9 Verification methods

9.1 Background

9.1.1 What is verification?

Forecast verification involves evaluating the quality of forecasts. Various methods exist to accomplish this. In all cases, the process entails comparing model-predicted variables with observations of those variables. The term validation is sometimes used instead of verification, but the intended meaning is the same. That said, the root word "valid" may imply to some that a forecast can either be valid, or invalid, whereas obviously there is a continuous scale that measures forecast quality. Thus, the term verification is preferable to many, and will be employed here. Special verification measures that are most applicable to ensemble predictions have been discussed in Chapter 7. There is an extensive body of literature on the subject of model verification, and students and researchers should read beyond the summary material in this chapter to ensure that they understand underlying statistical concepts and that they use the verification metrics that are most appropriate for their needs.

9.1.2 Reasons for verifying model simulations and forecasts

There are multiple motivations for evaluating the quality of model forecasts or simulations.

- Most models are under continuous development, and the only way modelers can know if routine system changes, upgrades, or bug fixes improve the forecast or simulation quality is to objectively and quantitatively calculate error statistics.
- For physical-process studies, where the model is used as a surrogate for the real atmosphere, the model solution must be objectively verified using observations, and if the observations and model solution correspond well where the observations are available, there is some confidence that one can believe the model where there are no observations. This is a necessary step in most physical-process studies.
- When a model is being set up for a research study or for operational forecasting, decisions must be made about choices for physical-process parameterizations, vertical and horizontal resolutions, LBC placement, etc. Objective verification statistics are employed for defining the best configuration.
- Forecasters learn, through using model products over a period of time, about the relative performance of the model for various seasons and meteorological situations. This

process can be made easier through the calculation of weather-regime-dependent and season-dependent verification statistics for the model.
- Objective decision-support systems, which utilize model forecasts as input, can benefit from information about the expected accuracy of the meteorological input data.
- Model-intercomparison projects, which compare model accuracy and skill in order to better understand the strengths and weaknesses of the participating models, are based on a foundation of objective model verification.

9.1.3 Some terminology related to forecast performance

It is useful to define some basic terminology that we will be employing. Further discussion of the following definitions can be found in Wilks (2006) and other general references on the subject.

- *Accuracy* – A measure of the average degree to which pairs of forecast values and observed values correspond. Scalar measures of accuracy summarize the overall quality of the forecasts in the form of a single number.
- *Bias* – A measure of the correspondence between the average of a forecast variable and the average of the observations.
- *Skill* – The accuracy of a forecast relative to a reference forecast.
- *Reference forecast* – This is an easily available, non-model-based data set that can be interpreted as a simple, minimal-skill forecast. See Section 9.3 for additional discussion of such reference forecasts.

9.2 Some standard metrics used for model verification

9.2.1 Accuracy measures for continuous variables

These measures apply to variables that are continuous in the sense that they can take on any value within a physically realistic range. For example, if temperature itself is the predictand, it represents a continuous variable. But, if the predictand is a binary "yes or no", regarding whether the temperature tomorrow will exceeds some threshold, it is a discrete variable (discussed in Section 9.2.2).

The MAE is the arithmetic average of the absolute difference between pairs of forecast and observed quantities. It is the average magnitude of the forecast error, and is defined as

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|x_k - o_k\right|,$$

where (x_k, o_k) is the k-th of n pairs of forecasts and observations. In order for the MAE to be zero, the difference between each forecast and observation pair must equal zero. Another scalar accuracy measure for continuous variables is the Mean-Square Error

(MSE), which is the average squared difference between the forecast and observation pairs. It is defined as

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(x_k - o_k)^2 .$$

Because the errors are squared, the MSE will be more sensitive to large errors than will the MAE. Sometimes, the square root of the MSE is used, such that $RMSE = \sqrt{MSE}$. This has the same physical dimensions as the forecast and observations. The above metrics represent both systematic and random components to the error.

An additional, commonly used measure of correspondence between observations and forecasts is the Anomaly Correlation (AC). As the name implies, it is designed to define similarities in the patterns of the departures (i.e., anomalies) of the observed and forecast variables from the climatological means. The AC can be calculated based on time series or spatial fields, and is designed to reward for good forecasts of the pattern (phase and amplitude) of the observed variable. See Wilks *et al.* (2006) for additional detail.

The bias is the same as the Mean Error (ME), such that

$$ME = Bias = \frac{1}{n}\sum_{k=1}^{n}(x_k - o_k) = \bar{x} - \bar{o} .$$

This is also known as the systematic error. Given that $\bar{o}$ is a simple way of defining the climatology of the variable (at least for the limited period of the verification), and $\bar{x}$ is the model climatology for the variable, the bias represents a comparison of the model and actual climatological values.

9.2.2 Accuracy measures for discrete variables

These measures apply when the verification question is defined in terms of a yes–no condition. For example, consider a precipitation forecast of whether the accumulated amount is above a specified threshold at a particular location. The observation at that point defines whether precipitation of that amount indeed occurred, a yes or no condition, and the forecast is also in the form of a yes or no. This problem can be illustrated with a 2 × 2 contingency table of the form shown in Fig. 9.1a. Of the n forecast–observation pairs, a represents the number of times that an observed event was correctly forecast (called hits), b is the number of times that no event occurred but the forecast was for an occurrence (called false alarms), c is the number of times that an observed event is forecast to not occur (called misses), and d is the number of times that an event was correctly forecast to not occur (called a correct negative). An example is shown in Fig. 9.1b of areas where a condition (e.g., 24-h accumulated precipitation above a threshold) is observed and forecast. Each area is defined in terms of the elements of the contingency table.

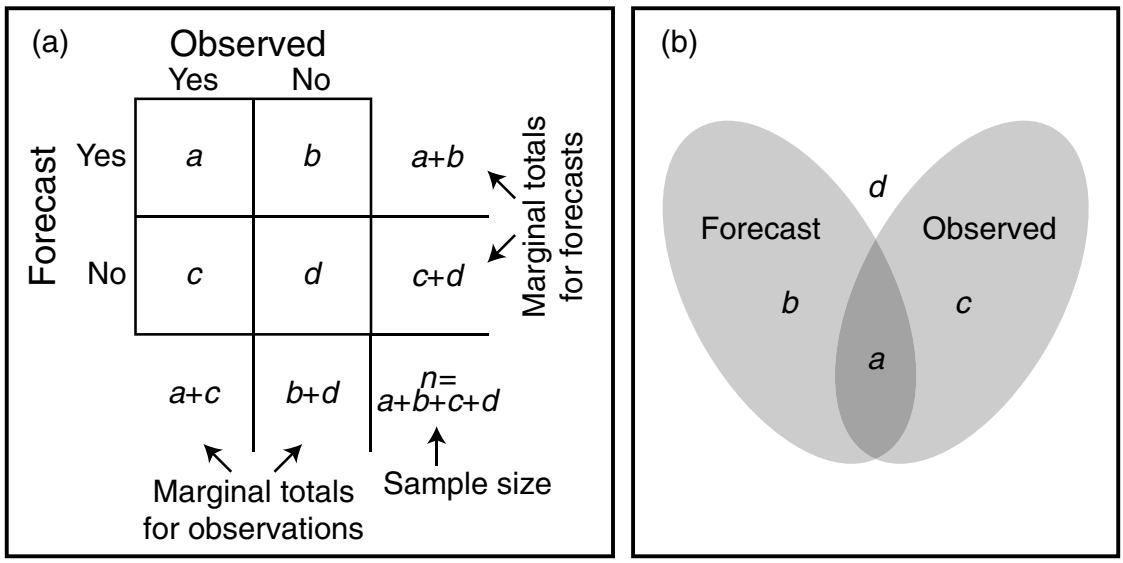

Fig. 9.1 Contingency table showing the four possible outcomes of a forecast of a discrete variable (a). Also shown is a schematic example of the observed and predicted areas where a variable (e.g., accumulated precipitation) exceeds a specific threshold.

Many forecast-accuracy measures are based on the components of the contingency table. The Proportion Correct is defined as

$$PC = \frac{a+d}{n}, \tag{9.1}$$

and represents the fraction of the forecasts that correctly anticipated the event or nonevent. A potential disadvantage of this score for some situations is that equal credit is given for correct positive or negative forecasts. If the forecast variable is the existence of sun in Cairo in summer, the correct forecast of sun is given equal credit as the more difficult correct prediction of the rare obscuring cloud. An alternative is the Threat Score (TS), which is useful when the yes-event to be forecast occurs much less frequently than the no-event. This is also termed the Critical Success Index, and is expressed as

$$TS = CSI = \frac{a}{a+b+c}. \tag{9.2}$$

The bias compares the average forecast and the average observation, and is defined as

$$B = \frac{a+b}{a+c}.$$

The False-Alarm Ratio,

$$FAR = \frac{b}{a+b},$$

is the fraction of yes forecasts that are wrong, and is different from the false-alarm rate

$$F = \frac{b}{b+d}$$

which is the ratio of false alarms to the total number of nonoccurrences of the event. The hit rate, which is also called the probability of detection, is defined as

$$H = POD = \frac{a}{a+c},$$

and represents the fraction of the event occurrences that were forecast.

9.2.3 Skill scores

As noted earlier, skill is defined as the accuracy of one forecast method relative to that of a reference forecast. The skill is usually represented as a Skill Score (SS), which is defined as a percentage improvement over the reference forecast. Mathematically, a SS can be defined as

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\,\%, \tag{9.3}$$

where A is the accuracy of a forecast, A_{ref} is the accuracy of a reference forecast, and A_{perf} is the accuracy of a perfect forecast. If $A = A_{perf}$, the skill score is 100%. If $A = A_{ref}$, the skill is zero, indicating no improvement relative to the reference forecast. If the forecast accuracy is less than that of the reference forecast, the skill score is negative.

