyewall/rainband structures as horizontal grid spacing was reduced from 81 to 1 km.

High resolution appears to be most useful for strongly forced convection (e.g., associated with fronts, drylines, or topography); without such strong forcing, high-resolution observations (which often are not available) are required to produce the detailed initialization needed to profitably apply increased model resolution. As noted by Brooks et al. (1992), although some studies have found improved structural definition of mesoscale convection as resolution increases, there is little evidence that smaller grid spacing improves convective forecasts at specific locations since significant position or timing errors often accompany even short-term predictions of convection.

**Circulations forced by topography and surface contrasts.** Both subjective and objective evaluations have found clear benefits in increasing resolution in regions
where orographic flows or diurnal circulations are important. Applying the Colorado State RAMS model at 40, 10, and 2.5 km for 12-h predictions in the Susquehanna River Valley of Pennsylvania, McQueen et al. (1995) found that the 2.5-km grid in combination with high vertical resolution produced far more realistic structures than predicted in the 10-km domain. Doyle (1997), running the Navy COAMPS model with horizontal grid spacings ranging from 45 to 2 km over the central California coast, demonstrated that the realism of the simulated orographically enhanced coastal jet and land-falling front improved with resolution. Martin (1998) evaluated the precipitation skill of the NCEP Aviation model (T126) and the NCEP Eta Model at 32-, 29-, and 10-km grid spacings for the heavy rainfall event of 23 February 1998 over the mountainous coastal zone of southern California. He found a progressive improvement in the precipitation intensity and spatial distribution as horizontal resolution increased. Colle and Mass (1998) showed that the realism of simulated downslope winds on the western sides of the Cascades improved as horizontal grid spacing was decreased from 27 to 1 km. Simulating gap flow through the Strait of Juan de Fuca using the MM5, Colle and Mass (2000a) found that sharpness of the Strait-exit transition became more realistic as the horizontal grid spacing was decreased from 12 to 1.33 km and the number of vertical levels was increased from 28 to 48. In a modeling study of the largest Pacific Northwest flooding event in recent years (5–9 February 1996), Colle and Mass (2000b) found a significant improvement in precipitation forecast skill as the grid spacing was decreased from 36 to 4 km. For most locations, there was little overall improvement going from 4 to 1.33 km, except along the upper windward slopes and in the immediate lee of major barriers, where precipitation amounts were enhanced with higher resolution. Simulating lake-effect snowbands using the The Pennsylvania State University–National Center for Atmospheric Research (Penn State–NCAR) Mesoscale Model version 5 (MM5), Ballentine et al. (1998) found that a horizontal resolution of approximately 20 km was needed to reproduce the general locations of the snowbands, and that decreasing grid spacing to 5 km resulted in substantial improvement (increase) in precipitation amounts. Rao et al. (1999) used two-way interactive domains with 1.6-, 0.4-, and 0.1-km horizontal grid spacing to show that less than 1-km grid spacing is required to realistically simulate the diurnal circulations of the Cape Canaveral region of Florida using the ARPS mesoscale model. Finally, Davis et al. (1999) showed that decreasing the grid spacing of the MM5 from 10 to 1.1 km improved the realism of the diurnal circulations produced by topography and varying land surface conditions over west-central Utah.

**Midlatitude cyclones and frontal zones.** Weygandt and Seaman (1994) examined the skill of the Penn State–NCAR mesoscale model for eight continental cyclogenesis events for horizontal grid spacings of 160, 80, and 27 km, making use of a feature-based skill score. Increasing resolution from 160 to 80 km improved forecast skill of synoptic-scale cyclone intensities and positions, as well as fixed and propagating mesoscale features. On the other hand, reducing horizontal grid spacing from 80 to 27 km only improved the verifications for geographically related phenomena such as lee troughs or cold air damming. Steenburgh and Mass (1996) demonstrated that decreasing grid spacing from 27 to 9 km allowed the MM5 to realistically simulate the structure of a bent-back occlusion and its interaction with Northwest coastal terrain during the inauguration day storm of 1993. Simulating a Pacific cold front with the MM5, Chien et al. (2001), found that decreasing grid spacing from 45 to 15 km did not greatly change the system’s synoptic structure; however, decreasing grid spacing down to 1.67 km was necessary to produce the observed narrow frontal structure, an associated convective rainband, and a gravity-current headlike structure near the front’s leading edge. In an MM5 simulation of a polar low that developed over the Bering Sea, Bresch et al. (1997) found that although a grid spacing of 6.7 km produced sharper mesoscale features than a 20-km grid, the central pressure and synoptic-scale structures were nearly unchanged. Zhang et al. (2000), simulating the 24–25 January 2000 snowstorm over the eastern United States, found that decreasing grid spacing from 20 to 10 km produced substantial benefits, while going from 10 to 3.3 km resulted in insignificant improvements. Benoit et al. (2000) coupled a hydrological model to the Mesoscale Compressible Community Model (MC2) atmospheric model for several days in April 1993, during which time a cyclone and associated warm front influenced their study region (southern Ontario). Using the hydrologic model and resulting streamflow as verification tools, they found a substantial improvement in hydrographs as atmospheric model resolution increased from 35 to 10 km, with a smaller enhancement as grid spacing was decreased to 3 km. Buckley and Leslie (2000) found that decreasing grid spacing from 50 to 10 km in the University of New South

In summary, these and other subjective studies suggest that moving to very high resolution (grid spacing less than 15 km) alone does not significantly improve the synoptic evolution of fronts and cyclones, but does produce better definition of frontal zones and other mesoscale features.

**Longer-period objective verification studies.** Only a handful of studies have examined the effects of model grid spacing on forecast accuracy for an extended period using objective measures such as skill score, root-mean-square (rms) error, or the like. For example, evaluating 24-h accumulated precipitation forecasts for October–December 1997 at nearly 800 sites in the intermountain west, White et al. (1999) found that 18-km MM5 forecasts outperformed the 29-km MesoEta and the far coarser Medium-Range Forecast (MRF) for heavier precipitation categories (19 mm and more). In contrast, for lighter precipitation amounts (12 mm and less), the lower-resolution MRF was the most skillful and the MM5 was the worst performer. Colle et al. (2000b) evaluated MM5 precipitation forecasts at 36-, 12-, and 4-km horizontal resolution over western Washington state for two winter seasons. They found a noticeable improvement in short-term precipitation forecast skill as grid spacing decreased from 36 to 12 km. In contrast, going from 12- to 4-km grid spacing only improved forecasts of heavy precipitation.

Nachamkin and Hodur (2000) verified twice-daily runs of the COAMPS model from 30 September 1998 to 31 December 1999 at radiosonde sites over a domain encompassing the Mediterranean Basin. They found that decreasing grid spacing from 81 to 12 km had little effect on rms error for temperature and relatively humidity, but produced noticeable improvements in the lower-tropospheric biases of these parameters. Wind speed errors, in terms of either rms error or bias, were nearly unchanged by increased resolution.

