Chapter 8

Ensemble Forecasting

8.1. BACKGROUND

8.1.1. Inherent Uncertainty of Dynamical Forecasts

It was noted in Section 1.3 that dynamical chaos ensures that the future behavior of the atmosphere cannot be known with certainty. Because the atmosphere can never be fully observed, either in terms of spatial coverage or accuracy of measurements, a fluid-dynamical model of its behavior will always begin calculating forecasts from a state at least slightly different from that of the real atmosphere. These models (and other nonlinear dynamical systems, including the real atmosphere) exhibit the property that solutions (forecasts) started from only slightly different initial conditions will yield quite different results for lead times sufficiently far into the future. For synoptic-scale weather predictions, "sufficiently far" is a matter of (at most) weeks, and for mesoscale forecasts this window is even shorter, so that the problem of sensitivity to initial conditions is of practical importance.

Dynamical forecast models are the mainstay of weather and climate forecasting, and the inherent uncertainty of their results must be appreciated and quantified if their information is to be utilized most effectively. For example, a single deterministic forecast of the hemispheric 500 mb height field two days in the future is at best only one member of an essentially infinite collection of 500 mb height fields that could plausibly occur. Even if this deterministic forecast of the 500 mb height field is the best possible single forecast that can be constructed, its usefulness and value will be enhanced if aspects of the probability distribution of which it is a member can be estimated and communicated. It is the purpose of ensemble forecasting to characterize the inherent uncertainty of dynamical forecasts, in a quantitative yet understandable way. Although much more attention has been devoted to initial-condition sensitivity in weather forecasts, the issue is also important for longer range forecasts and climate change projections (e.g., Deser et al., 2012; Hawkins et al., 2016).

8.1.2. Stochastic Dynamical Systems in Phase Space

Understanding the conceptual and mathematical underpinnings of ensemble forecasting requires the concept of a *phase space*. A phase space is an abstract geometrical space, each of the coordinate axes of which corresponds to one of the forecast variables of the dynamical system. Within the phase space, the "state" of the dynamical system is defined by specification of particular values for each of these forecast variables, and therefore corresponds to a single point in this (generally high-dimensional) space.

To concretely introduce the phase space concept, consider the behavior of a swinging pendulum, which is a simple dynamical system that is commonly encountered in textbooks on physics or differential equations. The state of the dynamics of a pendulum can be completely described by two variables: its

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00008-0

angular position and its velocity. At the extremes of the pendulum's arc, its angular position is maximum (positive or negative) and its velocity is zero. At the bottom of its arc the angular position of the swinging pendulum is zero and its speed (corresponding to either a positive or negative velocity) is maximum. When the pendulum finally stops, both its angular position and velocity are zero. Because the motions of a pendulum can be fully described by two variables, its phase space is two-dimensional. That is, its phase space is a phase plane. The changes through time of the state of the pendulum system can be described by a path, known as an *orbit*, or *trajectory*, on this phase plane.

Figure 8.1 shows the trajectory of a hypothetical pendulum in its phase space. That is, this figure is a graph in phase space that represents the motions of a pendulum and their changes through time. The trajectory begins at the single point corresponding to the initial state of the pendulum: it is dropped from the right with zero initial velocity (A). As it drops it accelerates and acquires leftward velocity, which increases until the pendulum passes through the vertical position (B). The pendulum then decelerates, slowing until it stops at its maximum left position (C). As the pendulum drops again it moves to the right, stopping short of its initial position because of friction (D). The pendulum continues to swing back and forth until it finally comes to rest in the vertical position (E).

The dynamics of a pendulum are simple both because the phase space has only two dimensions, but also because its behavior is not sensitive to the initial condition. Releasing the pendulum slightly further to the right or left relative to its initial point in Figure 8.1, or with a slight upward or downward push, would produce a very similar trajectory that would track the spiral in Figure 8.1 very closely, and arrive at the same place in the center of the diagram at nearly the same time.

The corresponding behavior of the atmosphere, or of a realistic mathematical model of it, would be quite different. The landmark paper of Lorenz (1963) demonstrated that solutions to systems of deterministic nonlinear differential equations can exhibit sensitive dependence on initial conditions. That is, even though deterministic equations yield unique and repeatable solutions when integrated forward from a given initial condition, projecting systems exhibiting sensitive dependence forward in time from very slightly different

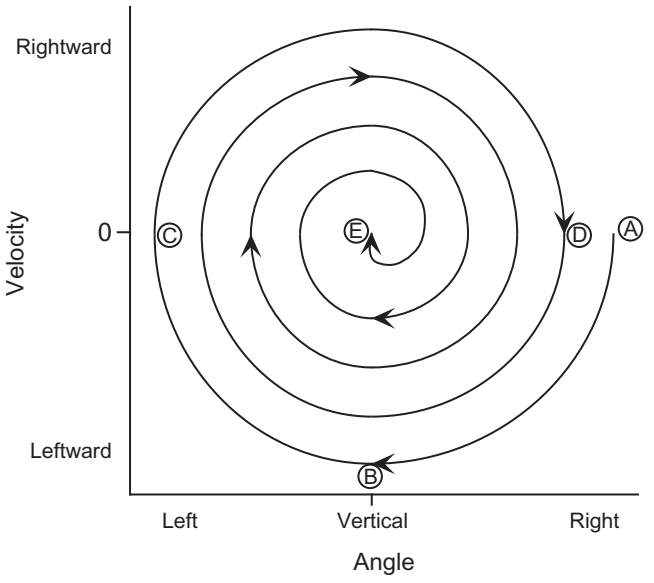

FIGURE 8.1 Trajectory of a swinging pendulum in its two-dimensional phase space or phase plane. The pendulum has been dropped from position (A) on the right, from which point it swings in arcs of decreasing angle. Finally, it slows to a stop, with zero velocity in the vertical position (E).

initial conditions eventually yields computed states that diverge strongly from each other. Sometime later Li and Yorke (1975) coined the name "chaotic" dynamics for this phenomenon, although this label is somewhat unfortunate in that it is not really descriptive of the sensitive-dependence phenomenon.

The system of three coupled ordinary differential equations used by Lorenz (1963) is deceptively simple:

$$\frac{dx}{dt} = -10x + 10y \tag{8.1a}$$

$$\frac{dy}{dt} = -xz + 28x - y \tag{8.1b}$$

$$\frac{dz}{dt} = xy - \frac{8}{3}z. \tag{8.1c}$$

This system is a highly abstracted representation of thermal convection in a fluid, where x represents the intensity of the convective motion, y represents the temperature difference between the ascending and descending branches of the convection, and z represents departure from linearity of the vertical temperature profile. Despite its low dimensionality and apparent simplicity, the system composed of Equation 8.1a–8.1c shares some key properties with the equations governing atmospheric flow. In addition to sensitive dependence, the simple Lorenz system and the equations governing atmospheric motion also both exhibit regime structure, multiple distinct timescales, and state-dependent variations in predictability (e.g., Palmer, 1993).

Because Equation 8.1a–8.1c involve three prognostic variables, the phase space of this system is a three-dimensional volume. However, not all points in its phase space can be visited by system trajectories once the system has settled into a steady state. Rather, the system will be confined to a subset of points in the phase space called the *attractor*. The attractor of a dynamical system is a geometrical object within the phase space, toward which trajectories are attracted in the course of time. Each point on the attractor represents a dynamically self-consistent state, jointly for all of the prognostic variables. The attractor for the simple pendulum system consists of the single point in the center of Figure 8.1, representing the motionless state toward which any undisturbed pendulum will converge. Figure 8.2, from

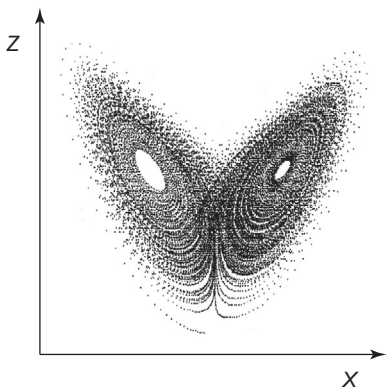

FIGURE 8.2 Projection of a finite approximation of the Lorenz (1963) attractor onto the x–z plane, yielding the Lorenz "butterfly." *From Palmer (1993). © American Meteorological Society. Used with permission.*

Palmer (1993), shows an approximate rendering of the Lorenz attractor, projected onto the x–z plane. The figure has been constructed by numerically integrating the Lorenz system forward for an extended time, with each dot representing the system state at a discrete time increment. The characteristic shape of this projection of the Lorenz attractor has come to be known as the Lorenz "butterfly." In a sense, the attractor can be thought of as representing the "climate" of its dynamical system, and each point on the attractor represents a possible instantaneous "weather" state. A sequence of these "weather" states then traces out a trajectory in the phase space, along the attractor.

Each wing of the attractor in Figure 8.2 represents a regime of the Lorenz system. Trajectories in the phase space consist of some number of clockwise circuits around the left-hand ($x < 0$) wing of the attractor, followed by a shift to the right-hand ($x > 0$) wing of the attractor where some number of counterclockwise cycles are executed, until the trajectory shifts again to the left wing, and so on. Circuits around one or the other of the wings occur on a faster timescale than residence times on each wing. The traces in Figure 8.3, which are example time series of the x variable, illustrate that the fast oscillations around one or the other wings are variable in number, and that transitions between the two wing regimes occur suddenly. The two traces in Figure 8.3 were initialized at very similar points, and the sudden difference between them that begins after the first regime transition illustrates the sensitive-dependence phenomenon.

The Lorenz system and the real atmosphere share the very interesting property of state-dependent variations in predictability. That is, forecasts initialized in some regions of the phase space (corresponding to particular subsets of the dynamically self-consistent states) may yield better predictions than others. Figure 8.4 illustrates this idea for the Lorenz system by tracing the trajectories of loops of initial conditions initialized at different parts of the attractor. The initial loop in Figure 8.4a, on the upper part of the left wing, illustrates extremely favorable forecast evolution. These initial points remain close together throughout the 10-stage forecast, although of course they would eventually diverge if the integration were to be carried further into the future. The result is that the forecast from any one of these

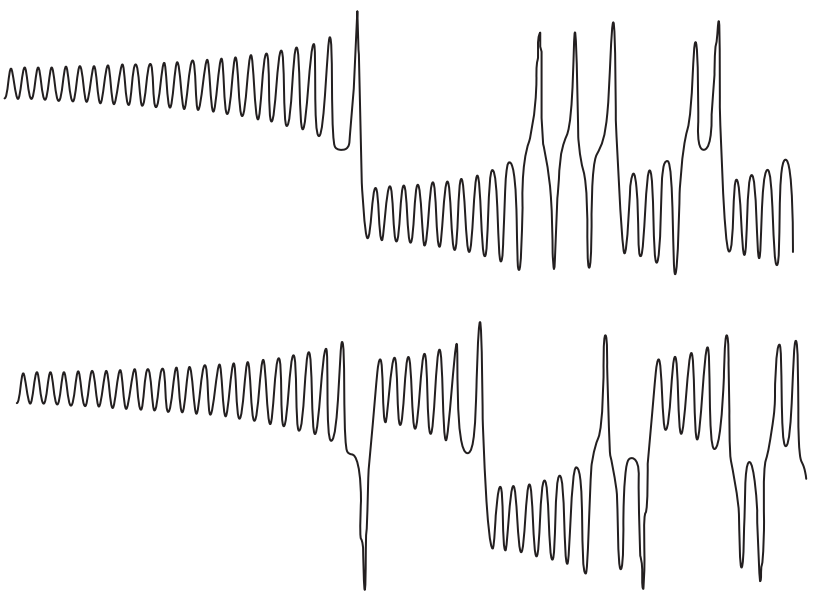

FIGURE 8.3 Example time series for the x variable in the Lorenz system. The two time series have been initialized at nearly identical values. *From Palmer (1993). © American Meteorological Society. Used with permission.*

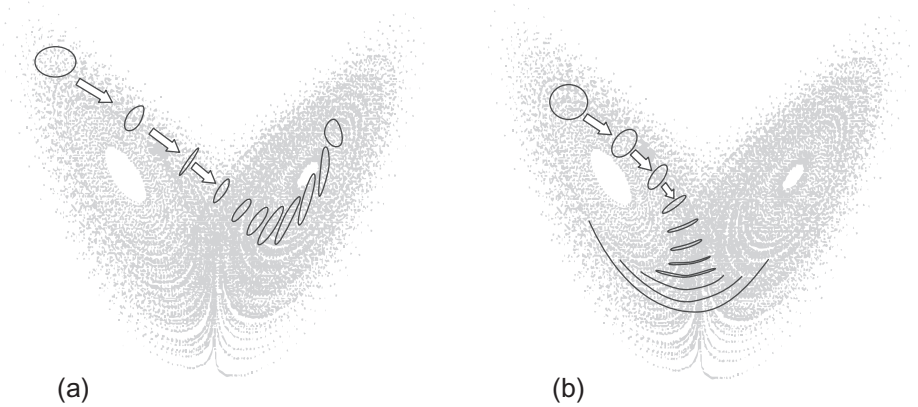

FIGURE 8.4 Collections of forecast trajectories for the Lorenz system, initialized at (a) a high-predictability portion of the attractor, and (b) a moderate-predictability portion of the attractor. Any of the forecasts in panel (a) would likely represent the unknown true future state well, whereas many of the results in panel (b) would correspond to poor forecasts. *From Palmer (1993). © American Meteorological Society. Used with permission.*

initial states would produce a good forecast of the trajectory from the (unknown) true initial condition, which might be located near the center of the initial loop. In contrast, Figure 8.4b shows forecasts for the same set of future times when the initial conditions are taken as the points on the loop that is a little lower on the left wing of the attractor. Here the dynamical predictions are reasonably good through the first half of the forecast period, but then diverge strongly toward the end of the period as some of the trajectories remain on the left-hand wing of the attractor while others undergo the regime transition to the right-hand wing. The result is that a broad range of the prognostic variables might be forecast from initial conditions near the unknown true initial condition, and there is no way to tell in advance which of the trajectories might represent good or poor forecasts.

The phase space of a realistic atmospheric model has many more dimensions than that of the Lorenz system. For example, the dimensions of the phase spaces of operational weather forecasting models are on the order of 10^9 (e.g., Lewis, 2014), each corresponding to one of the (horizontal gridpoints) $\times$ (vertical levels) $\times$ (prognostic variables) combinations represented. The trajectory of the atmosphere or a model of the atmosphere is more complicated than that of the Lorenz system, but the qualitative behavior is analogous in many ways, and the changes in the flow within a model atmosphere through time can still be imagined abstractly as a trajectory through its multidimensional phase space.

8.1.3. Uncertainty and Predictability in Dynamical Systems

The connection between uncertainty, probability, and dynamical forecasting can be approached using the phase space of the Lorenz attractor as a low-dimensional and comprehensible metaphor for the millions-dimensional phase spaces of realistic modern dynamical weather prediction models. Consider again the forecast trajectories portrayed in Figure 8.4. Rather than regarding the upper-left loops as collections of initial states, imagine that they represent boundaries containing most of the probability, perhaps the 99% probability ellipsoids, for probability density functions defined on the attractor. When initializing a dynamical forecast model we can never be certain of the true initial state, but we may be able to quantify that initial-condition uncertainty in terms of a probability distribution, and that distribution must be defined on the system's attractor if the possible initial states are to be dynamically consistent with the

governing equations. In effect, those governing equations will operate on the probability distribution of initial-condition uncertainty, advecting it across the attractor and distorting its initial shape in the process. If the initial probability distribution is a correct representation of the initial-condition uncertainty, and if the governing equations are a correct representation of the dynamics of the true system, then the subsequent advected and distorted probability distributions will correctly quantify the forecast uncertainty at future times. This uncertainty may be larger (as represented by Figure 8.4b) or smaller (Figure 8.4a), depending on the intrinsic predictability of the states in the initial region of the attractor. To the extent that the forecast model equations are not complete and correct representations of the true dynamics, which is inevitable as a practical matter, then additional uncertainty will be introduced.

Using this concept of a probability distribution that quantifies the initial-condition uncertainty, Epstein (1969c) proposed the method of *stochastic-dynamic prediction*. The historical and biographical background leading to this important paper has been reviewed by Lewis (2014). Denoting the (multivariate) uncertainty distribution as φ and the vector $\dot{\boldsymbol{X}}$ as containing the total derivatives with respect to time of the prognostic variables defining the coordinate axes of the phase space, Epstein (1969c) begins with the conservation equation for total probability, φ,

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot \left(\dot{\boldsymbol{X}} \varphi \right) = 0. \tag{8.2}$$

Equation 8.2, also known as the Liouville equation (Ehrendorfer, 1994a; Gleeson, 1970), is analogous to the more familiar continuity (i.e., conservation) equation for mass. As noted by Epstein (1969c),

> *It is possible to visualize the probability density in phase space, as analogous to mass density (usually* ρ*) in three-dimensional physical space. Note that* $\rho \geq 0$ *for all space and time, and* $\int\int\int (\rho/M) dx dy dz = 1$ *if* M *is the total mass of the system. The "total probability" of any system is, by definition, one.*

Equation 8.2 states that any change in the probability contained within a small (hyper-) volume surrounding a point in phase space must be balanced by an equal net flux of probability through the boundaries of that volume. The governing physical dynamics of the system (e.g., Equations 8.1a–8.1c for the Lorenz system) are contained in the time derivatives $\dot{\boldsymbol{X}}$ in Equation 8.2, which are also known as tendencies. Note that the integration of Equation 8.2 is deterministic, in the sense that there are no random terms introduced on the right-hand sides of the dynamical tendencies.

Epstein (1969c) considered that direct numerical integration of Equation 8.2 on a set of gridpoints within the phase space was computationally impractical, even for the idealized 3-dimensional dynamical system he used as an example. Instead he derived time-tendency equations for the elements of the mean vector and covariance matrix of φ, in effect, assuming multivariate normality (Chapter 12) for this distribution initially and at all forecast times. The result was a system of nine coupled differential equations (three each for the means, variances, and covariances), arrived at by assuming that the third and higher moments of the forecast distributions vanished. In addition to providing a (vector) mean forecast, the procedure characterizes state-dependent forecast uncertainty through the forecast variances and covariances that populate the forecast covariance matrix, the increasing determinant ("size") of which (Equation 12.6) at increasing lead times can be used to characterize the increasing forecast uncertainty. Concerning this procedure, Epstein (1969c) noted that "deterministic prediction implicitly assumes that all variances are zero. Thus the approximate stochastic equations are higher order approximations … than have previously been used."

Stochastic-dynamic prediction in the phase space in terms of the first and second moments of the uncertainty distribution does not yield good probabilistic forecasts because of the neglect of the higher

moment aspects of the forecast uncertainty. Even so, the method is computationally impractical when applied to realistic forecast models for the atmosphere. Full evaluation of Equation 8.2 (Ehrendorfer, 1994b; Thompson, 1985) is even further out of reach for practical problems.

8.2. ENSEMBLE FORECASTS

8.2.1. Discrete Approximation to Stochastic-Dynamic Forecasts

Even though the stochastic-dynamic approach to forecasting as proposed by Epstein (1969c) is out of reach computationally, it is theoretically sound and conceptually appealing. It provides the philosophical basis for addressing the problem of sensitivity to initial conditions in dynamical weather and climate models, which is currently best achieved through *ensemble forecasting*. Rather than computing the effects of the governing dynamics on the full continuous probability distribution of initial-condition uncertainty (Equation 8.2), ensemble forecasting proceeds by constructing a discrete approximation to this process. That is, a collection of individual initial conditions (each represented by a single point in the phase space) is chosen, and each is integrated forward in time according to the governing equations of the (modeled) dynamical system. Ideally, the distribution of these states in the phase space at future times, which can be mapped to physical space, will then represent a sample from the statistical distribution of forecast uncertainty. These Monte Carlo solutions bear the same relationship to stochastic-dynamic forecast equations as the Monte Carlo resampling tests introduced in Section 5.3.3 bear to the analytical tests they approximate. (Recall that resampling tests are appropriate and useful in situations where the underlying mathematics are difficult or impossible to evaluate analytically.) Lewis (2005) traces the history of this confluence of dynamical and statistical ideas in atmospheric prediction.

Ensemble forecasting is an instance of Monte Carlo integration (Metropolis and Ulam, 1949). Ensemble forecasting in meteorology appears to have been first proposed explicitly in a conference paper by Lorenz (1965):

> *The proposed procedure chooses a finite ensemble of initial states, rather than the single observed initial state. Each state within the ensemble resembles the observed state closely enough so that the differences might be ascribed to errors or inadequacies in observation. A system of dynamic equations previously deemed to be suitable for forecasting is then applied to each member of the ensemble, leading to an ensemble of states at any future time. From an ensemble of future states, the probability of occurrence of any event, or such statistics as the ensemble mean and ensemble standard deviation of any quantity, may be evaluated.*

Ensemble forecasting was first implemented in a meteorological context by Epstein (1969c) as a means to provide representations of the true forecast distributions to which his truncated stochastic-dynamic calculations could be compared. He explicitly chose initial ensemble members as independent random draws from the initial-condition uncertainty distribution:

> *Discrete initial points in phase space are chosen by a random process such that the likelihood of selecting any given point is proportional to the given initial probability density. For each of these initial points (i.e. for each of the sample selected from the ensemble) deterministic trajectories in phase space are calculated by numerical integration… Means and variances are determined, corresponding to specific times, by averaging the appropriate quantities over the sample.*

The ensemble forecast procedure begins in principle by drawing a finite sample from the probability distribution describing the uncertainty of the initial state of the atmosphere. Imagine that a few members

of the point cloud surrounding the mean estimated atmospheric state in phase space are picked randomly. Collectively, these points are called the ensemble of initial conditions, and each represents a plausible initial state of the atmosphere consistent with the uncertainties in observation and analysis. Rather than explicitly predicting the movement of the entire initial-state probability distribution through the phase space of the dynamical model, that movement is approximated by the collective trajectories of the ensemble of sampled initial points. It is for this reason that the Monte Carlo approximation to stochastic-dynamic forecasting is known as ensemble forecasting. Each of the points in the initial ensemble provides the initial conditions for a separate dynamical integration. At this initial time, all the ensemble members are very similar to each other. The distribution in phase space of this ensemble of points after the forecasts have been advanced to a future time then approximates how the full true initial probability distribution would have been transformed by the governing physical laws that are expressed in the dynamics of the model.

