7 Ensemble methods

7.1 Background

As we have seen in previous chapters, there is a variety of generally unavoidable sources of model error, including

- initial conditions,
- lateral-boundary conditions for LAMs,
- land/water-surface conditions,
- numerical approximations used in the dynamical core, and
- parameterizations of physical processes.

Each of these input data sets or modeling approaches introduces some error in the modeling process, and ensemble prediction involves performing parallel forecasts or simulations using different arbitrary choices for the above imperfect data or methods. The objective of defining the different conditions for each model integration is to sample the uncertainty space associated with the modeling process in order to define how this uncertainty projects onto the uncertainty in the forecasts. As a preliminary example of the sensitivity of model forecasts to the above factors, Fig. 7.1 illustrates an ensemble of 5-day track predictions for hurricane Katrina in 2005. The forecasts are based on the ECMWF ensemble prediction system. The tracks are strongly dependent on the specific errors in the input observations as well as the model configurations employed.

An ensemble of forecasts is more useful than an individual, deterministic forecast for the following reasons.

- The mean of the ensemble of forecasts is generally more accurate than the forecast from an individual ensemble member, when the statistics are computed over a number of forecasts.
- The difference (spread, variance) among the ensemble members can be an indication of the flow-dependent quantitative uncertainty in the ensemble-mean forecast, given a proper calibration.
- The Probability Distribution (or Density) Function (PDF) of the frequency distribution of a variable can provide information about extreme events, which is especially useful information from a practical standpoint (e.g., issuing weather warnings).
- The quantitative probabilistic products can be more effectively employed in decision-support software systems.

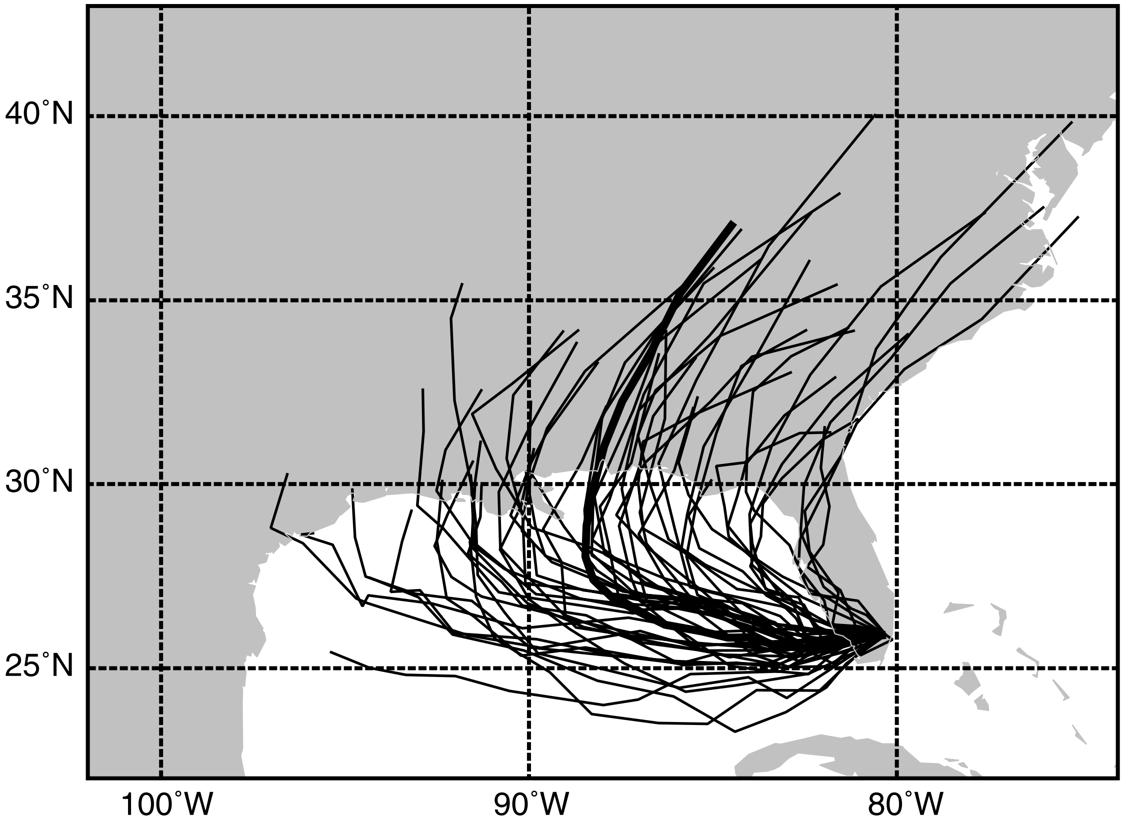


Fig. 7.1 An ensemble of track predictions for hurricane Katrina, initialized at 0000 UTC 26 August 2005, from the ECMWF ensemble-prediction system. The heavy line is the track forecasted by the ECMWF deterministic system. Adapted from Leutbecher and Palmer (2008).

The availability of stochastic forecast products is clearly a great advantage relative to the situation with a deterministic modeling system where a single realization of the future state of the atmosphere is produced, and the forecaster must guess at its veracity.

It should not be surprising that using ensembles of model simulations has led to improved forecast skill. Since the early 1960s it has been known that combining different forecasts from individual forecasters produces a group-mean probability forecast that is superior to the single probability forecast from the most skillful forecaster (Sanders 1963). These findings were confirmed through later studies by Sanders (1973), Bosart (1975), and Gyakum (1986). The recognition of the benefits of this statistical synthesis of human predictions has contributed to a similar process being applied to model predictions.

The potential of ensemble methods is also reflected in how forecasters have used model products for the last few decades. It has been well known by forecasters that when all available models are predicting a similar outcome, the probability is generally high that the predictions will verify reasonably well. In contrast, when the solutions from different models diverge significantly, there is more uncertainty. Thus, before the concept of ensemble prediction was formally established, forecasters were treating the available products as a multi-model ensemble and qualitatively relating the forecast spread to uncertainty.

Given that producing an ensemble forecast requires the parallel integration of many realizations of the modeling system, compromises of some sort need to be made in order to keep the process computationally tractable. Rather than use simpler and inevitably less-accurate physical-process parameterizations, the horizontal resolution is typically reduced in order to compensate for the cost of the multiple integrations. For the same model forecast area, we know that doubling the horizontal grid increment can allow the model to execute eight times faster. Thus, all other things being equal, halving the resolution will allow eight ensemble members to be run in the same amount of time. Quadrupling the horizontal grid increment will allow 64 ensemble members to be used. An issue that exists in parallel with that of computation speed is the general need for more memory by ensemble systems.

The sources of forecast error are a subject of this chapter because ensemble methods seek to sample the uncertainty associated with the sources in order to produce the ensemble. These same sources of error will also be considered in the next chapter on atmospheric predictability because that discussion must be based on the same concepts. The reader should consult Chapter 8 for additional information about that subject.

7.2 The ensemble mean and ensemble dispersion

7.2.1 The ensemble mean

One of the products of an ensemble-prediction system is the average of the members of the ensemble, which represents a forecast that can be interpreted in the same way as a deterministic (nonensemble) product. This ensemble mean, defined at the initial time or at any forecast lead time, is calculated simply by averaging together the gridded fields of the dependent variables from the ensemble members. The mean is typically more accurate than any arbitrarily chosen forecast from an individual member of the ensemble, when averaged over a number of forecasts. The averaging of the ensemble members appears to produce a nonlinear filtering that causes the unpredictable (random) aspects of the forecast to cancel each other, whereas the aspects of the forecasts on which the models agree are not removed in the averaging. Maps of the ensemble-average meteorological fields tend to be smoother than those from the individual members, especially for longer forecast lead times after the individual model solutions have diverged to a larger degree. Palmer (1993) has suggested that ensemble averaging will improve the forecast only up to the time when there is a change in meteorological regime – that is, when there is a bifurcation in the solutions of the members. For example, Fig. 7.2 shows a schematic of the model-state trajectories (dashed lines), for

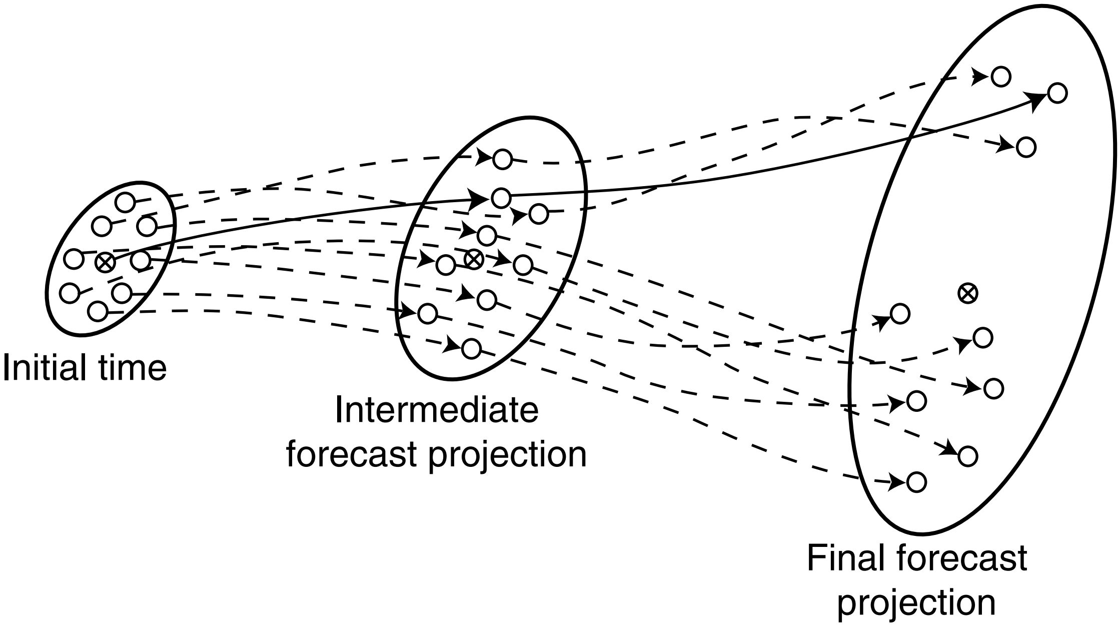


Fig. 7.2 Schematic of model-state trajectories for simulations from an eight-member ensemble initialized from perturbed initial conditions. See the text for details. Adapted from Wilks (2006).

a two-dimensional phase space.[1] The initial conditions for the ensemble members are shown with open circles, as are the solutions at two times during the forecast. At each time, the ensemble mean is defined with an "x". In this example, the eight-member ensemble is constructed by perturbing a control set of initial conditions, and thus the initial states are defined by different phase-space coordinates. At the intermediate forecast time, the model solutions differ more than they did at the initial time, but the error growth has been somewhat linear. Later, at the final time, the trajectories of two of the forecasts have diverged from the trajectories of the other six forecasts; a regime change has taken place. Such a bifurcation in a real ensemble forecast might correspond to some forecasts defining rapid cyclogenesis while others are producing cyclolysis (see Mullen and Baumhefner 1989). In the illustration of Fig. 7.2, the ensemble mean after the bifurcation corresponds with neither branch of the solution, and thus might not represent an especially accurate forecast. An example of a bifurcation in a real forecast is seen later in Fig. 7.13, where the ensemble members tend to be grouped into two different patterns of midtropospheric troughs.

[1] Phase space has a dimension corresponding to each dependent variable in the system, and the coordinates in the phase space define the value of each variable. Thus, a trajectory in phase space represents the temporal change in the state of the system. Sometimes, spatial independent variables may also be dimensions of the phase space.

The above figure also illustrates that the time-dependent behavior of a forecast that is initiated from the ensemble mean (the solid arrow) is different from the time-dependent behavior of the ensemble mean itself. The reason for this can be understood by recognizing that an atmospheric model is a highly nonlinear function that transforms a set of initial conditions into a set of forecast conditions (Wilks 2006). For a nonlinear function *f(x)*, with x the dependent variables and n the number of ensemble members,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right), \tag{7.1}$$

where the right side corresponds to the nonlinear function (the forecast model) applied to the mean and the left side is the average of the forecast fields. Stated another way, on average the best forecast does not result from the use of the best estimate of the initial conditions (the ensemble mean).