**Predictability and the impact of model resolution.** A number of theoretical and modeling studies have examined the relationship between model resolution and predictive skill. Early studies considered the growth and spread through wavenumber space of uncertainty and error, the most important source originating at shorter scales that are either unresolved/poorly resolved by models or inadequately described by the observational network. For example, using a two-dimensional form of the vorticity equation, Lorenz (1969) examined the effects of small initialization errors on predictability. He found that the period of useful forecast skill decreases with scale, with predictability of less than 1 h for spatial scales smaller than approximately 40 km—a very pessimistic finding for mesoscale prediction. Tennekes (1978) suggested that synoptic and large mesoscale systems act as two-dimensional turbulence and thus possess more inherent predictability (on the order of days) than cloud-scale convective systems, which act as three-dimensional turbulence with predictability on the order of hours or less.

Anthes et al. (1985), Anthes (1986), Paegle et al. (1997), Vukicevic and Errico (1990), Warner et al. (1997), among others, have found greater predictability and slower error growth at high resolution than suggested by the arguments of Lorenz (1969). Specifically, they suggest that lateral boundary conditions can significantly reduce error growth in local area models, since larger-scale information constantly sweeps through the inflow boundaries. Surface forcing (e.g., topography, land–water contrasts) can extend the period of predictability for small-scale features. Topographically driven mesoscale circulations often result from the interaction of synoptic-scale flow with fixed complex terrain, thus resulting in mesoscale predictability being controlled by synoptic-scale predictability. Diurnal circulations, driven by surface variability, are generally forecast skillfully by mesoscale models, and thus contribute to reduced error growth. With lateral boundaries and surface features slowing error growth, some of these studies (e.g., Paegle et al. 1997) have suggested that key sources of local error growth are small uncertainties in the larger spatial scales rather than uncertainties on the smaller scales.

Even away from topography and surface contrasts, substantial mesoscale predictability may exist. For example, Gall and Shapiro (2000) noted that since oceanic fronts usually result from synoptic-scale areas of frontogenesis, they may possess larger-scale predictability. Furthermore, mesoscale circulations such as fronts and tropical cyclones may dynamically “resist” the energy cascade and thus experience extended predictability (Anthes 1986). Enhanced predictability may also extend to rotating severe thunderstorms, in which helicity effects may suppress the turbulent energy cascade (Lilly 1986). Finally, atmospheric energy spectra often do not match the idealized spectra used in theoretical studies such as Lorenz (1969), which can result in increased predictability on the mesoscale (Lilly 1986; Anthes 1990).
GOALS. This study evaluates the forecast accuracy of a mesoscale model (MM5) run over the Pacific Northwest at 36-, 12-, and 4-km horizontal resolutions over a period of several years using objective verification tools. This evaluation makes use of a particularly dense mesoscale verification database created by combining all available regional data sources. Although the results are strictly applicable only to the western portion of Washington state, they should provide general guidance regarding forecast skill in a midlatitude coastal zone of substantial topography. In the discussion section, these results are placed in perspective and the nature of mesoscale model verification is examined.

THE UNIVERSITY OF WASHINGTON REAL-TIME MODELING AND VERIFICATION SYSTEM. UW MM5 Modeling System. Since October 1997 the University of Washington (UW) has run the MM5 twice daily (initialized at 0000 and 1200 UTC) for 48 h on a 36-km horizontal resolution grid that covers much of the eastern Pacific and western North America and a 12-km resolution nested grid encompassing Oregon, Washington, and southern British Columbia (Fig. 1). In addition, a 4-km nested domain has been run over western Washington for 24 h (from 12 to 36 h into the forecast) through November 1999, after which the 4-km nest was expanded to include all of Washington state. The 4-km domain was further expanded to include Oregon and the 36/12-km runs were extended to 60 h in August 2000. Thirty-three sigma levels are used in the vertical, with maximum resolution in the boundary layer. The nesting is one-way.

The UW MM5 forecast system has used the explicit moisture scheme of Hsie et al. (1984), with the improvements to allow ice-phase microphysics below 0°C (Dudhia 1989); the Kain–Fritsch cumulus parameterization (Kain and Fritsch 1990) was applied in the 36- and 12-km domains. The MRF planetary boundary layer parameterization (Hong and Pan 1996) was used in all domains. A 30-s land-use dataset from NCAR was utilized to initialize 13 surface categories, and sea surface temperature was derived from the real-time U.S. Navy Optimum Thermal Interpolation System (OTIS) analysis (30-km horizontal resolution). Snow cover was based on NCEP analyses through March 1998, after which the U.S. Air Force snow distribution was used. The terrain within the 4-km domain and some major geographic features are shown in Fig. 2a.

Initial and boundary conditions for the real-time forecasts were obtained from NCEP Eta Model initial conditions and forecast fields, which were available initially at 48-km horizontal resolution before being enhanced to 32 km in May 1999. The UW real-time MM5 is initiated as a “cold start,” with no preforecast spinup period or assimilation of additional observations. More information about the UW

---

1 In November 1999 the 4-km integration was also started 6 h earlier, i.e., from 6 to 36 h into the 36/12-km forecasts.
2 The 33 sigma levels were \( \sigma = 1, 0.99, 0.98, 0.97, 0.96, 0.94, 0.92, 0.90, 0.88, 0.86, 0.83, 0.80, 0.77, 0.74, 0.71, 0.68, 0.64, 0.60, 0.56, 0.52, 0.48, 0.44, 0.40, 0.36, 0.32, 0.28, 0.24, 0.20, 0.16, 0.12, 0.08, 0.04, 0.0\).
3 The Eta Model output was acquired from NCEP’s “221” grids when possible, defaulting to the lower-resolution (80 km) “104” grids.
real-time mesoscale forecasting system is found in Colle et al. (2000; or online at the UW MM5 Web site http://www.atmos.washington.edu/mm5rt/mm5.cgi).

**UW forecast verification system.** A real-time verification system was created to evaluate the skill of the regional MM5 forecasts. The basic approach has been to verify at the observation sites by interpolating model fields to the observation locations.\(^4\) For precipitation, this interpolation used the Cressman (1959) scheme, in which the model precipitation at the four grid points surrounding the observation site were weighted depending on the square of the distance of each grid point to the observation location (see Colle and Mass 2000b for further details). For the temperature verification, the difference between the model terrain height and the actual height of the observations had to be taken into account. To do this the temperature of the lowest model level (40 m), the model terrain height, and the model ground temperature were first bilinearly interpolated to the observation location. The interpolated model 40-m temperature was then “reduced” to the observation elevation using a standard lapse rate (6.5°C km\(^{-1}\)). This reduced free air temperature was then averaged with the model ground temperature to provide the model’s prediction of 2-m temperature. For wind speed, the 40-m model values were first bilinearly interpolated to the observation location, followed by a reduction to the standard 10-m height using a logarithmic wind profile:

\[
U(10m) = U(40m) \frac{\ln(10m/z_0)}{\ln(40m/z_0)},
\]

where \(z_0\), depending on land use was applied and \(U\) represents wind speed. For winds, no attempt was made to compensate for the difference between model and actual terrain heights.