Figure 8.5 illustrates the nature of ensemble forecasting in an idealized two-dimensional phase space. The circled X in the initial-time ellipse represents the single best initial value, from which a conventional deterministic dynamical integration would begin. Recall that, for a real model of the atmosphere, this initial point defines a full set of meteorological maps for all of the variables being forecast. The evolution of this single forecast in the phase space, through an intermediate forecast lead time and to a final forecast lead time, is represented by the heavy solid lines. However, the position of this point in phase space at the initial time represents only one of the many plausible initial states of the atmosphere consistent with errors in the analysis. Around it are other plausible states, which are meant to sample the probability distribution for states of the atmosphere at the initial time. This distribution is represented by the small

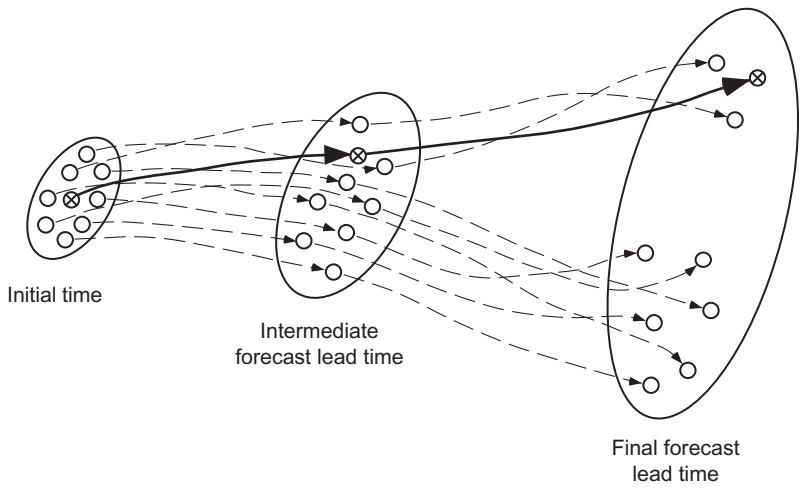

FIGURE 8.5 Schematic illustration of some concepts in ensemble forecasting, plotted in terms of an idealized two-dimensional phase space. The heavy line represents the evolution of the single best analysis of the initial state of the atmosphere, corresponding to the more traditional single deterministic forecast. The dashed lines represent the evolution of individual ensemble members. The ellipse in which they originate represents the probability distribution of initial atmospheric states, which are very close to each other. At the intermediate lead time, all the ensemble members are still reasonably similar. By the final lead time some of the ensemble members have undergone a regime change and represent qualitatively different flows. Any of the ensemble members, including the solid line, are plausible trajectories for the evolution of the real atmosphere, and there is no way of knowing in advance which will represent the real atmosphere most closely.

ellipse. The open circles in this ellipse represent eight other members of this distribution. This ensemble of nine initial states approximates the variations represented by the full distribution from which they were drawn.

The Monte Carlo approximation to a stochastic-dynamic forecast is constructed by repeatedly running the dynamical model, once for each of the members of the initial ensemble. The trajectories of each of the ensemble members through the phase space are only modestly different at first, indicating that all nine integrations represented in Figure 8.5 are producing fairly similar forecasts at the intermediate lead time. Accordingly, the probability distribution describing uncertainty about the state of the atmosphere at the intermediate lead time would not be a great deal larger than at the initial time. However, between the intermediate and final lead times the trajectories diverge markedly, with three (including the one started from the central value of the initial distribution) producing forecasts that are similar to each other, and the remaining six members of the ensemble predicting rather different atmospheric states at that time. The underlying distribution of uncertainty that was fairly small at the initial time has been stretched substantially, as represented by the large ellipse at the final lead time. The dispersion of the ensemble members at that time allows the nature of that distribution to be estimated, and is indicative of the uncertainty of the forecast, assuming that the dynamical model includes only negligible errors in the representations of the governing physical processes. If only the single forecast started from the best initial condition had been made, this information would not be available.

Reviews of recent operational use of the ensemble forecasting approach can be found in Buizza et al. (2005), Cheung (2001), and Kalnay (2003).

8.2.2. Choosing Initial Ensemble Members

Ideally, we would like to produce ensemble forecasts based on a large number of possible initial atmospheric states drawn randomly from the PDF of initial-condition uncertainty in phase space. However, each member of an ensemble of forecasts is produced by a complete rerunning of the dynamical model, each of which requires a substantial amount of computing. As a practical matter, computer time is a limiting factor at operational forecast centers, and each center must make a subjective judgment balancing the number of ensemble members to include in relation to the spatial resolution of the model used to integrate them forward in time. Consequently, the sizes of operational forecast ensembles are limited, and it is important that initial ensemble members be chosen well. Their selection is further complicated by the fact that the initial-condition PDF in phase space is unknown. Also, it presumably changes from day to day, so that the ideal of simple random samples from this distribution cannot be achieved in practice.

The simplest, and historically first, method of generating initial ensemble members was to begin with a best analysis, assuming it to be the mean of the probability distribution representing the uncertainty of the initial state of the atmosphere. Variations around this mean state can be easily generated, by adding random numbers characteristic of the errors or uncertainty in the instrumental observations underlying the analysis (Leith, 1974). For example, these random values might be Gaussian variates with zero mean, implying an unbiased combination of measurement and analysis errors. In practice, however, simply adding independent random numbers to a single initial field has been found to yield ensembles whose members are too similar to each other, probably because much of the joint variation introduced in this way is dynamically inconsistent, so that the corresponding energy is quickly dissipated in the model (Palmer et al., 1990). The consequence is that the dispersion of the resulting forecast ensemble

underestimates the uncertainty in the forecast as the trajectories of the initial ensemble members quickly collapse onto the attractor.

As of the time of this writing (2018), there are three dominant methods of choosing initial ensemble members, although a definitively best method has yet to be demonstrated (e.g., Hamill and Swinbank, 2015). Until relatively recently, the United States National Centers for Environmental Prediction (NCEP) used the *breeding method* (Ehrendorfer, 1997; Kalnay, 2003; Toth and Kalnay, 1993, 1997; Wei et al., 2008). In this approach, differences in the three-dimensional patterns of the predicted variables, between the ensemble members and the single "best" (control) analysis, are chosen to look like differences between recent forecast ensemble members and the forecast from the corresponding previous control analysis. The underlying idea is that the most impactful initial-condition errors should resemble forecast errors in the most recent ("background") forecast. The patterns are then scaled to have magnitudes appropriate to analysis uncertainties. These bred patterns are different from day to day and emphasize features with respect to which the ensemble members are diverging most rapidly. The breeding method is relatively inexpensive computationally.

In contrast, the European Centre for Medium-Range Weather Forecasts (ECMWF) generates initial ensemble members using *singular vectors* (Bonavita et al., 2012; Buizza, 1997; Ehrendorfer, 1997; Kalnay, 2003; Molteni et al., 1996). Here the fastest growing characteristic patterns of differences from the control analysis in a linearized version of the full forecast model are calculated, again for the specific weather situation of a given day. Linear combinations (in effect, weighted averages) of these patterns, with magnitudes reflecting an appropriate level of analysis uncertainty, are then added to the control analysis to define the ensemble members. There is theoretical support for the use of singular vectors to choose initial ensemble members (Ehrendorfer and Tribbia, 1997), although its use requires substantially more computation than does the breeding method.

The Meteorological Service of Canada and the U.S. NCEP both currently generate their initial ensemble members using a method called the *ensemble Kalman filter* (EnKF) (Burgers et al., 1998; Houtekamer and Mitchell, 2005). This method is related to the multivariate extension of conjugate Bayesian updating of a Gaussian prior distribution (Section 6.3.4). Here the ensemble members from the previous forecast cycle define the Gaussian prior distribution, and the ensemble members are updated using a Gaussian likelihood function (i.e., data-generating process) for available observed data assuming known data variance (characteristic of the measurement errors), to yield new initial ensemble members from a Gaussian posterior distribution. The initial ensembles are relatively compact as a consequence of their (posterior) distribution being constrained by the observations, but the ensemble members diverge as each is integrated forward in time by the dynamical model, producing a more dispersed prior distribution for the next update cycle. The UK Met Office uses a related technique known as the ensemble transform Kalman filter, or ETKF (Bowler et al., 2008; Bowler and Mylne, 2009; Wang and Bishop, 2003). Expositions and literature reviews for the EnKF are provided by Evensen (2003) and Hamill (2006).

Choice of ensemble size is another important, but not yet definitively resolved, question regarding the structuring of an ensemble forecasting system (Leutbecher, 2018). Fundamentally, this issue involves a trade-off between the number of ensemble members versus the spatiotemporal resolution of the dynamical model used to integrate them (and therefore the required computational cost for each member), in the face of a finite computing resource. At present, addressing this question involves both balancing the priorities related to the needs of different users of the forecasts and also consideration of which lead times to favor in optimizing forecast accuracy. Richardson (2001) and Mullen and Buizza (2002) concluded that the appropriate ensemble size depends on user needs, and in particular that

forecasts of rarer and more extreme events benefit from larger ensembles at the expense of dynamical model resolution. On the other hand, devoting more of the computational resource to improved model resolution is beneficial if forecast accuracy at shorter lead times (when overall uncertainty is lowest) has been prioritized, whereas the balance shifts toward larger ensembles of lower resolution integrations when medium- and long-range forecast accuracy is more important (Buizza, 2008, 2010; Ma et al., 2012; Machete and Smith, 2016; Raynaud and Bouttier, 2017). Ultimately, however, the balance should shift toward devoting new computational resources to increasing ensemble sizes. The reason is that error growth at the smallest scales (those newly resolved by an increased dynamical model resolution) both increases much faster than error growth at the larger scales of primary interest, and contaminates forecasts for the larger scales, so that a limit exists on the improvements to be achieved by increasing resolution (Lorenz, 1969; Palmer, 2014b).

8.2.3. Ensemble Mean and Ensemble Dispersion

One simple application of ensemble forecasting is averaging the members of the ensemble to obtain a single *ensemble mean* forecast. The motivation is to obtain a forecast that is more accurate than the single forecast initialized with the best estimate of the initial state of the atmosphere. Epstein (1969a) pointed out that the time-dependent behavior of the ensemble mean is different from the solution of forecast equations using the initial mean value, and concluded that in general the best forecast is not the single forecast initialized with the best estimate of initial conditions.

The first of these conclusions, at least, should not be surprising since a dynamical model is in effect a highly nonlinear function that transforms a set of initial atmospheric conditions to a set of forecast atmospheric conditions. In general, the average of a nonlinear function over some set of particular values of its argument is not the same as the function evaluated at the average of those values. That is, if the function $f(x)$ is nonlinear,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \tag{8.3}$$

To illustrate simply, consider the three values $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. For the nonlinear function $f(x) = x^2 + 1$, the left side of Equation 8.3 is 5 2/3, and the right side of that equation is 5. We can easily verify that the inequality of Equation 8.3 holds for other nonlinear functions (e.g., $f(x) = \log(x)$ or $f(x) = 1/x$) as well. For the linear function $f(x) = 2x + 1$ the two sides of Equation 8.3 are both equal to 5.

Extending this idea to ensemble forecasting, we might like to know the atmospheric state corresponding to the center of the ensemble in phase space for some time in the future. Ideally, this central value of the ensemble will approximate the center of the stochastic-dynamic probability distribution at that future time, after the initial distribution has been transformed by the nonlinear forecast equations. The Monte Carlo approximation to this future value is the ensemble mean forecast. The ensemble mean forecast is obtained simply by averaging together the ensemble members for the lead time of interest, which corresponds to the left side of Equation 8.3. In contrast, the right side of Equation 8.3 represents the single forecast started from the average initial value of the ensemble members. Depending on the nature of the initial distribution and on the dynamics of the model, this single forecast may or may not be close to the ensemble average forecast. The benefits of ensemble averaging appear to derive primarily from averaging out elements of disagreement among the ensemble members, while emphasizing

features that generally are shared by the members of the forecast ensemble. Empirically, ensemble means generally outperform single-integration forecasts, and theoretical support for this phenomenon is provided by Thompson (1977), Rougier (2016), and Christiansen, 2018).

Particularly for longer lead times, ensemble mean maps tend to be smoother than instantaneous states of the actual system, and so may seem unmeteorological, or more similar to smooth climatic averages. Palmer (1993) suggests that ensemble averaging will improve the forecast only until a regime change, or a change in the long-wave pattern, and he illustrates this concept nicely using the simple Lorenz (1963) model. This problem also is illustrated in Figure 8.5, where a regime change is represented by the bifurcation of the trajectories of the ensemble members between the intermediate and final lead times. At the intermediate lead time, before some of the ensemble members undergo this regime change, the center of the distribution of ensemble members is well represented by the ensemble average, which is a better central value than the single member of the ensemble started from the "best" initial condition. At the final forecast lead time the distribution of states has been distorted into two distinct groups. Here the ensemble average will be located somewhere in the middle, but near none of the ensemble members.

An especially important aspect of ensemble forecasting is its capacity to yield information about the magnitude and nature of the uncertainty in a forecast. In principle the forecast uncertainty is different on different forecast occasions, and this notion can be thought of as state-dependent predictability. The value to forecast users of communicating the different levels of forecast confidence that exist on different occasions was recognized early in the 20th century (Cooke, 1906b; Murphy, 1998). Qualitatively, we have more confidence that the ensemble mean is close to the eventual state of the atmosphere if the dispersion of the ensemble is small. Conversely, when the ensemble members are very different from each other the future state of the atmosphere may be more uncertain. One approach to "forecasting forecast skill" (Ehrendorfer, 1997; Kalnay and Dalcher, 1987; Palmer and Tibaldi, 1988) is to anticipate the accuracy of a forecast as being inversely related to the dispersion of the ensemble members. Operationally, forecasters do this informally when comparing the results from different dynamical models, or when comparing successive forecasts for a particular time in the future that were initialized on different days.

More formally, the *spread-skill relationship* for a collection of ensemble forecasts often is characterized by the correlation, over a collection of forecast occasions, between some measure of the ensemble spread such as the variance or standard deviation of the ensemble members around their ensemble mean on each occasion, and a measure of the predictive accuracy of the ensemble mean on that occasion. The idea is that forecasts entailing more or less uncertainty are then characterized by larger or smaller ensemble variances, respectively. The accuracy is often characterized using either the mean squared error (Equation 9.33) or its square root, although other measures have been used in some studies. These spread-skill correlations have sometimes been found to be fairly modest (e.g., Atger, 1999; Grimit and Mass, 2002; Hamill et al., 2004; Whittaker and Loughe, 1998), at least in part because of sampling variability of the chosen ensemble spread statistic, especially when the ensemble size is modest. Even so, some of the more recently reported spread-skill relationships (e.g., Sherrer et al. 2004; Stensrud and Yussouf, 2003) have been fairly strong. Figure 8.6 shows forecast accuracy, as measured by average root-mean squared error of ensemble members, as functions of ensemble spread measured by average root-mean squared differences among the ensemble members; for forecasts of 500 mb height over western Europe by the 51-member ECMWF ensemble prediction system for June 1997–December 2000. Clearly the more accurate forecasts (smaller RMSE) tend to be associated with smaller ensemble spreads, and vice versa, with this relationship being stronger for the shorter, 96-h lead time.

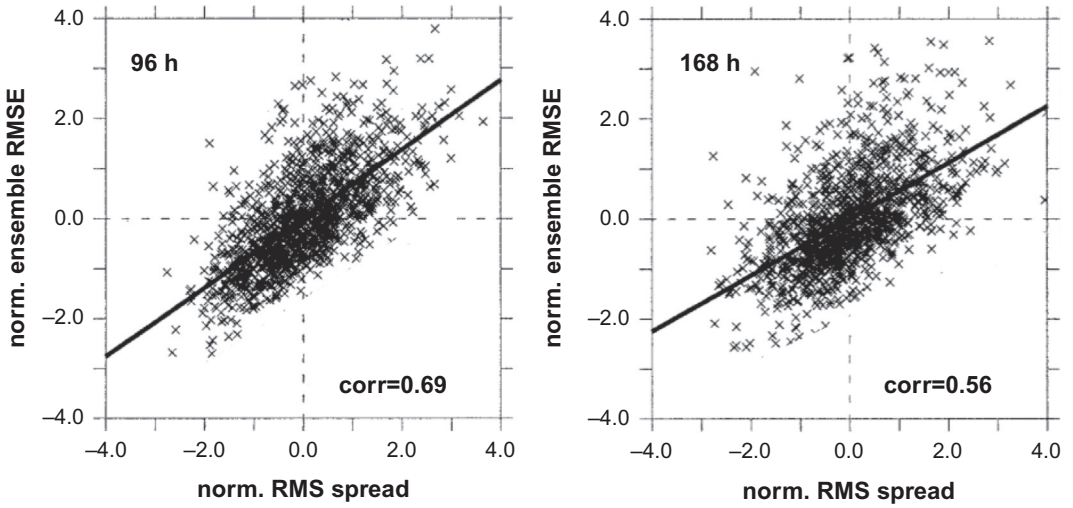

FIGURE 8.6 Scatterplots and correlations between forecast accuracy (vertical) and ensemble spread (horizontal) for ECMWF 500 mb height forecasts over western Europe, 1997–2000, at 96 h- and 168 h lead times. *Modified from Scherrer et al. (2004).* © *American Meteorological Society. Used with permission.*

8.2.4. Effects of Model Errors

Given a perfect dynamical model, integrating a random sample from the PDF of initial-condition uncertainty forward in time would yield a sample from the PDF characterizing forecast uncertainty. Of course dynamical models are not perfect, so that even if an initial-condition PDF could be known and correctly sampled from, the distribution of a forecast ensemble can at best be only an approximation to a sample from the true PDF for the forecast uncertainty (Hansen, 2002; Palmer, 2006; Smith, 2001).

Leith (1974) distinguished two kinds of model errors. The first derives from the models inevitably operating at a lower resolution than the real atmosphere or, equivalently, occupying a phase space of much lower dimension (Judd et al., 2008). Although still significant, this problem has been gradually addressed and partially ameliorated over the history of dynamical forecasting through progressive increases in model resolution. The second kind of model error derives from the fact that certain physical processes—prominently those operating at scales smaller than the model resolution—are represented incorrectly. In particular, such physical processes (known colloquially in this context as "physics") generally are represented using some relatively simple function of the explicitly resolved variables, known as a *parameterization*. Figure 8.7 shows a parameterization (solid curve) for the unresolved part of the tendency (dX/dt) of a resolved variable X, as a function of X itself, in the highly idealized Lorenz '96 (Lorenz, 2006) model. The individual points in Figure 8.7 are a sample of the actual unresolved tendencies, which are summarized by the regression function. In a realistic dynamical model there are a large number of such parameterizations for various unresolved physical processes, and the effects of these processes on the resolved variables are included in the model as functions of the resolved variables through these parameterizations. It is evident from Figure 8.7 that the parameterization (smooth curve) does not fully capture the range of behaviors for the parameterized process that are actually possible (scatter of points around the curve). Even if the large-scale dynamics have been modeled correctly, nature does not supply the value of the unresolved tendency given by "the" parameterized curve, but

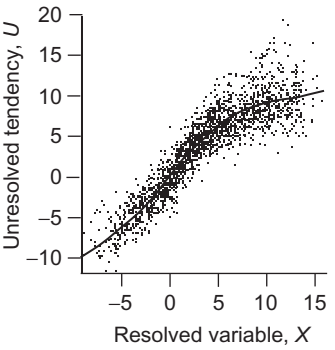

FIGURE 8.7 Scatterplot of the unresolved time tendency, U, of a resolved variable, X, as a function of the resolved variable; together with a regression function representing the average dependence of the tendency on the resolved variable. *From Wilks (2005).*

rather provides an effectively random realization from the point cloud around it. One way of looking at this kind of model error is that the effects of the parameterized physics are not fully determined by the resolved variables. That is, they are uncertain.

One way of representing the errors, or uncertainties, in the parameterized model physics is to extend the idea of the ensemble to include simultaneously a collection of different initial conditions *and* multiple dynamical models (each of which has a different collection of parameterizations). Harrison et al. (1999) found that ensemble forecasts using all four possible combinations of two sets of initial conditions and two dynamical model formulations differed significantly, with members of each of the four ensembles clustering relatively closely together, and distinctly from the other three, in the phase space. Other studies (e.g., Hansen, 2002; Houtekamer et al., 1996; Mullen et al., 1999; Mylne et al., 2002a; Stensrud et al., 2000) have found that using such *multimodel ensembles* (e.g., Kirtman et al., 2014) improves the resulting ensemble forecasts. The components of the Canadian Meteorological Center's operational multimodel ensemble share the same large-scale dynamical formulation, but differ with respect to the structure of various parameterizations (Houtekamer et al., 2009), in effect using different but similar parameterization curves of the kind represented in Figure 8.7, for different ensemble members.

A substantial part of the resulting improvement in ensemble performance derives from the multimodel ensembles exhibiting larger ensemble dispersion, so that the ensemble members are less like each other than if an identical dynamical model is used for all forecast integrations. Typically the dispersion of forecast ensembles is too small (e.g., Buizza, 1997; Stensrud et al., 1999; Toth and Kalnay, 1997), and so they express too little uncertainty about forecast outcomes (see Section 9.7).

Another approach to capturing uncertainties in the structure of dynamical models is suggested by the scatter around the regression curve in Figure 8.7. From the perspective of Section 7.2, the regression residuals that are differences between the actual (points) and parameterized (regression curve) behavior of the modeled system are random variables. Accordingly, the effects of parameterized processes can be more fully represented in a dynamical model if random numbers are added to the deterministic parameterization function, making the dynamical model explicitly stochastic (e.g., Berner et al., 2017; Palmer, 2001, 2012; Palmer et al., 2005b). Even if the system being modeled truly does not contain random components, adopting the stochastic view of unresolved, parameterized processes in a dynamical model may improve the resulting forecasts in terms of both ensemble spread and climatological fidelity (e.g., Buizza et al., 1999; Leutbecher et al., 2017; Ollinaho et al., 2017; Palmer, 2012; Tenant et al., 2011; Wilks, 2005).

The idea of stochastic parameterizations in dynamical models is not new, having been proposed as early as the 1970s (Lorenz, 1975; Moritz and Sutera, 1981; Pitcher, 1974, 1977). However, its use in

realistic atmospheric models has been relatively recent (Bowler et al., 2008; Buizza et al., 1999; Garratt et al., 1990; Leutbecher et al., 2017; Lin and Neelin, 2000, 2002; Sanchez et al., 2016; Williams et al., 2003). Particularly noteworthy is the first operational use of a stochastic representation of the effects of unresolved processes in the forecast model at ECMWF, which they called *stochastic physics*. Stochastic parameterization is still at a fairly early stage of development and continues to be the subject of ongoing research (e.g., Bengtsson et al., 2013; Berner et al., 2010, 2015; Christensen et al., 2017b; Frenkel et al., 2012; Neelin et al., 2010; Shutts, 2015).

Stochastic parameterizations also have been used in simplified climate models, to represent atmospheric variations on the timescale of weather, beginning the 1970s (e.g., Hasselmann, 1976; Lemke, 1977; Sutera, 1981), and in continuing work (e.g., Batté and Doblas-Reyes, 2015; Imkeller and von Storch, 2001; Imkeller and Monahan, 2002). Some papers applying this idea to prediction of the El Niño phenomenon are Christensen et al. (2017a), Penland and Sardeshmukh (1995), and Thompson and Battisti (2001).

8.3. UNIVARIATE ENSEMBLE POSTPROCESSING

8.3.1. Why Ensembles Need Postprocessing

In principle, initial ensemble members chosen at random from the PDF characterizing initial-condition uncertainty, and integrated forward in time with a perfect dynamical model, will produce an ensemble of future system states that is a random sample from the PDF characterizing forecast uncertainty. Ideally, then, the dispersion of a forecast ensemble characterizes the uncertainty in the forecast, so that small ensemble dispersion (all ensemble members similar to each other) indicates low uncertainty, and large ensemble dispersion (large differences among ensemble members) signals large forecast uncertainty.

