Chapter 8
From Dynamical Model Predictions to Seasonal Climate Forecasts

Simon J. Mason

Producing a seasonal climate forecast from a dynamical model involves a great deal more than simply running the model and viewing the results. The first problem is to decide which dynamical model(s) should be run given the practical constraints of computing resources. In this chapter the pros and cons of using the more computationally intensive fully coupled models compared to atmosphere-only models are discussed. After running a dynamical model, regardless of its complexity, corrections need to be made for systematic errors because the model's climatology and that of the observed climate are invariably different. Some simple procedures for correcting these systematic errors are assessed, but more sophisticated methods are advisable to adjust for spatial displacements of the model climate. Since the model predictions represent large spatial averages, and generally are presented as seasonal averages, downscaling may be required to make the forecast relevant for specific locations, and to provide more detailed information about the statistics of weather within the season. Commonly used spatial and temporal downscaling procedures are described. Some procedures for describing the uncertainty in the forecast are discussed (further details are provided in Chapter 9). Evidence is presented that forecasts can be improved by combining outputs from different models. Finally, the reliability of the forecast needs to be determined by verification of a historical set of forecasts. Verification procedures are discussed in Chapter 10.

8.1 Introduction

In Chapter 7, the procedures for constructing a statistical model for generating seasonal climate forecasts were described. Although dynamical models have been described in detail in Chapters 3 and 6, the procedures for using such models to

Simon J. Mason
International Research Institute for Climate and Society

A. Troccoli et al. (eds.), *Seasonal Climate: Forecasting and Managing Risk.*

produce forecasts are far from straightforward. These procedures are described in this chapter, beginning with a discussion of alternative methods of running the dynamical models (Section 8.2), followed by explanations of how to correct for systematic errors in the outputs of the models (Section 8.3) and to tune the predictions so that they become valid for specific locations (Section 8.4). Procedures for obtaining a probabilistic forecast from an ensemble of model predictions are then discussed in Section 8.5, and finally methods for combining predictions from different models are outlined (Section 8.6).

8.2 One-Tiered and Two-Tiered Forecasting

As discussed in Chapters 3 and 4, seasonal climate forecasting is premised upon feedbacks between the atmosphere and boundary conditions at and near the earth's surface. When producing seasonal climate forecasts using general circulation models (GCMs), there are a number of fundamentally different ways of modelling these interactions between the atmosphere and the lower boundary. These approaches range in complexity: in the simplest case, only the atmosphere is forecast using dynamical models while the boundary conditions are specified by persisting the most recently observed values; in the most complex case the atmosphere together with all the various components of the lower boundary thought to be of importance to atmospheric variability at seasonal timescales are modelled as fully interacting. These two extremes, as well as some intermediate options, are discussed in further detail below (Section 8.2.1), and arguments for and against the various levels of complexity in the modelling are considered in Section 8.2.2.

8.2.1 One- and Two-Tiered Forecasting Designs

The simplest method of dynamically modelling the climate system at seasonal timescales is to model only the atmosphere while specifying values for the various parameters of interest in the lower boundary. If forecasts of the atmosphere are to be made, future values for the boundary conditions have to be specified, and so these values have to be forecast prior to integrating the atmospheric model. A "two-tiered" forecast is thus required: forecasts of the boundary conditions are made first, followed by forecasts of the atmosphere with the forecast boundary conditions prescribed (Bengtsson et al. 1993).

Two-tiered forecasting systems invariably involve a system in which sea surface temperatures (SSTs) are forecast first, while procedures for forecasting the other components of the atmospheric boundary are not explicitly mentioned. Forecasts of

land-surface conditions, for example, generally are produced by coupling a land-surface model to an atmospheric model, even in two-tiered systems in which SSTs are prescribed. Forecasts of SSTs have involved methods from as simple as persistence of the latest observed conditions, through statistical forecasts and partial-ocean hybrid model forecasts, to basin forecasts from fully coupled models, or some combination of the above. Forecasts of land-surface conditions, including of the biosphere, remain relatively primitive compared to forecasts of the sea surface, primarily because of a paucity of observational data, and there are even substantial problems using the best estimates of the latest observed conditions (Anderson and Ploshay 2000).

Two-tiered approaches allow the boundary to influence the atmospheric variability over the period of model integration, but do not permit the atmosphere to feedback to the boundary. Rather than specifying the boundary conditions at the ocean surface and allowing no feedback from the atmosphere, highly simplified models of the oceans can be coupled to the atmospheric model. Although fully non-linear ocean models coupled to simplified atmospheric models, known as hybrid models (Barnett et al. 1993), have been popular, their counterpart models have not been widely used in seasonal forecasting of the atmosphere. Such slab ocean models would allow two-way heat fluxes between the atmosphere and ocean, but do not involve ocean circulation. This restricted feedback of the atmosphere to the ocean may have advantages over the standard two-tiered approaches, and such models deserve further attention.