The UW model verification system makes use of a collection of regional observations including the following:

1) National Weather Service/FAA Automated Surface Observing System (ASOS) sites
2) Fixed buoys and coastal marine (CMAN) stations
3) Washington State Department of Ecology sites
4) Hanford (U.S. Department of Energy) and Public Agriculture Weather System (PAWS) networks
5) Northwest Avalanche Center (NW) and Remote Automatic Weather Station (RAWS) networks
6) USDA Snow Telemetry (SNOTEL) network and NOAA Cooperative Observer (COOP) networks.
7) Washington State Department of Transportation Road Weather Information System (RWIS) sites.

These data sources result in approximately 150 observations within the 4-km domain covering the western half of Washington. The quality of these data varies, with the ASOS observations having the best calibration and smallest instrument error.\(^5\) The

\(^4\) An alternative verification approach is to interpolate the observations to a grid on which verification takes place. However, over the mountainous western United States, where observations are often sparse and gradients are large, such grid-based verification is often not viable.

\(^5\) The
Northwest MM5 verification sites for precipitation and other parameters are shown in Figs. 2b and 2c, respectively.

After collection, each observation is tested with a gross-error check, and temperatures are run through a spatial consistency check to ensure that each temperature observation is not radically different from values observed at nearby stations at approximately the same elevation. More details of the UW quality control system can be found at the Web site noted above.

After quality control, a variety of verification statistics are calculated, including rms error, mean absolute error (mae), mean error (me; forecast/observed), and bias (forecast/observed). Verification statistics are calculated every hour at all stations within all three domains. However, in this paper only 3-hourly statistics for the area encompassing the 4-km domain (essentially the western half of Washington state) are presented, using model data output from all three resolutions.

**MULTIYEAR COOL-SEASON PRECIPITATION VERIFICATION.** This section briefly reviews the long-term verification of the University of Washington MM5 precipitation forecasts, described in greater detail in Colle et al. (1999) and Colle et al. (2000). For the latter paper, 36-, 12-, and 4-km precipitation forecasts (see Fig. 1 for domains) of the UW real-time MM5 system were verified for a large portion of the winter seasons of 1997/98 and 1998/99. Approximately 125 precipitation verification locations are found in the 4-km verification domain, a number of which are located at high elevations (mainly SNOTEL and NW Avalanche Center reports). A bias score is calculated using a contingency table approach (see Colle et al. 1999 for more details) with

\[
\text{Bias} = \frac{F}{O}
\]

where \(F\) is the number of forecasts at an observation location with precipitation equal to or exceeding a given threshold amount, and \(O\) is the number of occurrences in which the observations meet or exceed the threshold. Colle et al. (2000) estimated precipitation undercatchment ranging from 0%–10% at low levels to 25%–40% at high elevations where strong winds and snow are frequent; thus, a “perfect” forecast is undoubtedly associated with a modest positive bias. The rms errors were also calculated at observation locations using

\[
\begin{align*}
\text{rms error} &= \sqrt{\frac{1}{N} \sum (F - O)^2} \\
\text{mae} &= \frac{1}{N} \sum |F - O| \\
\text{me} &= \frac{1}{N} \sum (F - O) \\
\text{bias} &= \frac{1}{N} \sum (F/O - 1)
\end{align*}
\]

\(^5\) According to the official ASOS manual, ASOS wind speed is accurate to ±2 knots or 5%, whichever is greater; wind direction is accurate to ±5° when the wind speed is blowing at 5 knots or more, the pressure sensor is accurate to ±0.02 in of Hg, and temperature is accurate to ±1.8°F.
where NOBS is the number of observations at that location reaching a certain precipitation threshold, $X_n$ is the observed precipitation at the site, and $P_n$ is the model precipitation interpolated to the observation location.

Figure 3 shows the forecast bias scores for the 24-h period from 20 to 44 h into the forecast, verified within the 4-km domain for all three grid resolutions as a function of precipitation threshold. Although all resolutions have reasonable (~105%) biases for the light thresholds, the 36-km grid spacing greatly underpredicts at higher thresholds (generally over windward slopes). A major improvement is apparent at 12-km grid spacing for nearly all thresholds; only for the heaviest precipitation amounts (greater than about 1.9 in. in 24 h) is a modest underprediction apparent. Decreasing grid spacing to 4 km produces an apparent overprediction at most thresholds.

The corresponding precipitation rms errors are shown in Fig. 4. Overall, the 12-km grid spacing results in the lowest rms errors, with the 36-km grid generally being the least skillful for moderate and large amounts. The 4-km grid spacing is superior only for the highest thresholds, and is the least skillful domain for light precipitation amounts.

There is considerable spatial variation in MM5 precipitation biases over the domain. To illustrate this fact, Fig. 5 presents model biases calculated by adding 20–44-h precipitation forecasts during the above-mentioned winter periods and dividing by the corresponding observed precipitation. With 36-km grid spacing, there is underprediction over a considerable portion of the domain, particularly over the windward slopes of the Olympics and the central Cascades (Fig. 5a). Not surprisingly, overprediction is found to the lee (generally northeast) of barriers such as the Olympics, since simulated downslope subsidence is attenuated by the smooth 36-km terrain. Decreasing grid spacing to 12 km significantly reduces the magnitude and extent of the underprediction problem, which is now limited to the southern (windward) slopes of the Olympics and a few major gaps in the central Cascades (Fig. 5b). With 12-km grid spacing another problem develops—overprediction over the windward slopes and crests of the Cascades, with a few locations receiving twice (or more) of the observed precipitation. At 4 km, underprediction is only found at a few stations upstream of the Olympics, and overprediction increases over some of the higher terrain (Fig. 5c). Although the 4-km domain appears to produce excessive amounts of precipitation over terrain, this overprediction occurs mainly for light or moderate intensities, and the 4-km domain is generally the most skillful resolution for the heaviest precipitation.

In fact, Colle and Mass (2000b), in a study of the...
record-breaking Northwest flood of 5–9 February 1996, found that MM5 precipitation amounts for the storm improved as model grid spacing decreased from 12 to 4 km. The tendency for overprediction of light to moderate precipitation intensities over terrain by the 4-km MM5 forecasts has been confirmed by coupled MM5/distributed hydrologic model forecasts over extended periods Westrick et al. (2001); verifying several consecutive months of runoff forecasts at a number of river gauges on the windward slopes indicated consistent overprediction.