In practice, the initial ensemble members are chosen in ways that do not randomly sample from the PDF of initial-condition uncertainty (Section 8.2.2), and errors in the dynamical models deriving mainly from processes operating on unresolved scales produce errors in ensemble forecasts just as they do in conventional single-integration forecasts. Accordingly, the dispersion of a forecast ensemble can at best only approximate the PDF of forecast uncertainty (Hansen, 2002; Smith, 2001). In particular, a forecast ensemble may reflect errors both in statistical location (most or all ensemble members being well away from the actual state of the atmosphere, but relatively nearer to each other) and dispersion (either under- or overrepresenting the forecast uncertainty). Often, operational ensemble forecasts are found to exhibit too little dispersion (e.g. Buizza, 1997; Buizza et al., 2005; Hamill, 2001; Toth et al., 2001; Wang and Bishop, 2005), which leads to overconfidence in probability assessment if ensemble relative frequencies are interpreted directly as estimating probabilities.

To the extent that ensemble forecast errors have consistent characteristics, they can be corrected through *ensemble-MOS* methods that summarize a historical database of these forecast errors, just as is done for single-integration dynamical forecasts. From the outset of ensemble forecasting (Leith, 1974) it was anticipated that use of finite ensembles would yield errors in the forecast ensemble mean that could be statistically corrected using a database of previous errors. MOS postprocessing (Section 7.9.2) is a more difficult problem for ensemble forecasts than for ordinary single-integration dynamical forecasts, or for the ensemble mean, because ensemble forecasts are susceptible to both the ordinary biases introduced by errors and inaccuracies in the dynamical model formulation, in addition to their usual underdispersion bias. Either or both of these kinds of problems in ensemble forecasts can be corrected using MOS methods.

Ultimately the goal of ensemble-MOS methods is to estimate a forecast PDF or CDF on the basis of the discrete approximation provided by a finite, m-member ensemble. If the effects of initial-condition errors and model errors were not important, this task could be accomplished by operating only on the ensemble members at hand, without regard to the statistical characteristics of past forecast errors. Probably the simplest such non-MOS approach is to regard the forecast ensemble as a random sample from the true forecast CDF and estimate cumulative probabilities from that CDF using a plotting-position estimator (Section 3.3.7). The so-called *democratic voting* method is a commonly used, although usually suboptimal, such estimator. Denoting the quantity being forecast, or verification, as V, and the distribution quantile whose cumulative probability is being estimated as q, this method computes

$$\Pr\{V \le q\} = \frac{1}{m}\sum_{k=1}^{m} I(q \le x_k) = \frac{\text{rank}(q) - 1}{m}, \tag{8.4}$$

where the indicator function $I(\bullet) = 1$ if its argument is true and is zero otherwise, and rank(q) indicates the rank of the quantile of interest in a hypothetical $m+1$ member ensemble consisting of the ensemble members x_k and that quantile. Equation 8.4 is equivalent to the Gumbel plotting position estimator (Table 3.2) and has the unfortunate property of assigning zero probability to any quantile less than the smallest ensemble member, $x_{(1)}$, and full unit probability to any quantile greater than the largest ensemble member, $x_{(m)}$.

Other plotting position estimators do not have these deficiencies. For example, using the Tukey plotting position (Wilks, 2006b) yields

$$\Pr\{V \le q\} = \frac{\text{Rank}(q) - 1/3}{(m+1) + 1/3}. \tag{8.5}$$

Katz and Ehrendorfer (2006) derive a cumulative probability estimator equivalent to the Weibull plotting position using a conjugate Bayesian analysis (Section 6.3.2) with a uniform prior distribution and a binomial likelihood for the ensemble members' binary forecasts of outcomes above or below q. However, the cumulative probability estimators in Equations 8.4 and 8.5 will still lead to inaccurate, overconfident results unless the ensemble size is large or the forecasts are reasonably skillful, even if the ensemble is free of bias errors and exhibits dispersion that is consistent with the actual forecast uncertainty (Richardson, 2001, see Section 9.7).

8.3.2. Nonhomogeneous Regression Methods

Direct transformation of a collection of ensemble forecasts using estimators such as Equation 8.5 will usually be inaccurate because of bias errors (e.g., observed temperatures warmer or cooler, on average, than the forecast temperatures) and/or dispersion errors (ensemble dispersion smaller or larger, on average, than required to accurately characterize the forecast uncertainty), which occur in general because of imperfect ensemble initialization and deficiencies in the structure of the dynamical model. Ordinary regression-based MOS postprocessing of single-integration dynamical forecasts through regression methods (Section 7.9.2) can be extended to compensate for ensemble dispersion errors also, by using an ensemble dispersion predictor in appropriately reformulated regression models. Adjusting

the dispersion of the ensemble according to its historical error statistics can allow information on possible state- or flow-dependent predictability to be included also in an ensemble-MOS procedure.

In common with other regression methods, *nonhomogeneous regressions* represent the conditional mean of the predictive distribution as an optimized linear combination of the predictors, which in this setting are generally the ensemble members. However, unlike the ordinary regression models described in Sections 7.2 and 7.3, in which the "error" and predictive variances are assumed to be constant (homogeneous), in nonhomogeneous regressions these variances are formulated to depend on the ensemble variance. This property allows the predictive distributions produced by nonhomogeneous regression methods to exhibit more uncertainty when the ensemble dispersion is large, and less uncertainty when the ensemble dispersion is small.

Nonhomogeneous Gaussian Regression

Nonhomogeneous Gaussian regression (NGR) was independently proposed by Jewson et al. (2004), and Gneiting et al. (2005), where it was named EMOS (for ensemble MOS). However, it is only one of many ensemble-MOS methods, so that the more descriptive and specific name NGR was applied to it by Wilks (2006b). The distinguishing feature of NGR among other nonhomogeneous regression methods is that it yields predictive distributions that are Gaussian.

Specifically, for each forecasting occasion t the Gaussian predictive distribution has mean and variance that are particular to the ensemble for that occasion,

$$y_t \sim N\left[\mu_t, \sigma_t^2\right], \tag{8.6}$$

where y_t is the quantity being predicted. The Gaussian predictive distribution has mean

$$\mu_t = a + b_1 x_{t,1} + b_2 x_{t,2} + \cdots + b_m x_{t,m} \tag{8.7}$$

and variance

$$\sigma_t^2 = c + d\, s_t^2, \tag{8.8}$$

where $x_{t,k}$ is the kth ensemble member for the tth forecast, s_t^2 is the ensemble variance,

$$s_t^2 = \frac{1}{m}\sum_{k=1}^{m} (x_{t,k} - \bar{x}_t)^2, \tag{8.9}$$

m is the ensemble size, and the ensemble mean is

$$\bar{x}_t = \frac{1}{m}\sum_{k=1}^{m} x_{t,k}. \tag{8.10}$$

The $m+3$ regression constants a, b_k, c, and d are the same for each forecast occasion t and need to be estimated using past training data. (Division by $m-1$ in Equation 8.9 will yield identical forecasts because the larger sample variance estimates will be compensated by a smaller estimate for the parameter d.) Usually the $x_{t,k}$ are dynamical-model predictions of the same quantity as y_t, but these can be any useful predictors (e.g., Messner et al., 2017).

Equation 8.7 is a general specification for the predictive mean, which would be appropriate when the m ensemble members are nonexchangeable, meaning that they have distinct statistical characteristics.

This is the case, for example, when a multimodel ensemble (i.e., an ensemble where more than one dynamical model is being used) is composed of single integrations from each of its m constituent models. Very often a forecast ensemble comprises exchangeable members (necessarily from the same dynamical model) having the same statistical characteristics, so that the regression coefficients b_k must be the same apart from estimation errors. In this common situation of statistically exchangeable ensemble members, Equation 8.7 is replaced by

$$\mu_t = a + b\,\bar{x}_t, \tag{8.11}$$

in which case there are four regression parameters a, b, c, and d to be estimated. Mean functions of intermediate complexity are also sometimes appropriate. For example, a two-model ensemble in which the members within each model are exchangeable would require two b coefficients operating on the two within-model ensemble means. Sansom et al. (2016) extend NGR for seasonal forecasting, where time-dependent biases deriving from progressive climate warming are important (e.g., Wilks and Livezey, 2013), by specifying

$$\mu_t = a + b_1\,\bar{x}_t + b_2 t \tag{8.12}$$

rather than using Equation 8.11.

Probabilistic NGR forecasts for the predictand y_t are computed using

$$\Pr\{y_t \le q\} = \Phi\left(\frac{q-\mu_t}{\sigma_t}\right), \tag{8.13}$$

where q is any quantile of interest in the predictive distribution, and $\Phi(\bullet)$ indicates the cumulative distribution function (CDF) of the standard Gaussian distribution. The predictive mean (Equation 8.7, or 8.11) corrects unconditional forecast bias (consistent over- or underforecasting) when $a \neq 0$, and corrects conditional forecast biases when the $b_k \neq 1/m$ in Equation 8.7 or $b \neq 1$ in Equation 8.11. The (square root of the) predictive variance (Equation 8.8) corrects dispersion errors, and further allows incorporation of any spread-skill relationship (i.e., positive correlation between ensemble spread and ensemble-mean error; e.g., Figure 8.6) into the forecast formulation. In Equation 8.8, ensembles exhibiting correct dispersion characteristics correspond to $c \approx 0$ and $d \approx 1$, larger values of d reflect stronger spread-skill relationships, and $d \approx 0$ indicates lack of a useful spread-skill relationship.

Unlike the situation for conventional linear regression, there is no analytic formula that can be used for NGR parameter fitting. An innovative idea proposed by Gneiting et al. (2005) is to estimate the NGR parameters a, b_k, c, and d by minimizing the average continuous ranked probability score (CRPS, Matheson and Winkler, 1976, see Section 9.5.1) over the training data. For Gaussian predictive distributions this is

$$\overline{\mathrm{CRPS}}_G = \frac{1}{n}\sum_{t=1}^{n}\sigma_t\left\{\frac{y_t-\mu_t}{\sigma_t}\left[2\Phi\left(\frac{y_t-\mu_t}{\sigma_t}\right)-1\right]+2\phi\left(\frac{y_t-\mu_t}{\sigma_t}\right)-\frac{1}{\sqrt{\pi}}\right\}, \tag{8.14}$$

where $\phi(\bullet)$ denotes the probability density function of the standard Gaussian distribution, n is the number of training samples, μ_t is defined by Equation 8.7 or Equation 8.11 as appropriate, and σ_t is the square root of the quantity in Equation 8.9. Both $c \ge 0$ and $d \ge 0$ are required mathematically, which can be achieved by setting $c = \gamma^2$ and $d = \delta^2$, and optimizing over γ and δ.

A more conventional parameter estimation approach, used by Jewson et al. (2004), is maximization of the Gaussian log-likelihood function

$$L_G = -\sum_{t=1}^{n}\left[\frac{(y_t-\mu_t)^2}{2\sigma_t^2} - \ln(\sigma_t)\right], \tag{8.15}$$

which formally assumes independence among the n training samples. Maximization of this objective function is equivalent to minimization of the average logarithmic score (Good, 1952), which is also known as the Ignorance score (Roulston and Smith, 2002, see Section 9.5.3). Gneiting et al. (2005) found that use of Equation 8.15 for estimation of NGR parameters yielded somewhat overdispersive predictive distributions, although Baran and Lerch (2016), Gebetsberger et al. (2017a), Prokosch (2013), Williams et al. (2014), and Williams (2016) have reported little difference in forecast performance between use of Equations 8.14 and 8.15 for estimating the NGR parameters.

Parameter estimates for many of the ensemble-MOS methods in addition to NGR can be computed using either CRPS or negative log-likelihood minimization. Both require iterative solutions. Maximum likelihood is typically less computationally demanding than CRPS minimization, although it is less robust to the influence of outlying extreme values (Gneiting and Raftery, 2007). Either approach can be modified by adding a penalty for lack of calibration (i.e., correspondence between forecast probabilities and subsequent predictand relative frequencies), Section 9.7.1, which may increase the economic value of the postprocessed forecasts to users (Wilks, 2018a).

Figure 8.8 shows an example NGR predictive distribution (solid), in relation to the raw $m = 35$-member ensemble (tick-marks on the horizontal axis, some of which represent multiple members), and a kernel density smoothing (Section 3.3.6) of them. The NGR distribution is shifted several degrees to the right relative to the raw ensemble, indicating that the training data used for fitting the parameters exhibited a cold bias. The dispersion of the NGR predictive distribution is wider than that of the raw ensemble, exemplifying correction of the typical underdispersion of raw ensemble forecasts. However, because the NGR predictive distribution is constrained to be Gaussian, it is incapable of representing the bimodality exhibited by the kernel density smoothing of the raw ensemble.

Wilks and Hamill (2007), Hagedorn et al. (2008), and Kann et al. (2009) have reported good results when postprocessing ensemble forecasts for surface temperatures (which are approximately Gaussian) using NGR, yielding substantial improvements in forecast skill over direct use of raw ensemble output (e.g., Equation 8.4 or 8.5).

More Flexible Nonhomogeneous Regression Distributions

By construction, the NGR formulation can yield only Gaussian predictive distributions. Some meteorological and climatological predictands are not Gaussian or approximately Gaussian, so that modifications to the nonhomogeneous regression framework are required to adequately represent such data. One possible approach is to subject the target predictand and its predictor counterparts in the ensemble to a Box-Cox ("power" or "normalizing") transformation (Equation 3.20), before fitting an NGR model (Hemri et al., 2015), which is intended to make the climatological distribution of the predictand y as nearly Gaussian as possible. The transformation exponent λ is then an additional parameter to be estimated.

In a special case of this approach, Baran and Lerch (2015, 2016) investigate use of nonhomogeneous lognormal regression, which is equivalent to NGR after the predictand y has been log-transformed. Predictive probabilities on the transformed scale are estimated using

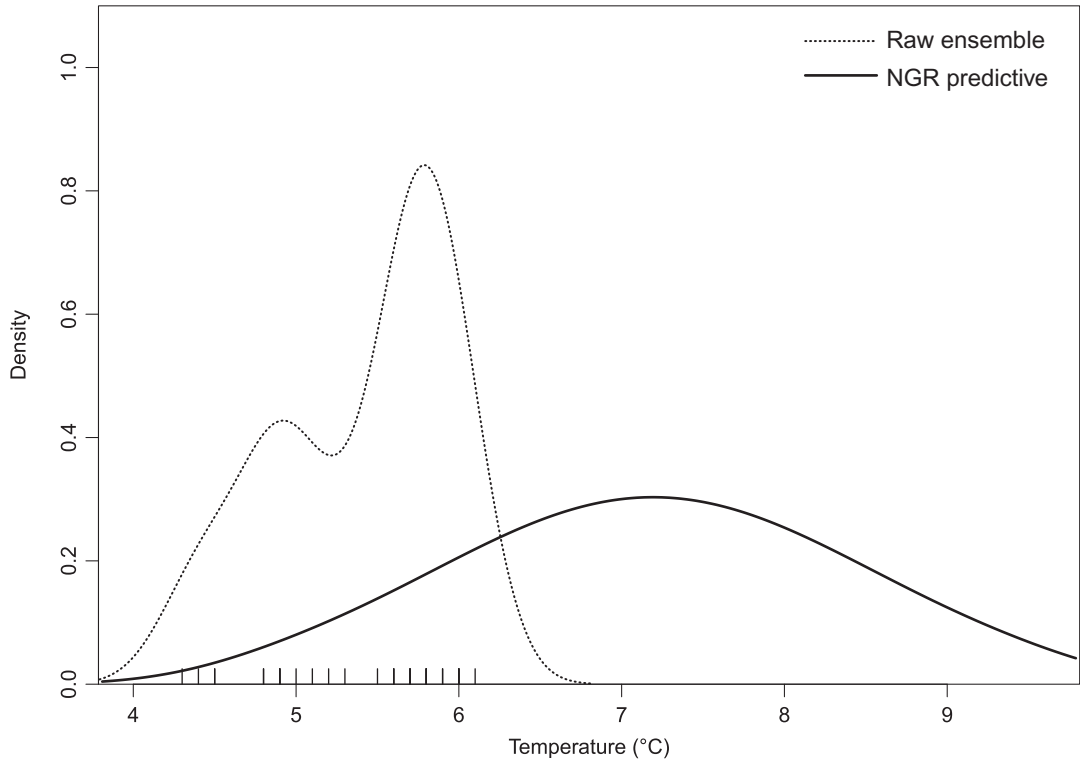

FIGURE 8.8 Example NGR (solid) predictive distribution, compared to the raw ensemble (tick-marks on horizontal axes) and smoothed predictive distribution based on it (dotted). *Modified from Taillardat et al. (2016). © American Meteorological Society. Used with permission.*

$$\Pr\{y_t \le q\} = \Phi\left[\left(\ln(q) - \ln\left(\frac{\mu_t^2}{\sqrt{\mu_t^2+\sigma_t^2}}\right)\right) \Big/ \sqrt{\ln\left(1+\frac{\sigma_t^2}{\mu_t^2}\right)}\right], \tag{8.16}$$

where σ_t^2 and μ_t are parameterized according to Equations 8.8 and 8.11 (for an exchangeable ensemble), respectively. The counterpart of Equation 8.14 for average lognormal CRPS is given in Baran and Lerch (2015).

Another possible approach to nonhomogeneous regression is to specify non-Gaussian predictive distributions. Messner et al. (2014a) modeled square-root transformed wind speeds using nonhomogeneous regressions with logistic (Section 4.4.4) predictive distributions,

$$\Pr\{y_t \le q\} = \frac{\exp[(q-\mu_t)/\sigma_t]}{1+\exp[(q-\mu_t)/\sigma_t]}. \tag{8.17}$$

Here the conditional means μ_t were modeled using Equation 8.11, and the corresponding (strictly positive) scale parameters were specified using

$$\sigma_t = \exp(c + d\, s_t). \tag{8.18}$$

Messner et al. (2014a) estimated the regression parameters in Equation 8.17 using maximum likelihood, and Taillardat et al. (2016) provide the expression necessary for minimum-CRPS estimation. Scheuerer and Möller (2015) investigate nonhomogeneous regressions with gamma-distributed predictive distributions and provide a closed-form CRPS equation.

In another approach to modeling positively skewed predictands in a nonhomogeneous ensemble regression setting, Lerch and Thorarinsdottir (2013) forecast probability distributions for maximum daily wind speed using generalized extreme-value (GEV) predictive distributions. In this case the prediction probabilities are given by

$$\Pr(y_t \le q\} = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{q-\mu_t}{\sigma_t}\right)\right]^{-1/\xi}\right\}, & \xi \ne 0 \\ \exp\left\{-\exp\left[-\left(\dfrac{q-\mu_t}{\sigma_t}\right)\right]\right\}, & \xi = 0 \end{cases} . \tag{8.19}$$

Lerch and Thorarinsdottir (2013) used Equation 8.11 to specify the location parameter μ_t but in their application found better results when the scale parameter depended on the ensemble mean according to

$$\sigma_t = \exp(c + d\,\overline{x}_t), \tag{8.20}$$

which ensures $\sigma_t > 0$ as required. Equation 8.20 does not allow explicit dependence on the ensemble spread, but does reflect the heteroscedastic nature of variables such as wind speeds and precipitation amounts which often exhibit greater variability for larger values. When $\xi > 0$ (found in the overwhelming majority of forecasts by Lerch and Thorarinsdottir, 2013), Equation 38.19 allows $y_t > \mu_t - \sigma_t/\xi$, so that GEV predictive distributions may yield nonzero probability for negative values of the predictand. These probabilities were quite low when maximum daily wind speeds in Lerch and Thorarinsdottir (2013) but could be more substantial if the predictand is daily precipitation.

Truncated Nonhomogeneous Regressions

The original NGR approach will often be appropriate when the predictand of interest exhibits a reasonably symmetric distribution, for example, temperature or sea-level pressure. When the distribution of the predictand is not symmetric but unimodal, and if any discontinuity in its distribution at zero is small, NGR for the transformed predictand or nonhomogeneous regressions with different predictive forms, as just described, may work well. However, strictly nonnegative predictands often exhibit large discontinuities in their probability densities at zero, which may prevent simple transformations from achieving adequate approximation to Gaussian shape, and which may be poorly represented by alternative conventional predictive distributions. This problem is especially acute in the case of relatively short-duration (such as daily) precipitation accumulations, which often feature a large probability "spike" at zero. Nonhomogeneous regression approaches based on truncation or censoring can be successful at addressing this issue.

Thorarinsdottir and Gneiting (2010) proposed using zero-truncated Gaussian predictive distributions (Equations 4.39–4.41) in a nonhomogeneous regression framework for forecasting positively skewed and nonnegative predictands. The average CRPS for the truncated Gaussian distribution (Thorarinsdottir and Gneiting, 2010; Taillardat et al., 2016), which is the counterpart to Equation 8.14, can again be used to fit the regression parameters. Having obtained these estimates, either Equation 8.7 or 8.11 is used to compute the location parameter μ_t for the current forecast, and Equation 8.8 is used to compute the scale parameter σ_t. Forecast probabilities can then be calculated using Equation 4.40.

Hemri et al. (2014) used zero-truncated Gaussian regressions following square-root transformation of the predictand. Scheuerer and Möller (2015) used nonhomogeneous regressions with truncated logistic predictive distributions (Equation 8.17). Junk et al. (2015) fit zero-truncated Gaussian regressions using training data restricted to near-analogs of the current forecast.

Observing that the heavier tails of the lognormal probability model better represent the distribution of stronger wind speeds whereas the truncated Gaussian model is better over the remainder of the distribution, Baran and Lerch (2016) investigated probability mixtures of nonhomogeneous lognormal and zero-truncated Gaussian predictive distributions, so that predictive probabilities were computed using

$$\Pr\{y_t \le q\} = w\Pr_{TG}\{y_t \le q\} + (1-w)\Pr_{LN}\{y_t \le q\}, \tag{8.21}$$

where the indicated truncated Gaussian (TG) and lognormal (LN) probabilities are specified by Equations 4.40 and 8.16, respectively. They found clearly better-calibrated predictive distributions derived from the mixtures as compared to the individual lognormal or truncated Gaussian distributions. A related model (Lerch and Thorarinsdottir, 2013) uses a regime-switching idea, where GEV predictive distributions (Equation 8.19) are used when the ensemble median is above a threshold θ, and truncated Gaussian predictive distributions (Equation 4.40) are used otherwise, with θ being an additional parameter requiring estimation. Baran and Lerch (2015) investigate a similar regime-switching model, where the predictive distribution is lognormal rather than GEV if the ensemble median is larger than θ.

Censored Nonhomogeneous Regressions

Censoring, in contrast to truncation, allows a probability distribution to represent values falling hypothetically below a censoring threshold, even though those values have not been observed (Section 4.4.3). In the context of ensemble postprocessing, the censoring threshold is generally zero, and any probability corresponding to negative predictand values is assigned to exactly zero, yielding a probability spike there. Specifying the parameters for these censored predictive distributions using regression models yields what is known in the economics literature as Tobit regression (Tobin, 1958). Censored formulations allow representation of a discontinuous probability spike at the lower limit of the predictive distribution, whereas truncated formulations are likely more appropriate when the variable in question is at least approximately continuous there.