A number of skill scores are based on the previously described 2 × 2 verification contingency table, and have the form of Eq. 9.3. One of the most-frequently used is called the Heidke Skill Score (HSS), and is based on the proportion correct (Eq. 9.1) as the accuracy measure (A, in Eq. 9.3). The reference accuracy measure, A_{ref}, is the proportion-correct value that would be obtained by random forecasts that are statistically independent of the observations. The expression for the HSS is

$$HSS = \frac{2(ad - bc)}{(a+c)(c+d) + (a+b)(b+d)},$$

where the derivation is described in Wilks (2006) and elsewhere. Analogously, the TS (Eq. 9.2) can be used as the basic accuracy measure, and the TS for random forecasts is used as the reference. This is called the Gilbert Skill Score (GSS) or the Equitable Threat Score (ETS), and is derived in Wilks (2006) as

$$GSS = ETS = \frac{a - a_{ref}}{a - a_{ref} + b + c}, \text{ where}$$

$$a_{ref} = \frac{(a+b)(a+c)}{n}.$$

Skill scores are also computed for continuous variables, using MAE, MSE, or RMSE, again based on Eq. 9.3. Climatology or persistence are generally used for the reference forecast. Using the MSE as an example, the accuracies for these references are

$$MSE_{Clim} = \frac{1}{n}\sum_{k=1}^{n}(\bar{o} - o_k)^2 \text{ and}$$

$$MSE_{Pers} = \frac{1}{n}\sum_{k=1}^{n}(o_{k-1} - o_k)^2,$$

where o_k is the observation, $\bar{o}$ is the climatological mean of the observed variable, and o_{k-1} is the previous value of the variable. Similar equations apply for MAE. For either metric, and for either reference forecast, the skill score, based on Eq. 9.3, can be written as

$$SS = \frac{MSE - MSE_{ref}}{0 - MSE_{ref}} = 1 - \frac{MSE}{MSE_{ref}}.$$

Many more skill scores, with various strengths and weaknesses, are described in Wilks (2006) and Gilleland *et al.* (2009).

9.3 More about reference forecasts and their use

The reference forecast defines a minimal-accuracy or zero-accuracy point on the scale – essentially a zero point on the accuracy scale. This is the forecast accuracy that can be achieved without running a model. Example reference forecasts include (1) a persistence forecast where it is assumed that present conditions prevail throughout the forecast period, (2) a diurnal-persistence forecast where it is assumed that the previous day's diurnal cycle is replicated, (3) a forecast based on seasonal climatological-average values of the forecast variables, and (4) the use of random forecasts. The first three approaches are self-explanatory. For the random forecast, the bootstrap technique of Efron and Tibshirani (1993) is an example of a method that can be used. Here, the available observations throughout the study period are repeatedly and randomly resampled (with replacement), to yield multiple synthetic samples (hundreds to thousands) of the same size as the set of observations that are used normally in the verification. These are the random forecasts, which are 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 data set. Each of the random forecasts is compared with the observations at each verification time, and the average verification score at each time is then used to define the error.

The accuracy of the random no-skill forecast or one of the other reference forecasts, and the estimates described later in Section 9.5.2 of the maximum (or perfect-model) accuracy

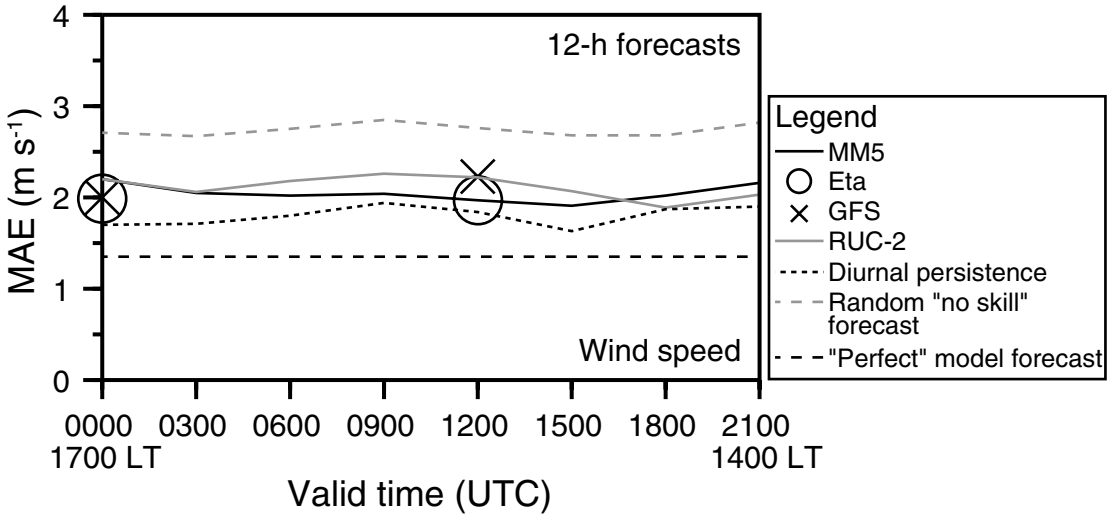

Fig. 9.2 The 12-h forecast MAE for 10-m AGL wind speed for the MM5, Eta, GFS, and RUC-2 models during the 3 February to 30 April 2002 period. Model forecasts from MM5 and RUC-2 were initiated every 3 h. Also displayed are the corresponding statistics from the diurnal persistence, random no-skill, and perfect-model forecasts. Adapted from Rife *et al.* (2004).

that is related to the existence of instrument and representativeness error, effectively produce a range within which falls the actual model-forecast accuracy. Figure 9.2 shows an example of the verification of 12-h wind-speed forecasts from four models, where the bounding perfect-model and no-skill curves are shown. Wind-speed MAEs from all models cluster around 2 m s^{-1}, within the range of about 1.5 m s^{-1} between the two bounding curves. The verification statistics for the Eta, GFS, and RUC-2 models were calculated for the mesoscale area spanned by the MM5 LAM.

9.4 Truth data sets: observations versus analyses of observations

Verification of model solutions can be performed using either observations or analyses of observations. For the latter approach, operational analyses can be used for near-real-time verification of forecasts, or reanalyses can be used for verification of retrospective simulations. See Chapter 6 for a review of how operational analyses are produced, and Chapter 16 for a description of reanalyses. Using analyses for verification is advantageous in situations where conventional *in-situ* data are sparse, either temporally or spatially. In such situations, variational assimilation of satellite data in the analysis process can constrain the analysis and compensate to some degree for the paucity of *in-situ* observations. One problem with the use of analyses is that they are model generated, so one model is being verified with the products from another model. This is especially troublesome when verifying a variable such as precipitation, where the model that created the analysis often has not assimilated any precipitation observations – that is, the analyzed precipitation is

entirely a creation of the model. Another issue is that only global-scale operational analyses or reanalyses are available for much of the world, so there is a clear scale mismatch if forecasts from high-resolution mesoscale models must be verified. The fine-scale details in the solution from the mesoscale model will not have counterparts in the analysis, and measures of accuracy will interpret the differences as errors. Thus, the mesoscale model will be penalized for successfully serving its intended purpose of providing information on the mesoscale. For verification using analyses, the forecast values can be interpolated to the grid points of the analysis, or vice versa. Because of the short distance between grid points, simple bilinear interpolation is often used.

In regions where observations are sufficiently dense, whether they be *in situ* or remotely sensed, it can be advisable to interpolate model forecasts or simulations to the observation points, and calculate the statistics there. *In-situ* observations include those from radiosondes, near-surface mesonetwork stations, aircraft-borne sensors, and rain and snow gauges. Remotely sensed data that can be compared directly with model output may be from wind profilers (Doppler radars that point approximately in the vertical), radial winds from scanning Doppler radars and lidars, satellite cloud-track winds, satellite water-vapor-track winds, and satellite-estimated precipitation. Interpolation of model values from the grid to observation locations can be done through simple linear interpolation. Near-surface wind observations are often at 10 m AGL, although this height can vary, and corresponding temperature and humidity observations are generally at 2 m AGL. When the lowest model computational level is above the elevation of the observations, Monin–Obukhov similarity theory can be used to extrapolate the wind, temperature, and humidity predictions to the height of the observations (Stull 1988).