In summary, based on hundreds of forecasts, the results from Colle et al. (1999, 2000) summarized above indicate that MM5 precipitation skill over western Washington improves as grid spacing decreases from 36 to 12 km. The 36-km grid spacing does not adequate-

**MULTIYEAR VERIFICATION OF SURFACE WIND, TEMPERATURE, AND PRESSURE.** In addition to precipitation, other surface parameters produced by the UW real-time mesoscale modeling system were also verified over a multiyear period.

**Wind.** Figure 6a summarizes the mean absolute errors for wind direction at all three resolutions for the UW MM5 runs initialized at 0000 UTC for the period from 29 September 1997 to 12 December 1999 and verifying within the 4-km domain. Wind direction is perhaps the best parameter for evaluating the skill of a mesoscale model over the Northwest: there are a large number of sites measuring this parameter and wind direction is greatly modulated by the region’s terrain and diurnal circulations. Wind direction errors were calculated by comparing the directions at the lowest model level (40 m) with near-
surface (nominally 10 m) observations, since under unstable or neutral conditions little directional shear is expected in the surface layer (Estoque 1973). Furthermore, only those observations with wind speeds greater than 5 knots were considered since wind direction measurements are often unreliable at low wind speeds. Between 16 792 (33 h) and 29 674 (21 h) verification pairs (observations and forecasts) were evaluated for each forecast hour. It is apparent that except for 0 h (initialization time), the 12-km wind directions were noticeably (~5°) better than those from the 36-km grid. Both 36- and 12-km wind direction errors decreased considerably during the first 3 h (as the model spun up) and then slowly increased in time. Superposed on the slow error increase is a diurnal signal in which wind direction errors are largest during the morning (~1500–1800 UTC, 7–10 AM PST) and smallest during the late afternoon and early evening (~0000–0300 UTC, 4–7 PM LST), when vertical mixing is greatest. The 4-km domain is the most skillful domain for the hours it is available (hours 12–36); however, the improvement over the 12-km wind direction is quite small, averaging about 1°.

The mean wind direction error for the 0000 UTC runs (Fig. 6b) shows a substantial improvement with resolution, decreasing from approximately 15°–24° in the 36-km domain to 10°–16° using 12-km grid spacing. There is an additional improvement by 1°–2° at 4-km resolution, with errors dropping below 10° at hour 24. These errors reflect the tendency for the direction of the MM5 model winds at all resolutions to be cyclonically rotated from the observed, that is, they tend to be too geostrophic. This tendency might be explained by excessive mixing in the vertical, a characteristic that several of the MM5 boundary layer parameterizations, including the MRF scheme used in the UW real-time effort, appear to possess. Interestingly, there is no suggestion of increasing mean error with time; if anything, mean direction error seems to decrease with forecast projection. Mean direction error

---

6 The only exception is at hour 12, when the 4-km grid is initialized and thus not spun up.
error also evinces considerable diurnal modulation, with large error during the night and morning hours (6–18 and 30–42 h, 10 PM–10 AM LST) and small error during the late afternoon and early evening (24–27 h, 45–48 h). Thus, mean direction errors are smallest during the portion of the day in which there is naturally more vertical mixing.

Mean absolute errors of 10-m wind speed for all three grids verified over the 4-km domain are shown in Fig. 7a. As noted previously, since the lowest model wind level is at 40 m, wind speeds were reduced to standard anemometer height (10 m) using a logarithmic profile. About 50 000 observation/model pairs were used to calculate each 3-hourly value. As with wind direction, there is a drop in 36- and 12-km errors during the first 3 h, followed by very slow error growth over the subsequent 2 days. At all hours, the wind speed error is substantial (~4 kt). The 12-km wind speed errors are always (but only slightly) less than those of the 36-km domain, while 4-km wind speeds are marginally better than the 12-km values for only 3 h (18, 21, and 24 h). There is some hint of diurnal modulation of wind speed errors, with lesser errors during the day and greater errors during the early morning hours.

The mean wind speed error (or bias) is shown in Fig. 7b. For the 36- and 12-km domains there is a tendency for excessive winds (by ~1 kt) during the night and early morning (6–15 h, 27–39 h) and weaker than observed winds (by ~0.5 kt) during the day (18–24 h, 24–48 h), with resolution helping slightly during the day but increasing the bias at night. The 4-km domain has little bias during the day, but enhanced positive bias at night. One contributor to the differing behaviors of the 36-, 12-, and 4-km domains is the subgrid-scale drag parameterization that was used for the lower-resolution domains over complex terrain but was not applied in the 4-km grid. This scheme makes $u^*$ proportional to the variance of the subgrid-scale height terrain and attempts to include the drag effects from unresolved topographic elements. Further details are found in Chien et al. (2001). In any case, excessive winds at night imply that the MM5 MRF boundary layer scheme produces too much mixing during that part of the day.

**Temperature.** The mean absolute error of near-surface (2 m) air temperature for the 0000 UTC runs is shown in Fig. 8a. As with wind errors, there is a decrease in error during the first 3 h, followed by a slow increase during the remainder of the period. Superposed on the slow error increase with time, a diurnal component is evident, with largest error during the warmest part of the day (hours 21–24, 45–48). The 36-km forecasts are marginally less skillful than those from the higher-resolution grids, particularly during the warmer afternoon periods. The 12- and 4-km temperature errors are very similar, with the 4 km being slightly better between 15 and 36 h into the forecast.

The mean error of the near-surface air temperature is found in Fig. 8b. At all resolutions there is a modest (~1°C) warm tendency during the night (model warmer than observed) and a similarly sized cold bias during the day. The 36-km domain appears to have the smallest warm bias during the day, but the largest cold bias at night.

**Sea level pressure.** Figure 9 shows the mae for sea level pressure for all three resolutions verified in the 4-km domain. Unlike wind and temperature, the sea level pressure mae increases steadily after initialization, growing from approximately 1 mb at 0 h to roughly 2.5 mb after hour 39. There is only a small difference between the resolutions.
in the errors among the various resolution grids, with the 36-km domain consistently being the least skillful.

Spatial variations of forecast accuracy for nonprecipitation parameters. In Colle et al. (2000) it was shown that domain-average precipitation statistics told only part of the story regarding model performance, since they found substantial spatial variations in precipitation forecast accuracy across the mountainous Pacific Northwest. This section explores the regional variations in forecast accuracy for wind direction and other parameters.