The most complex method of modelling the climate system at seasonal timescales is to model all components of the climate system thought to be relevant at seasonal timescales. Operational examples of such models involve separate models for the atmosphere and ocean that are run synchronously and interactively. Such "fully-coupled" models generate forecasts of the atmosphere and of the boundary conditions simultaneously, and so sometimes are referred to as "one-tiered" forecasting systems.

8.2.2 Advantages of One- and Two-Tiered Forecasting Designs

One-tiered forecasting systems, or fully coupled models, are widely acknowledged to represent the state-of-the-art in seasonal climate forecasting. However, comprehensive comparisons of one and two-tiered systems are lacking (see Graham et al. 2005 and Guérémy et al. 2005 for some preliminary results), and regardless of relative performances, there are advantages to two-tiered approaches that are likely to contribute to their continued use for the next several years at least. Some of these advantages are outlined in Sections 8.2.2.2–8.2.2.4 after a brief summary of the advantages of one-tiered systems (Section 8.2.2.1).

8.2.2.1 Advantages of One-Tiered Forecasting Designs

One-tiered forecasting systems represent the most comprehensive attempt to incorporate all the components of the climate system thought to be relevant for understanding atmospheric variability at seasonal to interannual timescales. Because they allow for feedbacks between the atmosphere and the other components of the climate system, coupled models should, theoretically, provide the most realistic representation of how the real climate system operates, and hence should be able to generate better forecasts than their two-tiered counterparts. An implicit assumption in two-tiered systems is that the atmosphere responds to SST forcing, but does not in turn affect the oceans. As indicated in Chapters 4 and 6, strong feedback between the ocean and the atmosphere occurs within the equatorial Pacific Ocean, for example, while atmospheric influence on tropical Indian Ocean variability appears to be stronger than the influence of the ocean on the atmosphere. Similarly, in the extra-tropics, pioneering research on ocean-atmosphere interaction over the North Pacific indicated that the ocean variability is more a response to atmospheric variability than vice versa.

In a two-tiered system, where the atmosphere is uncoupled from the ocean, unrealistic forcing of the model atmosphere can occur. For example, Indian monsoon rainfall in most uncoupled models is positively correlated to tropical Indian Ocean SSTs because of higher moisture fluxes, but in coupled models, and in the real world, negative correlations are evident because the ocean surface heats in response to changes in the trade winds (Wu and Kirtman 2005). The imposed forcing in two-tiered systems can therefore result in incorrect simulations, whereas the coupling permitted in one-tiered designs should result in a more realistic representation of observed climate variability. Although coupled models do not currently perform much better because of moisture flux problems (Wu et al. 2006), improvements in the model physics should result in more realistic simulations, whereas improvements in the physics of an uncoupled atmospheric model will not necessarily resolve the problem.

8.2.2.2 Computational Advantages of Two-Tiered Forecasting Designs

Fully coupled models require huge computational resources, and so currently are used for operational forecasting only at some of the so-called Global-Producing Centres (GPCs). Because of the computational costs, forecasts are compromised, either in the resolution of the model atmosphere and/or ocean, the ensemble size, the lead-time, the frequency of forecast production, and/or the generation of retrospective forecasts used for assessing forecast performance and calibrating for model errors. For example, of the seven models that constituted part of the DEMETER experiment (Palmer et al. 2004), only three have hindcasts extending

back more than 40 years, and these for only four initialization dates during the year and for a sufficient number of ensemble members to estimate the models' respective mean responses only in the tropics. Alternative savings involve coupling a global atmospheric model to a single-basin ocean-model, and prescribing sea temperatures elsewhere (Ineson and Davey 1997).

The computational advantages of two-tiered forecasting systems could permit the integration of the atmospheric model at higher resolutions than are possible when the same model is run in one-tiered mode, or the generation of a larger ensemble. In countries where only moderately powerful computing resources are available, the computational advantages enable two-tiered dynamical seasonal forecasts to be generated locally. These computational advantages are enhanced by a relatively weak improvement in forecast quality with increased spatial resolution in two-tiered systems compared to their one-tiered counterparts. Apparently the coupling of the ocean and atmosphere is modelled most effectively at high resolutions, whereas if the atmospheric model is uncoupled many of the benefits of improved resolution are lost.

Additional computational advantages can be achieved if no attempt is made to assimilate observed data into the atmospheric model. While there is some resultant loss of predictability from initial conditions in the first few weeks of the forecast, the loss of skill at longer lead-times is considered minimal, and is partly offset by avoiding problems associated with model drift (Chapter 6). The computational costs involved in data assimilation are substantial, and are an essential component of ocean forecasting (see Chapter 5), and so assimilation is dispensable only if no ocean model is to be run.