Scheuerer (2014) describes a nonhomogeneous regression model having zero-censored GEV predictive distributions, so that that any nonzero probability for negative y_t is assigned to zero precipitation. Assuming $\xi \neq 0$ in Equation 8.19 (with an obvious modification for $\xi = 0$), predictive probabilities are computed using

$$\Pr(y_t \le q\} = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{q-\mu_t}{\sigma_t}\right)\right]^{-1/\xi}\right\}, & q \ge 0 \\ 0, & q < 0 \end{cases}, \tag{8.22}$$

so that all probability for $y_t \le 0$ is assigned to zero. Scheuerer (2014) linked the censored GEV parameters to the ensemble statistics using the relationships

$$\mu_t = a + b_1\bar{x}_t + b_2\sum_{k=1}^{m} I(x_{t,k}=0) - \frac{\sigma_t}{\xi}[\Gamma(1-\xi)-1], \tag{8.23}$$

where $I(\bullet)$ is the indicator function whose value is 1 if its argument is true, and is zero otherwise, $\Gamma(\bullet)$ denotes the gamma function, and the predictive scale parameter is specified using

$$\sigma_t = c + \frac{d}{m^2}\sum_{k=1}^{m}\sum_{j=1}^{m}\left|x_{t,k} - x_{t,j}\right|. \tag{8.24}$$

Here division by $m(m-1)$ rather than m^2 might be preferred since the m terms in Equation 8.24 for which $k=j$ will be identically zero (Ferro et al., 2008; Wilks, 2018a). The distribution parameters were estimated by minimizing average CRPS, with the shape parameter ξ assumed independent of the ensemble statistics as was also done by Lerch and Thorarinsdottir (2013).

Using a similar approach, Scheuerer and Hamill (2015a) and Baran and Nemoda (2016) implement nonhomogeneous regressions using zero-censored shifted-gamma (i.e., Pearson Type III, Equation 4.55) predictive distributions for precipitation amounts, where both the shape α_t and scale β_t parameters are required to be positive, and the shift parameter $\zeta_t < 0$ in this application. As is also the case for the Scheurer (2014) zero-censored GEV model, any nonzero probability for negative values of y_t is assigned to $y_t = 0$, so that

$$\Pr\{y_t \le q\} = \begin{cases} F_{\gamma(\alpha_t)}\left(\dfrac{q-\zeta_t}{\beta_t}\right), & q \ge 0 \\ 0, & q < 0 \end{cases}, \tag{8.25}$$

where $F_{\gamma(\alpha)}$ denotes the CDF for the standard ($\beta = 1$) 2-parameter gamma distribution (Section 4.4.5) with shape parameter α.

Messner et al. (2014b) proposed nonhomogeneous regression for square-root transformed wind speeds, using censored logistic predictive distributions. Probabilities on the transformed scale are then computed using

$$\Pr\{y_t \le q\} = \begin{cases} \dfrac{\exp[(q-\mu_t)/\sigma_t]}{1+\exp[(q-\mu_t)/\sigma_t]}, & q \ge 0, \\ 0, & \text{otherwise}, \end{cases} \tag{8.26}$$

where the dependence of the mean and scale parameters on the ensemble statistics were formulated according to Equations 8.11 and 8.18, respectively. Stauffer et al. (2017) and Gebetsberger et al. (2017b) further extended this censored-logistic approach for (power-transformed) precipitation forecasts, where in some cases all ensemble members may forecast zero, by defining the logistic distribution parameters as

$$\mu_t = a + b_1 I\left(\sum_{k=1}^{m} x_k = 0\right) + b_2 \bar{x}_t \left[1 - I\left(\sum_{k=1}^{m} x_k = 0\right)\right], \tag{8.27a}$$

and

$$\ln(\sigma_t) = c + d\,\ln(s_t)\left[1 - I\left(\sum_{k=1}^{m} x_k = 0\right)\right]. \tag{8.27b}$$

Thus regression specifications for both mean and standard deviation parameters are invoked if at least one ensemble member forecasts nonzero precipitation, and fixed $\mu_t = a + b_1$ and $\sigma_t = \exp(c)$ are used if all ensemble members are dry. The logarithms in Equation 8.27b ensure that the predictive standard deviation is always positive. Equation 8.27b allows the logarithmic link for the standard deviation to be defined even when all ensemble members are zero, in which case $\sigma_t = 0$.

An alternative to censoring for representing the finite probability of zero precipitation is to construct a mixture model where one element of the mixture is a probability p_t for zero precipitation (Bentzien and Friederichs, 2012). In that paper this probability was specified using logistic regression (Section 7.6.2), and then combined with either a gamma or lognormal distribution F_t for the nonzero precipitation amounts, yielding the probability specification

$$\Pr\{y_t \le q\} = p_t + (1 - p_t) F_t(q \mid q > 0). \tag{8.28}$$

Bentzien and Friederichs (2012) also proposed specifying probabilities for large extremes using generalized Pareto distributions, which arise in extreme-value theory (Section 4.4.8) for the distribution of values above a high threshold, appended to the right tails of the gamma or lognormal predictive distributions.

8.3.3. Logistic Regression Methods

Hamill et al. (2004) first proposed use of logistic regression (Section 7.6.2) for ensemble postprocessing, using the ensemble mean as the sole predictor, in an application where only the two climatological terciles were used as the prediction thresholds, q:

$$\Pr\{y_i \le q\} = \frac{\exp\left[a_q + b_q \overline{x}_t\right]}{1 + \exp\left[a_q + b_q \overline{x}_t\right]}. \tag{8.29}$$

When fitting logistic regression parameters, the training-data predictands are binary (as indicated by the dots in Figure 7.20), being one if the condition in curly brackets on the left-hand side of Equation 8.29 is true and zero otherwise.

As indicated by the subscripts in Equation 8.29, separate regression coefficients must in general be estimated for each predictand threshold q in logistic regression. Especially when logistic regressions are estimated for large numbers of predictand thresholds, it becomes increasingly likely that probability specifications using Equation 8.29 may be inconsistent, implying negative probabilities for some outcomes. Wilks (2009) proposed extending ordinary logistic regression in a way that unifies the regression functions for all quantiles of the distribution of the predictand, by assuming equal regression coefficients b_q, and specifying the regression intercepts as a nondecreasing function of the target quantile,

$$\Pr\{y_t \le q\} = \frac{\exp\left[a_q\, g(q) + a_0 + b\overline{x}_t\right]}{1 + \exp\left[a_q\, g(q) + a_0 + b\overline{x}_t\right]}. \tag{8.30}$$

The function $g(q) = \sqrt{q}$ was found to provide good results for the data used in that paper. Equation 8.30 ensures coherent probability specifications and requires estimation of fewer parameters than do conventional logistic regressions for multiple quantiles. The parameters are generally fit using maximum likelihood (Messner et al., 2014a) using a selected set of observed quantiles, but once fit Equation 8.30 can be applied for any value of q. This approach has come to be known as extended logistic regression (XLR).

The mechanism of XLR can most easily be appreciated by realizing that the regression function in Equation 8.30 is linear when expressed on the log-odds scale:

$$\ln\left(\frac{\Pr\{y_t \le q\}}{1 - \Pr\{y_t \le q\}}\right) = a_q\, g(q) + a_0 + b\overline{x}_t. \tag{8.31}$$

Figure 8.9a shows example XLR probability specifications for selected predictand quantiles q, compared with corresponding results in when logistic regressions (Equation 8.29) have been fit individually for the same predictand quantiles in Figure 8.9b. The common slope parameter b in Equations 8.30 and 8.31 force regressions for all quantiles to be parallel in log-odds in Figure 8.9a; whereas the individual logistic regressions in Figure 8.9b cross, leading in some cases to cumulative probability specifications for smaller precipitation amounts being larger than specifications for larger amounts.

Of course the logistic regression function inside the exponentials of Equations 8.29 or 8.30 can include multiple predictors, in the form $b_1x_{t,1} + b_2x_{t,2} + \ldots + b_mx_{t,m}$ analogously to Equation 8.7, where the various x's may be nonexchangeable ensemble members or other covariates. Messner et al. (2014a) point out that even if one or more of these involves ensemble spread, these forms do not explicitly represent any spread-skill relationship exhibited by the forecast ensemble, but that the logistic regression framework can be further modified to do so (Equation 8.17).

Equations 3.30 and 3.31 can be seen as a continuous extension of the proportional-odds logistic regression approach (McCullagh, 1980) for specifying cumulative probabilities for an ordered set of discrete outcomes. Messner et al. (2014a) applied proportional-odds logistic regression to prediction of the climatological deciles of wind speed and precipitation, using both ordinary (homoscedastic) and nonconstant-variance (heteroscedastic) formulations. Hemri et al. (2016) applied the more conventional homoscedastic proportional-odds logistic regression to postprocessing ensemble forecasts of cloud cover. The cloud-cover predictand is measured in "octas," or discrete eighths of the sky hemisphere, so that the nine ordered predictand values are $y_t = 0/8,\ 1/8,\ \ldots,\ 8/8$. The (homoscedastic) proportional-odds logistic regression forecasts are then formulated as

$$\ln\left(\frac{\Pr\{y_t \le q\}}{1 - \Pr\{y_t \le q\}}\right) = a_q + b_1x_{t,1} + b_2x_{t,2} + \cdots + b_mx_{t,m}, \tag{8.32}$$

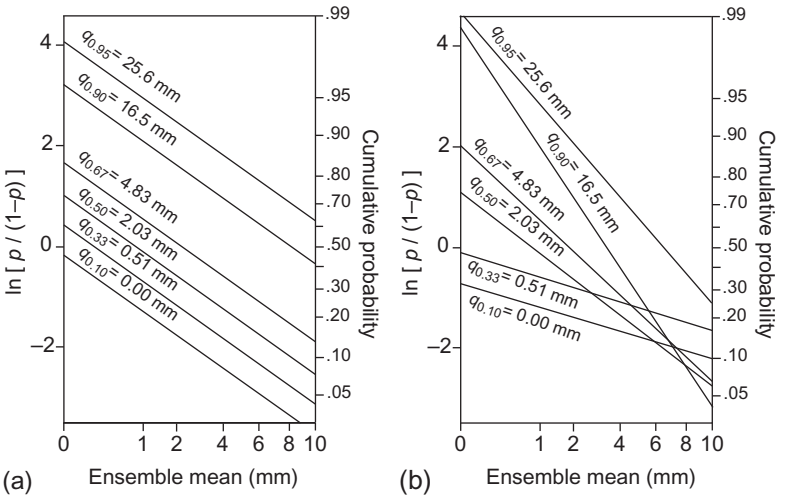

FIGURE 8.9 Logistic regressions plotted on the log-odds scale, for 28 November–2 December accumulated precipitation at Minneapolis, at 6–10 day lead time. Forecasts from Equation (8.30), evaluated at selected quantiles, are shown by the parallel lines in panel (a), which cannot yield logically inconsistent sets of forecasts. Regressions for the same quantiles, fitted separately using Equation (8.29), are shown in panel (b). Because these regressions are not constrained to be parallel, logically inconsistent forecasts are inevitable for sufficiently extreme values of the predictor. *From Wilks (2009).*

where $q = 0/8, 1/8, \ldots, 8/8$, as before the predictors $x_{t,k}$ might be nonexchangeable ensemble members and/or other statistics derived from the ensemble, and the intercepts are strictly ordered so that $a_{0/8} < a_{1/8} < \cdots < a_{8/8}$. The overall result is very much like that shown in Figure 8.9a, with regression functions that are parallel in the log-odd space and which have monotonically increasing intercepts a_q, but it differs in that intermediate functions between the plotted lines are not defined because of the discreteness of the predictand.

8.3.4. Bayesian Model Averaging and other "Ensemble Dressing" Methods

Bayesian model averaging (BMA), introduced as an ensemble postprocessing tool by Raftery et al. (2005), is the second of the two most commonly used ensemble-MOS methods, the other being nonhomogeneous regression (Section 8.3.2). Despite the name, it is not a fully Bayesian method in the sense of the approach presented in Chapter 6. In common with regression methods, BMA yields a continuous predictive distribution for the forecast variable y. However, rather than imposing a particular parametric form, BMA predictive distributions are mixture distributions (Section 4.4.9) or equivalently kernel density estimates (Section 3.3.6). That is, BMA predictive distributions are weighted sums of m component probability distributions, each centered at the corrected value of one of the m ensemble members being postprocessed. The BMA procedure is an example of the process sometimes referred to as "ensemble dressing," because it is the aggregate result of m probability distributions being metaphorically draped around each corrected ensemble member.

Construction of a BMA predictive distribution can be expressed in general as

$$f_{BMA}(y_t) = \sum_{k=1}^{m} w_k f_k(y_t), \tag{8.33}$$

where each w_k is the nonnegative weight associated with the component probability density $f_k(y_t)$ pertaining to the kth ensemble member $x_{t,k}$, and $\Sigma_k w_k = 1$. Figure 8.10 illustrates the process for a five-member ensemble of nonexchangeable members, where the component distributions are Gaussian, have been constrained to have equal variance, and the weights can be interpreted as probabilities that the respective members will provide the best forecast (Raftery et al., 2005). Figure 8.10 emphasizes that, even when the component densities are Gaussian, their weighted sum can take on a wide variety of shapes. In the case of Figure 8.10 the BMA density is bimodal, reflecting the bifurcation of this small ensemble into two groups.

In order for BMA and other ensemble dressing procedures to work correctly, the raw ensemble members must first be debiased, in order to correct systematic forecast errors exhibited in the available training data. Usually this initial debiasing step is accomplished using linear regressions, although more sophisticated methods could be used if appropriate. For nonexchangeable ensembles, separate regressions are fit for each ensemble member to reflect their different error statistics (Raftery et al., 2005), so that the bias correction is accomplished by centering each component distribution at mean

$$\mu_{t,k} = a_k + b_k x_{t,k}, k = 1, \ldots, m, \tag{8.34}$$

where as usual the regression coefficients minimize the average squared difference over the training period between the observed values y_t and the conditional regression specifications $\mu_{t,k}$. When the ensemble members are exchangeable, then these correction equations should be the same for each member, which can be accomplished by constraining the regression parameters in Equation 8.34 to be equal for all ensemble members (Wilson et al., 2007), or with a regression involving the ensemble mean as the sole predictor (Hamill, 2007; Williams et al., 2014),

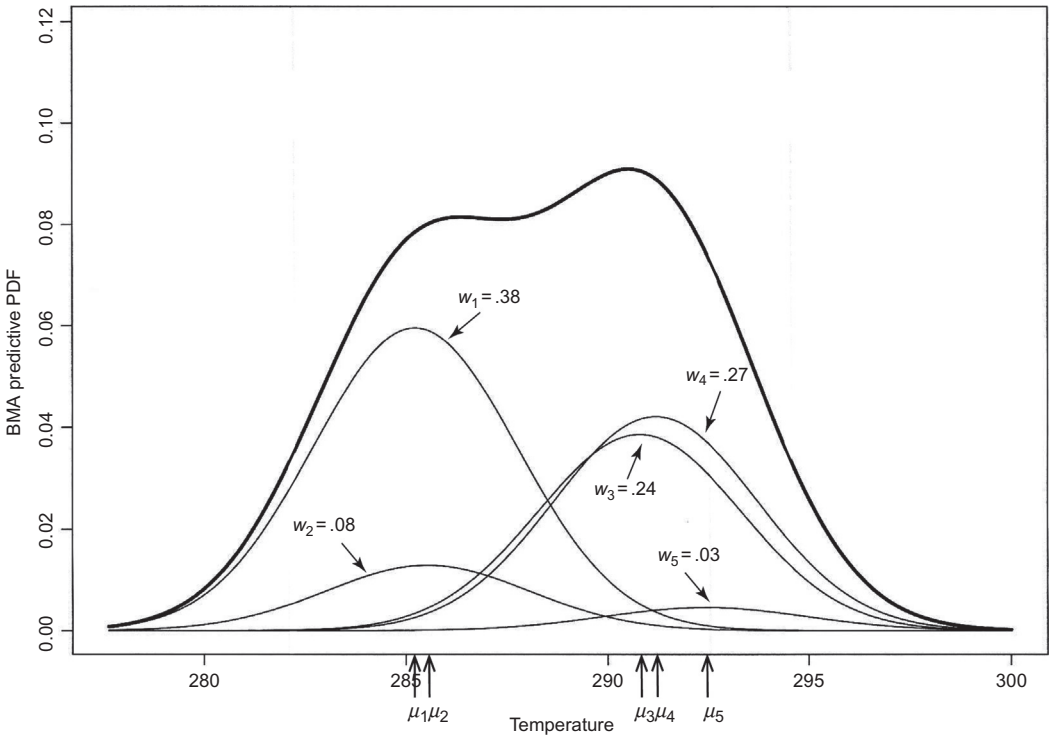

FIGURE 8.10 Illustration of a BMA predictive distribution (heavy curve) constructed from a bias-corrected ensemble of size $m = 5$. The magnitudes of the bias-corrected ensemble members are indicated by the arrows on the horizontal axis, and the weights for the five Gaussian components (light curves) correspond to their areas. The BMA predictive distribution is the weighted sum. *Modified from Raftery et al. (2005). © American Meteorological Society. Used with permission.*

$$\mu_{t,k} = (x_{t,k} - \overline{x}_t) + (a + b\overline{x}_t). \tag{8.35}$$

Very often Gaussian distributions are adopted for the component probability densities $f_k(y_t)$ in Equation 8.33, as illustrated in Figure 8.10. The corresponding component standard deviations σ_k and weights w_k are then usually estimated by maximizing the log-likelihood function

$$L_{\text{BMA}} = \sum_{t=1}^{n}\sum_{k=1}^{m}\left[\ln(w_k) - \frac{(y_t - \mu_{t,k})^2}{2\sigma_k^2} - \ln(\sigma_k)\right] \tag{8.36}$$

over the n available training samples. In this context maximum likelihood estimation is preferably computed using the Expectation-Maximization (EM) algorithm (Section 4.6.3), which is particularly convenient for fitting parameters of mixture distributions such as Equation 8.33 (e.g., Example 4.14). The Gaussian BMA predictive probabilities are then computed as

$$\Pr\{y_t \le q\} = \sum_{k=1}^{m} w_k \Phi\left(\frac{q - \mu_{t,k}}{\sigma_k}\right). \tag{8.37}$$

This equation reflects the fact that case-to-case differences in the spread of BMA predictive distributions derive from the dispersion of the corrected ensemble members, $\mu_{t,k}$, since the standard deviations σ_k of the component ("dressing") distributions are fixed.

When all ensemble members are exchangeable, then the standard deviations σ_k and weights w_k would be constrained to be equal (Wilks, 2006b; Wilson et al., 2007). Similarly if there are groups of exchangeable members within the ensemble (e.g., exchangeable members from each of several dynamical models), then these parameters (and also the debiasing means in Equation 8.34) would be equal within each group (Fraley et al., 2010).

As was also the case for the nonhomogeneous regression methods (Section 8.3.2), basing BMA post-processing on Gaussian component distributions may be inappropriate for predictands having skewed distributions, and/or those that can take on only nonnegative values. Duan et al. (2007) approach the first of these problems simply by forecasting Box-Cox (i.e., "power") transformed predictands (Equation 3.20).

The problem of nonnegative predictands, which is relevant especially for wind speed and precipitation amount forecasting, is somewhat more difficult. Sloughter et al. (2010) describe BMA forecasts for wind speed using gamma distributions (Section 4.4.5) for the component probability densities in Equation 8.33. They link the ensemble statistics to the parameters of these component gamma densities using Equation 8.34 for the means, and

$$\sigma_{t,k} = c + dx_{t,k}, \tag{8.39}$$

for the standard deviation, where $\mu_{t,k} = \alpha_{t,k}\beta_{t,k}$ and $\sigma_{t,k} = \beta_{t,k}\sqrt{\alpha_{t,k}}$ relate these regressions to the two gamma distribution parameters $\alpha_{t,k}$ and $\beta_{t,k}$ in Equation 4.45.

Baran (2014) proposed using truncated Gaussian component distributions in BMA for forecasting wind speeds, analogously to the nonhomogeneous regression approach of Thorarinsdottir and Gneiting (2010). The component probability density functions to be weighted in Equation 8.33 then have the same form as Equation 4.39, with the location parameters $\mu_{t,k}$ defined using Equation 8.34 and the scale parameter σ assumed to be the same for each ensemble member. Forecast probabilities are then computed using

$$\Pr\{y_t \le q\} = \sum_{k=1}^{m} w_k \left[\Phi\left(\frac{q-\mu_{t,k}}{\sigma}\right) - \Phi\left(\frac{-\mu_{t,k}}{\sigma}\right) \Big/ \Phi\left(\frac{\mu_{t,k}}{\sigma}\right)\right], \; q > 0. \tag{8.40}$$

A forecast precipitation distribution is usually more difficult to model, as it often consists of a discrete probability for exactly zero, and a continuous probability density for the nonzero amounts. Sloughter et al. (2007) describe BMA for such precipitation distributions, specifying the probability of exactly zero precipitation with the logistic regression

$$p_{t,k} = \frac{\exp\left[a_{0,k} + a_{1,k}x_{t,k}^{1/3} + a_{2,k}I(x_{t,k}=0)\right]}{1 + \exp\left[a_{0,k} + a_{1,k}x_{t,k}^{1/3} + a_{2,k}I(x_{t,k}=0)\right]}, \tag{8.41}$$

and a gamma distribution for transformed nonzero amounts, yielding the mixed discrete-continuous component distributions

$$f_k(y_t) = p_{t,k}I(y_t=0) + (1-p_{t,k})I(y_t>0)\frac{\left(y_t/\beta_{t,k}\right)^{\alpha_{t,k}-1}\exp\left(-y_t/\beta_{t,k}\right)}{\beta_{t,k}\Gamma(\alpha_{t,k})}. \tag{8.42}$$

Schmeits and Kok (2010) modified this approach slightly, specifying equal probabilities of zero precipitation for each ensemble member's component dressing distribution using the logistic regression

$$p_{t,k} = \frac{\exp\left[a_{0,k} + a_1 \sum_{k=1}^{m} x_{t,k}^{1/3}\right]}{1 + \exp\left[a_{0,k} + a_1 \sum_{k=1}^{m} x_{t,k}^{1/3}\right]}, \tag{8.43}$$

rather than Equation 8.41. Equations 8.41, and 3.43 indicate that Sloughter et al. (2007) and Schmeits and Kok (2010) found best results when working with cube-root transformed precipitation.

Figure 8.11 illustrates the construction of a BMA predictive distribution for precipitation using these mixed discrete and continuous component distributions. The heavy vertical bar at zero indicates that the weighted sum of the probabilities specified by $w_k\, p_{t,k}$ is approximately 0.37. The weighted component gamma distributions are shown as the light curves, each having area equal to $w_k\,(1 - p_{t,k})$, and their sum is the nonzero part of the predictive distribution indicated by the heavy curve. Note that this mixed discrete-continuous distribution form is different from both the truncated and censored distributions exemplified by Figs. 4.3 and 4.7, respectively, in that no part of the continuous distribution is defined for $y_{t,k} \leq 0$, but is similar in spirit to the regression mixture distribution (Equation 8.28) proposed by Bentzien and Friederichs (2012). Each $f_k(y_t)$ in Equation 8.42 will integrate to unit probability because the second term contains the scaling factor $(1 - p_{t,k})$. Probability calculations for this model are therefore

$$\Pr\{y_t \leq q\} = \sum_{k=1}^{m} w_k \left[p_{t,k} + (1 - p_{t,k}) F_{\gamma(\alpha_{t,k})}\left(q/\beta_{t,k}\right)\right],\ y_t \geq 0, \tag{8.44}$$

since $F_{\gamma(\alpha)}(y_t) = 0$ for $y_t = 0$, where $F_{\gamma(\alpha)}(y_t)$ is the CDF for the standard gamma distribution.