The mean absolute errors for wind direction at hour 24 for the 36- and 12-km MM5 forecasts are shown in Fig. 10. Only hours in which the observed winds were greater than 5 knots were used, since wind direction observations are unreliable at low speeds. Both resolutions possess large spatial variances in forecast accuracy. At 36 km, mean absolute wind direction errors are generally between 30° and 50° over Puget Sound, with larger errors (50+°) over the eastern slopes of the north Cascades as well as to the northeast of the Olympics from the eastern Strait of Juan de Fuca to Bellingham (Fig. 10a). Interestingly, a number of stations on the western slopes of the Cascades had relatively low errors (0°–30°). Increasing horizontal resolution to 12 km significantly reduces the errors to the northeast of the Olympic Peninsula, as well as lessening errors at a few southern Puget Sound sites to less than 30° (Fig. 10b). Enhancing resolution from 12 to 4 km provides nearly imperceptible improvements (not shown). The substantial differences in verification scores for several nearly collocated locations highlights the significance of local exposure and instrument error.

Mean Absolute Wind Direction Error at 24h for Winds >5kt

36 km

12 km

Fig. 10. The mae for wind direction (degrees) at hour 24 for the 36- and 12-km domains verified over western Washington state for the 0000 UTC forecast cycle. Only observations with wind speeds exceeding 5 kt were used.
Figure 11 further clarifies the effects of horizontal resolution on wind direction error by dividing observing sites into those that had lessened (left side) or increased (right side) mean absolute errors for the 36–12-km (top) and 12–4-km (bottom) resolution changes. Again, only wind observations in which the speed exceeded 5 knots were used. For the 36–12-km resolution increase, far more stations had improved than degraded verifications, with particular benefit over southeastern Vancouver Island, the eastern Strait of Juan de Fuca, and Puget Sound (Fig. 11, top). For the transition from 12 to 4 km (Fig. 11, bottom), there was more balance between improvement and degradation. Improvements were generally modest in magnitude and most prominent over the Strait of Juan de Fuca and Georgia Strait. Degradations were more localized and reached larger amplitudes, especially over and to the east of the Cascades. Wind speed and surface temperature show a similar geographical distribution of error change with resolution (not shown).

For all parameters, the region encompassing the southeastern corner of Vancouver Island and the eastern Strait of Juan de Fuca experienced the largest improvement in forecast skill as resolution was increased from 36 to 12 km, with noticeable improvements as grid spacing decreased to 4 km. High resolution contributes to better definition of the Olympic Mountains and the terrain of Vancouver Island, resulting in better simulations of the complex pressure (lee troughing) and wind patterns in the lee of those barriers. With the climatological wintertime flow off the Pacific being southwesterly, lee troughing to the northeast of the Olympics is persistent and often stationary, making timing errors less of a problem in this area.

Filtering of poor synoptic forecasts: Implications for mesoscale forecasting skill. Mesoscale model forecast errors originate in model deficiencies, inadequate mesoscale and synoptic initializations, and the inherent predictability limits of mesoscale flows. How much improvement in mesoscale accuracy might be expected from enhancements in the synoptic-scale forecasts that provide the initializations and boundary
conditions for the nested mesoscale model simulations? To gain some insight into this issue, the verification data can be filtered to include only periods in which synoptic-scale forecasts are relatively accurate. Colle et al. (2000) evaluated 24-h precipitation verification statistics for those times in which wind direction and speed errors at 850 mb were less than 30° and 5 m s\(^{-1}\), respectively, at both the Salem, Oregon, and Quillayute, Washington, radiosonde sites. Retaining only these relatively well-forecast periods (about 40% of the total) produced a positive impact on forecast skill that grew as the rainfall rate increased, reaching about a 25% improvement in rms errors for a threshold of 4 in. in 24 h. The effects of synoptic filtering were similar at all three resolutions (36, 12, and 4 km).

Considering both all forecasts and those forecasts for which 850-mb winds at Quillayute (a coastal station with little upstream blockage) were within 20° and 5 kt of observed for the 12- and 24-h forecasts, the mean absolute wind direction errors of the 24-h forecasts were calculated for those sites with surface wind speeds greater than 5 kt. Approximately 93% of the forecast hours were eliminated with this synoptic filter, illustrating how rarely the large-scale flow is correctly predicted. The results, shown in Fig. 12, indicate modest (−6°) improvement for the 36-km domain, decreasing to approximately 3° for the 4-km grid. These findings suggest that a very substantial increase in synoptic forecast skill is needed to produce dramatic improvements in point verifications as done in this study. Of course, a more sophisticated filter (e.g., allowing only forecasts with less than 1-h timing errors) could be applied, but only a very small percent of the total forecasts would be included.

**A CASE STUDY OF THE EFFECTS OF RESOLUTION: THE FRONTAL PASSAGE OF 20 OCTOBER 2000.** The influence of resolution on forecast skill and the difficulties produced by timing and position errors are illustrated by a frontal passage across the Northwest coast on 20 October 2000. This event not only produced a substantial coastal wind shift but also left a strong Puget Sound convergence zone in its wake. Figure 13a shows 40-m winds and sea level pressures in the 12-km MM5 domain at 12 h into the forecast initialized at 0000 UTC 20 October 2000. At that time, the front was offshore of the Northwest coast and was associated with a pronounced trough and cyclonic windshift. Figure 13b shows the observed and forecast coastal winds at Astoria, Oregon (near the Oregon/Washington border), during the period of frontal passage. It is evident that the forecast front at all three resolutions made landfall approximately an hour early, and that the sharpness of the front increased and became more realistic as the horizontal grid spacing decreased from 36 to 4 km.

Figures 14a–c presents 18-h forecasts of 1-h precipitation amount verifying at 1800 UTC 20 October for the 36-, 12-, and 4-km horizontal resolutions over northwestern Washington state. At this time the forecast front was moving over Puget Sound, with the coastal winds already having veered into the west. Figure 14d shows the observed radar reflectivity from Camano Island at 1729 UTC 20 October, near the midpoint of the 1-h accumulation period, while Fig. 14e presents the radar an hour later (1828 UTC). Since the forecast frontal passage was about an hour too fast, comparison to the latter radar image should provide a better evaluation of the fidelity of model structures.

The 4-km 18-h precipitation forecast for the 1-h period ending at 1800 UTC (Fig. 14c) shows a band of heavy precipitation extending northwest of the Olympics in a position that was nearly identical to one shown in the radar image valid 1 h later (1828 UTC, Fig. 14e). This enhanced rainfall was caused by the confluence of westward air from the Strait of Juan de Fuca and southerly flow over Puget Sound, coupled with upslope along the western slopes of the Cascades. The 4-km forecast also produced the observed rain shadow over central Puget Sound. A comparison to the “on-time” radar image (1729 UTC, Fig. 14d) was far less favorable. The 12-km precipitation forecast
(Fig. 14b) produced a generally similar solution, but with less realistic structures: the confluence band was too wide and weak, and the observed rain shadow over central Puget Sound was substantially attenuated. The 36-km forecast was profoundly inferior to those of the higher-resolution grids, with little evidence of the precipitation band or rain shadowing east of the Olympics (Fig. 14a).