7.2.2 Ensemble dispersion, spread, or variance

Because of nonlinear interactions in the fluid, and interactions among the different sources of error in the modeling system, errors tend to grow during an integration. In fact, the errors will continue to grow until the forecast and the true state of the atmosphere (e.g., depicted by an objective analysis) will be as dissimilar as two randomly chosen observed states of the atmosphere. In a deterministic forecast setting we have no way of defining the future error growth. So, we employ an ensemble approach to estimate it by perturbing different aspects of the modeling system (the initial conditions or the model) and interpreting the degree to which the model solutions diverge. This divergence in the solutions, called the dispersion of the ensemble, can be related to the uncertainty in the ensemble mean and is an important component of the forecast. Figure 7.3 displays two arbitrarily chosen ECMWF ensemble predictions of temperature for London, UK and illustrates that the sensitivity of the model atmosphere to uncertainties in the inputs can be very flow (meteorological-situation or -regime) dependent. Clearly the ensemble members have much greater spread on one day than the other.

This relationship between the ensemble dispersion and the error of the ensemble mean is sometimes called the spread–skill relationship. It can be quantified by associating the variance in the ensemble members about their mean with the accuracy of the ensemble mean itself, for each of a large number of ensemble forecasts. Any of the standard metrics discussed in Chapter 9, such as the Root-Mean-Square Error (RMSE), can be used to quantify the accuracy of the ensemble mean. Calibration of the ensemble (see Section 7.5) is necessary in order to quantitatively relate the spread to the uncertainty. Discussions of the ensemble spread–error relationship can be found in Grimit and Mass (2007), and in references cited therein.

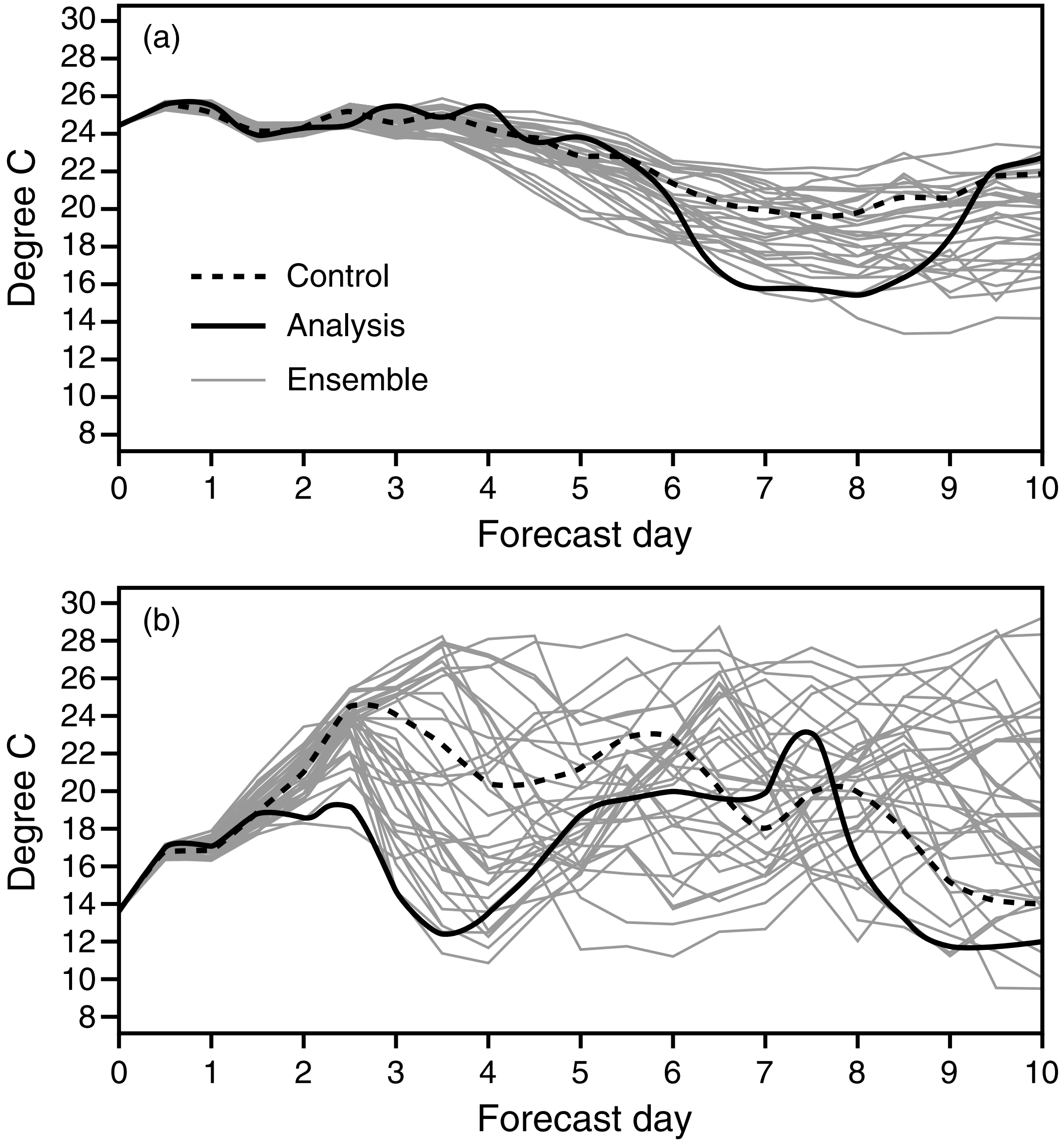


Fig. 7.3 Two ensemble forecasts of 2-m AGL temperature for London, UK, based on ECMWF forecasts separated by 1 year. Note the large difference in ensemble spread. Based on Buizza (2001).

7.3 Sources of uncertainty, and the definition of ensemble members

The list in Section 7.1 provides a brief summary of the sources of forecast uncertainty. This section will offer additional discussion of this subject, as well as of how the uncertainties can be represented in an ensemble-prediction system. The accepted terminology is that forecast error can be divided into initial-condition error and model error. The latter refers to all aspects of the modeling system other than initial conditions.

There are two general approaches for defining ensemble members. In one, the normal configuration of a model is used for a control simulation, and then those conditions (initial conditions or model specifications) are perturbed to create the remainder of the ensemble members. It is reasonable to assume that the control simulation will be the most accurate because it has been optimized in terms of the tuning of the model physics parameterizations and numerics, and has resulted from the generation of an optimal initial state. The perturbations to this control may be somewhat less skillful. In the other approach, entirely different models are used for the different members of the ensemble. These are called multi-model ensembles.

A number of single-model ensembles, which each employ variations in initial conditions, model physics, etc., can be combined to produce a multi-model ensemble. This is referred to as a multi-model *superensemble* (Krishnamurti *et al.* 1999, Palmer *et al.* 2004).

7.3.1 Initial-condition uncertainty

We have seen in Chapter 6 that errors in the observations that are used to define model initial conditions can result from a number of sources. Possible sources include instrument-calibration errors, improper siting of the instrument, representativeness error, and data-transmission errors. In addition, the use of a dynamic-balancing method can introduce errors. Lastly, initial-condition uncertainty cannot be disentangled from the model error because, as we have seen in Chapter 6, modern data-assimilation systems employ the model in various ways. For example, sequential data-assimilation systems use a short model forecast from the previous cycle in order to define the background field for the analysis process. Thus, errors in any aspect of the model formulation will influence the quality of the first-guess field, which in turn will impact the initial-condition error, especially where observations are sparse. Improving the accuracy of the model can thus provide more-accurate initial conditions.

It is intuitive that the sensitivity of forecasts to the initial conditions is flow dependent. Imagine a synoptic-scale situation that is dominated by a large semi-permanent anticyclone, for example associated with a Mediterranean-type climate in summer. The forecast will be relatively insensitive to the details of the initial-condition error because the situation is dominated by the large-scale planetary forcing associated with the general circulation and the land surface. In contrast, there are situations where the atmosphere is close to an instability threshold and small initial-condition differences can cause the state of the model atmosphere to follow either one path or another at the bifurcation.

Initial-condition error could be created by adding to a mean state a random error that is consistent with the uncertainty in the observations. However, it has been found that simply adding random numbers to initial conditions will create ensemble members that are very similar to each other. This should not be surprising, given the discussion in Chapter 6 about geostrophic adjustment. That is, perturbations that are imposed independently on the mass and windfield variables on small scales will be dispersed as inertia–gravity waves. In contrast, the methods that are used in practice in NWP to generate initial-condition uncertainty in ensembles produce perturbations that have dynamically consistent structures.

Three different approaches are used to define initial-condition uncertainty in ensembles.

- Ensemble-based data assimilation is used to define a sample of initial states, as described in Section 6.11.3. The Kalman filter combines (1) background fields that have been created with an ensemble of forecasts and (2) observations, with their associated errors.
- An approach is based on so-called bred vectors, and samples the dynamically most sensitive modes in the initial conditions. It consists of the following steps.
  1. Random perturbations are added to the dependent variables that define an initial state.
  2. Both the perturbed and unperturbed states are used as initial conditions for model simulations with a duration of 6 h to 24 h.
  3. The two simulated states are subtracted, and the gridded difference field is scaled so that its magnitude is similar to the error in a typical analysis.
  4. The scaled perturbation is added to a new initial-state estimate, and the perturbed and unperturbed initial states are again used for a pair of parallel model simulations.
  5. This process of perturbation growth and rescaling is repeated, where the bred vector is the perturbation that results after a few iterations.

  The bred patterns are different from day to day, and reflect the features with respect to which the ensemble members are diverging most rapidly. See Ehrendorfer (1997), Toth and Kalnay (1993, 1997), and Kalnay (2003) for additional information about the breeding method.
- Singular vectors (Buizza 1997, Ehrendorfer 1997, Molteni *et al.* 1996, Ehrendorfer and Tribbia 1997, Kalnay 2003), also called optimal perturbations, are obtained by using tangent-linear and adjoint models, and define the fastest-growing patterns for the prevailing weather situation of the day. Linear combinations of these patterns, with the magnitudes scaled according to the expected analysis uncertainty, are added to a control analysis to define the ensemble members.

7.3.2 Lateral-boundary-condition uncertainty for LAM ensembles

For LAMs, the model solution depends strongly on the LBCs, especially for longer integration times, so errors in the LBCs can contribute significantly to the model error. In a forecast setting, the LBC-related error in the LAM depends on both the error in the forecast from the large-scale model as well as errors introduced by the algorithm used to couple the two grids. If the large-scale model is an ensemble, the individual ensemble members can be used to provide the LBCs for the LAM ensemble members. If the large-scale forecast is not an ensemble, the LBCs need to be perturbed in such a way that the process estimates typical errors from the large-scale model. Thus, the time scale, space scale, and amplitude of the errors need to be estimated, and used in the generation of LAM ensemble members.

7.3.3 Surface-boundary-condition uncertainty

Errors in the calculation of the land-surface properties during a forecast can result from initial-condition error or model error.

- The initial conditions for the time-varying land-surface variables are defined by a LDAS, as described in Chapter 5, and errors in the LDAS calculations can result from both errors in the LSM that is the basis for the LDAS and errors in the estimates of the non-time-varying physical properties of the substrate such as the thermal conductivity and heat capacity of the dry substrate, leaf-area index, etc. Thus, as with the atmospheric model, land-surface initial-condition error is intertwined with model error.
- Model error is associated with the LSM that is integrated in parallel with the atmospheric model, as described in Chapter 5. As with the atmospheric models, errors can result from the numerical approximations to the differential equations, as well as from the parameterizations of the physical processes.

This source of error can be represented by defining uncertainties in the parameterizations of the processes, in the estimates of the initial conditions of the time-varying physical properties of the substrate (moisture and temperature profiles), and in the time-invariant physical properties of the substrate (pore space, specific heat and thermal conductivity of the dry substrate). Many studies have evaluated the uncertainty in the atmospheric structure that results from uncertainties in different aspects of the land surface. See Pielke (2001), Sutton *et al.* (2006), and Hacker (2010) for references.

7.3.4 Errors in the numerical algorithms

We have seen in Chapter 3 that numerical approximations to space and time derivatives introduce errors, and they, along with the physical-process parameterizations, contribute to the model error. Even though these dynamical-core errors are initially largest on the small scales near the truncation limit of the model, through nonlinear interactions they can affect the scales of mid-latitude high- and low-pressure systems within a couple of days of model integration. The typical way to include in an ensemble the uncertainty associated with the particular properties of the dynamical core is to use entirely different models for different ensemble members. Ensembles constructed in this way are called multi-model ensembles. Because of the limited number of models available for this approach, the size of such an ensemble is limited to perhaps 10 members. A recent approach to including model error in ensembles is the use of stochastic kinetic-energy backscatter methods, which represent upscale propagating energy caused by unresolved subgrid-scale processes (Shutts 2005, Berner *et al.* 2009).