8.2.2.3 Atmospheric Predictions from Improved Sea Surface Temperature Predictions in Two-Tiered Forecasting Designs

The quality of seasonal climate forecasts of the atmosphere is intricately related to the quality of the forecasts of the lower boundary forcing, particularly of SSTs. If coupled model forecasts of the lower boundary can be improved by using other forecasting methods, it may be possible to improve on the atmospheric forecasts by using these superior boundary forcings in a two-tiered scheme. For example, since forecasts of persisted SST anomalies are difficult to outperform at lead-times of less than about 3 months, prescribing SST anomalies at short lead-times may provide improved skill in two-tiered atmospheric predictions. While fully coupled models can outperform two-tiered systems in which SSTs are prescribed from simple statistical models, the two-tiered systems may perform at least equally as well as fully coupled systems when more skilful SST forecasts are used. More detailed research on the comparative performances of one- and two-tiered systems is required.

8.2.2.4 Research Value of Integrations with Controlled Boundary Conditions

Apart from the benefits of two-tiered forecasts in an operational setting, atmospheric GCM integrations uncoupled to ocean models can be of considerable research value. Some of the more valuable examples of such research are discussed in this section.

Atmospheric GCMs forced with observed SSTs have been analysed extensively. Such experiments attempt to provide estimates of the potential predictability (an indication of the upper limit of predictive skill) of the climate at seasonal and longer timescales. Typically estimates of potential predictability involve comparing the variability in the simulated atmospheric responses across different ensemble members (intra-ensemble variability) with the inter-annual variability of the ensemble mean to obtain an estimate of the contribution of the SST forcing to the total variability: if the intra-ensemble variability is small compared to the interannual, then the SSTs are evidently constraining the (model's) atmospheric variability, implying that there is predictability. Alternatively, if ensemble size is small, a more reliable approach may be to compare the interannual variability of the simulated atmosphere when forced with observed as against climatological SSTs. Other strategies include, for example, comparing the forecast distributions to the climatological distribution, or examining the distribution of the proportion of ensemble members exceeding the climatological median. However, all strategies are based on estimating how much of the atmospheric variability is forced, and how much is free internal variability. Detailed investigations of the potential predictability of the atmosphere were conducted as part of the PRediction Of climate Variations On Seasonal to inter-annual Time-scales (PROVOST; Branković and Palmer 2000; Palmer et al. 2000), and Dynamical Seasonal Prediction (DSP; Shukla et al. 2000) projects.

Differences in the skill of simulating observed atmospheric variability when a model is forced using persisted instead of observed SST anomalies can be used to diagnose the loss of predictability that results from having imperfect SST forecasts. In the Sahel, for example, where rainfall variability is strongly affected by SSTs in the tropical Atlantic Ocean, the weak persistence of SSTs from 1 month to the next effects poor forecast skill of seasonal rainfall over the region, but skill increases markedly with decreasing lead-time (Ward 1998).

Alternative experiments have considered the effects of prescribing SSTs in only one (or occasionally two) of the three main ocean basins, or in specific areas thought to have important influences on atmospheric variability. Such experiments are valuable in diagnosing model systematic errors and also forecast biases that may result from using incomplete forecasts of SSTs in operational settings. However, because of their artificial nature, coupled with the fact that the total oceanic impact on the atmosphere may not be a simple linear combination of the individual oceanic impacts, they cannot adequately provide answers to questions concerned with the influence of SSTs in specific areas on the global (or regional)

atmosphere. Other problems with these kinds of experiments result from the creation of artificial SST gradients at the edges of the domain of perturbed temperatures, even when the temperatures are reduced to climatology smoothly.

8.3 Systematic Model Error Correction

Regardless of how seasonal climate forecasts are made using atmospheric GCMs, substantial differences between the observed and model climates invariably are evident, and need to be corrected in order to provide reasonable forecasts. Definitions of various types of systematic error are provided in Section 8.3.1. Statistical tests for identifying errors in model output are detailed in Section 8.3.2, and are followed by a critique of commonly used methods for correcting for these errors (Section 8.3.3). Discussion on the correction of spatial errors in model output is provided in Section 8.3.4.

8.3.1 Systematic Model Errors

Systematic errors refer to any difference between the observed and the model climatology (implied definitions in the literature vary). The simplest form of systematic error is the mean bias: more generally, the central tendency of the model climatology differs from that for the observations. An example is shown in Fig. 8.1a, which compares observed[1] with simulated June–August precipitation rates for the 50-year period 1951–2000 averaged over a large area of eastern Africa (10°N–10°S, 3050°E). The precipitation was simulated using the ECHAM 4.5 model (Roeckner et al. 1996) at a resolution of about 2.8° and forced with observed SSTs, and the statistics were obtained using 24 ensemble members. The graph shows the frequencies of average precipitation rates over the 3-month period, and clearly indicates a bias in the model: simulated rates are consistently too high. This bias in the mean precipitation rate is known as an unconditional bias because the model rate is too high regardless of the actual simulated (or forecast) rate.