Three hours later (2100 UTC 20 January), there were substantial differences among the MM5 precipitation forecasts at the various resolutions (Fig. 15). The 4-km domain not only shows precipitation over the western slopes of the Cascades and Olympics, but also indicates a well-defined Puget Sound convergence zone precipitation band extending roughly east–west over southern Whidbey Island (Fig. 15c). Although this forecast position is somewhat south of the band observed by the radar at the nominal time (2033 UTC, Fig. 15d), it is virtually identical to the precipitation distribution shown 1 h later (2132 UTC, Fig. 15e). In contrast, the 12-km simulation shows only the slightest hint of a convergence zone east of Seattle, while the 36-km domain has no convergence zone feature at all.

This case, which reflects the model performance noted during many events, suggests major improvements in mesoscale structures as the grid spacing decreases from 36 to 12 km. Increasing resolution from 12 to 4 km produced further finer-scale improvements, increasing the definition and intensity of mesoscale features. Although subjective comparisons of observed and forecast structures suggest the value of resolution, objective evaluations using point verification and traditional skill scores (e.g., absolute or root-mean-square errors) can result in somewhat different conclusions, particularly if there are timing or positional errors of the mesoscale features.

To illustrate the effects of horizontal resolution and timing errors on objective point verification scores, a simple objective verification scheme was created in which precipitation during the 20 October event was

**Fig. 13.** (a) The 40-m winds and sea level pressure in the 12-km MM5 domain at 12 h into the forecast initialized at 0000 UTC 20 Oct. (b) Observed and model winds at Astoria, OR, during the period of frontal passage. The location of Astoria is indicated by a square box in (a) at the coastline.
verified at five locations over the Puget Sound lowlands (see Fig. 2a for the five positions). At each of these points the model 1-h precipitation for all three resolutions were compared to the radar reflectivities at two times: the on-time radar reflectivity at the midpoint of the 1-h collection period, and at the “displaced-time” radar reflectivity observed 1 h later. Use of the latter is an attempt to correct for the 1-h timing error in the simulated frontal passage. A value of 1 was given for each correct forecast (rain forecast and rain observed, no
rain forecast and no rain observed). This procedure was done for all three resolutions at three forecast times: 18 h (1800 UTC 20 October), 21 h (2100 UTC 20 October), and 24 h (0000 UTC 20 October). The results are presented in Table 1. The most obvious result was the significant increase in the number of correct forecasts as the resolution increased from 36 to 12 km, and the small improvement accompanying a further reduction in grid spacing from 12 to 4 km. The 36-km domain was penalized for spreading light precipitation over too

Fig. 15. The 21-h MM5 forecasts of 1-h precipitation amounts (in hundredths of an inch) verifying at 1800 UTC 20 Oct for the (a) 36-, (b) 12-, (c) 4-km grid spacing. WSR-88D radar reflectivity for the 0.5° elevation angle at (d) 2033 and (e) 2132 UTC 20 Oct 2000.
large an area. Using radar data 1 h later to compensate for the frontal timing error significantly improved the model verification at all resolutions, with the least improvement at 36 km.

Table 2 presents verification of mean and mean absolute precipitation errors (me, mae) for the 20 October event at eight observing sites within the Puget Sound lowlands for 3- or 6-h periods, time intervals sufficiently long to reduce the effects of the 1-h frontal timing error. For the mean precipitation error (essentially the bias), the 4-km domain was the best except for the 24–27-h period, when the 12-km domain had the smallest mean error. For the mean absolute error, the 12-km domain had the smallest error during the first three periods, while the 4-km domain was the most skillful from 24 to 33 h. This transition from 12- to 4-km superiority makes heuristic sense: during the initial period, while the front and the convergence zone were active and mobile, the position and timing errors of the more intense structures in the 4-km domain were more highly penalized than the more diffuse 12-km solution. During the latter two periods, the front had moved out of the region and the convergence zone has weakened, resulting in a more steady-state precipitation pattern that was dominated by orographic enhancement and lee rain shadowing, features that were best defined by the highest resolution.

**DISCUSSION AND CONCLUSIONS.** As reviewed in section 2 above, subjective evaluations and a limited number of short-term objective studies have generally found more realistic mesoscale structures and evolutions as grid spacing decreases into the single digits (km). This paper has attempted to extend these studies by presenting the results of an objective multiyear verification of the University of Washington real-time MM5 forecasts in which predictions from the 36-, 12-, and 4-km domains were verified at observation locations over western Washington state. These findings, and others presented in companion works (Colle et al. 1999, 2000), suggest clear improvements in precipitation, 10-m wind, 2-m temperature, and sea level pressure forecasts as grid spacing decreases from 36 to 12 km. It appears that transitioning from 36- to 12-km grid spacing allows the definition of the major mesoscale topographic features of the region and their corresponding atmospheric circulations, producing a beneficial effect on the verifications. In contrast, there are only small improvements in verification statistics as grid spacing decreased from 12 to 4 km. Decreasing grid spacing to 4 km provides more detail and structure (e.g., defining steeper orographic slopes) but has only a limited impact on traditional objective verification scores.

For several reasons, a coastal area of complex terrain such as the Pacific Northwest may be well suited for producing increased forecast skill as horizontal resolution increases. Most of the region’s mesoscale circulations are created by the interaction of the synoptic-scale flow with the mesoscale terrain; thus, mesoscale predictability is substantially controlled by longer-lived synoptic predictability. Defining upstream mesoscale circulations entering the domain is not as important as in the eastern and central United States where upwind complex terrain and surface variations produce a plethora of significant mesoscale structures. Upstream of the Pacific Northwest, a nearly uniform surface (the Pacific Ocean) results in fewer and less intense low-level mesoscale structures, and many of the mesoscale structures that do exist (e.g., frontal zones) are forced and controlled by synoptic-scale circulations. Deep, organized mesoscale convection is rare in the Northwest, and thus a difficult mesoscale forecast problems is avoided.