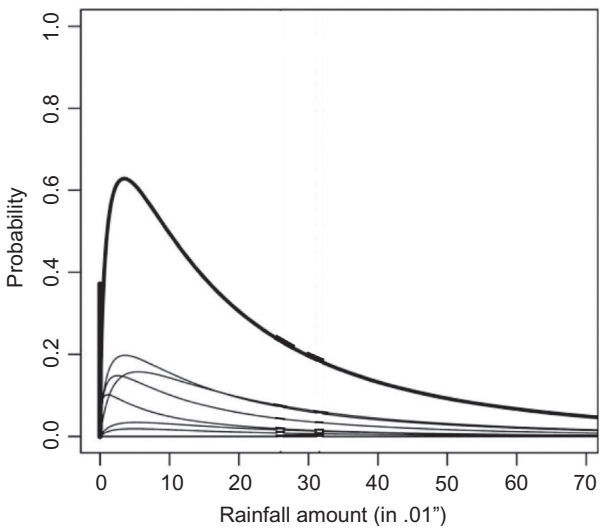

FIGURE 8.11 Example BMA predictive distribution (heavy vertical bar and curve) composed of component distributions that are mixtures of a discrete component at zero and gamma distributions for nonzero precipitation amounts (light curves). *Modified from Sloughter et al. (2007). © American Meteorological Society. Used with permission.*

Other Ensemble Dressing Methods

The BMA and other ensemble dressing approaches are closely allied to kernel density smoothing (Section 3.3.6), in which a probability distribution is estimated from a finite set of data by centering a characteristic shape (the kernel) at each data point (here, each corrected ensemble member), and summing or averaging these. Ensemble dressing as a statistical postprocessing idea was introduced by Roulston and Smith (2003), as a nonparametric method that will be described in Section 8.3.6. They proposed that the dressing kernels should represent the distribution of forecast errors around each ensemble member assuming that member provided the "best" of the individual members' forecasts on that occasion, in order that the dressing kernels should represent only uncertainty not already reflected by the ensemble dispersion.

Wang and Bishop (2005) extended the best-member dressing idea of Roulston and Smith (2003) as a parametric method by proposing use of continuous Gaussian kernels centered on each corrected ensemble member, all of which have the same variance

$$\sigma_D^2 = s_{\mu_t - y_t}^2 - \left(1 + \frac{1}{m}\right)\overline{s}_t^2 , \tag{8.45}$$

where the first term on the right-hand side indicates the error variance for the corrected ensemble-mean forecasts, and the second term approaches the average ensemble variance over the training period as the ensemble size m increases. Accordingly the second-moment constraint in Equation 8.45 can be viewed as reflecting a partition of the total error variance $s_{\mu_t - y_t}^2$ for the corrected ensemble-mean forecasts into uncertainty due to the average ensemble spread, $\overline{s}_t^2$, plus uncertainty σ_D^2 around each ensemble member. Probability forecasts are then computed using

$$\Pr\{y_t \le q\} = \frac{1}{m}\sum_{k=1}^{m} \Phi\left(\frac{q - \mu_{t,k}}{\sigma_D}\right), \tag{8.46}$$

where $\mu_{t,k}$ denotes the kth corrected ensemble member (Equation 8.34). Equation 8.46 amounts to a simplification relative to the calculation for BMA predictive probabilities (Equation 8.37), with all weights $w_k = 1/m$, and the common dressing variance estimated using Equation 8.45 rather than Equation 8.36. This Gaussian ensemble dressing approach is effective for the usual case of underdispersed ensembles, but it cannot reduce the predictive variances of overdispersed ensembles, and will fail by specifying negative dressing variance if the ensembles are sufficiently overdispersed. If the difference of the two terms in Equation 8.45 is positive but small, the mixture distribution will be noisy and unrealistic (Bishop and Shanley, 2008).

Fortin et al. (2006) note that different predictive mixture-distribution weights for each ensemble member should be appropriate in the best-member setting, depending on each member's rank within the ensemble, even if the raw ensemble members are exchangeable. The basic idea is that more extreme ensemble members are more likely to be the best member when the dynamical forecast model is underdispersive, whereas the more central ensemble members are more likely to be best when the raw ensemble is overdispersive. Fortin et al. (2006) model the mixture probabilities using beta distributions and show that the resulting postprocessed forecasts can correct both over- and underdispersion of the raw ensemble. Furthermore, for overdispersed ensembles the dressing kernels may be centered between their corrected ensemble members and the ensemble mean, at least for the more extreme ensemble members. Their method also allows different component distributional forms to be associated with the corrected ensemble members depending on their ranks.

Bröcker and Smith (2008) derived an ensemble dressing extension that they call affine kernel dressing (AKD). They proposed centering component Gaussian dressing distributions at the corrected values

$$\mu_{t,k} = a + b_1 x_{t,k} + b_2 \bar{x}_t, \tag{8.47}$$

and setting the variances for these dressing distributions as

$$\sigma_t^2 = c + b_1^2 d s_t^2, \tag{8.48}$$

where the parameters a, b_1, b_2, c, and d must be estimated from the training data. Equation 8.48 reduces to Equation 8.8, for $b_1 = 1$. The linkage of Equations 8.47 and 8.48 through the parameter b_1 allows AKD to correct both over- and underdispersion in the raw ensembles because the variance of the resulting predictive mixture distribution is

$$\sigma_{y_t}^2 = c + (1+d) b_1^2 s_t^2, \tag{8.49}$$

which can be smaller than the ensemble variance.

Bröcker and Smith (2008) also propose adding an additional "ensemble member," consisting of the climatological distribution for y, to the dressing procedure in order to make it more robust to the occasional particularly bad forecast ensemble. Including this climatological Gaussian distribution, with mean μ_C and standard deviation σ_C, AKD predictive probabilities are computed using

$$\Pr\{y_t \le q\} = \frac{1 - w_c}{m} \sum_{k=1}^{m} \Phi\left(\frac{q - \mu_{t,k}}{\sigma_t}\right) + w_c \Phi\left(\frac{q - \mu_c}{\sigma_c}\right). \tag{8.50}$$

Thus each of the actual ensemble members is given equal weight and equal dressing variance, although this variance changes from forecast to forecast depending on the raw ensemble variance (Equation 8.48). Bröcker and Smith (2008) found values for the weighting parameter for the climatological distribution w_C ranging from approximately 0.02 to 0.06 in their example, with larger values chosen for longer lead times.

Unger et al. (2009) calculate regression-based ensemble-dressing probabilities for ensembles with exchangeable members using Equation 8.46, but compute corrections to the individual ensemble members using the same correction parameters for each member,

$$\mu_{t,k} = a + b_1 x_{t,k}, \tag{8.51}$$

where the regression parameters a and b are fit through regression between the ensemble mean and the observations. Using results from regression, they set the standard deviation for the component Gaussian distributions to be

$$\sigma_D = \left[\frac{n}{n-2} s_y^2 \left(1 - \frac{r_M^2}{r_x}\right)\right]^{1/2} \tag{8.52}$$

where s_y^2 is the (climatological) sample variance of the predictand, r_M is the correlation between the ensemble means and the predictand, and r_x is the correlation between the individual ensemble members and the predictand, over the training period. Glahn et al. (2009b) and Veenhuis (2013) present a similar approach, using empirically based formulations for σ_D.

8.3.5. Fully Bayesian Ensemble Postprocessing Approaches

Although BMA (Section 8.3.4) has a grounding in Bayesian ideas, it is not a fully Bayesian procedure (Di Narzo and Cocchi, 2010). Truly Bayesian analyses are based on the relationship in Equation 6.1. In that equation, θ denotes the target of inference (in the present context the quantity being forecast, y), and x represents the available relevant training data.

Most implementations of Bayesian forecast postprocessing have assumed Gaussian distributions for the prior and likelihood distributions. This assumption is convenient because, if the variance associated with the Gaussian likelihood can be specified externally to the Bayesian analysis as a single value, the posterior distribution is Gaussian also and its parameters can be specified analytically (Section 6.3.4). Krzysztofowicz (1983) was apparently the first to use this framework for postprocessing weather forecasts, using Gaussian distributions to forecast a temperature variable y by postprocessing a single (i.e., nonprobabilistic) dynamical forecast x, although x in Equation 6.1 can as easily be regarded as the ensemble mean in the context of ensemble forecasting. Krzysztofowicz and Evans (2008) extended this postprocessing framework to a much broader range of possible distribution forms, by transformation to Gaussian distributions.

When the variable y to be postprocessed is an observed weather quantity, a natural choice for the prior distribution $f(y)$ is its climatological distribution. Long climatological records for y are generally available, and one strength of the Bayesian approach in the context of forecast postprocessing is that these long records can be brought naturally into the analysis even if the training data relating x and y are much more limited. In this context the likelihood encodes probabilistic information about past forecast errors within the training sample, characterizing the degree to which a forecast x reduces uncertainty about the predictand y. Accordingly, Equation 6.1 expresses a modification or updating of the prior information $f(y)$ in light of the past observed performance of the forecasts. However, the prior information need not necessarily be provided by the climatological distribution. For example, Coelho et al. (2004) use the forecast distribution from a statistical model as the prior, together with a likelihood encoding performance of a dynamical model, to combine the two predictive information sources through Bayes' Theorem.

Because the likelihood denotes conditional distributions for the (often, ensemble mean) forecast, given particular values of the observed variable y, it characterizes the discrimination (Section 9.1.3) of the forecasts in the training data set. Coelho et al. (2004) and Luo et al. (2007) estimate Gaussian likelihoods using linear regressions where the predictand is the ensemble mean, and the predictor is the observation y, so that the mean function for the Gaussian likelihood is $\mu_{\mathrm{L}} = a + by$, and the variance σ_L^2 is the regression prediction variance. Figure 8.12, from Coelho et al. (2004), illustrates the procedure for an example where the colder forecasts are nearly unbiased but the warmer forecasts exhibit a marked warm bias. Because the regression prediction variance σ_L^2 is specified as a point value external the Bayesian estimation, the resulting predictive distributions are also Gaussian with mean $\mu_{P,t}$ and standard deviation $\sigma_{P,t}$ for the tth forecast so that probabilities are calculated using

$$\Pr\{y_t \le q\} = \Phi\left(\frac{q - \mu_{P,t}}{\sigma_{P,t}}\right). \tag{8.53}$$

Luo et al. (2008) formulate the Gaussian predictive parameters as

$$\mu_{P,t} = \sigma_P^2\left[\frac{\mu_C}{\sigma_C^2} + \frac{b\,(\overline{x}_t - a)}{\sigma_L^2 + \overline{\sigma}_e^2}\right] \tag{8.54a}$$

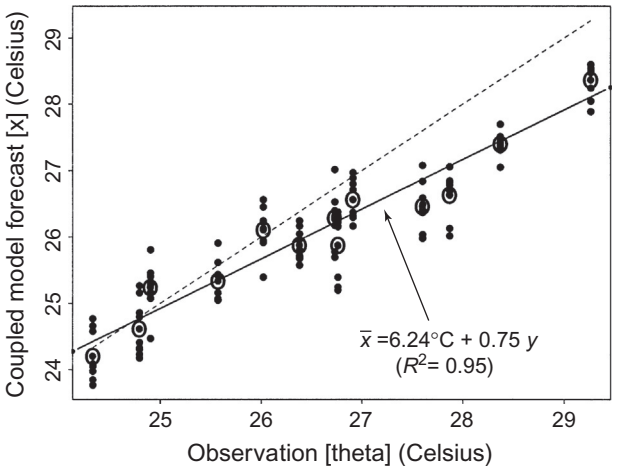

FIGURE 8.12 Individual ensemble members (small dots) and ensemble means (circles) as functions of observed December Niño 3.4 temperatures. Solid line is the weighted regression for the ensemble means defining the Bayesian likelihood. The dashed 1:1 line would correspond to perfect forecasts. *Modified from Coelho et al. (2004). © American Meteorological Society. Used with permission.*

and

$$\sigma_P = \left[\frac{\left(\sigma_L^2 + \overline{\sigma}_e^2\right)\sigma_C^2}{\sigma_L^2 + \overline{\sigma}_e^2 + b^2\sigma_C^2}\right]^{1/2}, \tag{8.54b}$$

where μ_C and σ_C^2 are the (prior) climatological mean and variance, and $\overline{\sigma}_e^2$ is the average ensemble variance.

In a method they call Forecast Assimilation, Stephenson et al. (2005) extend Gaussian-based Bayesian calibration using a multivariate normal likelihood distribution for individual nonexchangeable ensemble members, rather than the ensemble mean. Reggiani et al. (2009) implement Gaussian-based Bayesian recalibration (after transformation of the positively skewed hydrological variables) individually for each ensemble member, and then construct the predictive distribution as the average of the m individual predictive density functions, analogously to Equation 8.33 with $w_k = 1/m$. Hodyss et al. (2016) accommodate nonexchangeability of individual ensemble members by representing the likelihood as the product of conditional distributions for each member, but at the cost of requiring the estimation of a large number of regression parameters to define the likelihood. Siegert et al. (2016) describe a more elaborate Bayesian framework for ensemble-mean recalibration which does not have an analytic result for the posterior distribution, and so requires resampling from the final predictive distribution.

Friederichs and Thorarinsdottir (2012) describe a Bayesian approach to postprocessing peak wind speed forecasts, using the GEV (Equation 4.63) as the form of the likelihood distribution, where the location (ζ_t) and scale (β_t) parameters are specified as linear functions of covariates, while the shape parameter κ is assumed to be the same for all forecasts. They use noninformative distributions (independent Gaussian distributions with very large variances) for the prior and a computationally intensive parameter estimation procedure.

A different approach to Bayesian ensemble calibration was proposed by Satterfield and Bishop (2014). Their target of inference is the forecast error variance as predicted by the current ensemble variance, so that the procedure specifically seeks to represent the spread-skill relationship of the ensemble forecasts.

The Bayesian predictions summarized so far in this section are analogous to the regression methods reviewed in Section 8.3.2, in that the output is a single predictive distribution that in most cases is of a known parametric form. In contrast, Bishop and Shanley (2008), DiNarzo and Cocchi (2010), and Marty et al. (2015) formulate full Bayesian analyses for the ensemble-dressing setting, which is an alternative approach to the methods described in Section 8.3.4. Bishop and Shanley (2008) propose using the BMA mixture-distribution formulation as the Bayesian likelihood rather than the predictive distribution. DiNarzo and Cocchi (2010) employ a hierarchical Bayesian model in which a latent variable represents the "best member." Marty et al. (2015) combine the Krzysztofowicz and Evans (2008) Bayesian approach with conventional BMA. These approaches incorporate ensemble-variance information into the predictive distribution by construction, and so allow features such as bimodality in the raw ensemble to carry through to the postprocessed predictive distribution.

8.3.6. Nonparametric Ensemble Postprocessing Methods

Nonparametric ensemble postprocessing methods are wholly or mostly data based, in contrast to the methods described in previous sections which rely on prespecified mathematical forms. Hamill and Colucci (1997) proposed the earliest of these nonparametric methods, which estimates postprocessed ensemble probabilities on the basis of the verification rank histogram (Section 9.7.1). The verification rank histogram tabulates relative frequencies p_j in the training data that the observed value y_t is larger than j of its forecast ensemble members $x_{t,j}$, plus 1. For example, p_1 is the proportion of training-sample forecasts for which the observation was smaller than all ensemble members, and p_{m+1} is the proportion of forecasts where the observation is larger than all m ensemble members.

The Hamill and Colucci (1997) recalibration procedure operates on the rank histogram for the unconditionally debiased ensemble members

$$\widetilde{x}_{t,k} = x_{t,k} - \sum_{t=1}^{n} (\overline{x}_t - y_t), \tag{8.55}$$

and aims to achieve flatness (which is the ideal result) of the recalibrated rank histogram. When a quantile of interest is not outside the range of the ensemble, probabilities are estimated by linear interpolation:

$$\Pr\{y_t \le q\} = \sum_{j=1}^{k} p_j + p_{j+1} \frac{q - \widetilde{x}_{t,(k)}}{\widetilde{x}_{t,(k+1)} \widetilde{x}_{t,(k)}}, \; \widetilde{x}_{t,(k)} \le q \le \widetilde{x}_{t,(k+1)}. \tag{8.56}$$

The parenthetical subscripts denote that the ensemble members have been sorted in ascending order. When a quantile of interest is outside the range of the ensemble, the probabilities represented in p_1 and p_{m+1} must be extrapolated in some way. Because rank histogram recalibration tends to underperform more recently developed methods (e.g., Wilks, 2006b; Ruiz and Saulo, 2012) it is rarely used.

Quantile Regression

Bremnes (2004) introduced the use of quantile regression (Section 7.7.3) for ensemble postprocessing of continuous predictands. The predictand considered by Bremnes (2004) was precipitation amount, which has a mixed discrete (finite probability mass at zero) and continuous (nonzero precipitation amounts) distribution. He separately forecast the probability of exactly zero precipitation using probit regression

(which is very similar to logistic regression, Section 7.6.2), forecast selected nonzero precipitation quantiles with quantile regression, and combined the two according to the mathematical definition of conditional probability, yielding

$$\Pr\{y_t \le q_p\} = 1 - \frac{1 - p_{y>0}}{1 - \Pr\{y_t = 0\}}, \tag{8.57}$$

where $p_{y>0}$ is an appropriate cumulative probability derived from the quantile regression describing the nonzero amounts. For example, in order to calculate the median ($q_{.50}$) of the predictive distribution when the probability of zero precipitation is 0.2, the precipitation amount of interest would be the quantile of the predictive distribution for nonzero amounts corresponding to $p_{y>0} = 0.6$. When $p_{y>0} \le \Pr\{y_t = 0\}$ then $q_p = 0$ is implied. The method can of course be used, and indeed is more easily implemented, for strictly positive predictands where $\Pr\{y_t = 0\} = 0$.

For each preselected predictand quantile q_p an appropriate list of ensemble predictors $\boldsymbol{x}_t$ must be selected, as is also the case for ordinary least-squares regression, although Ben Bouallègue (2016) suggests use of lasso-penalization (Section 7.5.2) to choose predictor variables. Bremnes (2004) considered the five predictand quantiles $q_{.05}$, $q_{.25}$, $q_{.50}$, $q_{.75}$, and $q_{.95}$, and for each of the five quantile regressions he used the same $I = 2$ predictors, being the two quartiles of the ensemble. Figure 8.13 illustrates the resulting forecasts as functions of the 75th percentiles of the ensembles, for three levels of the ensemble 25th percentiles. These show greater forecast uncertainty (larger distances between the forecast quantiles) for increasing values of both but especially for $q_{.25}$.

Ensemble Dressing

Although ensemble dressing is usually implemented using parametric kernels (Section 8.3.4), it was originally proposed by Roulston and Smith (2003) as a nonparametric ensemble postprocessing approach. They "dressed" each corrected ensemble member using a random sample from the collection of errors that were defined relative to the ensemble member closest to the observation on each

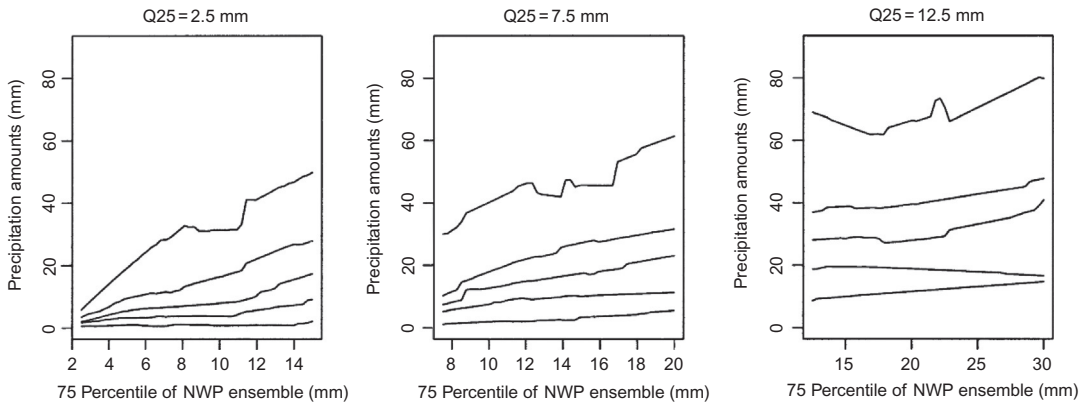

FIGURE 8.13 Quantile regression forecasts for (from top to bottom) the $q_{.95}$, $q_{.75}$, $q_{.50}$, $q_{.25}$, and $q_{.05}$ quantiles of the predictive distribution of nonzero precipitation, for three levels of the ensemble lower quartile, as functions of the ensemble upper quartile. *From Bremnes (2004). © American Meteorological Society. Used with permission.*

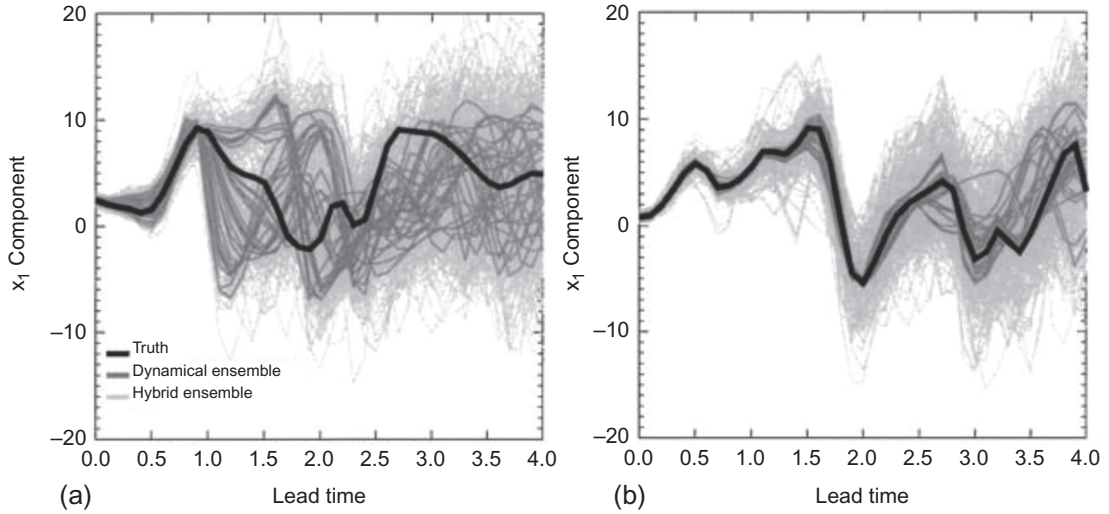

FIGURE 8.14 True evolution (heavy curves), 32 raw ensemble members (medium curves), and the dressed predictive distribution represented by 16 best-member trajectories around each raw ensemble member (light curves), for two forecast cases of the Lorenz '96 (Lorenz, 2006) system. Panel (a) shows a low-predictability case and panel (b) shows a high-predictability case. *From Roulston and Smith (2003).*

occasion (the "best member"), in an archive of past forecasts. Figure 8.14 shows example dressing distributions in the form of time trajectories for forecasts of the Lorenz '96 (Lorenz, 2006) system. Here there are 32 ensemble members (medium gray curves), each of which has been dressed using a sample of 16 best-member error trajectories (light curves), so that the predictive distribution is represented by a collection of 512 trajectories. Figure 8.14a shows a low-predictability case, Figure 8.14b shows a high-predictability case, and in both panels the heavy curves indicate the true trajectories. The same catalog of past best-member errors has been sampled in both examples, but the spread of the predictive distributions for the low-predictability case is larger because the underlying ensemble spread is larger. Messner and Mayr (2011) propose a similar resampling approach, where the discrete dressing kernels are derived from close analogs in the training data to the current ensemble members, following the method of Hamill and Whitaker (2006).