As well as indicating a mean bias, Fig. 8.1a indicates that the variance of the simulated precipitation rates is larger than the observed variance. Variance biases can occur even when the mean bias is minimal, as shown in Fig. 8.1b, which shows precipitation rates for March–May instead of June–August. Variance biases

[1] The New et al. (2000) gridded rainfall data were used. These data are based on station observations interpolated to a grid.

are also known as conditional biases because the model anomalies are consistently too strong (weak) when the model variance is larger (smaller) than the observed variance. Systematic errors in reproducing the shape of the climatological distribution can also occur: in Fig. 8.1c, the model's mean and variance are too high, while the skewness is too low. This example is for June–August precipitation averaged over part of southern Africa (20–30°S, 15°–25°E).

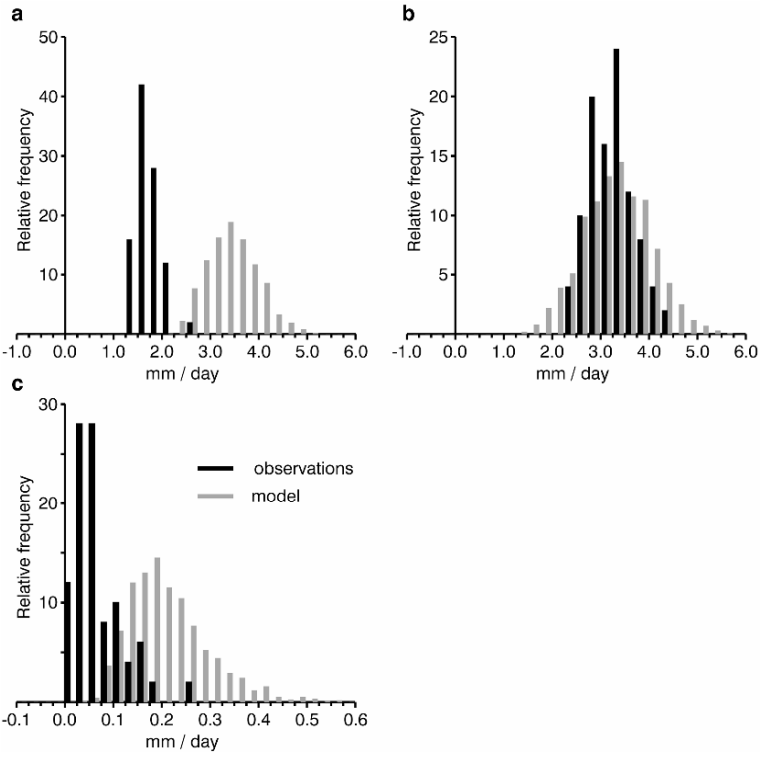

Fig. 8.1 Example of model systematic errors: (a) area-averaged June–August 1951–2000 observed (black) and simulated (grey) daily precipitation intensities over eastern Africa showing a mean bias; (b) area-averaged March–May 1951–2000 observed (black) and simulated (grey) daily precipitation intensities over eastern Africa showing a variance bias; and (c) area-averaged June–August 1951–2000 observed (black) and simulated (grey) daily precipitation intensities over southern Africa showing mean, variance, and shape biases

Any differences between the observed and model climatologies are symptomatic of differences in behaviour of the real and model atmospheres. However, these differences should be distinguished from predictive errors, which relate to differences in the observed and simulated/forecast climate for specific cases. Predictive errors relate to the skill of the model forecasts, and are not necessarily symptomatic of systematic errors. In the absence of any inherent predictability the individual forecasts will not generally correspond well with the observations, but the model climatology may be realistic.

8.3.2 Detecting Systematic Model Errors

Standard error statistics, such as the (root) mean-squared error and mean absolute error, measure differences between paired observations and simulations/forecasts. As a result, these metrics do not distinguish systematic model errors from predictive errors. Ideally these two forms of error should be distinguished. In the following sections, selected tests for systematic errors are described. A summary of the tests is presented in Table 8.1, where a few additional tests that are not discussed in the following text are mentioned. [See Sheskin (2007) for further details.] A selection of tests for predictive skill is provided in Chapter 10.