9.5 Special considerations

9.5.1 Orographic smoothing

Verification of a model solution is complicated by the fact that model orography is smooth compared to the actual orography. Unless envelope orography is used with a model, valleys are filled in and mountain ridges are lowered. Thus, a surface-based observation has a different elevation above sea level in the model compared to reality. As a practical example, assume that a model is used to forecast the winds at the 80-m AGL hub height of wind turbines that are located on a ridge near a coastline (see Fig. 9.3). And, assume that winds observed by anemometers at the hub height are available for model verification. The actual ridge is 100 m ASL, but the smoothness of the model terrain causes the ridge in the model to be only 20 m ASL. This raises the question about which model winds should be compared with the 180-m ASL observations. Based on the distance of the observation above sea level, 180-m ASL model winds (level 1) should be used. But, if distance above the local land surface is more physically relevant, the model winds at 100 m ASL (level 2) should be used. Note that this problem is not limited to wind prediction, and applies to temperature and humidity as well. In this example, the question takes on greater

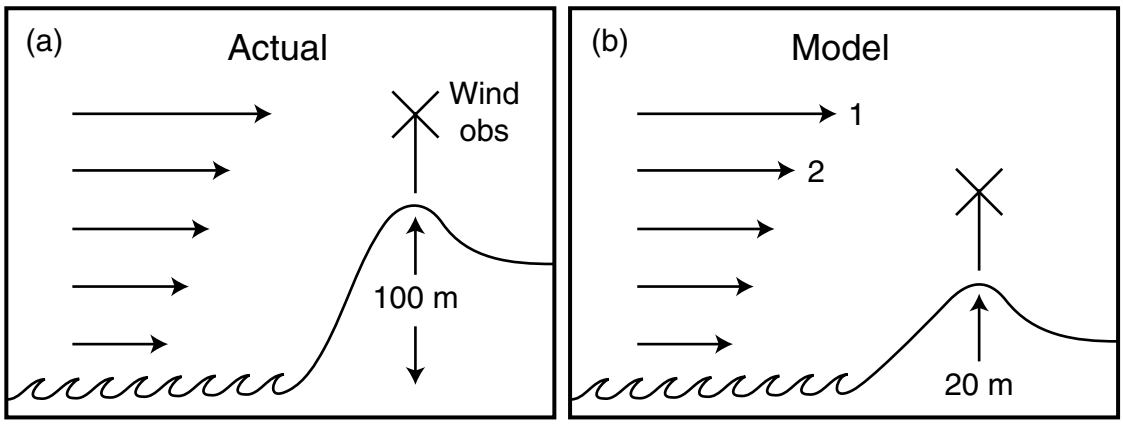

Fig. 9.3 Schematic showing a cross section of the orography near a coastline, based on the actual elevation (a) and that associated with the smoothed orography of a model (b). Wind observations are available at the top of an 80-m tall wind-turbine tower, and the issue is whether to compare the observation with model winds at the height ASL in the model (level 2) or the height ASL in nature (level 1).

importance when the prevailing large-scale wind is from the water. There is thus sometimes the question whether an observation should be assumed to apply at the correct location in the vertical (ASL) or at the correct distance above the model land surface. There is no easy answer to this question, but the modeler should be aware of the paradox because the choice will certainly influence verification statistics.

The above example might be interpreted as being somewhat isolated; however, in complex orography a systematic error will generally exist in forecasts of variables that have a significant climatological variation with height. For example, model forecasts will have a warm bias relative to observations located over mountains, where the surface elevation in the model is lower than in reality as a result of orographic smoothing. Similarly, wind speeds will be underforecast at higher elevations when the orography is smoothed.

9.5.2 The imperfect nature of verifying observations

It is well recognized that the errors produced by the model and the errors in the observations both contribute to the total error diagnosed in the verification process. The observations used for verification contain error associated with the accuracy of the instruments and the calibration error. In addition, there is another, somewhat less well-documented, cause of differences between the model solution and the observations that impacts the verification statistics. And, this error 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 instantaneous or time-averaged measurements at a point, whereas the model-predicted quantities represent spatial averages over

each model grid-box volume. In addition, the numerics and explicit diffusion smooth the grid-box information so that it, in fact, represents an even-larger area (see Fig. 3.36). Representativeness error can be understood through the following idealized example. Suppose there exists a perfectly known near-surface wind field over a 1-km^2 area. The wind at the center of this area is used 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 both the model-average and observation exactly characterize the wind field in their own way, the difference between the two will obviously not be zero. 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 fine-scale atmospheric structures, and the geographic extent of the sampling area (or size of the model grid box).

It is worthwhile estimating this error because it contributes to the maximum model accuracy (see definition above) that is practically achievable, given the properties of the forecasting systems and the verifying observations. A tractable approach is to use an extremely high-resolution model to define the variability within a larger grid-box area. For example, Rife *et al.* (2004) used the model described by Clark and Farley (1984) and Clark and Hall (1991, 1996) to estimate the representativeness error in a verification of MM5 mesoscale-model wind simulations in complex terrain. The Clark–Hall model was run for a real-data case over the complex terrain near Pinewood Springs, Colorado, where the highest resolution grid in a nest had an increment of approximately 50 m and encompassed a nearly 36-km^2 area. To estimate the representativeness error, the spatially averaged wind speed and direction were computed from the Clark–Hall model output within a stencil having the dimensions of a 1.33-km grid box of the MM5 model. There were 676 Clark–Hall model grid points for each MM5 grid box. Next, the point values of the speed and direction were determined at the stencil center. This process was repeated until the entire Clark–Hall model domain had 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) was computed to produce an estimate of the representativeness error.

Based on the above analysis, the representativeness errors for 10-m AGL wind speed and direction, under well-mixed boundary-layer conditions with this MM5 model resolution in complex terrain, are 1.15 m s^{-1} and 14.6°, respectively. This estimate is conservative because the Clark–Hall model with a 50-m grid increment underestimated the true amount of spatial variability that would exist in the near-surface wind field. Conventional cup and vane anemometers are generally accurate to within ±0.3 m s^{-1} and ±3° for wind speed and direction, respectively. 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 that the errors are additive).

The existence of representativeness error can extend beyond the influences of complex orography. For example, observations are made sufficiently far from natural obstructions, or those of human origin, such that the measured value of a variable is presumably not influenced by the obstacle. However, model grid-box-average values of surface properties, such as roughness length, are defined based on the average character of the surface over

the grid-box area. Because grid boxes often contain obstructions, the average roughness length is defined accordingly. Thus, the observed wind experiences a different roughness than does the model wind, and this representativeness problem can lead to a difference between observed and modeled winds that has nothing to do with model accuracy. Strassberg *et al.* (2008) calculate that this effect can lead to small but nontrivial artificial errors in the wind verification. See the problem at the end of the chapter regarding how similar landscape representativeness problems can lead to errors in the verification of near-surface temperature or humidity.

The results of the above analyses of representativeness of course apply only to a specific configuration of the forecast model, the spatial structure of the meteorological field being sampled, and the error properties of the observing system. A separate analysis would need to be performed for other situations. Nevertheless, the example illustrates the potential importance of these factors to the verification process. Note that this sum of the representativeness error and the instrument error can be viewed as an upper bound on the forecast accuracy. That is, this is the accuracy of a perfect model.

9.5.3 Special issues related to the verification of winds

There are a few special issues that should be kept in mind with respect to the verification of winds. One is related to the fact that, unlike the other variables discussed here, wind is a vector quantity. We thus have the option of comparing the observed and forecast wind in terms of (1) separate statistics for the u and v components, (2) separate statistics for the speed and direction, and (3) vector differences. On the synoptic scale in midlatitudes, or with verification of upper-air winds, individual verification of u and v can make physical sense if the components align with the zonal and meridional direction. That is, the components represent the direction of the mean wind, and the perturbations to the mean wind. Wind speed and direction are intuitive metrics because they are geometric attributes of the vector, and the two types of error can be easily visualized. Alternatively, the vector difference between the observed and forecast winds is a way of representing the error.

If wind speed and direction are used for verification, account should be taken of the fact that low mean-wind speeds are associated with highly variable directions because turbulence will dominate the measurement. Given that we wish to verify the mean (nonturbulent) wind direction, it is common practice to not include in the verification, wind directions that are associated with speeds of less than some threshold, such as 0.5 m s^{-1}. Also, direction-error calculations are complicated by the fact that the direction scale is periodic, and this needs to be remembered when differencing the modeled and observed values.

A point that was made in Chapter 3 is that models with Cartesian grids, which are defined on map projections, have u and v wind components defined in terms of the rows and columns of grid points. That is, the u component is parallel to the rows and the v component is parallel to the columns. Thus, the model-defined wind components at a particular latitude–longitude are not the same as those reported in an observation at the same point. The latter components, of course, are parallel to the local latitude and longitude lines. Thus, just as

with the model initialization process, the model and observed wind components need to be reconciled. In this case, if verification statistics are being computed at observation locations, the model wind components are transformed to the traditional geocentric components.

9.5.4 Response of the standard accuracy metrics to horizontal resolution

Conventional objective measures of forecast skill sometimes seem to show little improvement from increased horizontal resolution, in spite of the fact that a subjective assessment of model accuracy shows a definite positive impact. 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 skill as the horizontal grid spacing was decreased from 36 to 12 km. In contrast, there were only small improvements in the objective skill as the grid spacing was decreased from 12 to 4 km. However, in terms of subjective comparisons of observed and forecast structures, the coarser-resolution forecasts were often profoundly inferior to those from the highest-resolution grid. Similarly, Davis *et al.* (1999) showed that, in terms of conventional skill scores such as bias, MAE, and RMSE, a high-resolution (1.11-km grid increment) mesoscale model that was run operationally over the mountainous western USA provided only slightly better surface temperature forecasts than did the much coarser 80-km Eta model, with the two models exhibiting 10-m wind-field forecast errors of comparable magnitude. However, only the mesoscale model was able to accurately depict some important aspects of the observed locally forced circulations resulting from the regional orography and variations in other land-surface characteristics. Another study, for east-central Florida, compared objective skill 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. However, a detailed subjective analysis indicated that the mesoscale model exhibited considerably more skill in predicting the observed Florida sea breeze, which strongly determines temperature and the timing and location of thunderstorm initiation.

One well-known characteristic of some standard verification metrics is that they can reward smooth solutions. That is, if output from a high-resolution model is progressively smoothed, the accuracy metrics calculated from the output may show progressively greater skill. Thus, the use of higher horizontal resolution, or the use of numerical methods that have small truncation error, can lead to poorer model verification. This is in spite of the fact that these model properties should lead to a more-realistic representation of fine-scale structures in the model solution. Figure 9.4 illustrates the common situation where model forecasts can have both phase and amplitude errors. The solid line represents the observed wind speed in a jet that is oriented perpendicular to the page. The dashed line shows a forecast from a high-horizontal-resolution model, where the correct amplitude of the jet is retained, but the maximum is displaced to one side. The dot-dash line and the dotted line show solutions from models that have less horizontal resolution, and therefore produce a smoother solution. The RMSE between the observed and forecast wind speed is greater for the model solution that better retains the correct amplitude of the feature.