It is reasonable to ask whether the limited extents of the 12- and 4-km domains had a significant impact on the verification scores. Would our conclusions have changed significantly if much larger inner domains were used? Certainly, a number of earlier studies (Anthes et al. 1985; Warner et al. 1997; Paegle et al. 1997) have shown the importance of lateral boundary conditions in restraining error growth in limited-area models. To address this issue using the UW MM5 system, Steed et al. (2000) varied the size of the inner (4 km) domain from its current boundary to one encompassing most of the 12-km region (5.3 times larger area) for a collection of five events. Such a large increase in the size if the inner domain had very little effect on the 0–48-h surface forecasts of temperature,

---

<table>
<thead>
<tr>
<th>TABLE 1. The 1-h precipitation verified at five points over Puget Sound. The table shows the number of correct predictions (rain vs no rain) for 18 h (valid 1800 UTC 20 Oct), 21 h (2100 UTC 20 Oct), and 24 h (0000 UTC 20 Oct) MM5 forecasts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 km</td>
</tr>
<tr>
<td>On-time radar</td>
</tr>
<tr>
<td>Displaced-time radar</td>
</tr>
</tbody>
</table>

<sup>They also found that varying the lateral boundary update frequency from 36 s to 1 h had little effect on model forecasts.</sup>
An interesting aspect of the above verification results for wind and temperature is the slow growth in model error, far less than the 1.5–2.5-day doubling times found for large-scale flows by Lorenz (1982) and Simmons et al. (1995). Such slow error growth has been noted in regional modeling experiments discussed by Anthes et al. (1985), Anthes (1986), Vukicevic and Errico (1990), Warner et al. (1997), and Paegle et al. (1997). More recently, Cook (1998), evaluating the skill of the MesoEta (29-km grid spacing) and the University of Utah real-time MM5 forecasts (18 km) over a 3-month period, found slow error growth in surface winds and temperature. In an analysis of real-time ensemble results, Wandishin et al. (2001) noted that the spread of the various members (which they equated to model error) increased slowly with forecast time (through 48 h) for precipitation and instability parameters, but increased more rapidly for 500-mb height, a parameter that reflects the evolution of the synoptic flow.

There are a number of potential explanations for the slow error growth noted in mesoscale forecasting studies. As suggested by Anthes et al. (1985) and others, both the lateral boundary conditions of limited-area models and surface forcing (e.g., topography, land–water contrasts) constrain error growth. As noted above, over regions of orography many of the mesoscale features are produced by the interaction of the synoptic-scale flow with topography and thus mesoscale error growth is controlled by the slower error growth of the synoptic scale. In addition, local terrain can slow error growth further, since some features (such as windward precipitation enhancement and lee rain shadowing) can be insensitive to small changes in synoptic-scale flow. Another less sanguine explanation for slow error growth is that errors arise so rapidly that they become saturated within the first hours, thus resulting in slower subsequent error growth in time. For example, at higher resolution, model structures are stronger and better defined; thus even small timing and placement errors produce substantial forecast errors. The ability of a high-resolution model to verify well is also seriously degraded (compared to larger-scale forecasts) by data scarcity and representativeness problems, as well as deficiencies in the physical parameterizations of the planetary boundary layer and surface heat and moisture budget parameterizations on smaller scales.

In the results presented above, the mean absolute errors of sea level pressure grew rapidly, in contrast to the slower growth of wind and pressure errors. Perhaps this rapid loss of skill expresses the deterioration of the synoptic forecasts without the benefit of the deterministic contributions from terrain-induced and diurnal circulations. Furthermore, sea level pressure observations are only available at a limited number of surface observing sites that are preferentially located over the lower elevations. Such pressure observations generally fail to sample the strong orographic pressure signals produced by terrain (e.g., the trough on the lee slopes of the Olympics), and thus are more representative of the larger-scale flow.

For several reasons, the objective verification results for the UW MM5 real-time forecasts, which suggest minimal forecast improvement as grid spacing decreases from 12 to 4 km, may be overly pessimistic. First, as shown in the case study in section 7, inevitable errors in timing or position are amplified as resolution increases. Increased horizontal resolution generally produces better defined mesoscale structures, greater amplitude of troughs and other features, and larger gradients. Thus, even if structures are more realistic in a high-resolution domain, the existence of timing errors will result in lower objective skill scores (e.g., biases, mean and root-mean-square errors, threat scores, etc.) as resolution increases. Over the eastern Pacific, frontal timing errors are often quite large; evaluating MM5 and Eta trough timing errors over the eastern Pacific for the 1997–2000 cool seasons, Colle et al. (2001) found mean absolute timing errors of 2–5 h.

### Table 2. The 3- and 6-h precipitation verification over Puget Sound (mm). For the MM5 forecast initialized at 0000 UTC 20 Oct. The best performance in each period is highlighted in bold.

<table>
<thead>
<tr>
<th>Period</th>
<th>ME 36 km</th>
<th>ME 12 km</th>
<th>ME 4 km</th>
<th>MAE 36 km</th>
<th>MAE 12 km</th>
<th>MAE 4 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–21</td>
<td>−1.46</td>
<td>−1.26</td>
<td>−1.09</td>
<td>1.78</td>
<td>1.66</td>
<td>2.01</td>
</tr>
<tr>
<td>18–24</td>
<td>−1.71</td>
<td>−0.64</td>
<td>−0.046</td>
<td>2.07</td>
<td>1.75</td>
<td>2.10</td>
</tr>
<tr>
<td>24–27</td>
<td>−0.35</td>
<td>0.09</td>
<td>−0.63</td>
<td>1.40</td>
<td>0.67</td>
<td>0.90</td>
</tr>
<tr>
<td>24–30</td>
<td>0.442</td>
<td>0.83</td>
<td>−0.28</td>
<td>1.74</td>
<td>0.98</td>
<td>0.70</td>
</tr>
<tr>
<td>30–33</td>
<td>0.80</td>
<td>0.56</td>
<td>0.05</td>
<td>0.80</td>
<td>0.59</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Figure 16 illustrates the effects of a modest timing error on sea level pressure verification scores associated with a trough passage for grid spacings of 36, 12, and 4 km (all domains being the same size). It is assumed that the forecast trough at all resolutions was 40-km farther east than observed, and that the trough was better defined and more intense in the higher-resolution domains. Calculating mean absolute errors at 9 “observation” locations produces values of 4.19, 4.81, and 5.25 mb for the 36-, 12-, and 4-km domains, respectively. Thus, the verification scores declined with resolution even though the 4-km structure was more realistic. Is the more accurately defined 4-km trough a less skillful prediction? As discussed below, the answer to this question depends on the needs of the user. Finally, the effects of timing errors on model verification can be lessened by temporal averaging, as done by Colle et al. (2000) using 24-h precipitation amounts.

Position error is a closely related issue. If the forecast structure of an important mesoscale feature (such as an orographically forced leeside convergence zone or a convective system) is displaced spatially, a higher-resolution simulation will inevitably produce a poorer objective verification, even if smaller grid spacing produces a more realistic structure. Since small directional errors in synoptic-scale winds are not unusual, significant (10–50 km) displacement errors of important orographically forced Northwest mesoscale circulations (e.g., the Puget Sound convergence zone or lee rain shadows) are often observed. If both 12- and 4-km grids have identically positioned (but misplaced) convergence lines, and the 4-km domain has a sharper, stronger, and more realistic structure, is the 4-km forecast really inferior—as suggested by traditional verification approaches?