Table 8.1 Statistical tests, and respective distributional assumptions, for identifying systematic errors. All tests assume independence of the samples. Additional details of these tests can be obtained from Conover (2001) and Sheskin (2007)

Systematic Error	Test	Additional assumptions
All	Kolmogorov-Smirnov	
	Fisz-Cramér-von Mises	
	Relative entropy	
Central-tendency		
Mean	Student's *t*	Equal variance; normality
Median	Mann-Whitney *U*	Equal variance; similar shape
	Median	
Spread	*F*	Normality
	Siegel Tukey, David's, Mood's	Equal central-tendency; symmetry
	Moses	Similar shape

The standard test for systematic model errors is the two-sample Kolmogorov-Smirnov test, which compares the cumulative distributions derived from the model and the observed climatologies.[2] The test compares the maximum vertical difference between these two empirical distributions, D, against a null distribution for the statistic; if the maximum vertical distance is large, the two distributions are likely to be different, and so the model climatology does not match that for the observations. The null distribution for D, and for all the other statistics discussed in this section, depends upon the number of cases used to construct the empirical cumulative distributions, and so depends upon the number of years and the ensemble size. Systematic errors can be identified more robustly given large numbers of cases.

The two-sample Kolmogorov-Smirnov test does not distinguish between different forms of systematic error. Separate tests are available for identifying mean- and variance biases, while biases in skewness and higher order moments (collectively referred to as errors in the shape of the distribution) are not widely

[2] Alternatives include the Fisz-Cramér-von Mises test (the integral of the squared differences between the two cumulative distributions) and relative entropy (Elmore 2005).

used. Mean biases are commonly identified using Student's *t*-test, which compares the differences in the climatological means. The test is highly sensitive to distributional assumptions (the observed and model climatologies should both be Gaussian), and so alternative tests are required that are not sensitive to these assumptions. The alternative tests compare differences in medians rather than means, since the median is not strongly influenced by the presence of a few extreme values, and so they test for a bias in the central tendency rather than strictly testing for a mean bias. The Mann-Whitney *U*-test is the most frequently used alternative to the *t*-test. The *U*-test effectively calculates the probability that a randomly sampled observation is larger (or smaller) than a randomly sampled forecast. This probability should be 0.5 if there is no bias in the central tendency of the model climatology. Strictly, the *U*-test should not be used if there is a variance bias, in which case the median test is preferable. The median test calculates the proportion of observations (or simulations) above the pooled median, and is free of any assumptions about other forms of systematic error. Again the proportion should be 0.5 if there is no mean bias.

Tests for variance bias (or, more generally, dispersion bias) are numerous. The most commonly used is the *F*-test, which compares the ratio of the variances of the observations and simulations to Fisher's *F* distribution. The ratio should be 1.0 if there is no variance bias, but the test is highly sensitive to distributional assumptions, and should probably be used infrequently. Unfortunately, there is no obvious alternative test to use; there are of the order of 100 candidate tests, but virtually all of them carry some distributional assumptions. A Moses-type test, which is designed to compare the frequencies of extreme values in two samples, can be recommended if the assumption that there are no errors in the shape of the distribution is reasonable. There are a number of variations on this test, but the core idea is to compare the central tendencies of measures of dispersion of random sub-samples of the observations and simulations (Kössler 1999). If there is no dispersion bias, the central tendencies will be similar.

Although these tests are used widely when considering climate change simulations, in seasonal climate forecasting systematic errors are usually removed using a simple statistical correction (Section 8.3.3) and are then ignored, and so the tests are rarely applied. As long as there is some predictive skill, forecast accuracy need not be adversely affected by such errors. If the model's atmosphere is responding to anomalous boundary forcing in the correct direction (for example, the model indicates unusually dry conditions when unusually dry conditions occur) then this variability is believable regardless of any conditional and unconditional biases.

8.3.3 Correcting Systematic Model Errors

Although the terms are often used in different ways in the climate literate, a distinction is sometimes drawn between "calibrated" model output, which has

been corrected for systematic errors, and "recalibrated" model output, which has been corrected for model skill in addition to systematic errors. The procedures described in this section perform model calibration. Some model recalibration schemes are discussed later (see also Chapter 9, Section 9.3).

Removal of systematic errors usually involves application of the generalized formula:

$$\hat{z}_o = g_o[g_m[z_m]]^{-1}, \tag{8.1}$$

were z_m is the modelled value of the parameter of interest, $\hat{z}_o$ is the calibrated value, g_m is a function that transforms the modelled values onto a new distribution, and g_o is a function that transforms the observed values onto a distribution that is assumed to be the same as that for g_m.