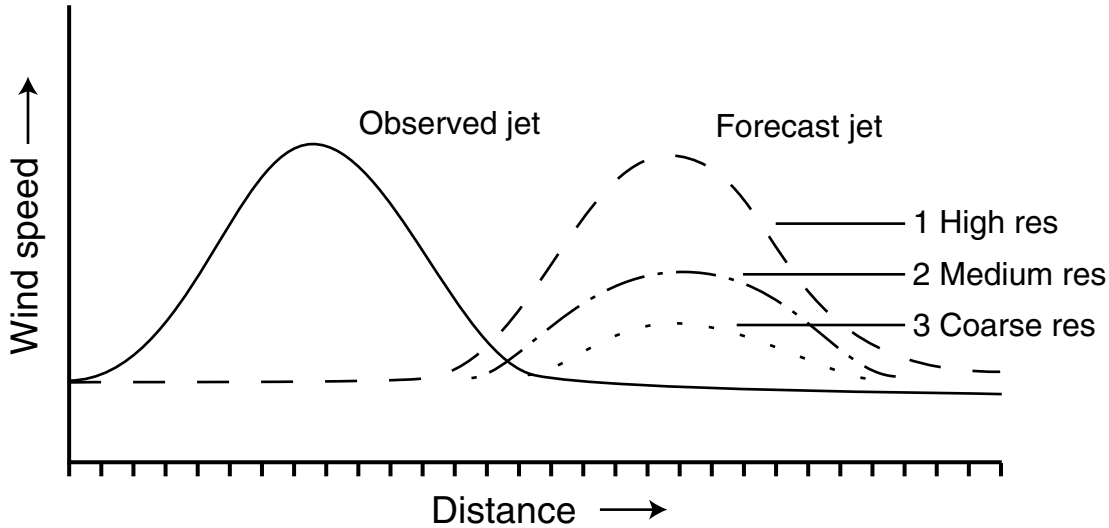

Fig. 9.4 Illustration of how a smooth forecast can lead to better verification statistics than a forecast with greater amplitudes in structural features. See text for details.

Another way to look at this influence of fine-scale spatial structures on the standard verification metrics is in the context of the decomposition of the definition for MSE. Murphy (1988) demonstrates that

$$MSE = (\bar{x} - \bar{o})^2 + \sigma_x^2 + \sigma_o^2 - 2\sigma_x\sigma_o r_{xo}\,.$$
$$\text{(1)} \qquad \text{(2)} \quad \text{(3)} \qquad \text{(4)}$$

Term (1) is the mean error, or bias; term (2) is the variance of the forecasts; term (3) is the variance of the observations; and term (4) contains r_{xo}, the coefficient of correlation between the forecasts and observations. Thus, all other things being equal, high-resolution forecasts or verification fields (observations) with larger variance, will lead to larger MSEs.

9.6 Verification in terms of probability distribution functions

Because the extremes in weather (temperature, precipitation) are often the most important situations that must be forecast with models, it is useful to use verification methods that provide specific information about how model accuracy varies for different values of the predictand. A simple approach would be to simply plot the frequency distributions of the observed and forecast variables for a point, based on a long series of forecasts. This will provide information about how well the model-forecast climatology verifies relative to the actual climatology for extreme values of the variable, which is important, but it does not quantify how accurately the extremes are forecast. To accomplish the latter, the joint distribution of the observed and forecast values of a variable can be plotted. For example, Fig. 9.5a

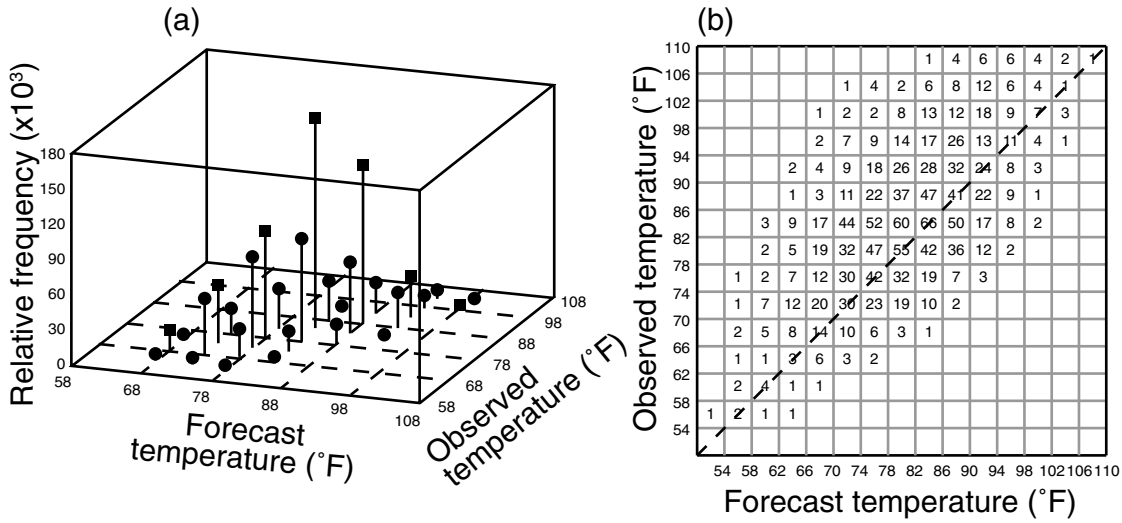

Fig. 9.5 Examples of joint distributions of observed and forecast 2-m AGL temperatures. A bivariate histogram shows the joint distribution of observed temperatures and 24-h model forecasts of temperature, for one location in the warm season (a). Another type of display is also shown (b), based on a different data set, where the numbers plotted are the scaled frequencies of occurrence. Panel (a) is adapted from Murphy *et al.* (1989).

contains a bivariate histogram that shows the distribution of forecast temperatures for different values of the observed temperature. Similar information is provided in the simpler display in Fig. 9.5b, for a different set of forecasts. Alternatively, scatter plots showing observed and forecast values could be used to reveal errors in different parts of the distribution. No matter how the information is displayed, it can be used to better understand how well a model predicts the extreme values in the PDF, so that the modeling system can be improved as needed, and so that forecasters can develop a better knowledge of model strengths and weaknesses relative to forecasts of weather extremes.

9.7 Verification stratified by weather regime, time of day, and season

Model-verification statistics can be stratified by time of day, season, forecast duration, and weather regime. The resulting situation-dependent model-performance statistics can reveal information that is useful for isolating model shortcomings. For example, separating the precipitation-prediction skill by season provides insight into the model calculations of convective versus stable precipitation. Calculation of time-of-day-dependent statistics for near-surface variables can distinguish errors in the boundary-layer parameterization that manifest themselves differently during daytime, nocturnal, and day–night-transition regimes. Figure 9.6 shows examples of the verification of a mesoscale model by season and time of day. In Fig. 9.6a is the 2-m AGL temperature RMSE for forecasts of 24-h duration initialized at 3-h intervals, aggregated for all seasons. The location is the southwestern USA.

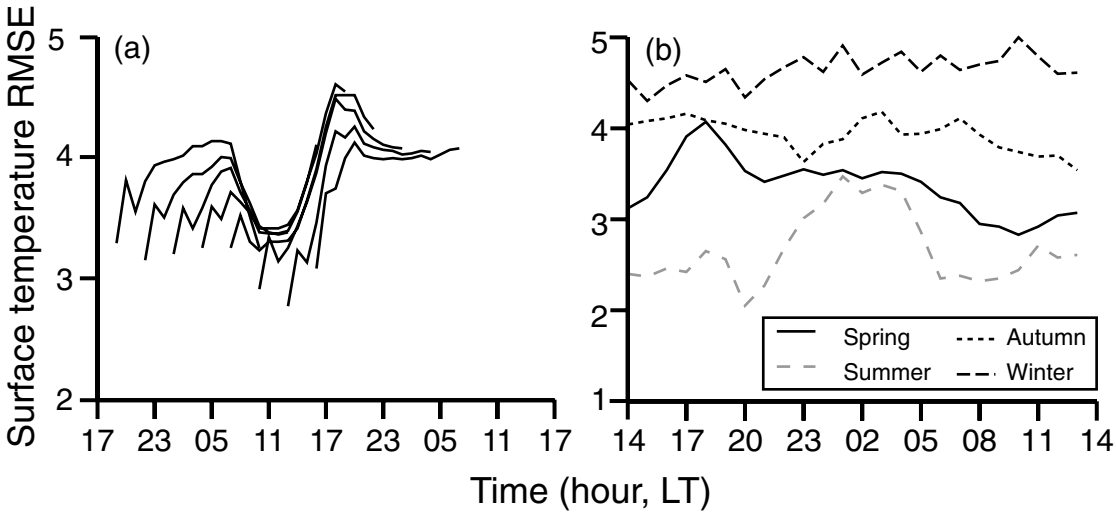

Fig. 9.6 Examples of the verification of a mesoscale model by season and time-of-day (local). For a region in the southwestern USA, the 2-m AGL temperature RMSE is shown for forecasts of 24-h duration, initialized at 3-h intervals, and aggregated for all seasons (a). Also, for Alaska, the temperature RMSE is shown for the 10–12-h segment of mesoscale forecasts, calculated separately for each season and for different times of the day (b). Adapted from Liu *et al.* (2008b).

Regardless of when the forecast is initialized, the RMSE is a minimum in the late morning and early afternoon, and increases rapidly during the late afternoon and evening. This is potentially valuable evidence for identifying aspects of the boundary-layer parameterization that need improvement. In a second example (Fig. 9.6b), the temperature RMSE for the 10–12-h lead time of forecasts is shown separately for each season and for different times of the day. There is a clear trend for the errors to be greater in the winter. There is no diurnal variation in the cold-season RMSE because the forecasts are for central Alaska. But, the warm-season forecasts display an obvious error maximum during the nighttime hours.