Individual Ensemble-Member Adjustments

Another nonparametric approach to ensemble postprocessing involves transforming each of the ensemble members individually, leading to a corrected ensemble of finite size m. This has been termed the member-by-member postprocessing (MBMP) approach (Van Schaeybroeck and Vannitsem, 2015). MBMP adjustments can be expressed as

$$y_{t,k} = (a + b\bar{x}_t) + \gamma_t(x_{t,k} - \bar{x}_t), \tag{8.58}$$

where each raw ensemble member $x_{t,k}$ maps to a distinct corrected counterpart, $y_{t,k}$. The parameters a and b define unconditional and conditional bias corrections, and the parameter γ_t controls a forecast-dependent expansion or contraction relative to the ensemble mean. Various definitions for these parameters have been proposed.

Eade et al. (2014) estimate the parameters in Equation 8.58 using

$$a = \overline{y} - b\overline{\overline{x}}, \tag{8.59a}$$

$$b = \frac{s_{y_t}}{s_{\overline{x}_t}} r_{y_t,\overline{x}_t}, \tag{8.59b}$$

and

$$\gamma_t = \frac{s_{y_t}}{s_t}\sqrt{1 - r^2_{y_t,\overline{x}_t}}, \tag{8.59c}$$

where s_{y_t} is the climatological predictand standard deviation, the correlations in Equation 8.59 relate the predictand and the ensemble means, and

$$\overline{\overline{x}} = \frac{1}{n}\sum_{t=1}^{n} \overline{x}_t \tag{8.60}$$

is the average of the ensemble means over the training data. Equations 8.59a and 8.59b are the ordinary least-squares regression coefficients relating the ensemble mean to the predictand y, and Equation 8.59c varies from forecast to forecast according to the ensemble standard deviation s_t in the denominator. Doblas-Reyes et al. (2005) and Johnson and Bowler (2009) have proposed an equivalent model assuming that both the forecast ensembles and the observations have been centered, so that $a = 0$ in Equation 8.59a.

Van Schaeybroeck and Vannitsem (2015) allow the "stretch" coefficient γ_t to depend on the ensemble spread according to

$$\gamma_t = c + \frac{d}{\delta_t}, \tag{8.61}$$

where

$$\delta_t = \frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{k=1}^{m}\left|x_{t,j} - x_{t,k}\right| \tag{8.62}$$

is the average absolute difference among pairs of the uncorrected ensemble members, and both c and d are constrained to be nonnegative. Thus Equation 8.58 can be viewed as a nonparametric counterpart to the nonhomogeneous regressions described in Sections 8.3.2 (Schefzik, 2017). Van Schaeybroeck and Vannitsem (2015) describe estimation of the parameters a, b, c, and d in two ways. The first is using the method maximum likelihood, assuming the errors of the corrected ensemble-mean forecasts are exponentially distributed, with mean δ_t, and so maximizing

$$L_{\text{exp}} = -\sum_{t=1}^{n}\left[\frac{\overline{y}_t - y_t}{\delta_t} + \ln(\delta_t)\right] \tag{8.63}$$

with respect to a, b, c, and d. Here

$$\overline{y}_t = \frac{1}{m}\sum_{k=1}^{m} y_{t,k} \tag{8.64}$$

is the average of the individually corrected ensemble members for forecast case t. Alternatively, the parameters can be estimated by minimizing the ensemble CRPS (Equation 9.83).

Williams (2016) uses the nonparametric adjustment procedure in Equation 8.58, but defines the "stretch" coefficient as

$$\gamma_t = \frac{\sqrt{d + cs_t^2}}{s_t} = \sqrt{c + d/s_t^2}. \tag{8.65}$$

With this formulation, method-of-moments estimators can be derived for the required parameters, the computation of which will be relatively fast. The two bias-correction parameters a and b are again the least-squares regression parameters defined in Equations 8.59a and 8.59b. The method-of-moments estimates for the "stretch" parameters are

$$c = \frac{\mathrm{cov}\left(s_t^2, y_t^2\right) - 2ab\,\mathrm{cov}\left(s_t^2, \overline{x}_t\right) - b^2\,\mathrm{cov}\left(s_t^2, \overline{x}_t\right)}{\mathrm{Var}\left(s_t^2\right)} \tag{8.66a}$$

and

$$d = s_{y_t}^2 - c\overline{s}^2 - b^2 s_{\overline{x}_t}^2, \tag{8.66b}$$

where

$$\overline{s}^2 = \frac{1}{n}\sum_{t=1}^{n} s_t^2 \tag{8.67}$$

is the average ensemble variance over the training period, and

$$s_{\overline{x}_t}^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(\overline{x}_t - \overline{\overline{x}}\right)^2 \tag{8.68}$$

is the variance of the ensemble means over the training period. Again both c and d are constrained to be nonnegative. Equations 8.61 and 8.65 were found to perform similarly in Wilks (2018a).

Figure 8.15 illustrates the MBMP adjustment method. Figure 8.15a (solid line) shows the debiasing function defined by the parameters a and b in Equations 8.59a and 8.59b, indicating that ensembles with moderate and large means were positively biased in the training data, and that ensembles having small ensemble means were negatively biased. Figure 8.15b shows a hypothetical $m = 5$ member ensemble (filled circles) to be corrected. The dotted arrows originating at the raw ensemble mean in Figure 8.15b locate the corrected ensemble mean in Figure 8.15c. The corrected ensemble members (open circles) in Figure 8.15c retain their distributional shape but have been expanded away from the corrected ensemble mean, by the factor $\gamma_t = 1.5$ relative to the raw ensemble in Figure 8.15b, so that the indicated corrected ensemble range in Figure 815c is larger than the ensemble range of the example underdispersive raw ensemble in Figure 8.15b by the factor 1.5.

Because the identities of the individual ensemble members are preserved in these adjustments, different forecast variables that have been subjected to independent postprocessing will continue to exhibit the rank correlation structures inherited from the raw ensemble. Thus correlations in the raw ensemble among different variables at a single location, and spatiotemporal correlations for a given variable, will all be carried forward to the postprocessed ensemble. Therefore MBMP can also serve as the basis for a multivariate postprocessing algorithm, as described in Section 8.4.3. On the other hand, postprocessing predictands such as wind speed or precipitation may be problematic without constraints requiring all

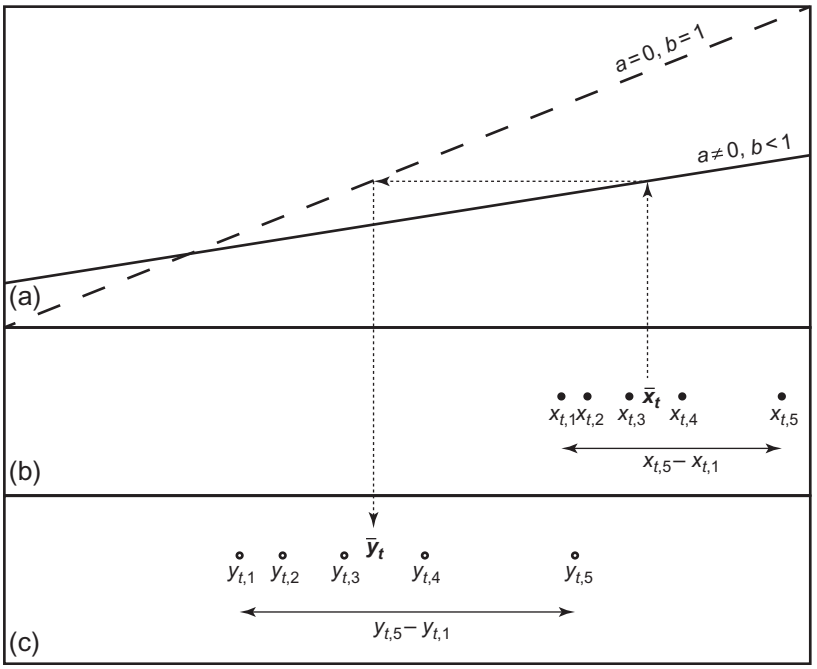

FIGURE 8.15 Illustration of the MBMP adjustment method, for (b) a hypothetical $m = 5$ member underdispersive and positively biased raw ensemble (filled circles), which has been transformed to (c) the corrected ensemble (open circles) using the debiasing function (solid line) in (a) and the "stretch" coefficient $\gamma_t = 1.5$. *From Wilks (2018a).*

postprocessed ensemble members to be nonnegative; because the smaller members of an underdispersed ensemble may be transformed to negative values, and Equations 8.61 or 8.65 will be undefined for precipitation ensembles in which all members are zero.

8.4. MULTIVARIATE ENSEMBLE POSTPROCESSING

The ensemble postprocessing methods outlined in Section 8.3 pertained to scalar predictands at single lead times. These univariate methods are adequate for many purposes, but in some settings forecasts of multiple predictands exhibiting realistic correlations structures are essential. Depending on the application, such multivariate forecasts might involve two or more different but correlated predictand variables, forecasts for a given variable at many spatially correlated locations simultaneously, forecasts for a given variable at multiple lead times into the future, or various combinations of these possibilities. For example, correct portrayal of the joint spatial and temporal correlation structure of forecast precipitation is essential for realistic streamflow and flood forecasting (e.g., Clark et al., 2004; Li et al., 2017), and realistic portrayal of the space-time correlation structure of wind speed forecasts is necessary in such applications as optimal wind power decision making (e.g. Pinson, 2013) and aircraft routing (Chaloulos and Lygeros, 2007). This section describes ensemble postprocessing approaches designed to capture these correlations in multivariate predictive distributions, which can be broadly classified as parametric methods, copula methods, and analog methods.

8.4.1. Parametric Methods

Schuhen et al. (2012) proposed extension of NGR (Section 8.3.2) to postprocessing of ensemble forecasts for northward (u) and eastward (v) wind components, jointly using bivariate normal (Equation 4.31) predictive distributions. Extending Equation 8.11, bivariate mean vectors are defined as linear functions of the respective ensemble means,

$$\mu_u = a_u + b_u \overline{u} \tag{8.69a}$$

and

$$\mu_v = a_v + b_v \overline{v}, \tag{8.69b}$$

and extending Equation 8.8 the respective postprocessed variances are defined as linear functions of the respective ensemble variances,

$$\sigma_u^2 = c_u + d_u s_u^2 \tag{8.70a}$$

and

$$\sigma_v^2 = c_v + d_v s_v^2. \tag{8.70b}$$

Schuhen et al. (2012) defined the correlation parameter as a function of the ensemble-mean angular wind direction θ, although not the correlation within the ensemble, using

$$\rho_{uv} = r \cos\left[\frac{2\pi}{360^\circ}(k\theta + \varphi)\right] + s, \tag{8.71}$$

where r, s, φ, and k are parameters to be estimated. Thus even with the simplification of Equation 8.71, the approach requires estimation of 12 parameters in Equations 8.69–8.71. Figure 8.16a shows an example result from Equation 8.71 (dashed curve) for the Sea-Tac (Seattle) airport together with the training data. Figure 8.16b shows an example forecast, with the dark contours representing the postprocessed bivariate normal distribution, in relation to the original 8-member ensemble (dots) and the 90% probability ellipse of the bivariate normal distribution fit to the raw ensemble (gray curve). Bias, mainly in the u wind speed, and overall underdispersion of the raw ensemble have been corrected.

Baran and Möller (2017) extended the idea of bivariate NGR to jointly postprocess bivariate forecasts of (scalar) wind speed (x_w) and temperature (x_T), with the marginal distribution of wind speed represented as zero-truncated Gaussian (Equation 4.39). Accordingly, Equation 4.31 for the bivariate normal PDF is modified to read

$$f(x_w, x_T) = \frac{I(x_w \geq 0)}{\Phi\left(\frac{\mu_w}{\sigma_w}\right) 2\pi\sigma_w\sigma_T\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_w-\mu_w}{\sigma_w}\right)^2 + \left(\frac{x_T-\mu_T}{\sigma_T}\right)^2 - 2\rho\left(\frac{x_w-\mu_w}{\sigma_w}\right)\left(\frac{x_T-\mu_T}{\sigma_T}\right)\right]\right\}. \tag{8.72}$$

In this case, eight parameters were required to jointly correct the two means, extending Equation 8.11, and another eight parameters were required to specify the standard deviations, for a total parameter count of 16.

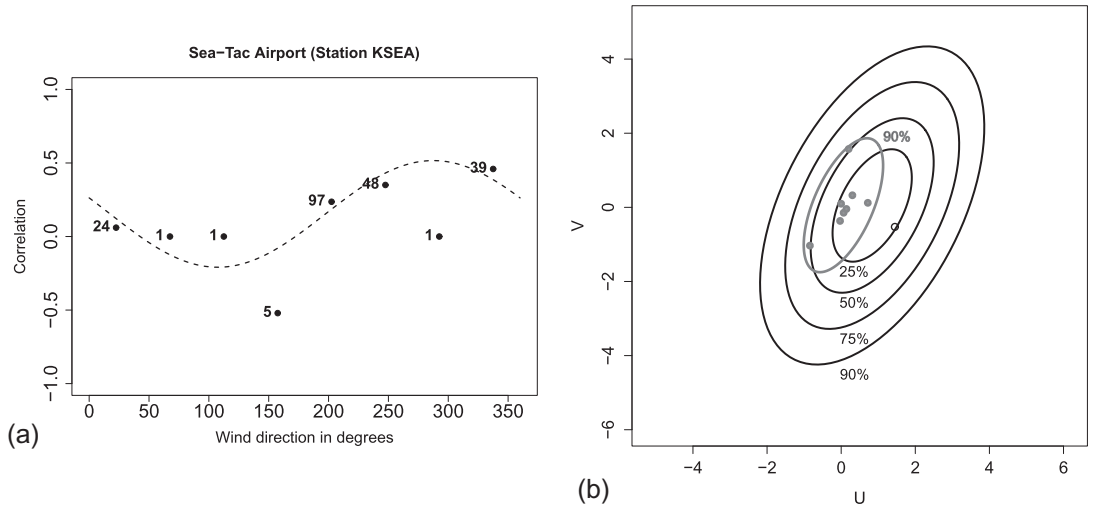

FIGURE 8.16 (a) Example correlation function (Equation 8.71) fit to bivariate normal wind component correlations as a function of ensemble-mean wind direction, with training data and sample sizes. (b) Example bivariate normal predictive distribution (dark contours), based on an 8-member ensemble (light dots). Gray ellipse is the 90% probability ellipse fit to the raw ensemble members, and open circle is the eventual verifying observation. *From Schuhen et al. (2012). © American Meteorological Society. Used with permission.*

Bivariate Gaussian distributions can also be used to extend the BMA and allied approaches (Section 8.3.4), in which case the PDFs f_k in Equation 8.33 are bivariate normal. As was also the case for univariate BMA, the means must first be debiased before being used to center the bivariate normal kernels that are used to build up the predictive distribution. Figure 8.17 shows an example of the result, where the circles locate the eight uncorrected ensemble members, the plusses locate their bias-corrected counterparts, and the triangle shows the corresponding observed wind vector. Because the extension of the debiasing Equation 8.34 to the bivariate setting required estimation of six parameters for each of the eight nonexchangeable ensemble members, the approach is highly parameterized. Schölzel and Hense (2011) employ this approach in a climate change setting. Baran and Möller (2015) extended the bivariate BMA idea to joint forecasts of temperature and scalar wind speed using the bivariate Gaussian distributions with zero-truncated wind speed component (Equation 8.72) as the dressing kernels. Berrocal et al. (2007) described a higher dimensional multivariate BMA approach, in which temperature forecasts at multiple locations are constructed using multivariate normal (Chapter 13) dressing kernels. The number of required parameters was kept to a manageable level by assuming particular mathematical forms for the spatial covariances (Gel et al., 2004). Hemri et al. (2013) represented (power-transformed) river runoff forecasts for a particular gage at multiple lead times using an equivalent approach.

8.4.2. Copula Methods

That large numbers of parameters need to be estimated for parametric multivariate postprocessing models, even in the low-dimensional settings described in Section 8.4.1, suggests that nonparametric methods may often be preferred for specifying the statistical dependence among the multiple predictands. *Empirical copulas* provide a flexible and attractive means to do so. For multivariate ensemble postprocessing, copulas allow each of multiple forecast variables to be postprocessed separately, using

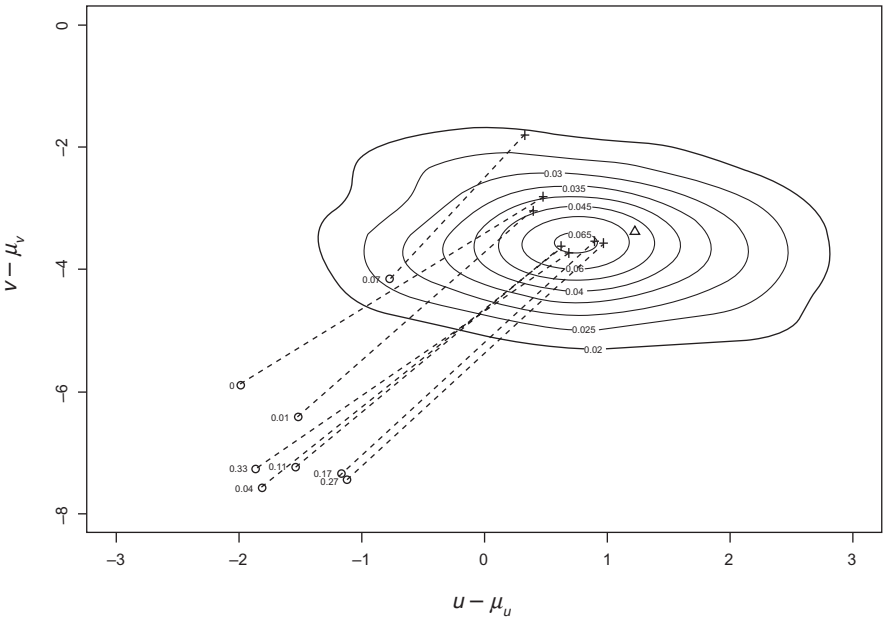

FIGURE 8.17 Example bivariate BMA forecast distribution for u and v wind components. The raw (nonexchangeable) ensemble members are shown as circles with associated weights, and the dashed lines connect to bias-corrected ensemble members (plusses). Triangle shows the eventual observed wind vector. *From Sloughter et al. (2013). © American Meteorological Society. Used with permission.*

a method such as NGR (Section 8.3.2) or BMA (Section 8.3.4), with the results assembled subsequently into a joint distribution with appropriate correlation structure.

Copulas are structures that connect multivariate distribution functions to their individual marginal distributions. In effect, a copula provides a "dependence template" onto which samples drawn independently from a collection of univariate marginal distributions can be arranged, in a way that will reflect a specified dependence structure. Empirical copulas (e.g., Bárdossy and Pegram, 2009; Rüschendorf, 2009) derive their structure from independent rank transformations of a sample of training data in each of the dimensions of the multivariate data space. Schefzik et al. (2013) proposed using uncorrected ensembles to provide the structure of the empirical copulas, calling the method *empirical copula coupling* (ECC), which allows the nature of the relationships among the prognostic variables to change from forecast to forecast. The underlying assumption is that, even though the raw ensemble may exhibit bias and dispersion errors, the dynamical model (at least approximately) correctly represents the statistical relationships among the forecast variables.

An empirical copula can be constructed for any data dimension K, but it is easiest to visualize and understand for bivariate ($K = 2$) data. Consider a hypothetical $m = 5$ member raw ensemble of bivariate data, x, shown in the matrix (or, equivalently the data table) in Figure 8.18a. The process of constructing the copula template begins by separately transforming each of the two variables to their ranks within the ensemble,

$$r_{i,k} = \sum_{j=1}^{m} I\left(x_{j,k} \le x_{i,k}\right), \quad i = 1,\ldots,m;\ k = 1,\ldots,K, \tag{8.73}$$

where again $I(\bullet)$ denotes the indicator function. These ranks then populate the matrix of $[R]$ of ranks in Figure 8.18b. For example, the first ($i = 1$) ensemble member for the first ($k = 1$) variable, $x_{1,1}$ is the

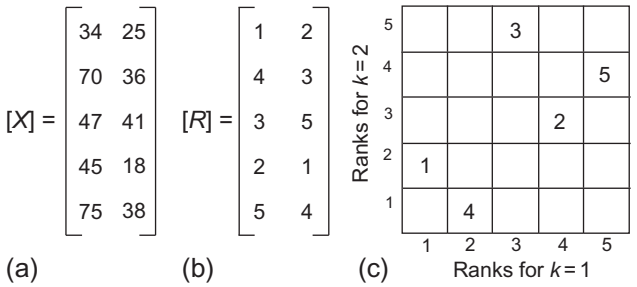

FIGURE 8.18 (a) Hypothetical bivariate ensemble $[X]$ of size $m = 2$ that is transformed to ranks in (b) according to Equation 8.73. (c) Latin square representation of the hypothetical matrix of ranks in (b). Horizontal axis pertains to the first column and vertical axis pertains to the second column of $[R]$. The integers in the interior of the square are the row indices. *Modified from Wilks (2015).*

smallest of the five x's, and so transforms to $r_{1,1} = 1$. Similarly the ($i = 5$) fifth ensemble member for the second ($k = 2$) variable, $x_{5,2}$ is the fourth-smallest among the ensemble for the second variable, and so transforms to $r_{5,2} = 4$. Schefzik et al. (2013) motivate the dependence template of ranks $[R]$ using a latin hypercube of dimension K, which for $K = 2$ is a Latin square, having one entry only in each row and column (Figure 8.18c). Another way to look at Figure 8.18c is as a scatterplot of the rank-transformed bivariate data. The arrangement of the five row index labels (the 'data points') within the Latin square suggests a positive correlation, and indeed the Spearman rank correlation (Equation 3.34) between the two columns of Figure 8.18a is 0.6.