In the simplest case, g_m and g_o are functions that centre the data to have a mean of zero (i.e. $g[z] = z - \overline{z}$, where g is a transformation function applied either to the model or the observed values, z is a model or observed value, and $\overline{z}$ is the corresponding climatological mean). In this case Eq. (8.1) simply subtracts the difference in the sample means between the model and the observations from the model climatology, thus removing the mean bias. An alternative option, which is suitable when correcting for variables with a zero bound (such as precipitation), scales by the ratio of the observed and simulated means (i.e. $g[z] = z/\overline{z}$). This scaling affects the variance (but not the shape) of the bias-corrected model climatology, unlike the centring procedure.

Scaling assumes that any errors in the variance are simply a function of the mean bias (i.e. that the coefficients of variation for the model and observed climatologies are identical). Since this assumption is frequently invalid, corrections for both mean and variance are generally made by standardizing the data (i.e. $g[z] = \left(z - \overline{z}\right)/s$, where s is the climatological standard deviation). Standardization is a widely used procedure that successfully removes mean and variance biases, but can be problematic when used on data with a zero bound, and/or when there are systematic errors in the shape of the model's climatological distribution. These problems occur because it is generally implicit that application of Eq. (8.1) implies application of the formula

$$\hat{z}_o = F_o[F_m[z_m]]^{-1}, \tag{8.2}$$

where F_m is a cumulative distribution function for the model data and F_o is a cumulative distribution function for the observations. Specifically, when data are standardized, F_m represents the normal distribution function fitted to the model data, and returns the quantile associated with the corresponding standard normal deviate; the corresponding quantile from the normal distribution fitted to the observed

data is then used to obtain the transformed value.[3] This procedure works only to the extent that the normal distribution provides a good fit to both sets of data, otherwise errors in estimating the quantiles of the two distributions can result in unreasonable transformations. Consider the effects of standardizing the June–August precipitation data for southern Africa, described above: the model and observed data are plotted as empirical distribution functions in Fig. 8.2, and the fitted normal distribution functions are superimposed. For model precipitation rates of less than about 0.1 mm/day (the driest 5–10% of cases), the transformed precipitation is negative, as illustrated by the corresponding vertical legs of the dotted line.

Unless both the model and the observed data are normally distributed, standardization should not be performed. Instead more appropriate distribution functions should be applied in Eq. (8.2). While the empirical distribution functions could be used, the function for the model data isknown better than for the observations because of the larger sample size provided by the multiple ensemble members. The relatively poor representation of the empirical distribution function for the observations can create problems particularly when transforming extreme values. The alternative is to use a fitted distribution other than the normal distribution. The two-parameter gamma distribution is an attractive option for data that are positively skewed and zero-bound, and its parameters are easy to estimate

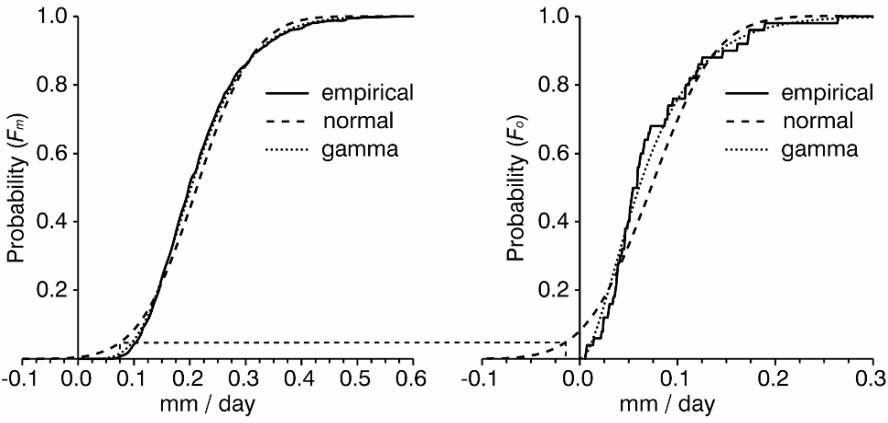

Fig. 8.2 Examples of empirical and fitted distribution curves for area-averaged June–August 1951–2000 simulated (left) and observed (right) daily precipitation intensities over southern Africa. The dotted line represents a transformation of model precipitation by standardization to remove mean and variance biases

[3] The conversion from a deviate on the standard normal distribution is redundant, but the application of the cumulative normal distribution function in Eq. 8.2 is implicit, as evident when standardization is viewed graphically, as in Fig. 1.2.

when the skewness is not too marked (Wilks 2005). Fitted gamma distributions for the model and observed data in Fig. 8.2 are shown, and the improvement in the estimation of the quantiles over the fitted normal distributions is evident not just in the tails of the distribution.