Calculation of verification statistics for different weather regimes can also be revealing of model strengths and weaknesses. The process involves using some method to classify different weather regimes that prevailed during a large number of model forecasts. On global scales, the regimes might be extremes in a global cycle such as ENSO. In this case, forecast skill during the El Niño phase could be compared with the skill during the La Niña phase. On regional scales, cluster analysis methods (Wilks 2006) or the methods of self-organizing maps (Cassano *et al.* 2006, Seefeldt and Cassano 2008) can be used to aggregate weather patterns into different regimes, and the forecast error statistics can be calculated separately for each of the regimes. Or, more manual and subjective methods can be used to identify weather regimes, before the separate statistics are calculated. As an example of this concept, Fig. 9.7 shows the 2-m AGL temperature bias, computed separately for two large-scale weather regimes, based on mesoscale-model (MM5) summer-season forecasts for Greece, in the vicinity of Athens. There are two dominant wind regimes in this season. When the northerly Etesian winds are strong, they sweep across the peninsula and inhibit the development of a sea breeze. When the Etesian flow is weak, the

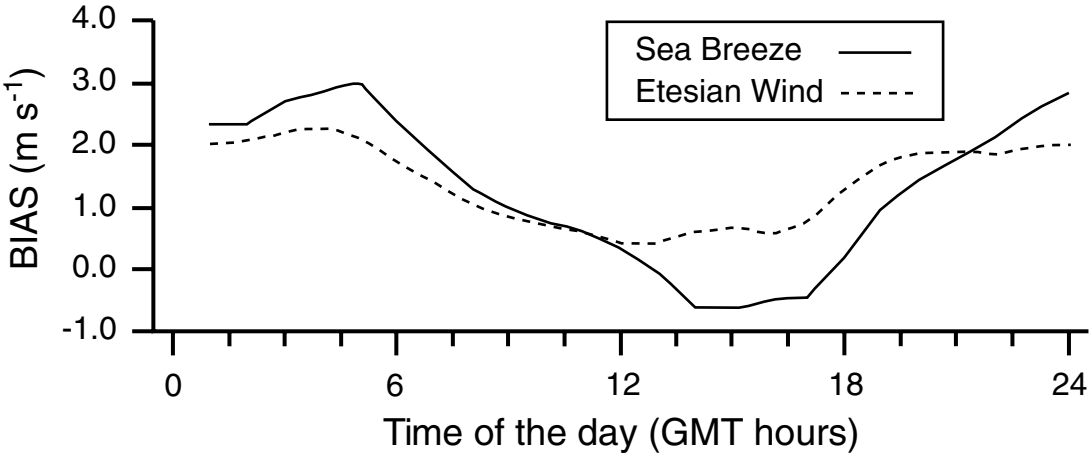

Fig. 9.7 The 2-m AGL temperature bias, computed separately for two large-scale weather regimes in the vicinity of Athens, Greece, based on three months of mesoscale-model summer-season simulations. One regime involves strong northerly Etesian winds and the other pertains to situations where the sea breeze dominates during weak Etesian forcing. Provided by Andrea Hahmann, Risø.

sea breezes from the north and south can dominate, penetrating to the center of the peninsula. Even though the daily-mean temperature bias is not much different for the two regimes, the diurnal amplitude of the bias is over twice as large when the sea breeze dominates, likely indicating a significant contribution by model errors in some aspect of the simulated land–atmosphere interaction.

9.8 Feature-based, event-based, or object-based verification

The most useful information in weather forecasts is often related to changes or events, such as abrupt shifts in temperature or wind speed associated with frontal passages. Thus, model forecast verification can be especially meaningful if it is performed in terms of how well events are forecast. The terms objects, features, or events are used interchangeably in the literature.

An application of this approach to wind events is described in Rife and Davis (2005). Figure 9.8 illustrates an event in a time series of hourly wind observations. In this study, an event is defined as a 2-h change in the wind speed that exceeds one standard deviation from the 1-year average value at a given station and time of day. Two verification metrics are used. For one, a set of events is defined in a time series of observations, where $o_t - o_{t-2\Delta t}$ is defined as the event in the observations, and $\Delta t = 1\text{h}$. For each observed event, the following quantity is calculated,

$$\frac{\sigma_o}{\sigma_x}\left(\frac{x_t - x_{t-2\Delta t}}{o_t - o_{t-2\Delta t}}\right),$$

where $x_t - x_{t-2\Delta t}$ is the change in the model solution for the location and time period of the observed event. The individual observed and forecast event magnitudes are normalized

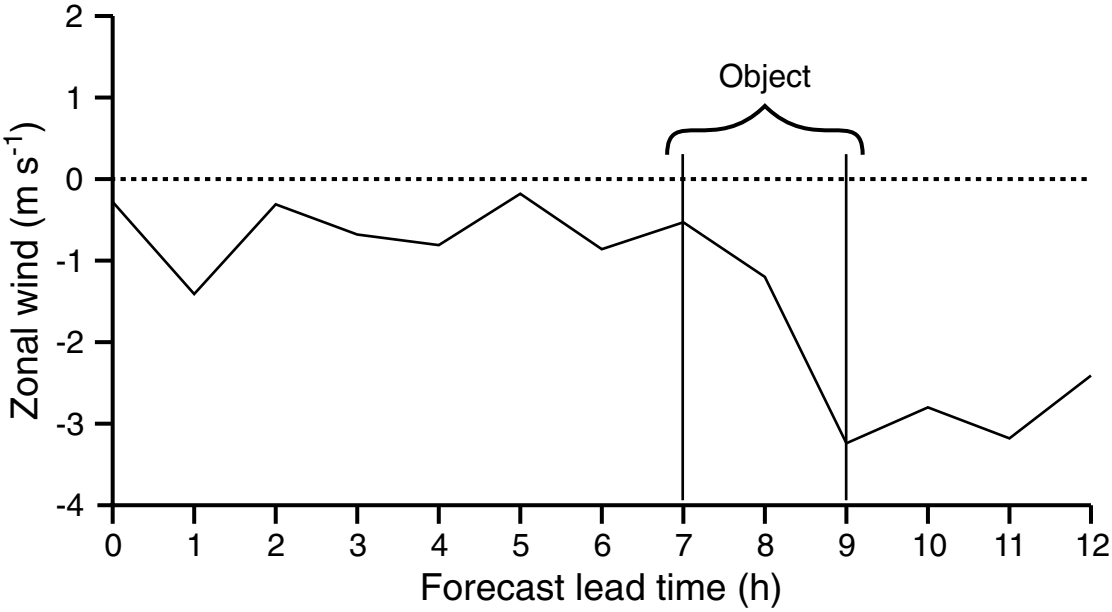

Fig. 9.8 Example of an object, or event, defined in terms of the speed of the zonal wind. Adapted from Rife and Davis (2005).

by the respective variances in the two time series for each location. This ratio is calculated for each station, where the value is +1 for a perfect forecast. In the second approach, again based on time series of observed and forecast values of a variable, both forecast and observed features are defined within every 12-h period. The resulting binary data set, of the existence of one or more features within each period, can be used to populate a contingency table of the sort shown in Fig. 9.3a, allowing the calculation of the many accuracy metrics that are based on discrete variables.

The verification of convective precipitation is well known to be especially problematic, and it lends itself to the use of feature-based methods. Other approaches in which analyses and forecasts of precipitation fields are overlaid, and the overlap regions used to compute scores (Section 9.2.2 above, Wilks 2006), sometimes do not adequately represent the accuracy or value of a forecast. Thus, alternative feature-based approaches have been developed that provide better metrics (Nachamkin *et al.* 2005, Ebert and McBride 2000, Davis *et al.* 2006a,b). Davis *et al.* (2006a,b) should be consulted for a summary of feature-based verification procedures applied to precipitation. In summary, the general approach involves (1) identifying features in the observed (e.g., radar-based analyses) and forecast precipitation fields using thresholds of precipitation amount; (2) describing the geometric properties of the features (e.g., number, location, shape, orientation, size, average precipitation intensity in the feature); (3) comparing the relative attributes of the observed and forecast features; and (4) associating features in the forecast and observed fields, where possible. Figure 9.9 illustrates the benefit of this method for verifying precipitation forecasts. Shown are different examples of observed and forecast areas of precipitation. Forecasts (a)–(d) all have identical basic verification statistics, with POD = 0, FAR = 1,