7.3.5 Errors in the physical-process parameterizations

As with the above method for representing model error associated with the dynamical core, errors related to physical-process parameterizations can also be represented with a multi-model ensemble. In addition, some modeling systems allow the user to choose from a list of options for each of the parameterizations. Even though some combinations of parameterizations are incompatible with each other, it is possible to create a significant number of ensemble members simply by varying the parameterizations used with a single dynamical core. And, as we have seen in Chapter 4 it is possible to vary uncertain parameters within a particular parameterization in order to create multiple ensemble members. Alternatively, Buizza *et al.* (1999) model random errors associated with physical-process parameterizations by multiplying a dependent-variable's time tendency from the parameterization by a random number that is sampled from a uniform distribution between 0.5 and 1.5. This method increases the ensemble spread and improves the skill of the probabilistic prediction. See Teixeira and Reynolds (2008) for an additional example of the use of a stochastic parameterization. Lastly, there are many studies that compare the use of initial-condition and model-physics uncertainties in ensemble systems (e.g., Stensrud *et al.* 2000, Clark *et al.* 2008a).

7.3.6 Multi-model ensembles

The use of multi-model ensembles is appealing when various models are already being routinely run operationally by different modeling centers for either weather or climate prediction. In these situations, the normally daunting challenge of creating a multi-model ensemble prediction is simply a matter of reconciling the outputs to a common grid for quantitative processing. Or, forecasters often view products from all the available models and qualitatively synthesize the information.

Regarding the latter approach, it was mentioned earlier that, for decades, forecasters have related the degree to which forecasts from different models agree to the overall uncertainty in the products. Fritsch *et al.* (2000), Woodcock and Engel (2005), and many others discuss the concept of consensus forecasting, which can involve the synthesis of forecasts made by humans as well as forecasts from different operational models.

Multi-model ensembles have been used especially extensively for seasonal prediction (Feddersen *et al.* 1999, Rajagopalan *et al.* 2002, Stefanova and Krishnamurti 2002, Barnston *et al.* 2003, Palmer *et al.* 2004, Robertson *et al.* 2004, Doblas-Reyes *et al.* 2005, Feddersen and Andersen 2005, Hagedorn *et al.* 2005, Hewitt 2005, Stephenson *et al.* 2005, Krishnamurti *et al.* 2006b), and climate prediction on longer time scales (Section 10.5.4 in Meehl *et al.* 2007, Section 16.1.6 in this text).

In its simplest form, a multi-model ensemble forecast can be produced by simply averaging the individual members using equal weights. However, more-complex methods for combining the model solutions are used, for example as described in Clemen (1989), Robertson *et al.* (2004), and Stephenson *et al.* (2005).

7.4 Interpretation and verification of ensemble forecasts

The ensemble mean and the individual members of the ensemble can be evaluated employing standard metrics that are also used for deterministic predictions, while other measures are used for the probabilistic aspects of the forecasts. This section summarizes some of these methods. Additional information on this subject can be found in Wilks (2006) and the Appendix of McCollor and Stull (2008a).

7.4.1 Ensemble – mean predictions

The predicted ensemble means can be evaluated using any of the conventional accuracy and skill metrics described in Chapter 9, which focusses on verification of nonprobabilistic model solutions. The accuracy measures include such familiar quantities as bias, RMSE, and Mean-Absolute Error (MAE). In contrast, the skill score is a way of comparing the accuracy of one forecast method with that of a reference forecast (Eq. 9.3). If the accuracy is no better than that of the reference forecast, the skill score is zero. This is relevant to ensemble prediction because a deterministic forecast can be used as the reference, and its accuracy can be compared with that of the ensemble mean. Lu *et al.* (2007) use the skill score in a comparison of the mean-absolute error of a time-lagged ensemble forecast (Section 7.6) with that of a deterministic counterpart.

The *Taylor diagram* (Taylor 2001) is used in atmospheric science to graphically summarize how well statistical properties of observed and forecast patterns match. Because it is easy to plot the statistical properties of multiple forecasts on the same diagram, it is used to display the performance of individual ensemble members as well as the ensemble mean. Figure 7.4 shows the form of the diagram, where the radial distance from the origin is proportional to the standard deviation of a pattern (of a forecast variable, in this application), and the azimuthal position is related to the correlation of the pattern with a reference field (the verification field). Plotted on these diagrams are points associated with the forecast fields, and a point corresponding to the analyzed field (open circle), where the latter has a correlation coefficient of 1.0 because it is perfectly correlated with itself.

In this example from Delle Monache *et al.* (2006a), the ensemble modeling system consists of coupled meteorological and air-quality models, and the forecast variable being plotted is ozone concentration. The numbers correspond to indices of forecasts of individual ensemble members, and the open square represents the ensemble mean. The objective is to graphically quantify the relationship between the forecast and verification ozone fields in terms of the two noted statistical metrics. Taylor (2001) shows how the forecast's Centered RMSE (CRMSE, the RMSE after the bias has been removed) can be plotted as well (the dashed line), because of its mathematical relationship to the original two statistical measures (the graph coordinates). The CRMSE is the distance between the two points that represent a forecast and the analysis, where the closeness of the two points is proportional to the accuracy of the ensemble member. In this case, the CRMSE coordinate associated with the ensemble mean ozone concentration is plotted. This CRMSE of the ensemble mean is smaller than the CRMSE of any of the ensemble members. Note that the bias of an ensemble must be represented separately, because the Taylor diagram shows the CRMSE, not the RMSE.

7.4.2 Probabilistic predictions

The following approaches are traditional ones used to evaluate the statistical properties of ensemble predictions.

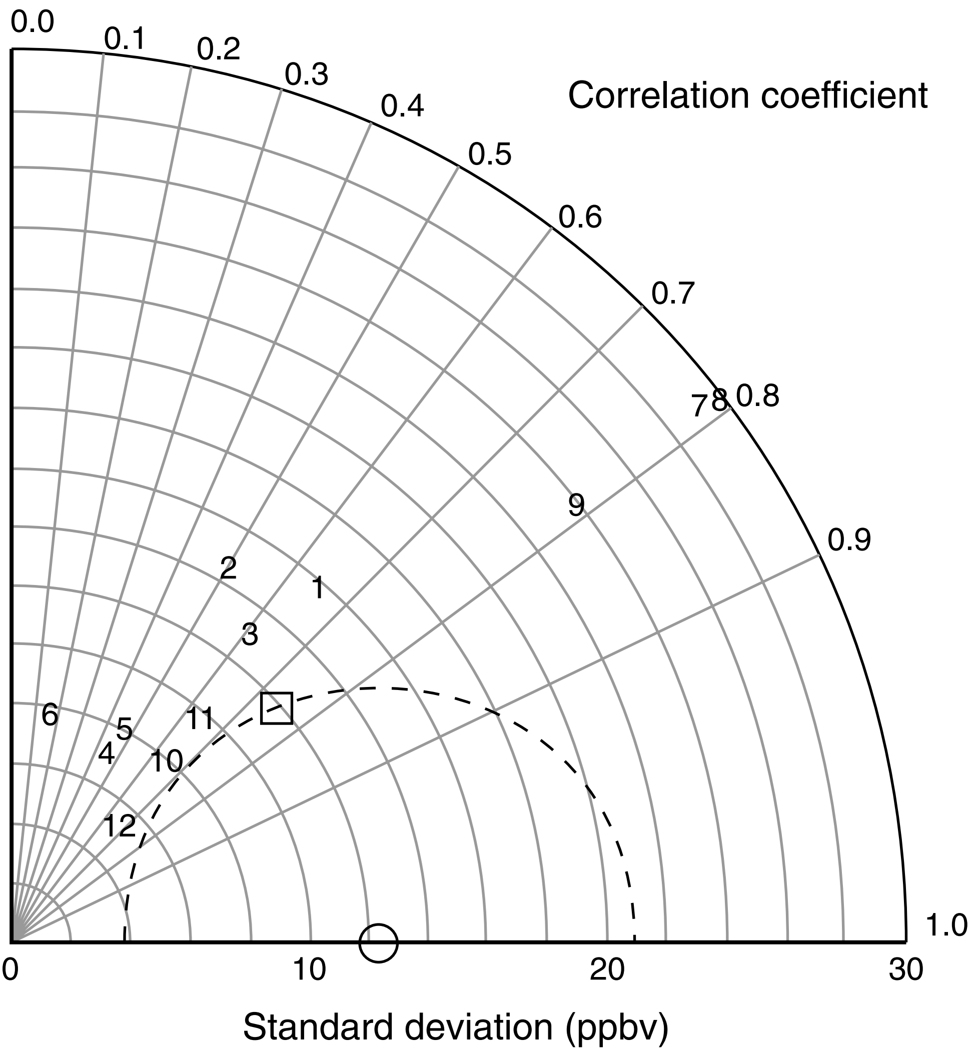


Fig. 7.4 Taylor diagram, showing the statistical performance of ensemble members in a forecasting system that consisted of coupled meteorological and air-quality models. The forecast variable is ozone concentration, where the open circle corresponds to the analysis, the open square is the ensemble mean, and the numbers are indices of individual ensemble members. See the text for details. Adapted from Delle Monache *et al.* (2006a).

Reliability diagrams

Reliability is an important attribute of ensemble forecasts of dichotomous events – ones that either occur or do not occur at a grid point or over an area – and reliability graphs are a device for easily visualizing the quality of probabilistic forecasts. Such discrete events include the existence of temperatures below freezing, or 3-h accumulated rainfall above a threshold value. Consider a set of ensemble forecasts of event E that are performed during a period of time, where, for each forecast, E is predicted to occur with probability p, for $0.0 < p < 1.0$. For the subset of the ensemble forecasts for which the forecast probability of occurrence is p_f, the observed frequency of occurrence is calculated to be p_o. For a perfect forecasting system, $p_f = p_o$. Figure 7.5 shows characteristic forms of the reliability graph (also called an *attributes diagram* or *calibration function*). Figures 7.5a and b illustrate unconditional biases, where the ensemble overpredicts (a) and underpredicts (b) the probability for all situations; that is, the sign of the bias is the same for all forecasts. Figure 7.5c shows a situation where the probabilities are predicted reasonably well in all situations. The plots in Figs. 7.5d and e show conditional biases in the model prediction of the event probabilities. In the former case, the model underpredicts the probability for low-probability situations and overpredicts it for high-probability situations. Here, it is said that the model forecasts have poor resolution (in a statistical sense). That is, the observed probability is similar over the full range of predicted probabilities. In contrast, in Fig. 7.5e there is good resolution because forecasts are able to identify situations with a variety of different probabilities, even though of course the forecasts are conditionally biased. See Wilks (2006) for additional discussion of this subject.

As an illustration of an actual reliability graph, Fig. 7.6 pertains to an ensemble of seasonal predictions, where the event is above-average 2-m temperature for the period February through April in the tropics. The panel on the left (a) is based on an ensemble that used a single model, where the ensemble was created by perturbing the atmosphere and ocean initial conditions. A number of other single-model-ensemble seasonal simulations were created, and all had reliability diagrams with a similar conditional bias. But, when the single models were combined into a multi-model superensemble, a considerably improved reliability resulted (b). This ensemble modeling was part of the Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction (DEMETER), and employed models from seven institutions in Europe. See Chapter 16 for more information.

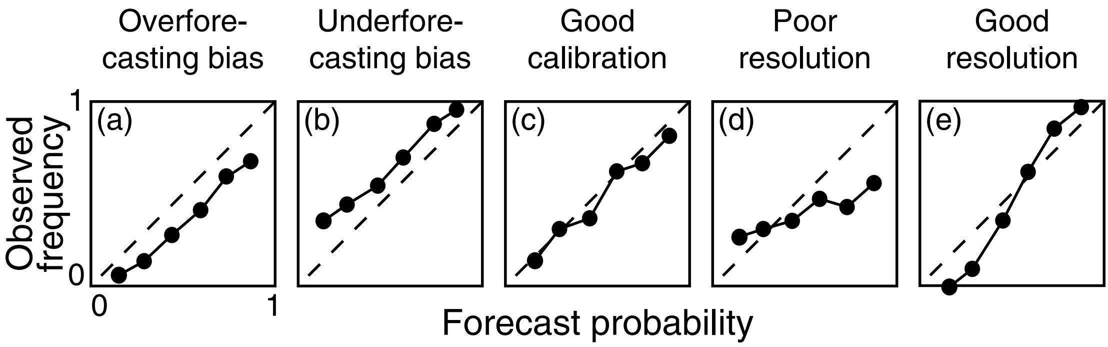


Fig. 7.5 Reliability graphs (also called attributes diagrams or calibration functions) for a variety of situations, showing different patterns of bias in predicted probabilities. The abscissa is probabilities predicted by an ensemble system over a large number of cases, and the ordinate is the corresponding conditional probabilities based on observations. See the text for details. Adapted from Wilks (2006).