The procedure of fitting appropriate distribution functions[4] and applying Eq. (8.2) requires methods for estimating the distribution parameters. In most cases, the simplest procedure is to use the method of moments: for a given distribution the mean and variance of the distribution can be calculated analytically in terms of the distribution parameters, and so these parameters can be set to give a distribution with the same mean and variance as the sample data. These parameter estimates can be sensitive to outliers, and so a more robust procedure, known as L-moments, has been developed based on order-statistics (Hosking 1990). A more popular approach, however, is to use maximum likelihood estimation: the parameter values that maximise the likelihood of yielding the sample data are obtained. In a few cases, such as with the normal distribution, these values can be derived easily, but for most distributions they have to be obtained using iterative procedures.

8.3.4 Correcting Spatial Errors in Model Output

One aspect of systematic error that has not been addressed in Section 8.3.3 is the problem of spatial errors in model output; climate features in the model are often displaced, as shown by example in Fig. 8.3. The figure compares the first principal components of ensemble-mean forecasts of October–December precipitation for eastern Africa from the ECMWF model (Palmer et al. 2004) and of observed rainfall for the same period (New et al. 2000). While the model successfully forecasts rainfall variability over much of the region to the east of about 30°E, the main mode of variability, which involves region-wide anomalously wet or dry conditions, is displaced to the west by about 15°. Such displacements can result in poor predictions if they are not corrected.

If climate features in the model are displaced relative to the observations, even by only short distances, comparing the model output at any grid with the corresponding observations using the types of methods described in Section 8.3.2 is inappropriate a priori. Instead, the spatial structure of the model output requires correction prior to correcting any systematic errors in the climatological distributions for individual gridpoints. Standard methods for correcting such spatial errors involve multivariate statistical techniques that typically address mean and dispersion

[4] Different distributional forms could be used for the model and the observed data if their distributions have different shapes.

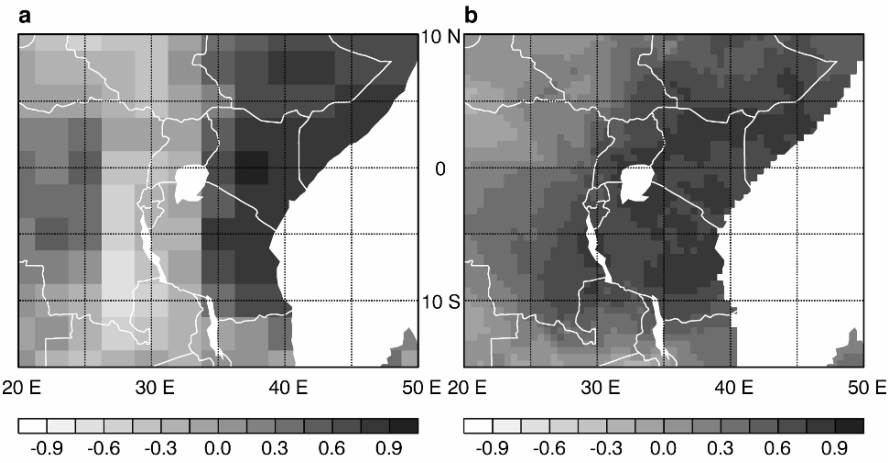

Fig. 8.3 Correlations with the respective first principal components of (a) forecast precipitation and (b) observed precipitation over eastern Africa for October–December 1961–2000. The precipitation forecasts are from the August runs of the ECMWF model generated as part of the DEMETER project (Palmer et al. 2004)

biases[5] at the same time. In practice, most techniques used to correct for spatial errors address forecast skill, and so perform model recalibration rather than simply model calibration, as distinguished in the previous section. A sample of spatial correction techniques is presented in this section. [See von Storch and Zwiers (1999) and Wilks (2005) for further details.]

The two most widely used statistical techniques for correcting spatial systematic errors are extensions to multiple linear regression, namely maximum covariance analysis (MCA) and canonical correlation analysis (CCA). The procedures are essentially identical to those described in Section 7.4.2.5 of Chapter 7, and so are discussed only briefly here. The idea is to use the model predictions as the predictors in a statistical prediction model. In both MCA and CCA spatial patterns of precipitation variability, for example, in the model are identified that have similar temporal variability to spatial patterns in the observations. Since the similarities are defined only in terms of the temporal variability there is no explicit attempt to match the spatial patterns. Consequently, in practice, MCA and CCA may be able to identify a feature of the climate such as the PNA pattern whose temporal variability may be predicted well because of a realistic modelled response to El Niño conditions, but which may be displaced in the model (as in Fig. 8.3). Both procedures will effectively transform the model's imperfect PNA prediction to a more realistic prediction of PNA variability.

[5] Non-linear statistical downscaling techniques such as neural networks could theoretically correct for shape biases in addition to mean and variance biases.