Having constructed the empirical copula $[R]$, it can be used to arrange independent samples from the K independently postprocessed univariate marginal distributions, each of size m, in a way that preserves the intervariable relationships encoded in the ranks of $[R]$. Denote these samples, which might be random draws from an NGR (Equation 8.6) or BMA (Equation 8.33) predictive distribution, or derived through some other postprocessing method (e.g., Flowerdew, 2014; Roulin and Vannitsem, 2012), as $y_{i,k}$, $i = 1, \ldots, m$; $k = 1, \ldots, K$. Adopting the conventional notation of parenthetical subscripts to indicate sorted data (so that, for example, $y_{(1),k}$ and $y_{(m),k}$, are the smallest and largest postprocessed samples, respectively, for the kth variable), the arrangement of these ranked values duplicating the multivariate relationships encoded in R is achieved by ordering the sorted data $y_{(i),k}$, according to the ranks $r_{i,k}$,

$$\widetilde{y}_{i,k} = y_{\left(r_{i,k}\right),k}, \quad i = 1, \ldots, m, \; k = 1, \ldots, K, \tag{8.74}$$

which organizes the sorted samples $y_{(i),k}$ onto the empirical copula R in Figure 8.18b and c. The resulting matrix of postprocessed and correlated ensemble members for the example in Figure 8.18 is then

$$\left[\widetilde{Y}\right] = \begin{bmatrix} y_{\left(r_{1,1}\right),1} & y_{\left(r_{1,2}\right),2} \\ y_{\left(r_{2,1}\right),1} & y_{\left(r_{2,2}\right),2} \\ y_{\left(r_{3,1}\right),1} & y_{\left(r_{3,2}\right),2} \\ y_{\left(r_{4,1}\right),1} & y_{\left(r_{4,2}\right),2} \\ y_{\left(r_{5,1}\right),1} & y_{\left(r_{5,2}\right),2} \end{bmatrix} = \begin{bmatrix} y_{(1),1} & y_{(2),2} \\ y_{(4),1} & y_{(3),2} \\ y_{(3),1} & y_{(5),2} \\ y_{(2),1} & y_{(1),2} \\ y_{(5),1} & y_{(4),2} \end{bmatrix}. \tag{8.75}$$

The rank correlation between the two columns in Equation 8.75 is also 0.6 because the reordered postprocessed data have exactly the same rank structure as those in Figure 8.18c. In higher dimensional settings the matrices in Figure 8.18 and Equation 8.75 are extended to include K columns, and all $K(K-1)/2$ pairwise correlations among the K columns of $[R]$ are reproduced for the corresponding columns of $\left[\widetilde{Y}\right]$.

Schefzik et al. (2013) proposed several ECC variants, which differ in the way the postprocessed variables are chosen before they are reordered by the copula. The first of these is based on equidistant quantiles (ECC-Q) from the individual univariate predictive distributions, F_k. That is, each of the univariate postprocessed predictive distributions is sampled systematically by evaluating the respective quantile functions at the points specified by

$$y_{i,k} = F_k^{-1}\left(\frac{i}{m+1}\right), \;\; i = 1,\ldots,m, \;\; k = 1,\ldots,K, \tag{8.76}$$

the argument of which is the Weibull plotting position estimator (Table 3.2).

Because the sampling of each postprocessed distribution F_k in Equation 8.76 is systematic and not random, the ECC-Q method is limited to producing postprocessed ensembles of the same size m as the original ensemble. The second empirical copula coupler is based on independent random (ECC-R) samples from the K univariate postprocessed predictive distributions,

$$y_{i,k} = F_k^{-1}(u_{i,k}), \;\; i = 1,\ldots,m, \;\; k = 1,\ldots,K, \tag{8.77}$$

where the $u_{i,k}$ are independent standard uniform random numbers. Because Equation 8.77 samples each univariate postprocessed distribution F_k randomly, repeated implementation of the ECC is not redundant, so that Equation 8.77 can be repeatedly reevaluated, reordered according to the copula template, and pooled. Accordingly the size of a multivariate ECC-R postprocessed ensemble can be any integer multiple of the original ensemble size m, and so can represent the forecast distribution more smoothly and is better able to sample its extremes.

Figure 8.19 illustrates the ECC approach, operating on an $m = 50$-member ensemble representing a $K = 4$-dimensional (temperatures and pressures at Berlin and Hamburg) multivariate forecast. Figure 8.19a shows the scatterplot matrix of the raw ensemble, with histograms on the diagonal representing the four marginal distributions. Figure 8.19b shows the corresponding plot after each of the four forecast variables has been independently postprocessed using BMA. The marginal distributions on the diagonal have been transformed, but the pairwise scatters indicate near zero correlation. Figure 8.19c shows the result after the BMA-postprocessed forecasts in Figure 8.19b are reordered using ECC. The marginal distributions in Figure 8.19b and c are identical, and the rank correlation structures in Figures 8.19c reproduce those in Figure 8.19a.

If the underlying dynamical model represents the statistical dependences among the forecast variables poorly, or if the ensemble size is too small to sample them adequately, then ECC methods may not perform well. Alternatively empirical copulas can be constructed using the *Schaake shuffle* (Clark et al., 2004). This method operates by drawing samples from the historical climatological record to populate the $x_{i,k}$ values used to compute the ranks in Equation 8.73, which in turn are used to reorder samples $y_{i,k}$ from the individually postprocessed forecast variables in Equation 8.74. Because this resampling from the climatological record is not constrained by the original ensemble size, the size of the final ensemble (corresponding to the number of rows in the matrices $[X]$ and $[R]$ in Figure 8.18) can be of any size, up to the size of the relevant climatology.

A disadvantage of the Schaake shuffle is that the climatology is sampled independently of the raw ensemble. To the extent that there may be case-to-case differences in the correlation structure among the K forecast variables (e.g., Clark et al., 2004; Demargne et al., 2014; Verkade et al., 2013) copulas derived from random sampling of the climatology will not reflect them. This problem can be addressed by sampling observations from the historical climatology that are most similar to the current forecast

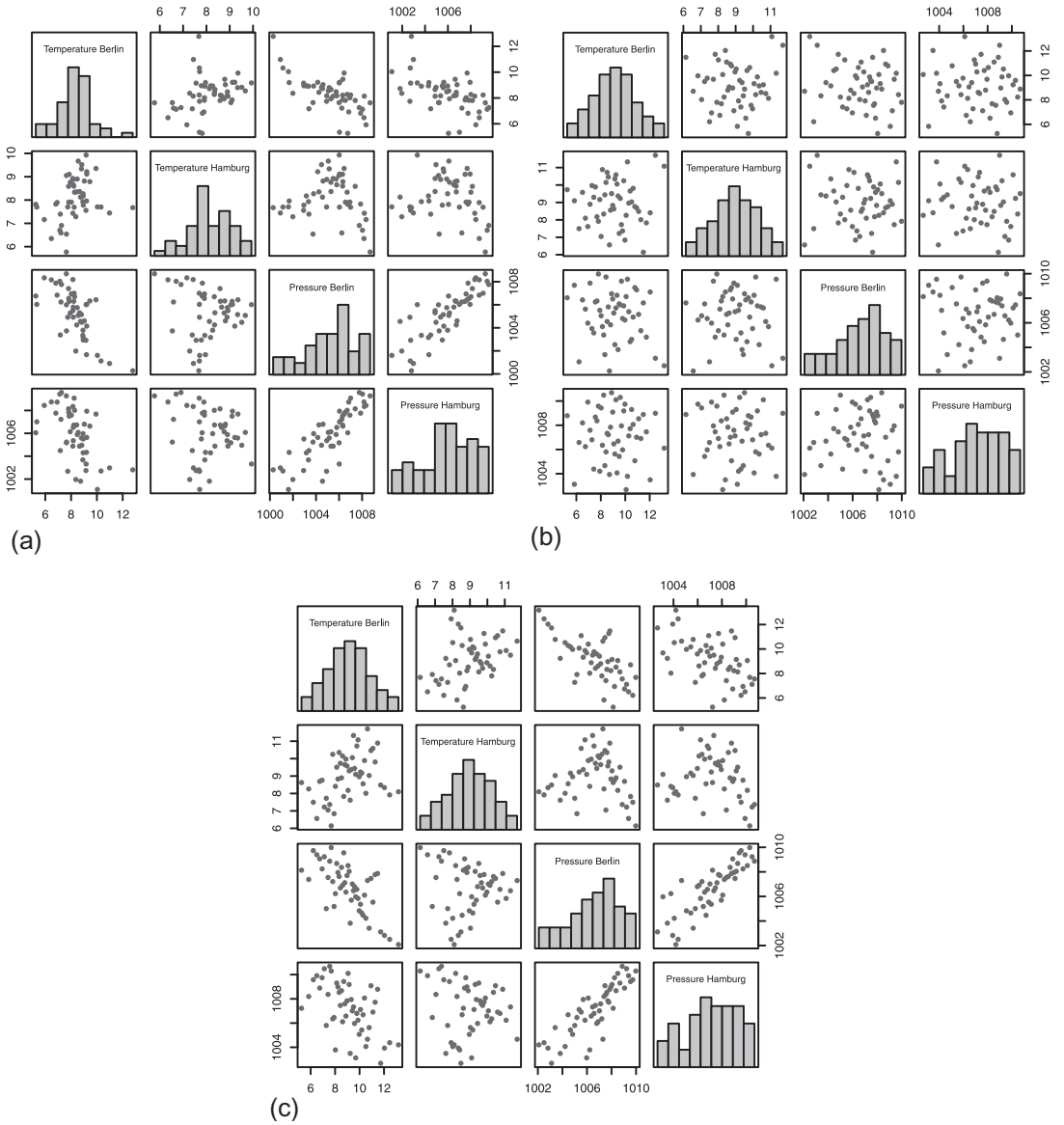

FIGURE 8.19 Illustration of the ECC approach, operating on a $K = 4$ dimensional (Berlin temperature, Hamburg temperature, Berlin pressure, Hamburg pressure) and $m = 50$-member ensemble forecast. (a) Scatterplot matrix of the raw, uncorrected ensemble, (b), scatterplot matrix after independent BMA postprocessing of the four forecast variables, and (c) the postprocessed forecasts after ECC reordering. *From Schefzik et al. (2013).*

(Clark et al., 2004), or sampling dates for which past forecasts are most similar to the current forecast (Schefzik, 2016; Scheuerer et al., 2017; Scheuerer and Hamill, 2018).

Application of empirical copula methods to precipitation fields, where many ensemble members may be exactly zero, may lead to inconsistent representation of the spatiotemporal variability that derives from random assignments to the lowest ranks (Wu et al., 2018).

This section has emphasized empirical copulas, which are computationally fast, and can accommodate very high dimension K. Copulas can also be constructed parametrically, most commonly on the basis of the multivariate Gaussian distribution (Chapter 12). These *Gaussian copulas* require estimation of a K-dimensional correlation matrix using training data, and then sampling from the resulting multivariate Gaussian distribution (Section 12.4.1) to define the values $x_{i,k}$ in Equation 8.73. The utility of this approach may be limited by the larger numbers of correlation parameters that must be estimated for large K, unless a theoretical model for the correlation structure (e.g., Equation 10.23 in a setting where the K forecast dimensions pertain to a time sequence) can be reasonably assumed. Hemri et al. (2015) and Möller et al. (2013) provide examples of the use of Gaussian copulas in ensemble postprocessing.

8.4.3. Member-by-Member Postprocessing

The nonparametric MBMP method presented in Equation 8.58 operates as a univariate postprocessing method that separately transforms individual members of an ensemble of scalar forecasts. However, because the procedure is a monotonic transformation, the resulting transformed ensemble members are ordered in the same way as their counterparts in the original raw ensemble. Accordingly, multivariate collections of MBMP-transformed ensembles inherit the rank correlation (Equation 3.34) structure of the original raw ensemble, so that the independently transformed ensembles can be assembled into a plausible multivariate forecast, in a way that is closely aligned with the ECC approach (Schefzik, 2017).

Figure 8.20 illustrates the result, for bivariate forecasts of 2 m temperatures, and the corresponding coldest temperatures during the previous 6 h. Figure 8.20a shows the result when these two temperature variables have been separately adjusted using Equation 8.58. The correlation between the two variables is realistically strong, and nearly all of the points (black) are above the dashed 1:1 line, with only a few physically implausible points (gray) below, even though the two MBMP adjustments were computed independently. Figure 8.20b shows the corresponding result when the points are drawn randomly from NGR (Section 8.3.2) distributions that were also fit independently to the two sets of temperature forecasts. In this case the correlation is nearly zero, and a large fraction of the points are below the 1:1 line.

Pinson (2012) describes a similar but less general approach, applied to bivariate (u, v) forecast wind vectors.

8.4.4. Analog Methods

A rather different approach to multivariate ensemble postprocessing searches an archive of past forecasts of the same kind that are currently available and chooses one or more of these previous forecasts as analogs for each ensemble member. The collection of observed weather states or events following these analogs is then assembled as a discrete ensemble representing a sample from the forecast PDF.

Roulston et al. (2003) used this approach to obtain probabilistic forecasts of time series of wind speeds having realistic temporal correlation characteristics, to support decision making in wind power management. Hamill et al. (2006, 2015) and Hamill and Whitaker (2006) implemented this idea for areal forecasts of precipitation amounts that appropriately capture the spatial correlations. However, the result is an estimate of a high-dimensional forecast distribution that is difficult to communicate concisely. Figure 8.21 shows one approach to conveying a portion of this information, namely, maps of exceedance probabilities ($\times 100$) for three precipitation amounts. The format is informative but incomplete, since for example it does not indicate or allow computation of such quantities as probability of at least 10 mm at

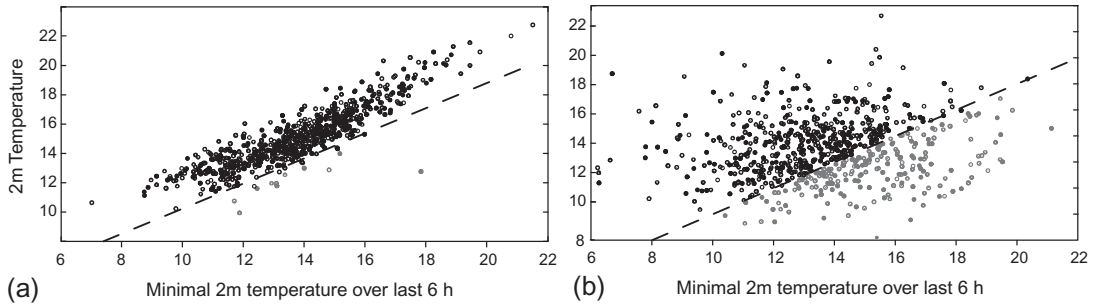

FIGURE 8.20 Scatterplots of ensemble-forecast surface temperatures versus corresponding coldest forecast temperatures in the previous 6 h (a) after MBM postprocessing, and (b) sampled from independent univariate NGR distributions. Values below the dashed 1:1 line, shown in gray, are physically implausible. *Modified from Van Schaeybroeck and Vannitsem (2015).*

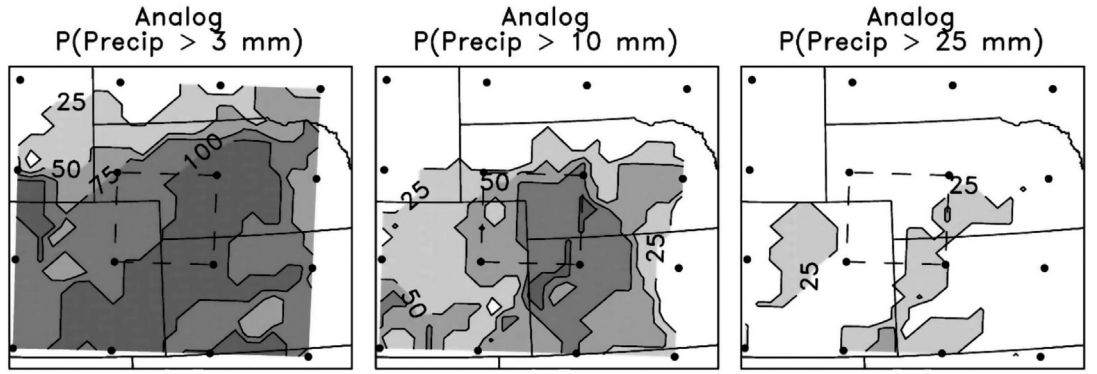

FIGURE 8.21 Exceedance probabilities (×100) for three precipitation amounts, derived from historical observations following forecasts analogous to a current ensemble forecast. *Modified from Hamill and Whitaker (2006). © American Meteorological Society. Used with permission.*

one location given at least 10 mm at another, or the probability that at least 3 mm will occur everywhere in the domain.

8.5. GRAPHICAL DISPLAY OF ENSEMBLE FORECAST INFORMATION

A prominent attribute of ensemble forecast systems is that they generate large amounts of multivariate information. As noted in Section 3.6, the difficulty of gaining even an initial understanding of a new multivariate data set can be reduced through the use of well-designed graphical displays. It was recognized early in the development of what is now ensemble forecasting that graphical display would be an important means of conveying the resulting complex information to forecasters (Epstein and Fleming, 1971; Gleeson, 1967), and operational experience is still accumulating regarding the most effective means of doing so. This section presents a selection of some graphical devices that are in current use for displaying ensemble forecast information.

Perhaps the most direct way to visualize an ensemble of forecasts is to plot them simultaneously. Of course, for even modestly sized ensembles each element (corresponding to one ensemble member) of such a plot must be small in order for all the ensemble members to be viewed simultaneously. Such collections are called *stamp maps*, because each of its individual component maps is sized approximately like a postage stamp, allowing only the broadest features to be discerned. For example, Figure 8.22 shows 51 stamp maps from the ECMWF ensemble prediction system, for surface pressure over western Europe ahead of a large and destructive winter storm that occurred in December 1999. The ensemble consists of 50 members, plus the control forecast begun at the "best" initial atmospheric state, labeled "deterministic predictions." The subsequently analyzed surface pressure field, labeled "verification," indicates a deep, intense surface low centered near Paris. The control forecast missed this important feature completely, as did many of the ensemble members. However, a substantial number of the ensemble members did portray a deep surface low, suggesting a potentially actionable probability for this destructive storm, 42 h in advance. Although fine details of the forecast are difficult if not impossible to discern from the small images in a stamp map, a forecaster with experience in the interpretation of this kind of display can get an overall sense of the outcomes that are plausible, according to this sample of ensemble members. A further step that sometimes is taken with a collection of stamp maps is to group them objectively into subsets of similar maps using a cluster analysis (see Chapter 16).

Part of the difficulty in interpreting a collection of stamp maps is that the many individual displays are difficult to comprehend simultaneously. Superposition of a set of stamp maps would alleviate this difficulty if not for the problem that the resulting plot would be too cluttered to be useful. However, seeing each contour of each map is not necessary to form a general impression of the flow, and indeed seeing only one or two well-chosen pressure or height contours is often sufficient to define the main features, since (especially away from the surface) typically the contours roughly parallel each other. Superposition of one or two well-selected contours from each of the stamp maps often does yield a sufficiently uncluttered composite to be interpretable, which is known as the *spaghetti plot*.

Figure 8.23 shows four spaghetti plots for the 5520-m contour of the 500 mb surface over North America, as forecast (a) 72, (b) 96, (c) 120 and (d) 144 h after the initial time of 0000 UTC, 15 March 2018. In Figure 8.23a the ensemble members generally agree quite well, and even with only the 5520-m contour shown the general nature of the flow is clear. At the 96-h lead time (Figure 8.23b) the ensemble members are still generally in close agreement about the forecast flow, except over the northeast Pacific. The 500 mb field over most of the domain would be regarded as fairly certain except in this area. At the 120- and 144-h lead times (Figure 8.23c and d) there is still substantial agreement about (and thus relatively high probability would be inferred for) the developing cutoff low over the eastern Pacific, but the forecasts throughout the domain have begun to diverge quite strongly, suggesting the pasta dish for which this kind of plot is named. Spaghetti plots have proven to be quite useful in visualizing the evolution of the forecast flow, simultaneously with the dispersion of the ensemble. The effect is even more striking when a series of spaghetti plots is animated, which can be appreciated at some operational forecast center websites.

It can be informative to condense the large amount of information from an ensemble forecast into a small number of summary statistics, and to plot maps of these. By far the most common such plot, suggested initially by Epstein and Fleming (1971), is simultaneous display of the ensemble mean and standard deviation fields. That is, at each of a number of gridpoints the average of the ensemble members is calculated, as well as the standard deviation of the ensemble members around this average.