Another point deals with observational data density. The objective verification scores are highly dependent on the quality and density of the observational network. With insufficient data density, important mesoscale structures can be missed or poorly represented in analyses, resulting in different (and often better) verifications for lower-resolution domains relative to high-resolution grids. Furthermore, a sparse observational network increasingly undermines verification as model resolution is increased. This sensitivity to verifying data was demonstrated in Colle et al. (2000), which showed a dramatic change in precipitation verification when the high-elevation SNOTEL data were not considered. Since even the relatively dense Northwest surface observation network (Fig. 2) is sparse over mountain and foothills locations, many of the smaller-scale structures produced by the 12- and 4-km domains are not observed adequately or at all. Radar verification using the National Weather Service WSR-88D network has not greatly aided model verification in the western United States due to beam blockage by terrain, frequent brightband contamination, and problems with $Z-R$ relationships (Westrick et al. 1999).

A crucial question underlying this discussion deals with the basic goals of mesoscale verification. As noted by several contributors at the 1998 Workshop on Mesoscale Model Verification (Davis and Carr 2000), a
satisfactory approach to mesoscale verification has been elusive. Traditional objective approaches based on verification at fixed observing locations or grid points are greatly influenced by timing and spatial errors, as well as deficiencies of the observing network used for verification (e.g., data scarcity, instrument error). Furthermore, the goals and measures of verification are inevitably dependent on the needs of the user; a satisfactory mesoscale forecast for one might be a complete failure for another. For example, a farmer needing to make a decision about irrigation might not care when precipitation fell, as long as it occurred during the day following planting. On the other hand, there are other applications (e.g., a weather forecast for the opening ceremonies of a new baseball stadium) where exact timing and position are critical. Thus, the value of a high-resolution numerical forecast can vary according to its application, and one verification system cannot satisfy the needs of all users.

It is clear that additional approaches should join the current verification toolbox. For example, one could verify the maximum wind or 1-h rainfall predicted by a forecast model over the subsequent day at a location. Time-averaged or spatially averaged parameters as well as model variability could be evaluated. Temporal or spatial shifting of model fields could be used to verify model structures. If suitable objective verification approaches can be devised, it may be possible to demonstrate increased value of high-resolution NWP.

High-resolution mesoscale model forecasts can increase the skill of human forecasters even if phase, position, or other errors mar the scores of objective verification approaches. Mesoscale model output can serve as a superb educational tool that can acquaint human forecasters with the complex evolution of three-dimensional mesoscale structures and how such structures are modulated or controlled by the synoptic-scale flow. It allows forecasters to develop conceptual models of regional circulations that can be applied even when the synoptic-scale models, and embedded mesoscale models, are problematic. High-resolution model output can also provide a “heads up” regarding the possible mesoscale phenomena on any given day. For example, the model might not be able to predict the exact position or intensity of a squall line, but might appraise the forecaster to the possibility that such a feature might be expected.

To illustrate the potential benefit of high-resolution mesoscale predictions to forecasters, Fig. 17 (provided by Lieutenant Commander Mike Angove, formerly of the U.S. Navy Detachment at Whidbey Island Naval Air Station) shows wind speed, wind direction, and ceiling errors at Whidbey Island Naval Air Station before and after they gained operational access to the UW real-time MM5 forecasts (late winter 1997). One notes that the errors during the winter of 1997/98 appear to be reduced from those of the previous winter, when only synoptic-scale models (grid spacing of 80 km) were available. The availability of the mesoscale model output proved to be a valuable educational tool for Whidbey forecasters; for example, it revealed to them how the Olympic Mountain rain shadow rotates with the synoptic-scale flow and the crucial importance of leeside troughing for their wind forecasts.

The results of this study and others mentioned above have obvious implications for the future of numerical weather prediction. First, there appears to be a point of diminishing returns for forecast skill when traditional verification methods are applied. In the northwestern United States this point seems to be around 10–15 km, with coarser resolution being unable to properly resolve crucial mesoscale features produced by terrain and surface contrasts. However, in some areas of terrain (e.g., the narrow Columbia River Gorge), ultra-high resolution (0.5–1-km grid spacing) may well be necessary to forecast essential mesoscale features. Over the eastern half of the United States where topographic relief is less dramatic, the horizontal resolution of diminishing returns appears to be considerably larger (20–40 km). These observations suggest that a “one size fits all” approach, in which all regions of a country are run with the same resolution, may not be an efficient use of computer resources. Second, the demonstrable benefits of increasing resolution may vary spatially.

![Fig. 17. Terminal Aerodrome Forecast (TAF) error trends for selected parameters. Here P1 represents the initial MM5 familiarization period, and P2 represents the period in which MM5 was used operationally.](image-url)
and temporally. For example, the value of high resolution might be more obvious near substantial mesoscale terrain, in areas with a dense observation network, or when the synoptic-scale flow is either well forecast or slowly varying. Third, the value of resolution may be underestimated when normal verification procedures (e.g., rms, biases, threat scores) are applied at fixed grid boxes or observation locations. For example, although the structures of important mesoscale features often become more realistic (stronger, better defined) as resolution increases, objective verification scores are profoundly degraded by even small timing and spatial errors. Nontraditional means of verification (e.g., verifying mesoscale structures) may be needed to demonstrate the benefits of high horizontal resolution. Finally, high-resolution output can be a valuable educational tool for improving conceptual models and the predictions of human forecasters.

As horizontal resolution approaches the point of diminishing returns, increasing emphasis should be given to evaluating short-term ensemble predictions in which a collection of forecasts are made by varying model initializations and physics. Although some positive results have been noted (Stensrud et al. 1999; Hou et al. 2001; Grimit and Mass 2002), the value of short-term mesoscale ensembles is still unproven. The optimal approach to high-resolution numerical prediction may well be a hybrid: a collection of ensemble runs providing probabilistic information about larger-scale evolution and forecast reliability, and a limited number of high-resolution runs being forced by the most skillful ensemble members or some type of weighted ensemble mean. Short-term ensembles may still demand relatively high resolution, since some features (e.g., topographic circulations) may require small grid spacing to realistically simulate crucial structures.

ACKNOWLEDGMENTS. This research was supported by grants from the Office of Naval Research (Grant N00014-98-1-0193) and NSF/USWRP (Grant ATM-9612876). The paper was greatly improved by the suggestions of Kelvin Droegemeier, Jim Steenburgh, and two anonymous reviewers. Support for the computational resources required for this project were provided by the Northwest Mesoscale Modeling Consortium, a group of federal, state, and local agencies interested in high-resolution numerical weather prediction over the Pacific Northwest. Use of the MM5 was made possible by the Microscale and Mesoscale Meteorology Division of the National Center for Atmospheric Research (NSF), which is supported by the National Science Foundation.

REFERENCES