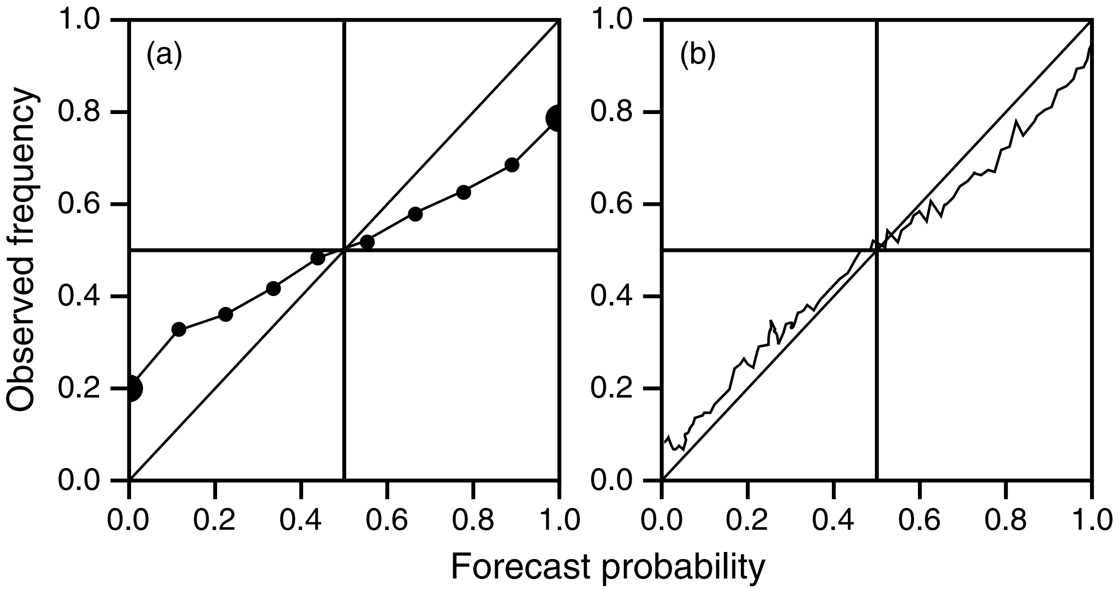


Fig. 7.6 Reliability diagram, where the predicted event is above-average 2-m temperature for the period February through April in the tropics. The panel on the left (a) is based on an ensemble that used a single model, where the ensemble was created by perturbing the atmosphere and ocean initial conditions. A number of other single-model-ensemble seasonal simulations were created, and all had reliability diagrams with a similar conditional bias. But, when the single models were combined into a multi-model superensemble, a considerably improved reliability resulted (b). Adapted from Palmer *et al.* (2005).

Rank histograms

Rank histograms, or verification rank histograms or Talagrand diagrams, are used to display the relationship between observations and forecasts from individual ensemble members. That is, they define the bias of probabilistic predictions. For a specific variable and the location of an observation, take an ensemble forecast of that variable at that location, and rank-order the forecasts from each of the members. Then define the $n + 1$ intervals bounded by the n ordered forecast values. Figure 7.7 shows an example schematic with four ensemble members and five intervals, for a forecast variable P. For this location, and at time $t_{forecast}$, the observed P (X obs) is lower than any of the forecast Ps, and the observation is thus in interval I_1. If we follow a similar process for all other pairs of observations and forecasts at this time, we can calculate the total number of observations in each of the five intervals, or ranks, and plot a histogram of the frequency. This will provide a graphical view of how the ensemble of forecasts relates to the observations. A nonuniformity in the histogram's distribution will reveal systematic errors in the ensemble. Figure 7.8 shows four problems with an ensemble, which can be defined with rank histograms. In this hypothetical eight-member ensemble, the rank histogram has nine ranks, or intervals. In panels (a) and (b), many of the observations fall near the edge of the distribution of forecasts in the ensemble, or outside of the distribution entirely, corresponding to

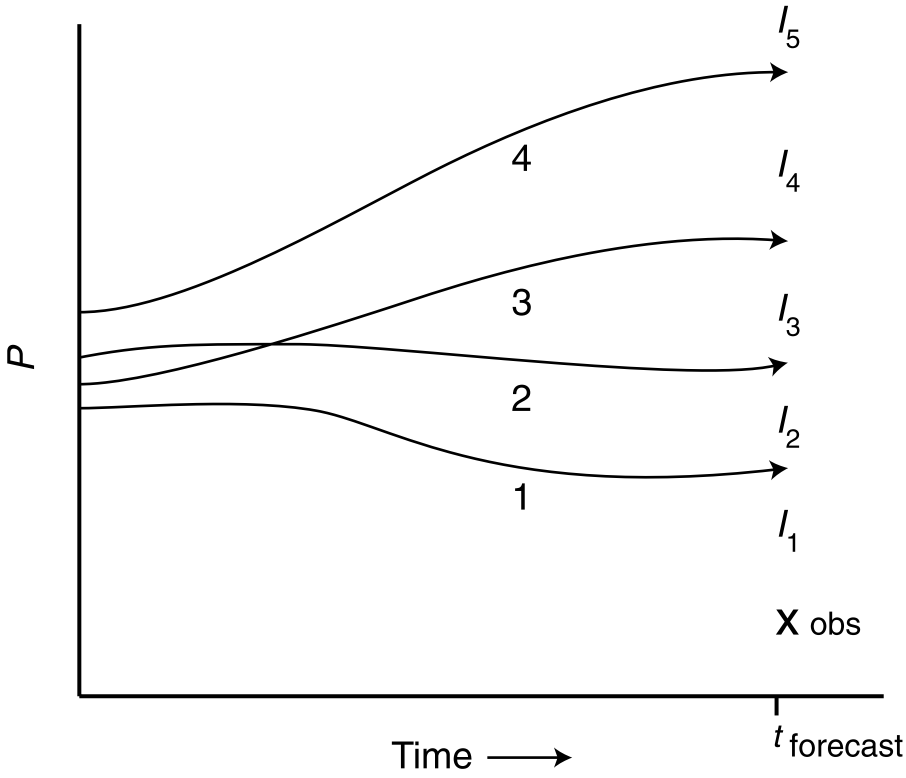


Fig. 7.7 Schematic illustrating how a rank histogram is constructed. The trajectories show the time evolution of a forecasted variable *P* at the location of an observation, from a four-member ensemble. The values of *P* at the forecast time define the intervals (*I*) against which an observation of *P* (X obs) is compared. Each observation of *P* is assigned to one interval in the rank (even though the *P* values of the intervals will be a function of the observation location), and the resulting fractions of the total observations that fall into each rank are plotted in the histogram.

an overforecasting bias and an underforecasting bias, respectively. For example, in the situation depicted in panel (a), the most common situation was for all the ensemble members to forecast a value larger than the observation. Panel (c) shows a desirable rank-uniform distribution where observations are equally likely to fall anywhere within the distribution of ensemble members. In panel (d), many observations fall outside of, or near the edge of, the range of the forecasts from the ensemble members. Because the histogram is symmetric, there is no bias. In this case, the ensemble spread is too small, or in other words the ensemble is underdispersive. This situation is common, where the ensemble fails to always encompass the observations. Stated another way, the spread in the ensemble is less than the difference between forecasts and the validating analysis. The opposite situation is indicated in panel (e) where the ensemble is overdispersive. The ordinate can also be plotted as a probability, where the probability for each rank is defined by dividing the total number of times the verification occurred in the rank (the frequency) by the total number of forecast–observation pairs. These types of diagrams are further discussed in Anderson (1996), Talagrand *et al.* (1997), Hamill and Colucci (1997, 1998), Hamill (2001), and Wilks (2006).

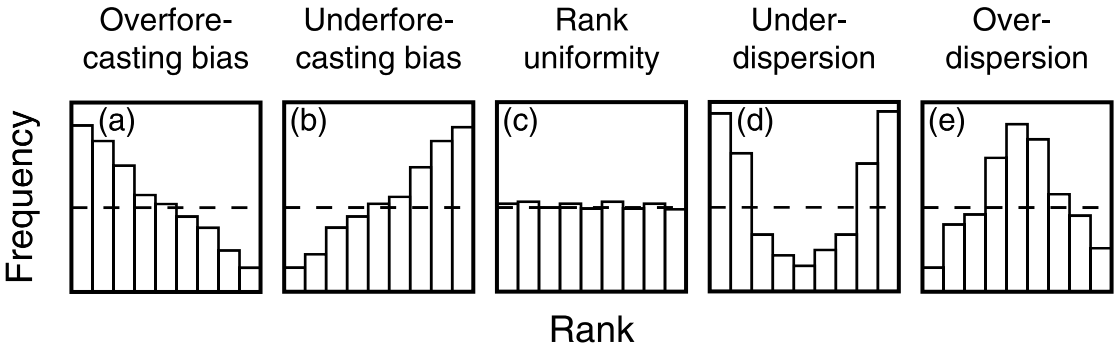


Fig. 7.8 Five rank histograms corresponding to different relationships between the ensemble members and the observations. The horizontal dashed line defines a perfectly rank-uniform distribution. See the text for details. Adapted from Wilks (2006).

Relative Operating Characteristic (ROC) diagrams

The *ROC diagram*, used to evaluate probability forecasts of binary predictands, has the false-alarm rate (F) as the abscissa and the hit rate (H) as the ordinate, where F and H are defined in Section 9.2.2. Such a forecast might be whether the daily precipitation amount exceeds 1 cm, or whether the maximum daily temperature exceeds 30 °C. Wilks (2006) describes how to convert probabilistic forecasts, in this case from ensemble systems, into 2 × 2 contingency tables from which F and H can be calculated. The pairs of F and H are used to define points on the diagram, and along with the (0.0, 0.0) and (1.0, 1.0) points represent a curve. Better forecasts have a low F and a high H, so more-accurate ones have points in the upper-left. The area under the curve has a maximum possible value of unity, corresponding to a perfect forecast. The diagonal corresponds to an unskilled forecast, and the associated area would be 0.5. Forecasts with ROC areas of ~0.75 or higher are considered to be good. Figure 7.9 shows an example ROC curve.

Brier scores and Brier skill scores

The Brier Skill Score (BSS, Jollife and Stephenson 2003, Wilks 2006) is based on the Brier Score (BS), which assesses the accuracy of probabilistic predictions. The BS is defined as

$$BS = \frac{1}{n}\sum_{j=1}^{n}(p_j - o_j)^2 ,$$

and calculates the average squared difference between forecast probabilities (p_j) and observational outcomes (o_j), for n forecast-event pairs, where o is zero if the event does not occur and unity if it does occur. This expression is completely analogous to that presented in Section 9.2.1 for the Mean-Square Error (MSE) that is used for nonprobabilistic

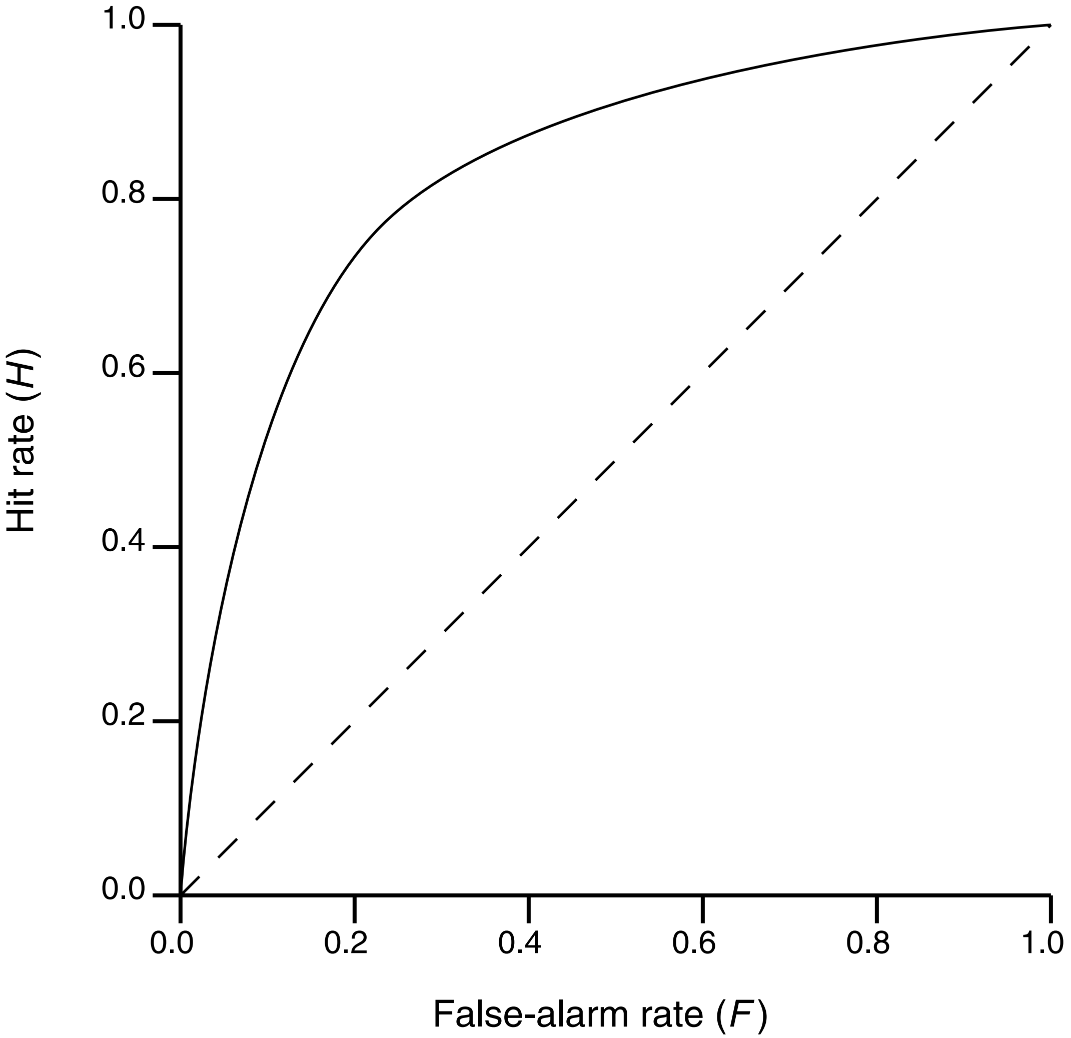


Fig. 7.9 An example Relative Operating Characteristic (ROC) diagram. The solid line is a ROC curve for a good forecast, the dashed line defines a random forecast, and the area under the ROC curve (fraction of 1.0) represents the overall performance in terms of this metric.