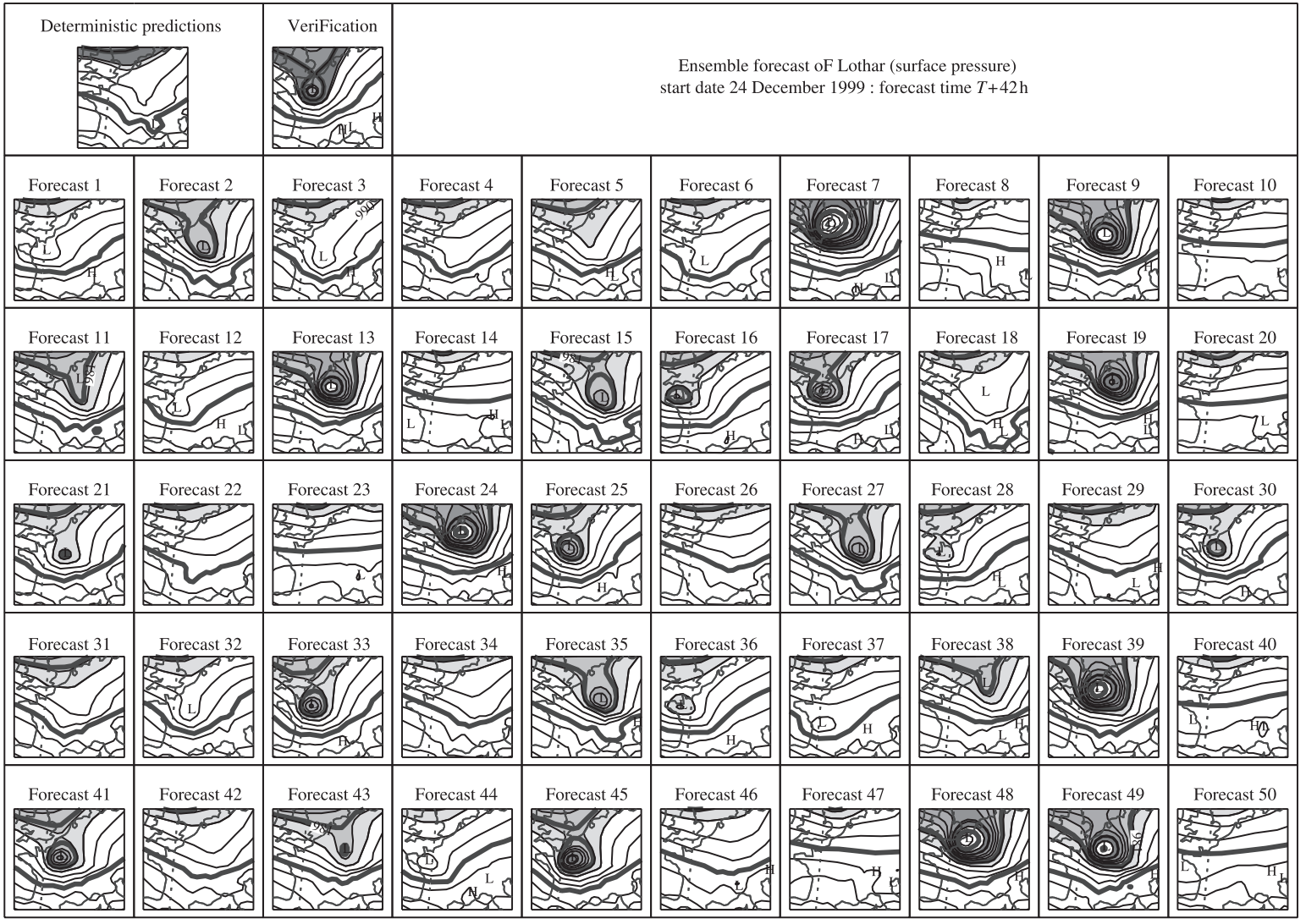

FIGURE 8.22 Stamp maps from the 51-member ECMWF ensemble forecast for surface pressure over western Europe. The verification shows the corresponding surface analysis 42h later during winter storm Lothar. *From Palmer et al. (2005a).*

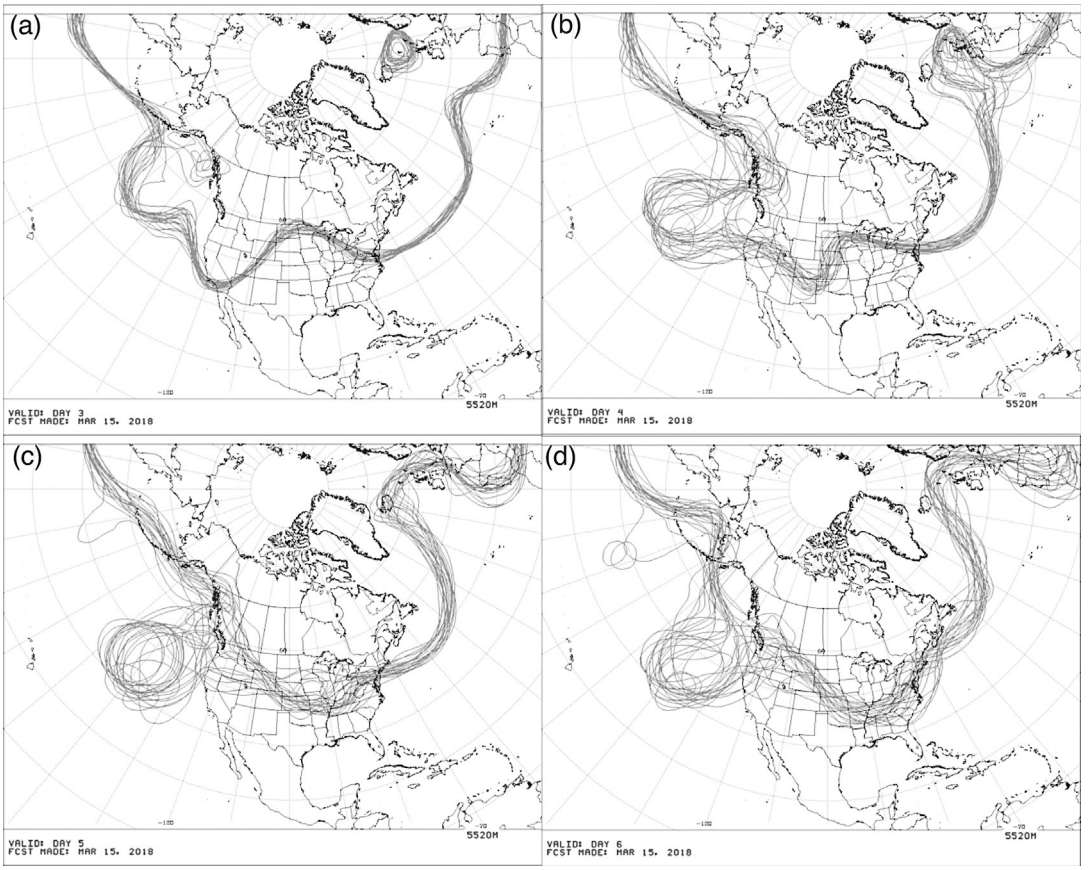

FIGURE 8.23 Spaghetti plots for the 5520-m contour of the 500 mb height field over North America, showing forecasts for (a) 72 h, (b) 96 h, (c) 120 and (d) 144 h after the initial time of 0000 UTC, 15 March 2018. *From www.cpc.ncep.noaa.gov*.

Figure 8.24 is one such plot, for a 96-h forecast of 500 mb heights over much of the northern hemisphere, valid at 0000 UTC, 19 March 2018. Here the solid contours represent the ensemble-mean field, and the shading indicates the field of ensemble standard deviations. Large standard deviations indicate that the possible trough or cutoff low over the northeastern Pacific, corresponding to the dispersion in the spaghetti plot in Figure 8.23b, is fairly uncertain.

Gleeson (1967) suggested combining maps of forecast u and v wind components with maps of probabilities that the forecasts will be within 10 knots of the eventual observed values. Epstein and Fleming (1971) suggested that a probabilistic depiction of a horizontal wind field could take the form of Figure 8.25. Here the lengths and orientations of the arrows indicate the means of the forecast distributions of wind vectors, and the probability is 0.75 that the true wind vectors will terminate within the corresponding ellipse. It has been assumed in this figure that the uncertainty in the wind forecasts is described by the bivariate normal distribution, and the ellipses have been drawn as explained in Example 12.1.

Ensemble forecasts for surface weather elements at a single location can be concisely summarized by time series of boxplots for selected predictands, in a plot called an *ensemble meteogram*.

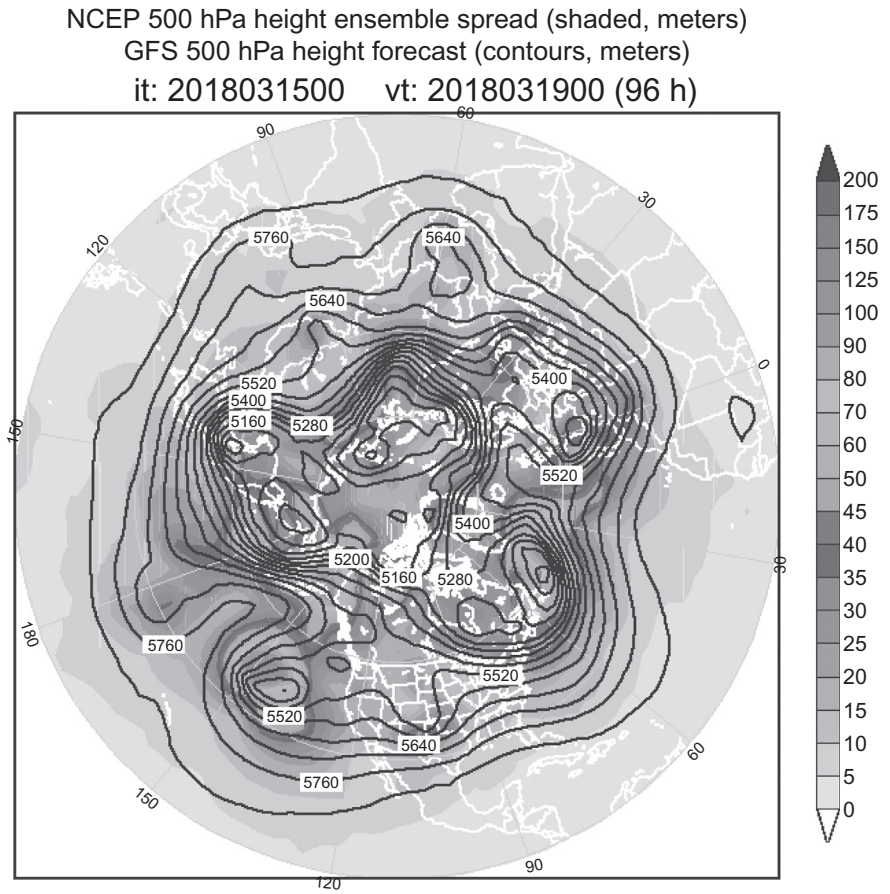

FIGURE 8.24 Ensemble mean (solid) and ensemble standard deviation (shading) for a 96-h forecast of hemispheric 500mb heights, valid 0000 UTC, 19 March 2018, corresponding approximately to Figure 8.23b. *From http://www.emc.ncep.noaa.gov/.*

Each of these boxplots displays the dispersion of the ensemble for one predictand at a particular forecast lead time, and jointly they show the time evolutions of the forecast central tendencies and uncertainties, through the forecast period. Figure 8.26 shows an example from the Japan Meteorological Agency, in which boxplots representing ensemble dispersion for four weather elements at Tsukuba are plotted at 6-hourly intervals. The plot indicates greater uncertainty in the cloud cover and precipitation forecasts, and the increasing uncertainty with increasing lead time is especially evident for the temperature forecasts.

Figure 8.27 shows an alternative to boxplots for portraying the time evolution of the ensemble distribution for a predictand. In this *plume graph* the shadings indicate the PDF of ensemble dispersion, as a function of time into the future for forecast surface temperatures at Hamburg, Germany. The ensemble can be seen to be quite compact early in the forecast, and expresses a large degree of uncertainty by the end of the period.

Finally, information from ensemble forecasts is very commonly displayed as maps of ensemble relative frequencies (Equation 8.4) for dichotomous events, which may be defined according to a threshold

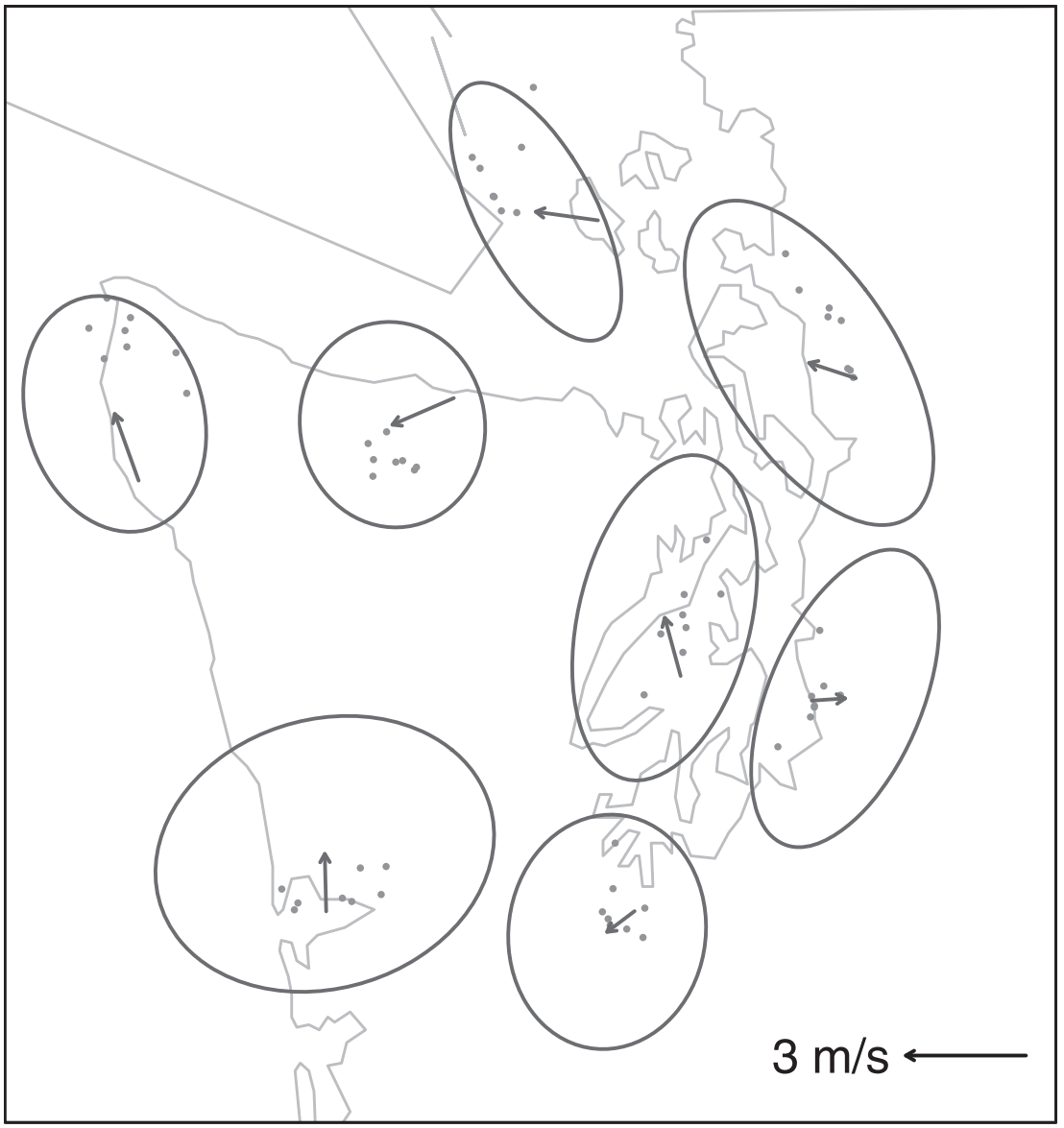

FIGURE 8.25 Seventy-five percent probability ellipses for bivariate NGR surface wind vector forecasts at stations in northwestern Washington State, valid at 0000 UTC 20 Oct 2008, at a lead time of 48 h. Arrows indicate the mean vectors, the bases of which locate the forecast locations. The eight raw ensemble members are shown as gray dots. *Modified from Schuhen et al. (2012). © American Meteorological Society. Used with permission.*

for a continuous variable. Such maps are often labeled as portraying "probability," but unless the underlying ensemble members have first been subjected to postprocessing using methods such as those described in Section 8.4.2 or 8.4.3, this description is generally not justified. Figure 8.28 shows an example of a very common plot of this kind, for ensemble relative frequency of >5 mm of precipitation at a lead time of 48 h.

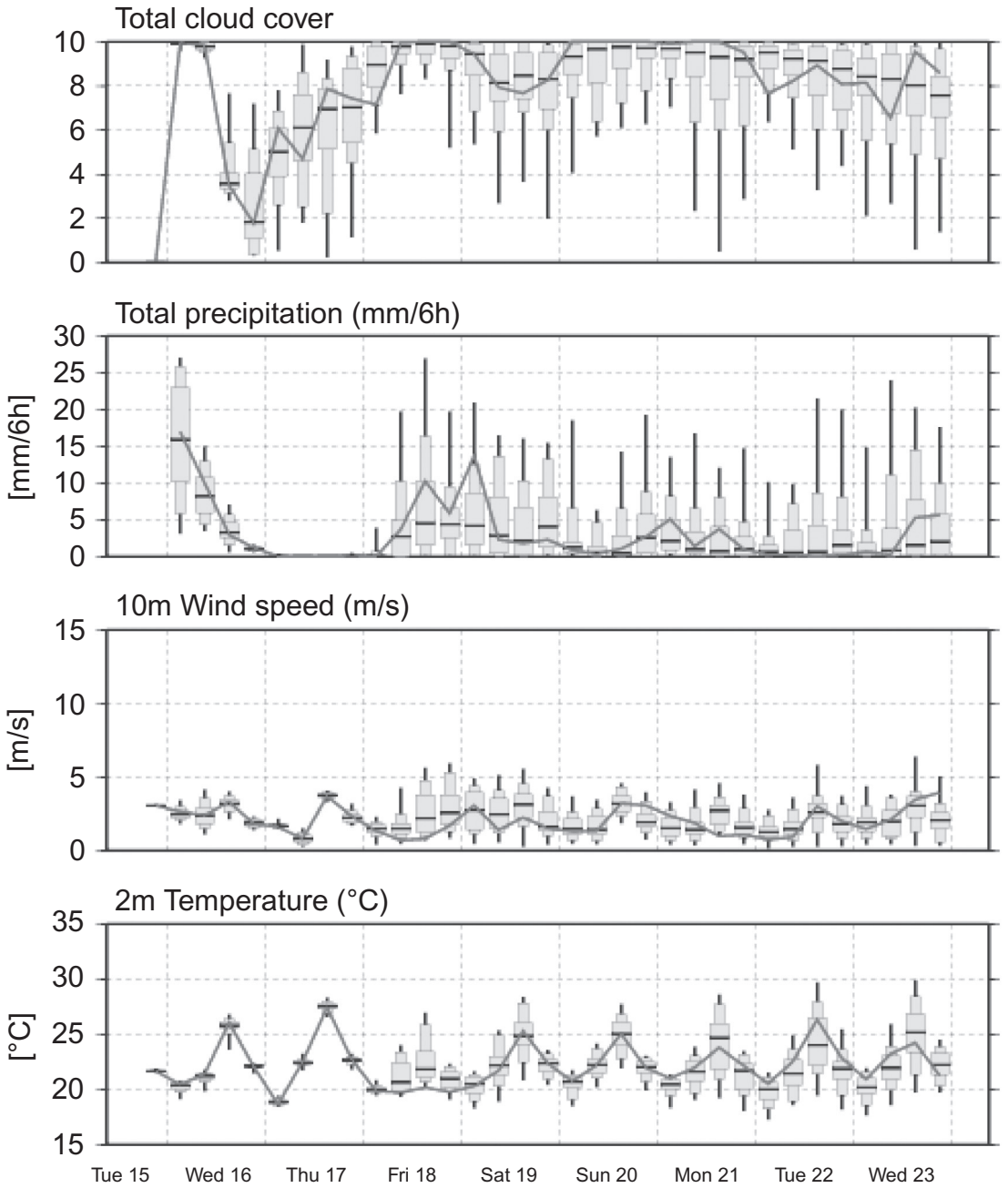

FIGURE 8.26 Ensemble meteogram for Tsukuba, Japan, from a Japan Meteorological Agency forecast ensemble begun on 1200 UTC, 15 June 2010. Wider portions of boxplots indicate the interquartile ranges, narrower box portions show middle 80% of the ensemble distributions, and whiskers extend to most extreme ensemble members. Solid line shows the control forecast. *From gpvjma.ccs.hpcc.jp*.

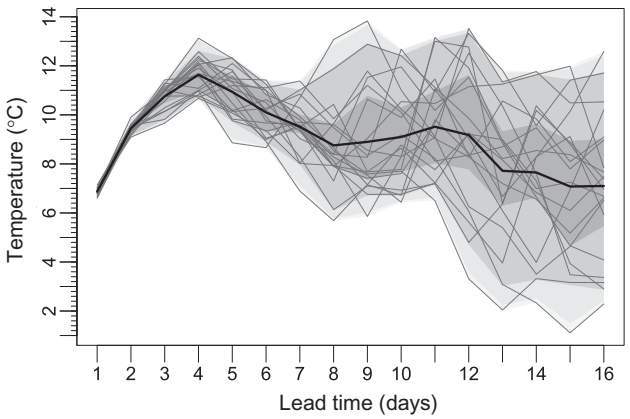

FIGURE 8.27 A plume graph, indicating probability density as a function of time, for a 16-day forecast of surface temperature at Hamburg, Germany, initiated 27 October 2010. The black line shows the ensemble mean, the light lines indicate the individual ensemble members, and the gray shadings indicate 50%, 90% and 99% probability intervals. *From Keune et al. (2014). © American Meteorological Society. Used with permission.*

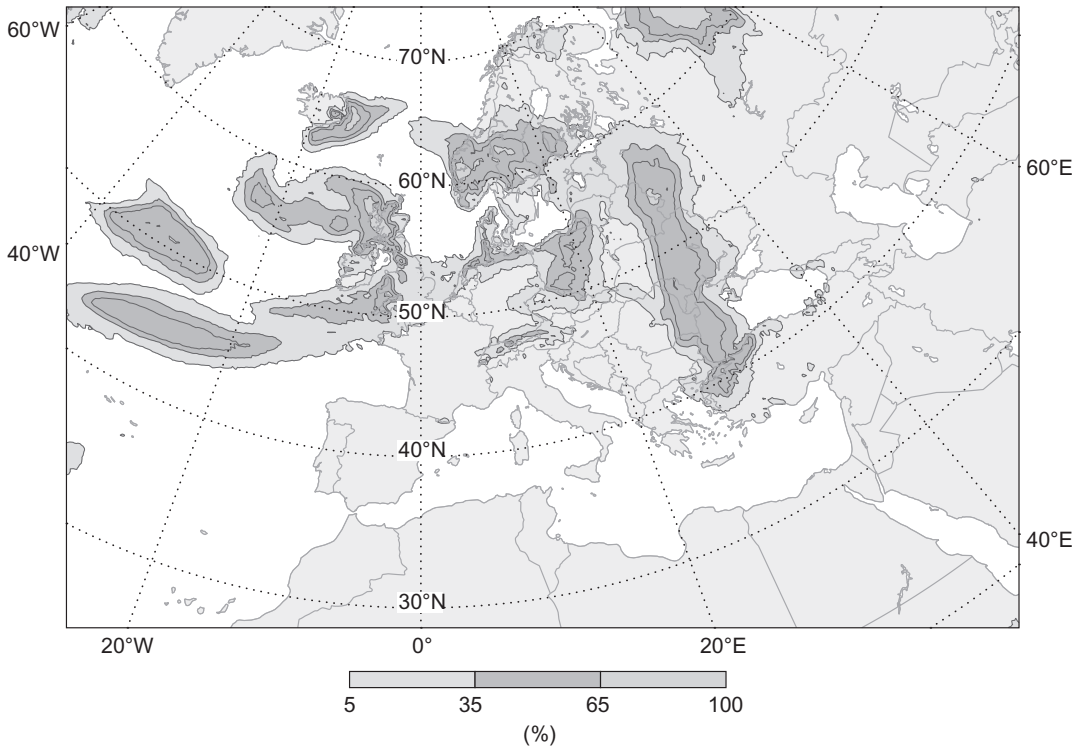

FIGURE 8.28 Ensemble relative frequency for accumulation of >5 mm precipitation over Europe, 48 h ahead of 27 July 2017. *From Buizza and Richardson (2017).*

8.6. EXERCISES

8.1. For the nonhomogeneous Gaussian regression model with $a=-0.5$, $b=1.4$, $c=1.5$, and $d=3.0$, compute the 10th and 90th percentiles of the predictive distribution if the ensemble consists of exchangeable members, and
(a) the ensemble mean is 5.25°C and ensemble variance is 0.25°C^2.
(b) the ensemble mean is 5.25°C and ensemble variance is 0.10°C^2.

8.2. The five bias corrected, nonexchangeable ensemble members shown in Figure 8.10 are 285.2, 285.6, 290.8, 291.2, and 292.5 K. Using the fact that each of the Gaussian BMA dressing kernels have standard deviation 2.6 K, compute $\Pr\{y \leq 293\,\text{K}\}$, which was the verifying observation.

8.3. Figure 8.9a shows XLR functions in the form of Equation 8.31, where the right-hand side is $0.836\sqrt{q} - 0.157 - 1.222\ \sqrt{\bar{x}_t}$. What is the probability of precipitation between 10 mm and 25 mm if the ensemble mean is
(a) 0 mm?
(b) 5 mm?
(c) 20 mm?

8.4. A five-member ensemble of temperature forecasts is NGR-transformed to yield a Gaussian predictive distribution with mean $\mu_1=27.0$°C and standard deviation $\sigma_1=1.5$°C. Wind speed forecasts for the same location and lead time are similarly postprocessed to yield a zero-truncated Gaussian predictive distribution with location parameter $\mu_2=1.5$ m/s and scale parameter $\sigma_2=1.5$ m/s. Use the ECC-Q approach together with the empirical copula structure in Figure 8.18 to compute a bivariate ensemble, jointly for the postprocessed temperatures and wind speeds.