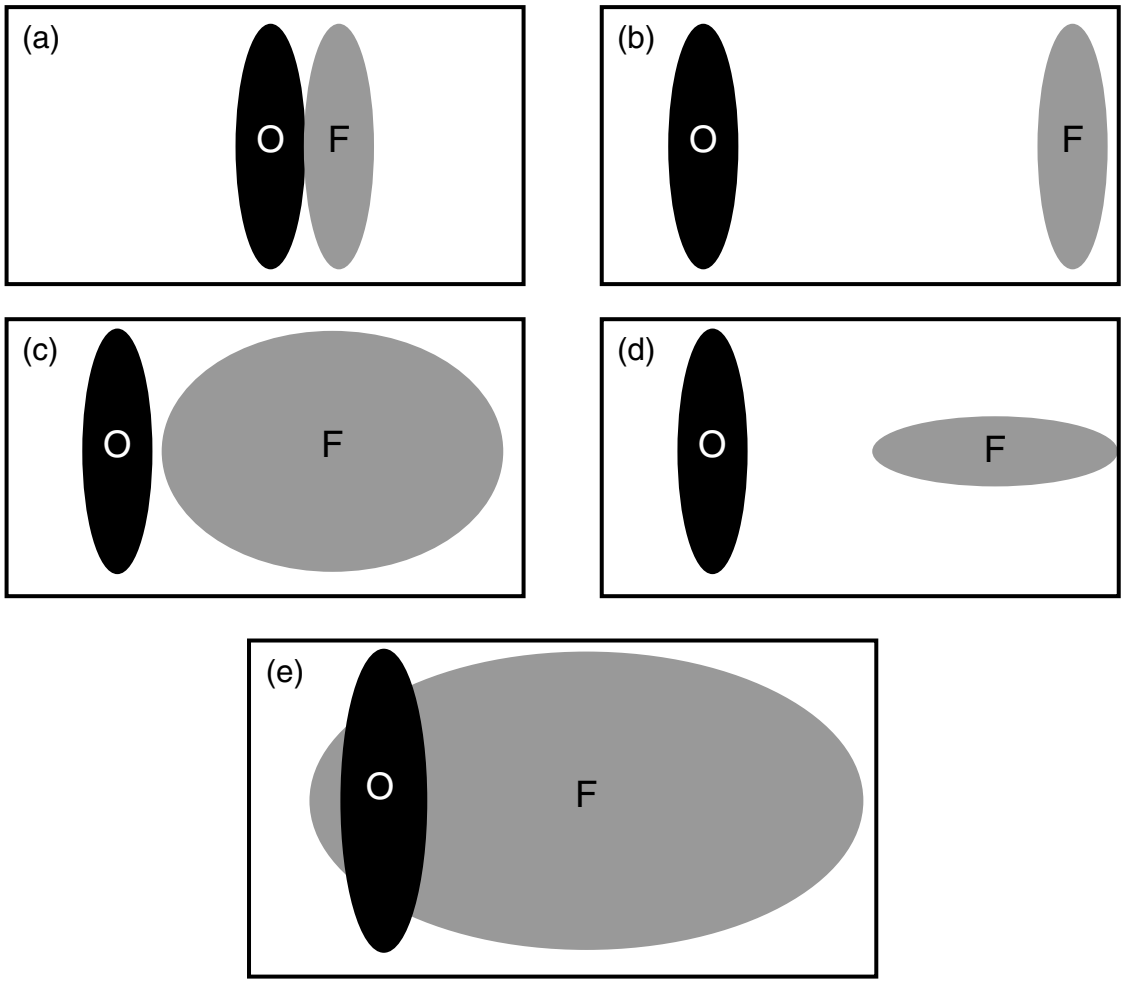

Fig. 9.9 Schematic example of different combinations of forecasts and observations. From Davis *et al.* (2006a).

and CSI = 0 because the forecast and observed precipitation areas do not overlap (the area of "a" in Fig. 9.1b is zero). However, there clearly are differences in the forecast accuracy or value. Forecast (e) has some skill based on these metrics, but would likely not be judged as the best of the five. However, feature-based methods that compare the distances between features, their orientation, and area would provide a better ranking.

An example is shown in Fig. 9.10 of the use of feature-based methods to compare the skill of a few models at predicting summer-season precipitation in an area of complex orography in the southwestern USA. In particular, the method described in Davis *et al.* (2006a,b) will be employed here to compare the accuracy of short-range precipitation forecasts from the MM5, NAM, and RUC models. The horizontal grid increments are 10 km for MM5, 12 km for NAM, and 13 km for RUC. The precipitation forecasts are compared with the NCEP Stage-IV analyses, which are constructed by compositing

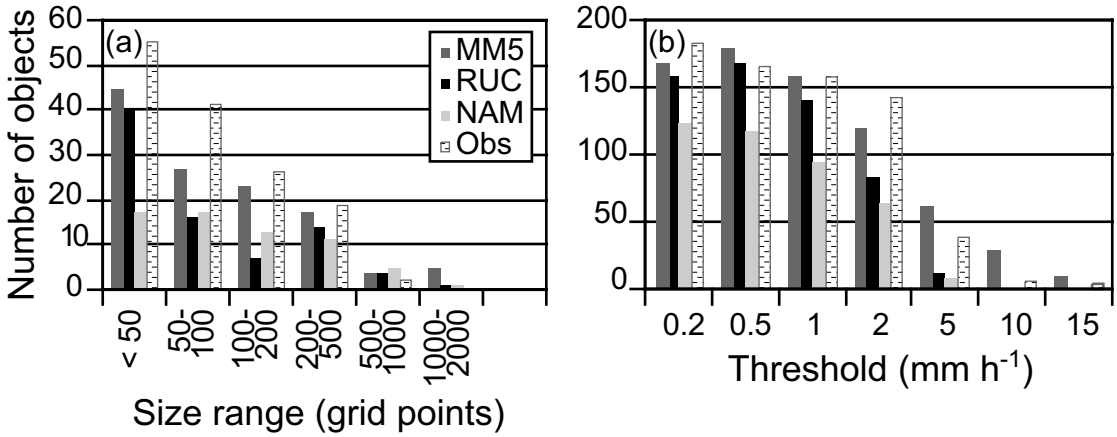

Fig. 9.10 Comparison of the number of observed and forecast precipitation features (based on closed isohyets) in each feature size range (number of grid boxes) (a) and for different precipitation thresholds (b), for three models, and observations based on radar and rain-gauge data. See text for details. Provided by Yubao Liu, NCAR.

WSR-88D radar and gage estimates (Fulton *et al.* 1998, Lin and Mitchell 2005). Precipitation features in the analyses and model forecasts are defined using thresholds of precipitation amount. Any closed isohyet defines a feature, whether or not it represents a maximum or minimum within the surrounding field. The forecasts were verified for the entire month of August 2005, a period during which the North American monsoon was active in the southwestern USA. Given that there were eight forecast cycles per day, this verification is based on over 200 forecasts, most of which have some rainfall.

The comparison here will be limited to the number of forecast and observed rainfall objects (1) having different area coverage (size) and (2) defined by different rainfall thresholds. Figure 9.10a shows the size distributions of the observed and forecast features for the 2-mm h^{-1} intensity threshold. This measure reflects whether the models capture the degree to which the rainfall occurs in a scattered versus contiguous pattern. Figure 9.10b compares the number of observed features for each rain-rate threshold, with those in the model forecasts. For both measures, there are clear differences among the models.

9.9 Verification in terms of the scales of atmospheric features

A model solution should approximately preserve the observed spatial and temporal spectra of the dependent variables. Thus, the spectral power for the atmosphere and the model solution have been compared in numerous studies. Of course, unlike other verification criteria, model-simulated features do not need to be in phase with observed features in order for the model to verify well in this context. Rather, the model solution simply needs to contain the features on the correct scales. This type of verification can:

- help the modeler better understand the explicit and implicit spatial and temporal filters in a model;

- provide information about whether the lower-boundary forcing in the model is imparting the correct scales of motion in the lower troposphere and boundary layer;
- define the model's true resolution;
- illustrate the amount of fine-scale information that is contained in the initial conditions, provided by a data-assimilation system; and
- define the time required for the model to spin up scales of motion that are not in the initial conditions.

9.9.1 Temporal spectra

It is common to partition the temporal spectrum into three regions: periods longer than diurnal (super-diurnal), periods that are approximately diurnal, and periods that are shorter than diurnal (sub-diurnal). Features with longer-than-diurnal periods may be viewed as synoptic scale or planetary scale, and therefore reasonably representable by global or regional models that have typical horizontal resolution. Diurnal time scales of course are related in some way to the heating cycle. For example, in the wind field a diurnal signal could be related to stability-related momentum mixing, mountain-valley circulations, coastal circulations, etc. Provided that the model reasonably represents the land-surface and boundary-layer processes, features with these time scales should verify well. Motions with sub-diurnal time scales include mesoscale features or circulations that are not forced by the diurnal heating cycle. They can result from orographic or other landscape forcing, perhaps far upstream, or from nonlinear interactions. Comparison of the observed and model-simulated spectral power in each of these three regions is a way of verifying the ability of the model to simulate these types of features.

Figure 9.11a shows how the observed temporal spectrum, which must be represented by a model, depends on geographic location. Illustrated is the spectral power for time series of the observed zonal wind at three locations in Slovenia. The higher-elevation station on a mountain (M) is exposed to the synoptic-scale flow, more than are the other stations, so there is greater power in the longer time scales. The coastal station (C) has power maxima on approximately the diurnal time scale (12 h and 24 h peaks), as a result of thermally forced coastal circulations. And the valley station (V), which is protected by the orography from synoptic-scale features and has weak diurnal forcing, shows the flattest spectrum, the lowest overall power, and the largest fraction of the total power in the sub-diurnal range. A model that properly represents the various prevailing processes should replicate this spectrum, and have roughly the same percentage of the spectral power in each of the bands. This type of verification thus has the potential to reveal model strengths and weaknesses in an interesting way. Figure 9.11b shows how well zonal-wind power spectra from the ERA-40 reanalysis with a 125-km grid increment (see Chapter 16 for more information on this analysis) and the ALADIN model with a 10-km grid increment compare with the observed spectrum for the coastal station. Both the reanalysis and the model underestimate the power at all sub-diurnal scales, and overestimate the power on the synoptic scales. The relatively coarse-resolution ERA-40 analysis misses most of the diurnal component, but the ALADIN model captures its approximate amplitude.