predictions. The BS ranges from zero to one, with lower values indicating better forecasts. The BSS is defined as

$$BSS = \frac{BS - BS_{ref}}{BS_{perf} - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{7.2}$$

where $BS_{perf} = 0$. The BSS is unity for a perfect forecast, and zero or negative for unskillful forecasts relative to the reference forecast.

Rank probability skill score

The Rank Probability Score (RPS) describes the quality of categorical probabilistic forecasts for any number of event categories. The BS can be regarded as the special case of an RPS with two forecast categories. The Rank Probability Skill Score (RPSS) is defined analogous to Eq. 7.2, where the RPS_{ref} is based on the climatological probabilities. See Wilks (2006) for the mathematical definition of the RPS and Weigel *et al.* (2007) for a comparison of the RPSS and BSS. An example of the use of the RPSS is found in the next section.

7.5 Calibration of ensembles

The *calibration* of ensemble forecasts of weather and climate is a post-processing step that removes the bias from the first moment (the ensemble mean) and possibly the higher moments. This is necessary in order to:

- provide greater accuracy in the ensemble mean,
- provide improved estimates of the probabilities of extreme events, and
- represent ensemble spread in terms of quantitative measures of the uncertainty in the forecast of the ensemble mean.

Figure 7.10 illustrates the calibration process in terms of its influence on the PDF. Both the mean and the spread of the distribution have been adjusted in the process.

Like many other statistical corrections that are applied to model output to remove systematic errors, a history of high-quality observations and ensemble forecasts is required to calibrate the operational forecasts. Historical, archived operational ensemble forecasts are not ideal for this purpose because models are continually being updated, and thus the required calibration changes as well. Ideally, reforecasts that use the current version of the ensemble system to recreate forecasts for a significant historical period should be used for the calibration. Discussion of ensemble reforecasts can be found in Hamill *et al.* (2004).

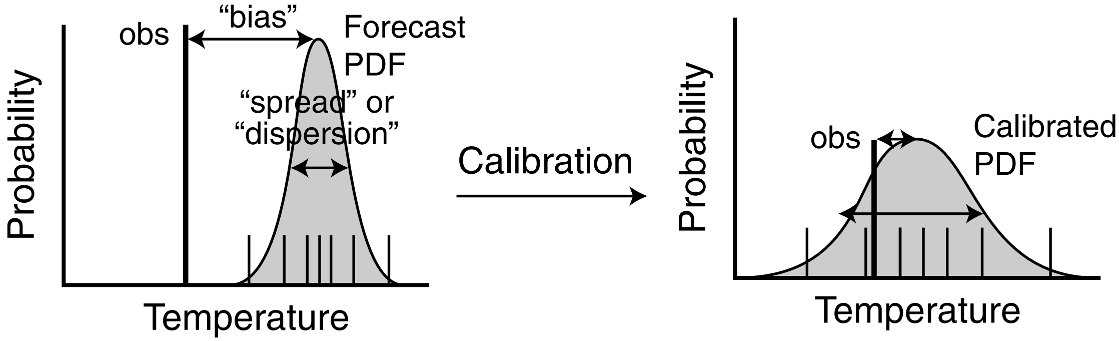


Fig. 7.10 Schematic showing how the calibration process adjusts a PDF. Provided by Thomas Hopson, NCAR.

There are several ensemble-calibration techniques. Hamill and Colucci (1997, 1998) employ the information in the verification rank histogram to interpret and correct the ensemble forecasts. It is important to note that a calibration performs better when it is done for as specific a set of conditions as possible. For example, calibrations can be dependent on weather-regime, forecast lead-time, geographic-area, season, etc. Also, separate calibrations are sometimes performed for different amounts of ensemble spread, for example as quantified by the standard deviation of the ensemble members. That is, the ensemble forecasts are divided into different bins based on their standard deviation, and separate rank histograms and calibrations are constructed for each group.

Eckel and Walters (1998) used a subset of a long history of Medium-Range Forecast model (MRF) ensemble predictions as a training data set with which they calibrated that model. They then used a complementary period for forecast verification. Both uncalibrated and calibrated forecasts were verified using the RPSS, which employed climatology as the reference forecast, to assess the benefit of the calibration of the MRF-ensemble Quantitative Precipitation Forecasts (QPF). Figure 7.11 shows the RPSS, based on calibrated and uncalibrated two-week forecasts. In one approach, the forecasts were calibrated using the "weighted ranks" method of Hamill and Colucci (1997, 1998). In another, an uncalibrated "democratic voting" method was used, where each ensemble member gets an equal vote regarding the occurrence of precipitation above some threshold. In the uncalibrated approach, the total number of ensemble members for which the precipitation exceeds the threshold, divided by the number of ensemble members, defines the probability. The skill score for a climatology-based forecast is zero, so positive scores beat climatology while negative ones do not. In this case, calibration of the forecasts extended the predictability by about one day. Other examples of the many calibration methods available are found in Bremnes (2004), Doblas-Reyes *et al.* (2005), Raftery *et al.* (2005), Roulston (2005), and Weigel *et al.* (2009). Weigel *et al.* (2009) address the issue of whether calibrated single-model ensembles are superior to multi-model ensembles.

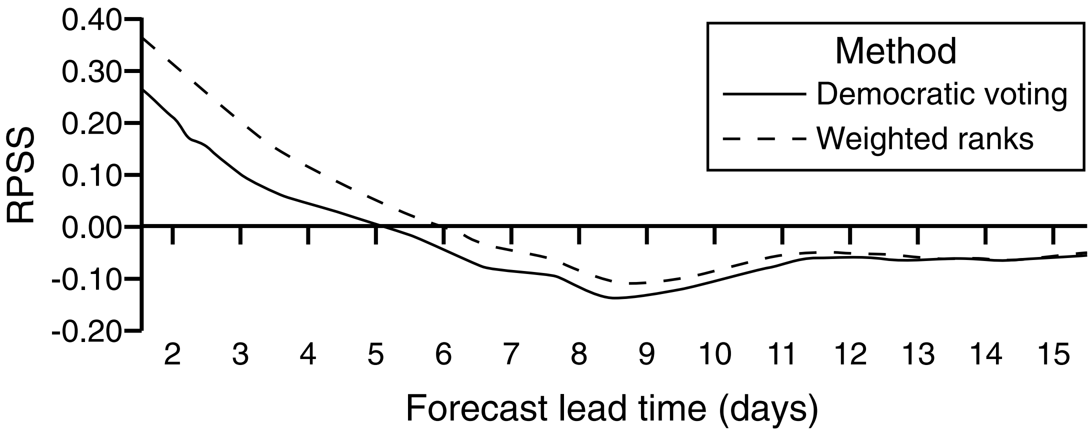


Fig. 7.11 The RPSS for all forecast lead times, based on two methods of defining the probability from the ensemble: the uncalibrated "democratic voting" method and the "weighted ranks" calibration method of Hamill and Colucci (1997, 1998). Adapted from Eckel and Walters (1998).

7.6 Time-lagged ensembles

Because contemporary forecasting systems typically involve the use of a production cycle in which new forecasts are initiated every 1 h to every 12 h, multiple forecasts are available for the same times when forecasts overlap. Figure 7.12 illustrates the concept. In the top example of a time-lagged ensemble, forecasts are initiated every 6 h as part of a normal deterministic operational system. These forecasts can be combined to form an ensemble, say for the time in the future corresponding to the dotted line. Only the initial conditions are different in this ensemble – different by an amount equal to the changes in the initial states between the 6-h initialization times. The bottom half of the diagram depicts the traditional approach for creating an ensemble, where multiple forecasts are initialized at the same time using different initial conditions or model configurations. The time-lagged ensemble can be created at no additional cost because the forecasts are part of a normal forecasting system, in contrast to the traditional ensemble approach where multiple forecasts are performed from the same initial time.

This approach was first proposed and evaluated by Hoffman and Kalnay (1983) and Dalcher *et al.* (1988). Lu *et al.* (2007) employ the RUC model with a 1-h update cycle in tests of very-short-range (1–3 h), time-lagged ensemble forecasts, where two approaches were used to create the ensemble mean. In one, the forecast values from each of the

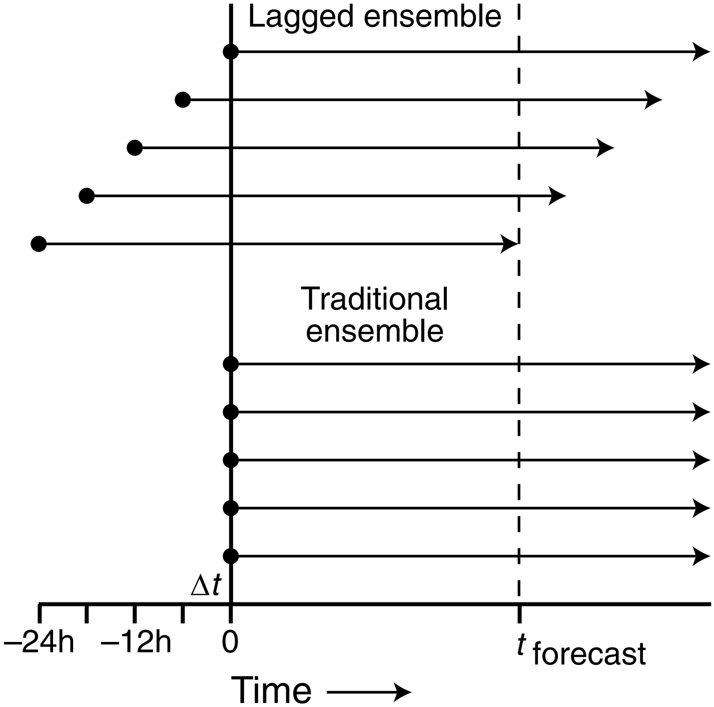


Fig. 7.12 Schematic showing forecasts in time-lagged (top) and traditional (bottom) approaches to ensemble prediction. See text for details.

ensemble members were equally weighted. In the other, different weights were used for the different time-lagged forecasts. Even though both methods provided improved forecasts relative to the deterministic-system products, the method with unequal weights was superior. Yuan *et al.* (2009) describe the verification of a multi-model time-lagged forecast system that employed the WRF and MM5 models running on a 1-h cycle. Because the cycle frequency was so high, a large number of forecasts were valid at the same time and thus it was possible to utilize a large ensemble. The ensemble-mean QPF had greater skill than did the forecasts from the deterministic NAM model running at the same horizontal grid increment (12 km), but other aspects of the verification were less positive. In Yuan *et al.* (2008) a larger time-lagged multi-model ensemble, based on MM5, WRF, and RAMS with a 3-km grid increment, was tested for hydrological applications. They concluded that such time-lagged ensemble systems can provide valuable ensemble-mean QPFs and probabilistic QPFs for water-management applications. For longer-range global-model forecasting, Buizza (2008) compared the performance of the 51-member low-resolution (T399L62) ECMWF traditional ensemble system with that of a higher-resolution (T799L91) 6-member lagged ensemble, for a 7-month period. The cycle frequency was 12 h, and the initialization times of the lagged forecasts spanned 60 h. The 51-member ensemble was superior to the lagged ensemble in terms of probabilistic measures, but the ensemble-member-weighted lagged ensemble had similar skill in predicting the ensemble mean out to forecast-day 4. Lastly, Delle Monache *et al.* (2006a) describe encouraging results from the use of an 18-member lagged-ensemble system for air-quality applications. Additional discussion of the lagged-ensemble method is found in Mittermaier (2007).

The relationship between forecast uncertainty and the spread of time-lagged ensemble members has a foundation in the way that forecasters have used operational models for decades. When consecutive forecasts from a series of forecasts in a cycle predict the same outcome at a particular verification time, the forecasters are confident in the model solution. On the other hand, models occasionally signal uncertainty by producing different outcomes in consecutive forecasts in the cycle. This results in the forecasters having less confidence in the products.

7.7 Limited-area, short-range ensemble forecasting

Mesoscale ensemble modeling with LAMs is becoming more prevalent in the research and operational communities. As noted earlier in Section 7.3.2, the existence of LBCs affects the error that must be sampled in the generation of the ensemble. Because the limited-area ensemble systems are used for producing shorter-range forecasts than are the global ensemble models, the process is often referred to as Short-Range Ensemble Forecasting (SREF). Eckel and Mass (2005) provide a summary of challenges posed by SREF, compared to medium- and longer-range ensemble forecasting with global models.

- Near-surface variables exhibiting fine-scale structures are important forecast quantities, but they are less predictable and their errors may saturate for short forecast lead times, thus limiting the use of an ensemble. Error saturation means that the forecast has no skill in the sense that it is completely uncorrelated with the verification field (see Fig. 8.1, and related discussion).

- Model error is poorly understood and difficult to quantify, and has a larger impact on near-surface variables for short-range forecasts (Stensrud *et al.* 2000).
- Methods for generating optimal initial-condition perturbations (e.g., the breeding method) were developed for the medium range, where nonlinear error growth generates a large spread of solutions. It is unclear how to generate initial-condition perturbations for SREFs, where error growth is initially linear (Gilmour *et al.* 2001).
- The use of LAMs may result in insufficient ensemble dispersion, even when the LBCs are perturbed (Nutter 2003).
- Very-high resolution may be needed in order to capture variability at small scales.

For additional examples of mesoscale ensemble prediction with LAMs, see Marsigli *et al.* (2005) and Holt *et al.* (2009).

7.8 Graphically displaying ensemble-model products

Probabilistic forecast information must be displayed in ways that are meaningful to both model developers as well as to the ultimate users of the forecast information. Even though the needs of those two groups, in terms of appropriate graphical products, are somewhat different, there is the commonality that more creativity is required than for the display of the state variables themselves. The following subsections review some of the common types of displays.

7.8.1 Spaghetti plots

One of the greatest challenges in interpreting the spread among ensemble members is graphically synthesizing the vast amount of information in an easily interpretable way. The use of small individual maps that show a particular variable field from each of the ensemble members is one approach (see Fritsch *et al.* 2000, Legg *et al.* 2002, Palmer 2002, and Buizza 2008 for examples). However, these "stamp maps" are small and can be difficult to interpret when details are important. An alternative is to define a meteorologically important and graphically simple aspect of the variable field, and display that for each of the ensemble members in the same image. Figure 7.13 shows an example of this approach that uses a single contour (5520 m) of the 500-hPa height field. Even though the entire field cannot be visualized, the shape and position of this contour reveal how the pattern evolves with time, and more importantly how it differs among the 17 ensemble members in this case. At the 12-h lead time (a), some of the ensemble members are beginning to produce a trough over central Canada, although they are all in good agreement elsewhere. Thus, except in the region of this trough the forecast would be interpreted with confidence. By 36 h (b), this trough development has continued in some of the members, leading to a

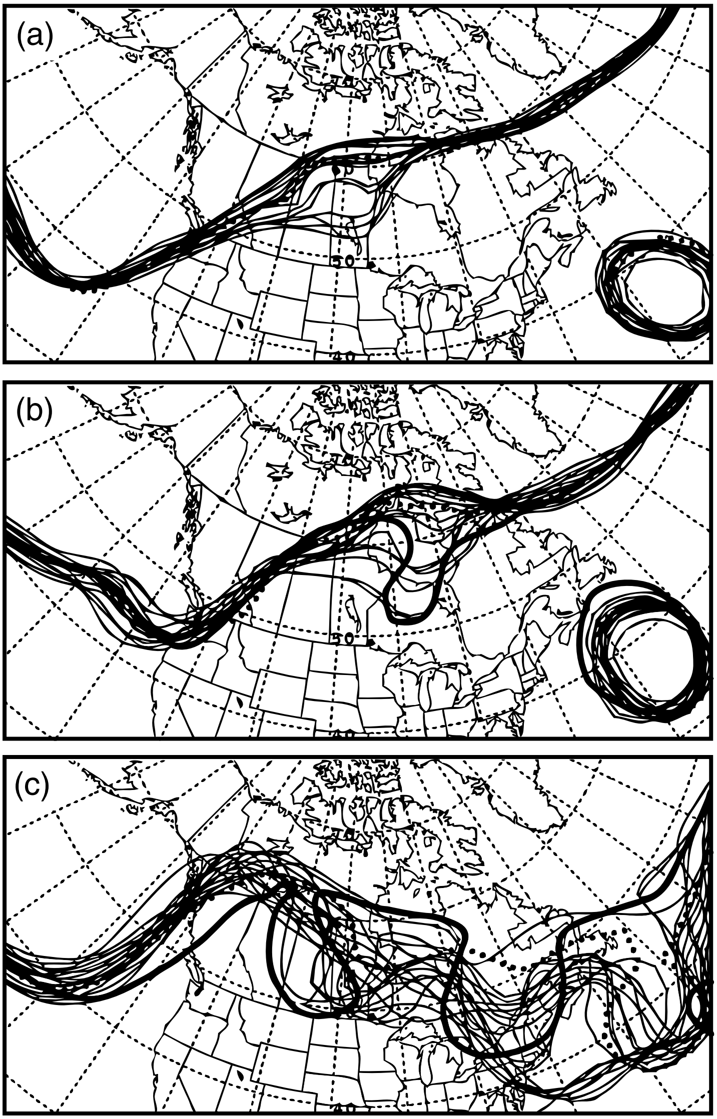


Fig. 7.13 Spaghetti plots of the 5520-m contour of the 500-hPa height field over North America, based on an ensemble forecast by NCEP, where 12 h (a), 36 h (b), and 84 h (c) lead times are shown. The light solid lines are the contours associated with each of the 17 ensemble members, the heavy solid lines in panels (b) and (c) are based on the verifying analyses, and the dotted lines show the control forecast. From Toth *et al.* (1997).

growing degree of uncertainty about the solution in this area. The control simulation (dotted line) shows no evidence of the trough, even though the verifying analysis (heavy solid line) indicates a large-amplitude feature. At the 84-h lead time (c), all members tend to agree with respect to the location and amplitude of the trough in the East Pacific, but over the continent and West Atlantic there is much scatter among all the members and much reason to be suspicious about the forecast accuracy.

7.8.2 Meteograms, or box plots

Ensemble forecasts of a single variable at a single location (or an average over an area) can be displayed, where the value of the variable is plotted as a function of time for each ensemble member. The spread of the members and the mean can easily be visualized as a function of forecast lead time. The ultimate user of the forecasts (not the forecaster) often has a particular variable of concern (e.g., precipitation rate) and a specific location (e.g., a city or watershed), so this type of plot makes more sense than a more-complicated mapping for a large area. An example of this type of display is shown in Fig. 7.14, in the form of plots of an ensemble of 6-month forecasts of El Niño SST anomalies for the NINO3 region in the eastern Pacific, based on an ECMWF coupled ocean–atmosphere model. The model solutions progressively diverge throughout the simulation period. Figure 7.3, shown previously, is another example of such a display, for near-surface temperature in London, UK.

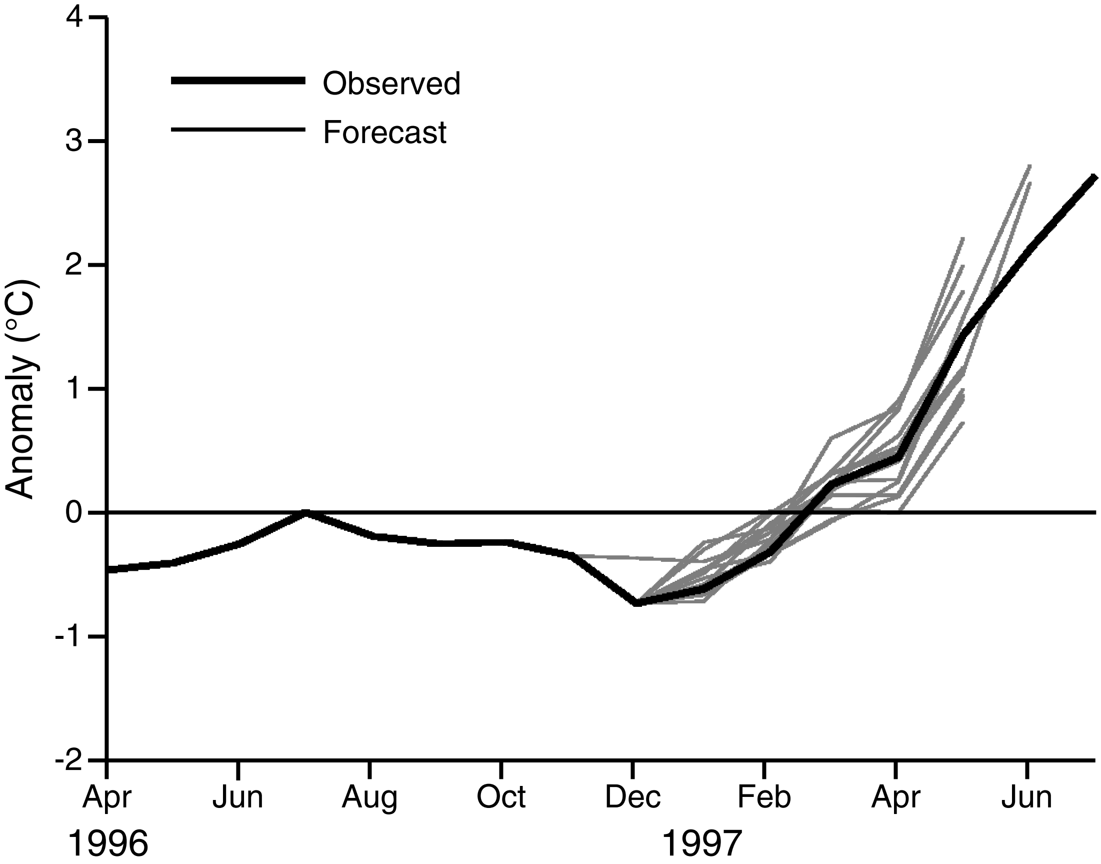


Fig. 7.14 An ensemble of 6-month forecasts of spatially averaged El Niño SST anomalies for the NINO3 region in the eastern Pacific, based on an ECMWF coupled ocean–atmosphere model. The simulation begins in December 1996. The solid line indicates observed average SSTs. Adapted from Palmer (2002).