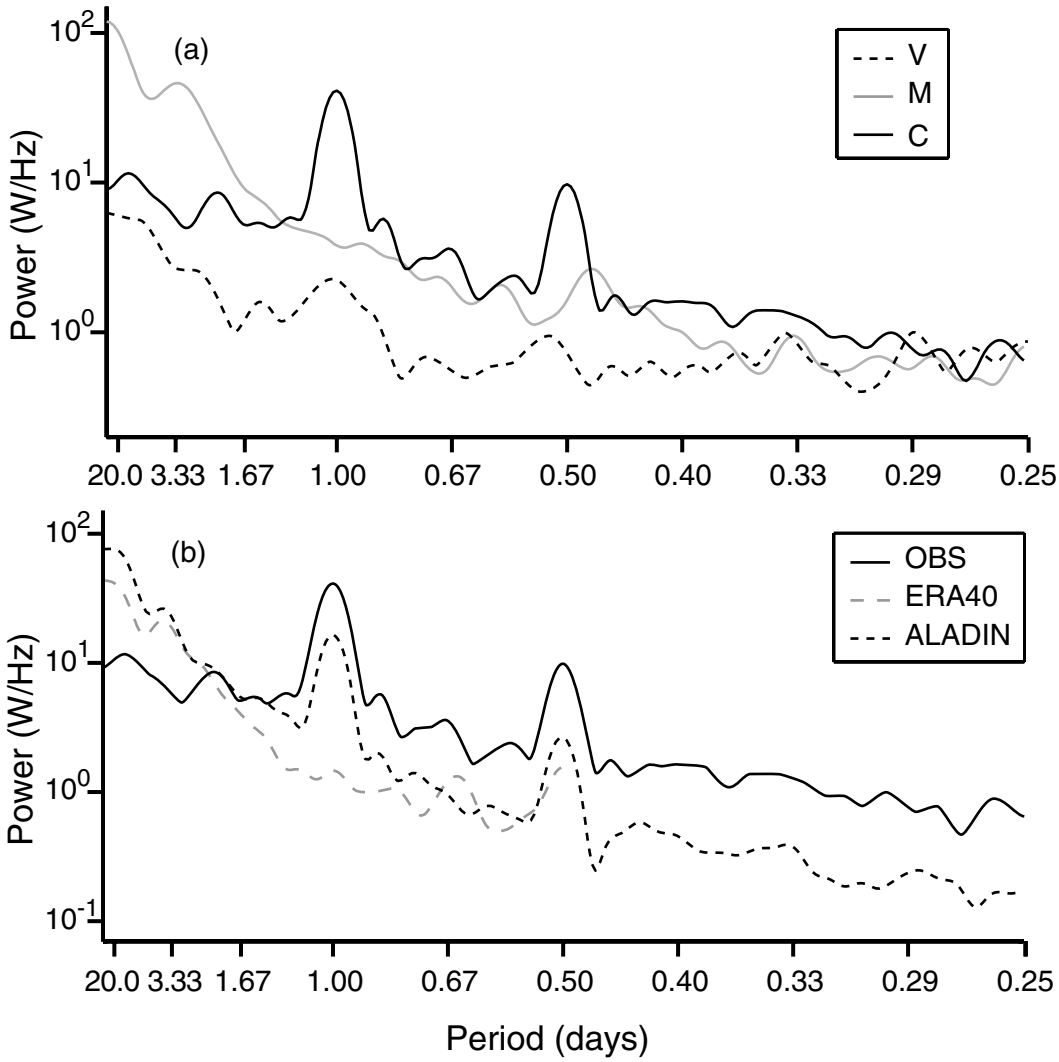

Fig. 9.11 Observed spectral power in the time series of zonal wind speed at three locations: a mountain station (M), a valley station (V), and a coastal station (C) (a). Also shown, along with the observed spectrum for the coastal station, are spectra associated with the ERA40 reanalysis (125-km grid increment) and the ALADIN model (10-km grid increment) (b). Adapted from Žagar *et al.* (2006).

9.9.2 Spatial spectra

The above type of verification of model temporal spectra is made possible and convenient by the general availability of frequent measurements at surface stations. However, there is no equivalently good source of dense observations for use in comparisons of observed and model-based spatial spectra. Thus, the model solution is compared with field-program observations and theoretical solutions, for example of the shape of kinetic-energy spectra. As with the verification of the model temporal spectra, this spatial verification will also confirm the degree to which the model is faithful to the dynamics of the atmosphere.

See Skamarock (2004) for additional information about the characteristics of the kinetic-energy spectrum to be expected on different space scales. In summary, global-scale models should reproduce the large-scale, k^{-3} slope of the spectrum. In mesoscale and cloud-scale model solutions, the slope should be $k^{-5/3}$. Examinations of model spectra have been used to verify possible negative impacts of explicit or implicit damping mechanisms (Laursen and Eliasen 1989). When model resolutions span the global scale and the mesoscale, as is the case with high-resolution global models, the verification of the existence of the slope transition in the kinetic-energy spectrum is a test that the model is faithful to the atmospheric dynamics (Koshyk and Hamilton 2001). And, analyses of kinetic-energy spectra have been used to verify the ability of a model to represent scales near the $2\Delta x$ limit of resolution (Bryan *et al.* 2003, Lean and Clark 2003, Skamarock 2004). This latter type of verification defines the *effective resolution* of the model. Figure 9.12 illustrates the concept of examining the kinetic-energy spectrum for this purpose. The sloping straight line is the anticipated spectrum, where the slope depends on the wavenumber range (k). Because of explicit and/or implicit dissipation in the model, there is some wavelength above the $2\Delta x$ limit where the model spectrum shows kinetic energy that is less than the expected value. This has been defined as the effective resolution because the model dissipation causes the kinetic energy to be unrealistic. The specific shape of the high-wavenumber part of the spectrum is dependent on the model. See Skamarock (2004) for illustrations of actual model spectra that demonstrate this concept.

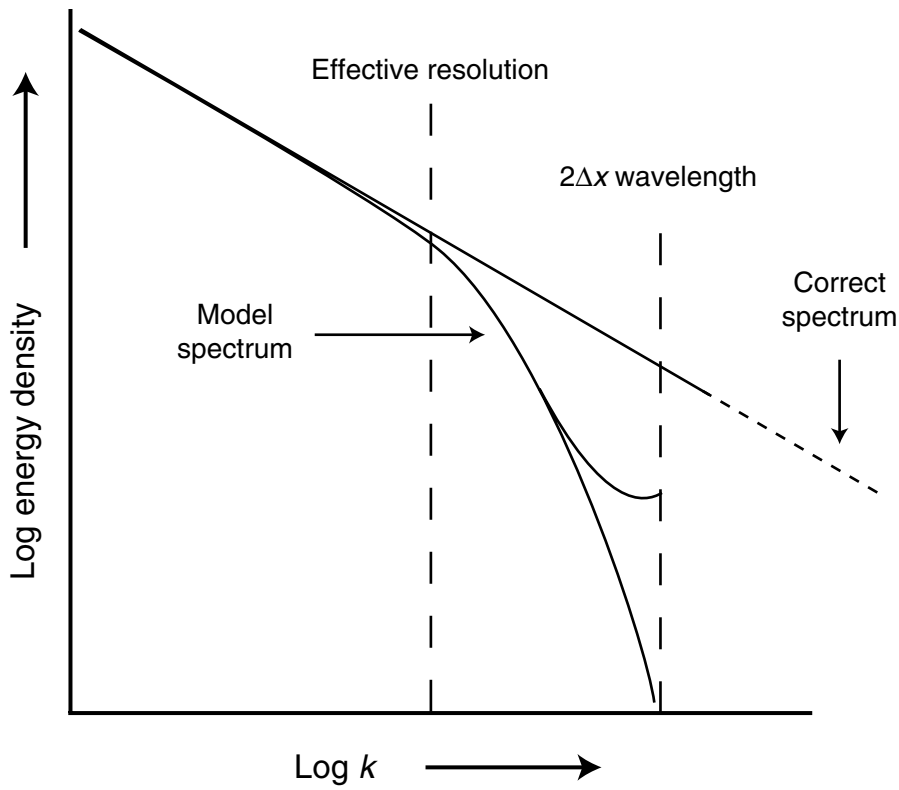

Fig. 9.12 Schematic of the theoretical and model kinetic-energy spectrum, showing the effective resolution. Two examples are shown of the decay in the kinetic energy at the high-wavenumber end of the model spectrum. Adapted from Skamarock (2004).

9.9.3 Variance

Comparing the observed spatial variance among observations with the corresponding variance based on the model output for the same locations is another way of estimating the realism of the simulated spatial structure. Figure 9.13 shows a comparison of the spatial variance for observed, and 12-h model forecast, 10-m AGL winds for a region of complex orography. The variances for a coarse-resolution global model (GFS) and for two resolutions of a mesoscale model (MM5) are shown. In each case, the model-forecast winds