7.8.3 Probability-of-exceedance plots

The PDF from properly calibrated ensemble forecasts can be interpreted in terms of the probability of an event occurring at a particular grid point. For example, 30 ensemble members will produce an equal number of estimates of the wind speed at a grid point at a particular forecast lead time. The number of members for which the speed exceeds a certain threshold can be used to define the probability that the speed will be exceeded at that point (this is the uncalibrated democratic voting method defined in Section 7.6). The resulting gridded field of probabilities can be contoured. An example of this type of plot is shown in Fig. 7.15. Based on a 50-member ECMWF ensemble, the map shows the probability that wind gusts will exceed 50 m s^{-1} at the 42-h lead time of a forecast of a severe synoptic-scale storm that devastated parts of Europe on 26 December 1999. Similar exceedance plots

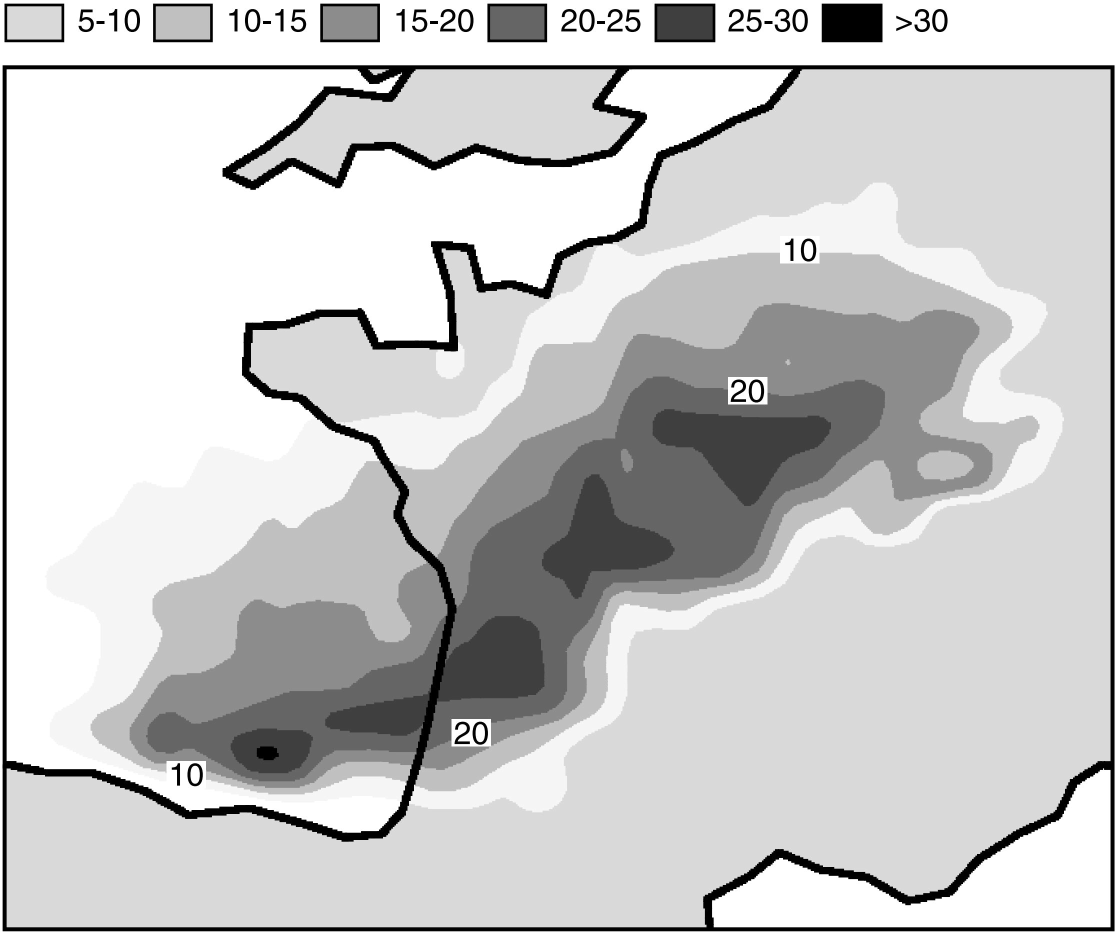


Fig. 7.15 The probability (percentage) that wind gusts will exceed 50 m s^{-1} at the 42-h lead time of a forecast of a severe synoptic-scale storm that devastated parts of Europe on 26 December 1999. This is based on a 50-member ECMWF ensemble. Adapted from Palmer (2002).

are used widely, where Delle Monache *et al.* (2006c) contains other examples. These types of products could, of course, be based on a PDF that has been adjusted through calibration.

7.8.4 Plots of some metric of ensemble variance

A number of measures are available for reflecting, in a single number, the spread of an ensemble, averaged over a model computational domain or defined at a point. For example, variance can be plotted as a function of forecast lead time in order to represent the spread of the ensemble as a function of time. Figure 7.16 is an example of this type of plot, and shows the variance in the 850-hPa specific humidity as a function of forecast lead time for a 19-member physics ensemble and a 19-member initial-condition ensemble that were run with MM5 for the same case of a long-lived mesoscale convective system.

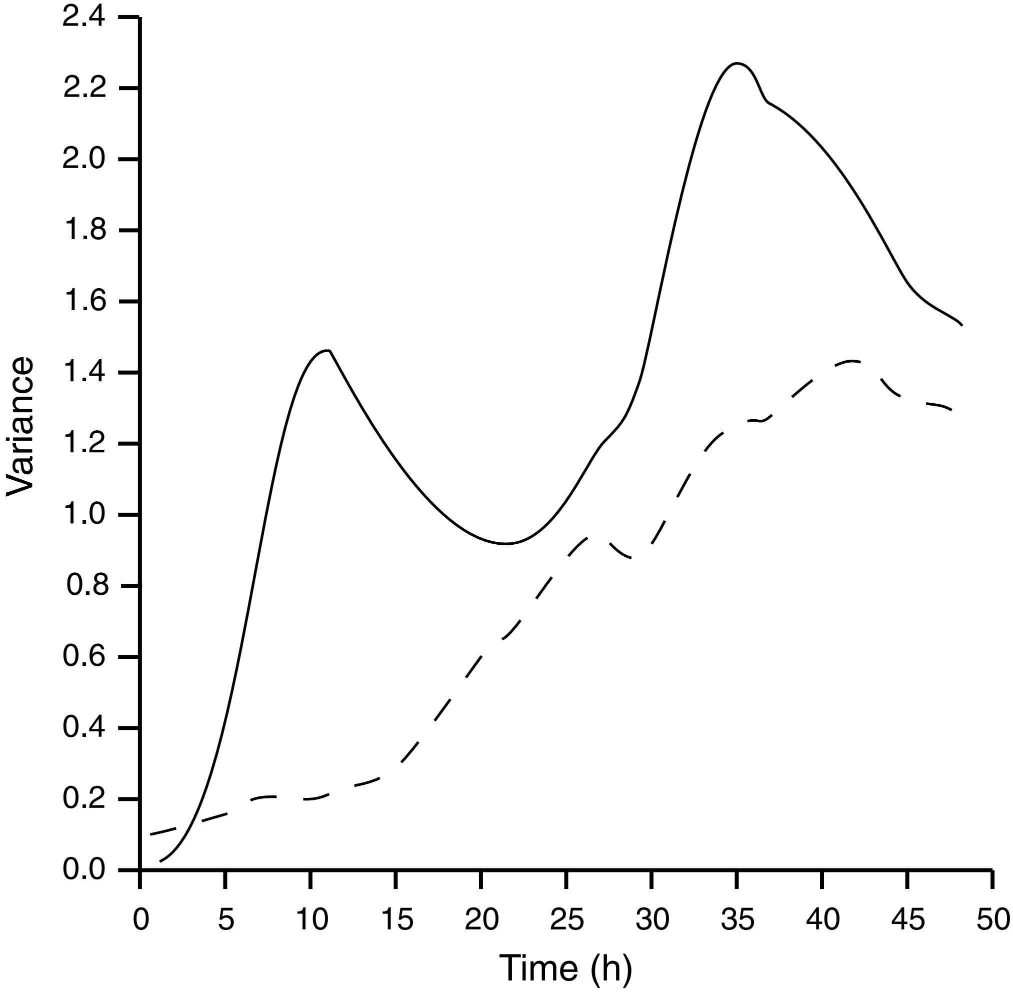


Fig. 7.16 The variance in the 850-hPa specific humidity as a function of forecast lead time for a 19-member physics ensemble (solid line) and a 19-member initial-condition ensemble (dashed line) that were run with MM5 for the same case of a long-lived mesoscale convective system. Adapted from Stensrud *et al.* (2000).

7.8.5 Ensemble plots from coupled special-applications models

Special-applications models that use atmospheric model output fields will be discussed in Chapter 14. This section shows example displays of variables produced by the secondary models that have used ensemble products from atmospheric models as input. For example, gridded ensemble forecasts or simulations may be used as input to air-quality models or plume-dispersion models that calculate the transport of gases or aerosols released into the atmosphere. Figure 7.17 illustrates stamp maps of dosages from plumes of gas whose transport and diffusion have been calculated using the Second-order Closure Integrated PUFF (SCIPUFF) plume model (Sykes *et al.* 1993), which has employed gridded meteorological products from a 12-member ensemble that is based on the MM5 regional model.

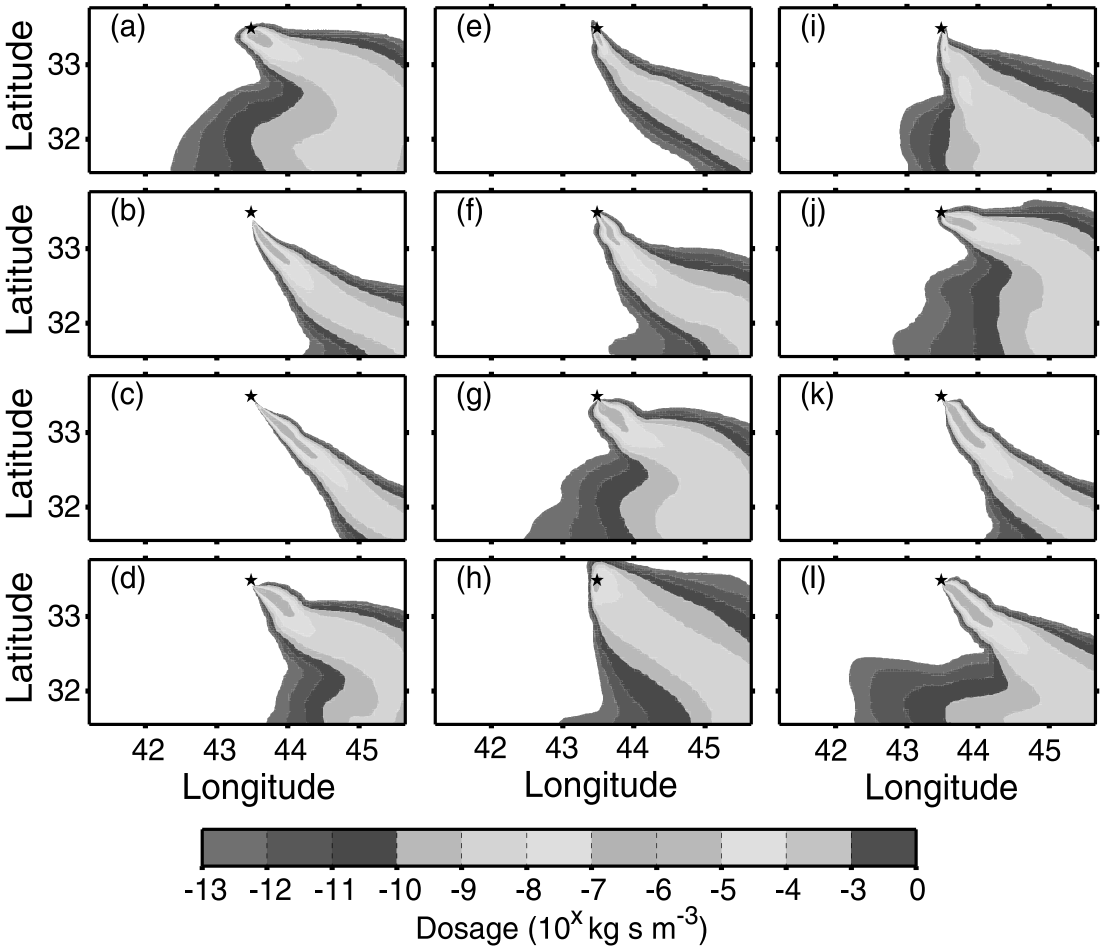


Fig. 7.17 Stamp maps of dosages from plumes of gas whose transport and diffusion have been calculated using the SCIPUFF plume model (Sykes *et al.* 1993), which has employed gridded meteorological products from a 12-member ensemble that is based on the MM5 regional model. From Warner *et al.* (2002).

The dosages show considerable sensitivity to the input from the meteorological ensemble. Also shown (Fig. 7.18) is a probability-of-exceedance plot that is based on the dosages in Fig. 7.17.