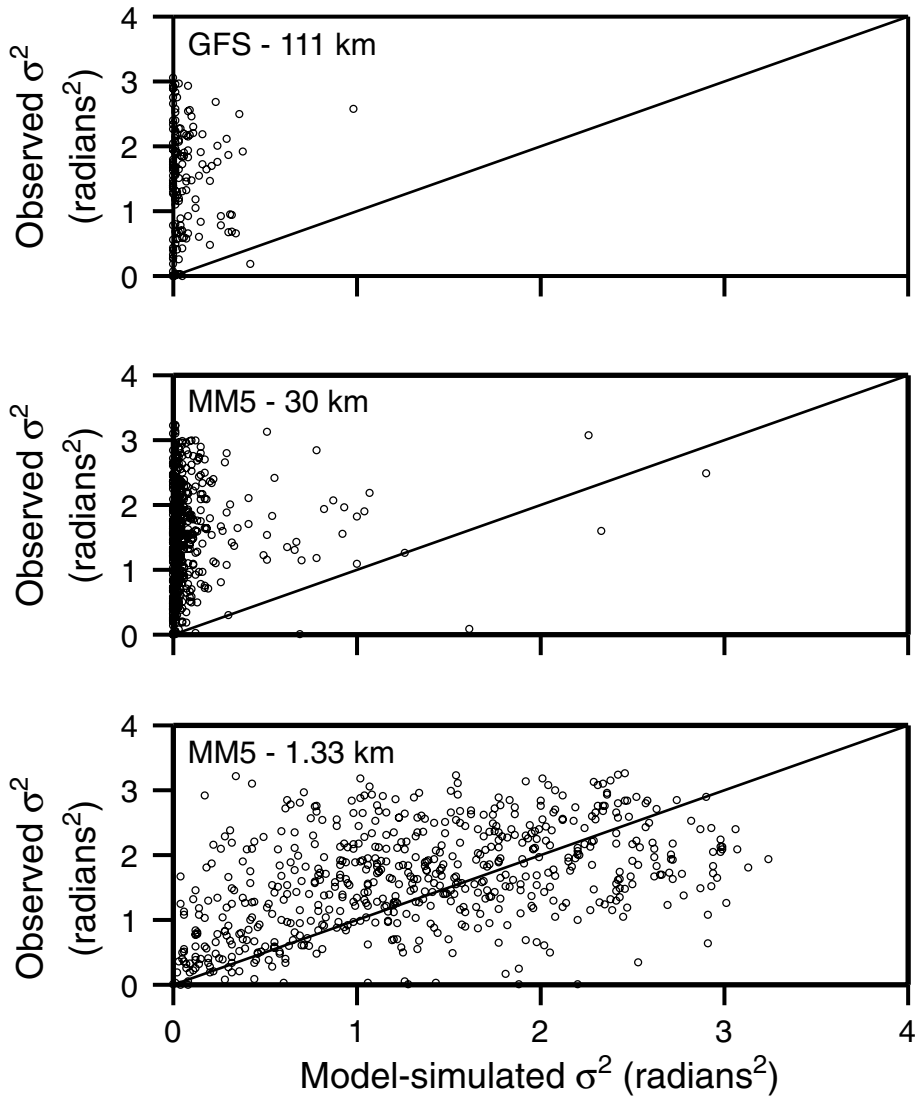

Fig. 9.13 Comparison between the observed spatial variance of 10-m-AGL wind direction and the corresponding variances from 12-h forecasts from the GFS model (111-km grid increment) and the MM5 model (30-km and 1.33-km grid increments) during a 3 February to 30 April 2002 study period. Each point corresponds to a single observation time. Adapted from Rife *et al.* (2004).

were interpolated to the locations of the observations before the variance was calculated. The analysis illustrates the benefit of the higher horizontal resolution toward replicating the observed variance.

9.10 The use of reforecasts for model verification

Chapter 10 defines reforecasts as retrospective forecasts that have been produced with a fixed version of a model, which is the same as what is currently used operationally. Verification statistics that are calculated from a long series of reforecasts provide a more meaningful description of model biases and other shortcomings than would statistics that are based on only a short recent history of operational forecasts. Additionally, rerunning forecasts with the current version of the model means that the verification statistics apply specifically to the current model, a situation that would not be the case if verification statistics were calculated from archived operational forecasts that were based on an evolving model. References on this subject are Hamill *et al.* (2004, 2006), Hamill and Whitaker (2006), and Glahn (2008).

9.11 Forecast-value-based verification

Assessment of the value of ensemble probabilistic forecasts was discussed in Chapter 7, but the importance of this process is worth repeating in the context of model verification. Forecast models are developed and employed because it is anticipated that there is some value to the products that they provide. This value associated with the numerical products can be defined in terms of money saved by business and government, lives saved, and public convenience. Thus, given that these are the ultimate goals of using the models, it is reasonable to verify the models in this value-centric context. That is, these values can serve as metrics that demonstrate, for example, the relative merits of different models. However, it is generally quite challenging to quantify the monetary value of forecasts. And, verification methods that are aimed at improving model performance must be based on accuracy that is expressed in terms of physical forecast variables. Nevertheless, for any operational model application, it is instructive for the modeler to at least qualitatively consider how the forecast value could be defined. There is a substantial body of literature on this subject, and Wilks (2006) and McCollor and Stull (2008b) are among many sources of information about the mathematical basis behind the assessment of forecast value.

9.12 Choosing appropriate verification metrics

There are obviously many choices of metrics available for assessing the accuracy or skill of model forecasts or simulations, and it is often difficult to decide which ones are the best for a particular situation. The reader should refer to the general references at the end of the

chapter, to better understand the sometimes subtle differences in the properties of the different metrics. There are also choices with respect to the most appropriate variables on which to focus. In both cases, the answers depend on the ultimate use of the model output.

In terms of the relevant variables, if the most important use of the model output is to provide input to a flash-flood-prediction system, obviously hourly precipitation rate is most important, with an emphasis on verifying the higher-rate thresholds. Forecasting the precipitation in the correct watershed is important, so the geometry of the landscape can inform the selection of acceptable displacement criteria in feature-based verification methods. If the model is used to provide input to an air-quality model, errors in the low-level winds and boundary-layer depth contribute to errors in the boundary-layer ventilation. And, the depth and strength of surface-based inversions would be important aspects of the forecast. Ideally, if the atmospheric model is being used to provide input to specialized models, such as the above examples for flood and air-quality prediction, the verification should also be in the context of the ultimate variables – e.g., ozone concentration, water discharge in a river, etc.

9.13 Model-verification toolkits

Model verification can be made easier by taking advantage of toolkits available for this purpose. Some are model specific, and others are not. For example, the WRF model has the Model Evaluation Toolkit (MET, supported by NCAR). A free software environment for statistical computing and graphics, which can be used for verification of any model, is provided by the "R Project for Statistical Computing". Available on-line for virtually all the toolkits are the software itself, manuals, announcements of conferences for users to compare applications, frequently asked questions, newsletters, etc. Regardless of the model that is being used, it is worth inquiring about the availability of verification support services such as these.

9.14 Observations for model verification

Observations are, of course, required for both model initialization and verification. However, the availability of observations varies greatly for different parts of the world. Section 6.2 summarizes the observation platforms that can provide *in-situ* observations over land, as well as remotely sensed data that can be processed with retrieval algorithms to produce state variables or precipitation rates. Global data are archived by operational and research centers throughout the world, such as NCEP, ECMWF, the United Kingdom Meteorological Office (UKMO), NASA, the European Space Agency (ESA), and NCAR. The data are often available on-line, at no cost. As an example, Advanced Data Processing data sets are available from the NCAR Computational and Information Systems Laboratory's Data Support Section. The data represent a global synoptic set of hourly surface and

6-hourly data reports, operationally collected by NCEP. The surface data set includes mostly SYNOP[1] and METAR[2] land reports, but a few ship observations also exist. The upper-air data consist mainly of radiosonde soundings.

Because precipitation is a variable that often receives special attention for model verification, some special sources of such data are worth mentioning. Gauge data are, of course, useful, except they are only available over land. And their spatial distribution is far from uniform, and their locations tend to be biased toward more-populated lower elevations rather than the higher elevations where the precipitation is generally greater. Other data sets are based on a merger of satellite and gauge data. For example, the Tropical Rainfall Measurement Mission (TRMM) product (product 3B43; Huffman *et al.* 2007) combines precipitation estimates from multiple satellites (retrievals from measurements in the microwave and infrared regions of the spectrum) as well as gauge-based analyses on a 0.25° × 0.25° grid that extends from 50° N to 50° S for the period from 1998 to the present. For verification of model climatologies, the Global Precipitation Climatology Centre (GPCC; Beck *et al.* 2005) data set provides gauge-based monthly precipitation totals from 1901 to the present on a 0.5° × 0.5° global grid, but only over land. And, some national weather services produce analyses based on weather-radar data. For example, the US NCEP, 4-km, Stage-IV multi-sensor precipitation analysis is constructed using WSR-88D radar estimates, corrected by available gauge measurements (Lin and Mitchell 2005, Fulton *et al.* 1998).

In addition to national data networks, there are numerous regional mesonetworks whose data are often available in real time for no cost at a central repository. Examples in the USA are the Oklahoma (Brock *et al.* 1995) and MesoWest (Horel *et al.* 2002) mesonetworks.

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert (2009). Intercomparison of spatial forecast verification metrics. *Wea. Forecasting,* **24**, 1416–1430.

Jolliffee, I. T., and D. B. Stephenson (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester, UK: Wiley and Sons Ltd.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*. San Diego, USA: Academic Press.

PROBLEMS AND EXERCISES

1. Explain sources of representativeness error in addition to the influence of orography on the low-level windfield. For example, how can differences between local and

[1] SYNOP reports are observations that are made at internationally agreed upon times, every 3, 6 or 12 hours, by meteorological observers. Specific practices are prescribed by the World Meteorological Organization, and adhered to by all national meteorological services.

[2] METAR reports are near-surface observations of the standard meteorological variables that are made hourly, or between hours when special observations are warranted, often at airports. METAR codes are regulated by the World Meteorological Organization.

grid-box-average landscape properties lead to artificial errors in the temperature and humidity fields?
2. On what factors do the spatial and temporal spectra of model solutions depend? What determines how well these properties of the model solution verify against the real atmosphere?
3. Why are there peaks in the power at both 12 h and 24 h for the coastal station winds in Fig. 9.11?
4. For the precipitation observation–forecast pairs in Fig. 9.9, how would you subjectively rank the forecasts in terms of accuracy? Explain your choices.
5. Is the so-called representativeness error really an error?
6. Describe different types of meteorological feature or events that could be used for forecast verification.
7. Using the general references above, summarize which accuracy and skill metrics are most appropriate for different purposes.