Atmospheric-model forecasts are also routinely used for estimating the future (e.g., next day) demand for electricity. The near-surface temperature, cloud cover, etc., are used as input to energy-demand models, and forecasts of the demand are plotted as a function of forecast lead time. When ensemble atmospheric models are employed, stochastic energy-demand products are produced, which can be plotted as lines on a meteogram or in the

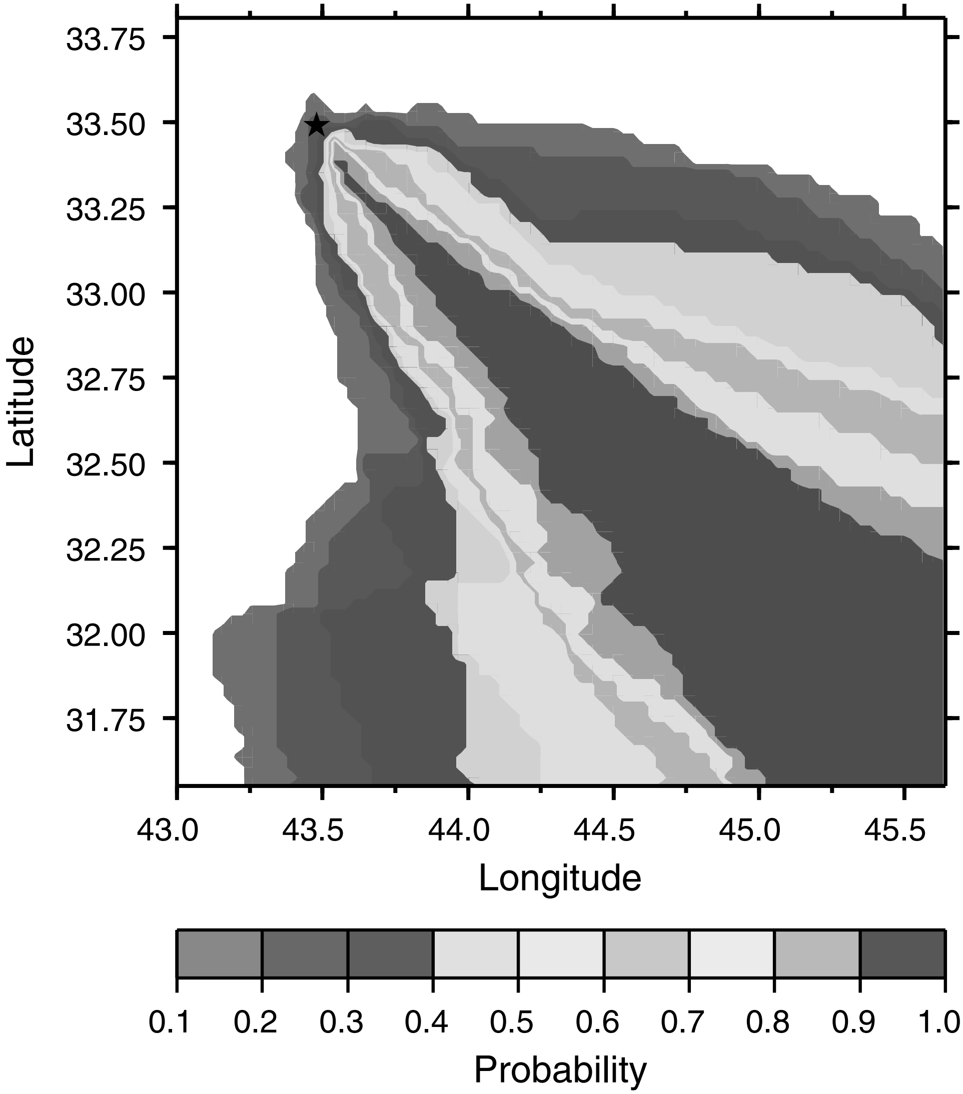


Fig. 7.18 Probability that the dosage of a plume of gas will exceed a specified threshold, based on the ensemble of plume predictions shown in Fig. 7.17. The star in the upper left shows the location of the release. From Warner *et al.* (2002).

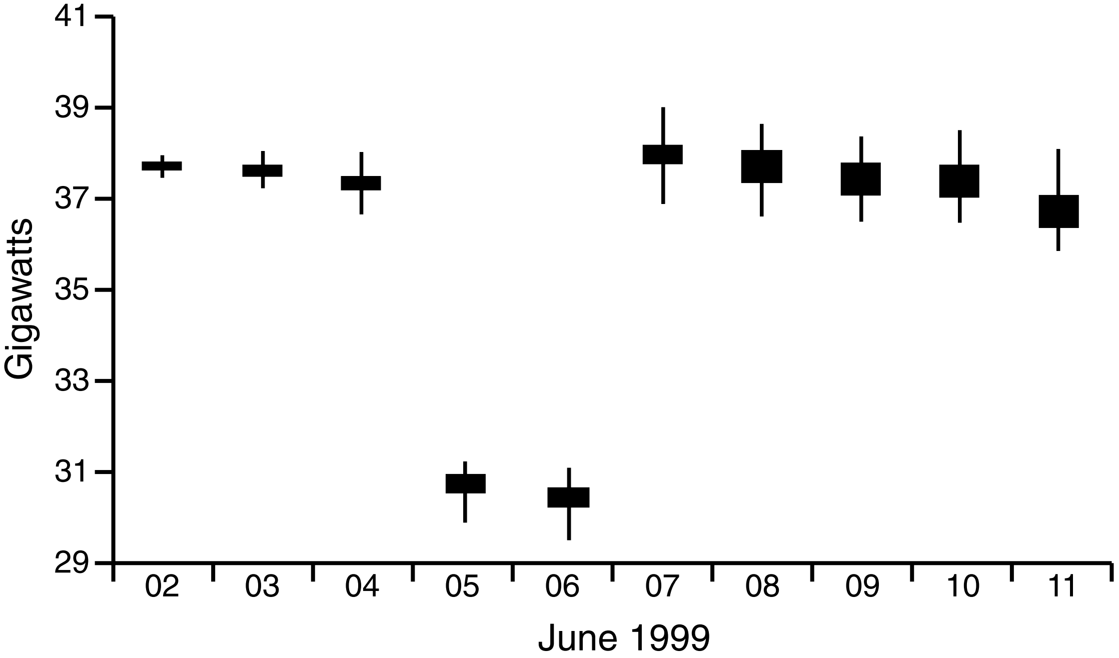


Fig. 7.19 An example of an "electricity-gram", which displays the uncertainty in a 10-day forecast of electric-energy demand for England and Wales. The meteorological input to an energy-demand model was provided by an ECMWF ensemble-prediction system. All of the predicted demand values lie within the vertical whiskers, and the middle 50% of the predictions fall within the box. Provided by James W. Taylor, University of Oxford.

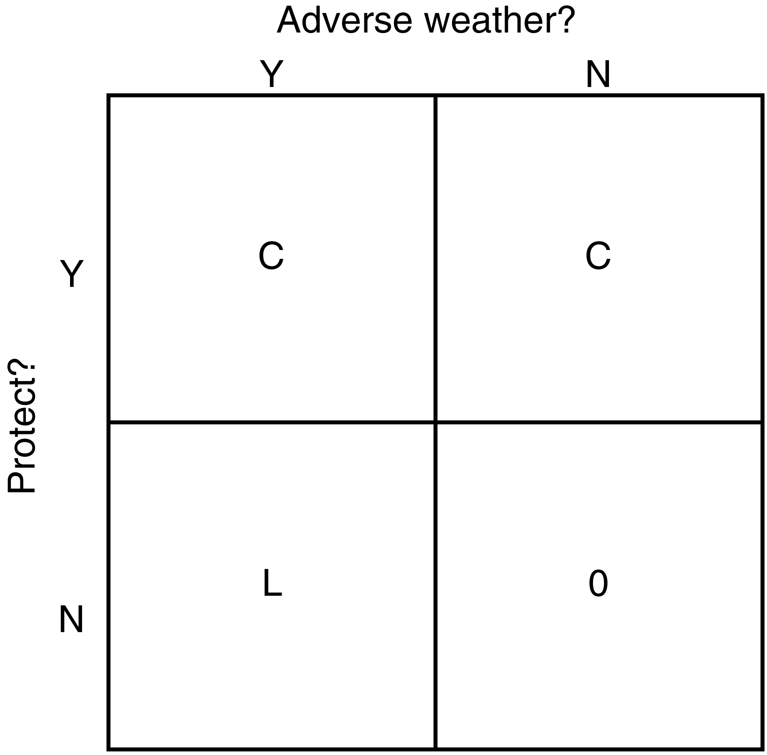
form of the "electricity-gram" in Fig. 7.19. Here, forecasted average-daily demand for England and Wales is plotted in terms of the total range of the predictions and the range of the middle 50% of the demand forecasts.

7.9 Economic benefits of ensemble predictions

One of the stated motivations for performing ensemble predictions is that the probabilistic information can have greater value for making decisions than would the output from a deterministic prediction. Such value can be defined in the context of societal, environmental, or economic impacts, even though the economic benefits are more easy to quantify. Unfortunately, the skill and dispersion measures for the forecasts, discussed above, do not provide direct information about the value of the forecast information. This value, in fact, is dependent on the weather sensitivity and decision-making process of a forecast user group.

There are different frameworks for assessing value, where one of the most common is the *cost–loss model*. Given an uncertain prediction of whether an event will or will not occur, a decision maker has the option of choosing to either protect against the occurrence of the weather event or not protect against it. This is the simplest decision problem because there are only two possible actions (protect, not protect) and two possible outcomes (the event occurs, the event does not occur). Examples of the many potential events of economic consequence include sub-freezing temperatures that can damage agricultural crops, daily precipitation in excess of an amount that can produce flooding, heavy snowfall that can impact highway or air travel, or damaging wind speeds. A decision to protect against the event will incur a cost (C), whether or not the event actually occurs. A decision to not protect will result in a loss (L) if the event occurs. Figure 7.20 summarizes the cost–loss consequences of the different outcomes.

It is assumed that probabilistic forecasts for the dichotomous weather event are available, and, if their quality is sufficiently good, decisions with better economic outcomes will be possible. Assume that a calibrated ensemble forecast predicts that the probability of an event occurring is p. The optimal decision about whether to protect or not to protect will be the one yielding the smallest expected expense. If the decision is made to protect, the expense will be C with a probability of 1.0. If no protective action is taken, the



Protect?	Adverse weather? Y	Adverse weather? N
Y	C	C
N	L	0

Fig. 7.20 The costs (C) and losses (L) associated with the four possible combinations of the occurrence of adverse weather and the decision to protect against it. A decision to protect against the event will incur a cost (C), whether or not the event actually occurs. A decision to not protect will result in a loss (L) if the event occurs. There is no economic effect (0) if the event does not occur and no protective action is taken.

probability-weighted expense will be pL. Therefore, protecting against the risk will result in the smallest expense whenever

$$1.0C < pL,$$

or

$$\frac{C}{L} < p.$$

Thus, protecting against an event is the optimal action when the predicted probability of its occurrence is more than the ratio of the cost to the loss. If the predicted probability is less than the cost/loss ratio, not protecting is the least-expensive action. Because the costs and losses are very strongly dependent on the particular situation, the threshold for protection will differ. The above discussion is only applicable if $C < L$, because otherwise the protective action could not result in a gain.

The economic value (V) of forecasts can be defined using an expression that is similar to that employed for skill scores for meteorological forecasts. Where E is an expected expense,

$$V = \frac{E_{forecast} - E_{climate}}{E_{perfect} - E_{climate}}.$$

The quantity $E_{climate}$ is a default that represents the minimum of the expenses resulting from always protecting or never protecting. Always protecting incurs a constant cost, C, while never protecting results in losses $\bar{o}L$, where $\bar{o}$ is the climatological probability of the event occurring. With perfect forecasts, the protective action would only take place when the event was going to occur, so $E_{perfect} = \bar{o}C$. See Wilks (2006) and McCollor and Stull (2008b) for a complete discussion of the calculation of this value score.

The cost–loss decision model has been frequently applied to assess the economic benefits of ensemble prediction, for many specific applications that require decisions. These applications include hydroelectric reservoir operations (McCollor and Stull 2008a,b), medium-range flood prediction (Roulin 2007), temperature forecasts for the energy sector (Stensrud and Yussouf 2003), precipitation predictions (Mullen and Buizza 2002, Yuan *et al.* 2005), severe-weather forecasts (Legg and Mylne 2004), and air-quality prediction (Pagowski and Grell 2006). Additional discussion of the economic value of ensemble predictions is found in Richardson (2000, 2001).

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge, UK: Cambridge University Press.

Leutbecher, M., and T. N. Palmer (2007). Ensemble forecasting. *J. Computational Phys.*, **227**, 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774.

Palmer, T. N., G. J. Shutts, R. Hagedorn, *et al.* (2005). Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193, doi: 10.1146/annurev.earth.33.092203.122552.

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

PROBLEMS AND EXERCISES

1. Choose a nonlinear function and demonstrate the correctness of the inequality in Eq. 7.1.
2. Given that we know from Chapter 9 that smoother forecasts verify better with conventional statistics (e.g., RMSE, MAE), and that the ensemble mean must be smoother than the solution from individual members, is it true that the smoothness of the ensemble mean contributes to its superior performance?
3. Access the web site of an operational ensemble prediction system and observe how the spread of the ensemble members varies from day to day, both within the same forecast and from one forecast to another in the cycle.
4. When ensemble modeling systems are coupled with special-application models, such as discussed in Section 7.8.5, describe how the system might be calibrated in terms of the variables predicted by the coupled model rather than the meteorological variables.