Whichever approach is used for correcting systematic spatial errors, the size of the domain(s) used requires consideration. If the objective is simply to correct for the displacement of climate features in the model, then forecasts only from nearby areas should be considered. However, multiple CCA or MCA corrections would then be necessary, and these would have to be blended somehow. Using larger domains helps to avoid artificial spatial noise in the corrected fields, and is computationally more efficient, but the statistical correction procedures are likely to identify teleconnection patterns, and so are no longer conducting purely spatial correction. Whether or not the identification of teleconnection patterns is undesirable is an open question, and the general question of domain selection requires further research.

8.4 Statistical Downscaling

A typical gridpoint in a GCM used to make seasonal predictions represents an area of about 50,000–100,000 km^2, which is invariably much coarser than the spatial scales at which opportunities to apply seasonal climate forecasts exist. The GCM output therefore needs to be "downscaled" to resolutions and/or locations commensurate with user-requirements. Downscaling involves the translation of a forecast to a spatial (and/or temporal) resolution that is finer than that at which the forecasts are produced. Reasons for performing downscaling are discussed in more detail in Section 8.4.1, and some examples of spatial downscaling using statistical models are provided. An introduction to some statistical techniques to downscale seasonal forecasts to finer temporal resolutions is given in Section 8.4.2. Dynamical methods of downscaling using limited area models are not discussed.

8.4.1 Spatial Downscaling

Since GCMs are designed to represent planetary scale processes, those processes that operate at spatial scales smaller than the model resolution have to be parameterized. Computational constraints make it impractical to operate GCMs at resolutions that would permit more realistic reproductions of regional climate, and even if computational resources were available, careful re-parameterization of the models would be required (parameterizations are tuned to work at specific model resolutions). Apart from the inevitable errors that arise from the imperfect representation of the real world because of the discretization of space (and time) within GCMs, downscaling is required even for a model that reproduces the observed climate perfectly because of the detailed spatial variability of climate. Such issues are discussed in further detail in the following paragraphs.

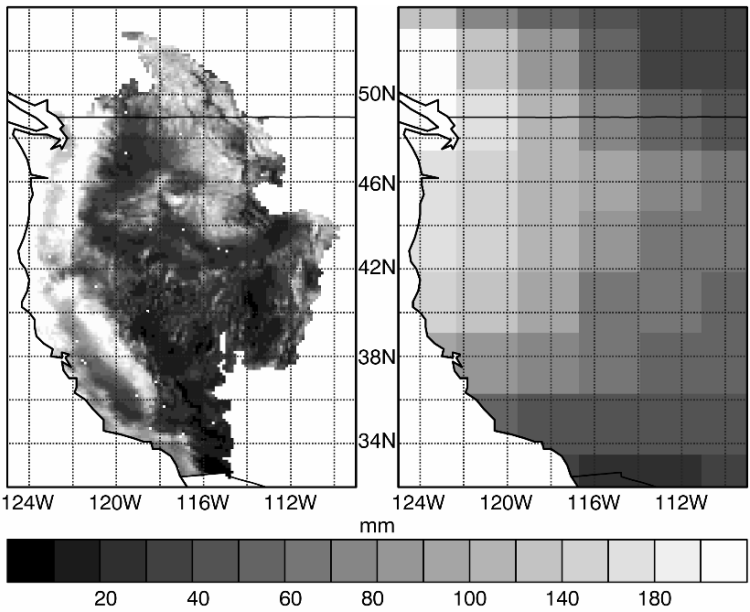

Fig. 8.4 Observed (a) and simulated (b) January–March mean precipitation for 1950–1999

Gridded model output represents an average of an essentially arbitrary area, and so, even if the model reproduces the area-averaged climatology realistically, there may be substantial systematic "errors" when these forecasts are interpreted as representative of specific locations. For example, Fig. 8.4 illustrates averaged January–March precipitation totals for 1950–1999 over part of North America (Fig. 8.4a) together with simulated precipitation totals using the ECHAM 4.5 model averaged over the same period (Fig. 8.4b). The observed data were obtained from the Surface Water Modeling Group at the University of Washington (Maurer et al. 2001, 2002). This dataset is derived from station data spatially interpolated to a grid resolution of 0.125° latitude × longitude over land, which should be compared with the approximately 2.8° resolution of the ECHAM model data. Apart from any errors in the reproduction of the broad-scale climate features, the variability of climate within any of the GCM grids is obvious, and so, at a minimum, GCM grid averages would have to be rescaled to become representative for any specific location.

Detailed spatial variability of mean climate not only affects the systematic "errors" for specific locations, but also translates into detailed variability in the predictability of climate. As a simple illustration, the correlations between the NIÑO3.4 index and observed January–March precipitation over part of North America are shown in Fig. 8.5. Within short distances large differences in the correlation are evident, and imply that GCM output could give highly misleading forecasts for sub-grid areas even after correcting for systematic errors. In addition, because seasonal

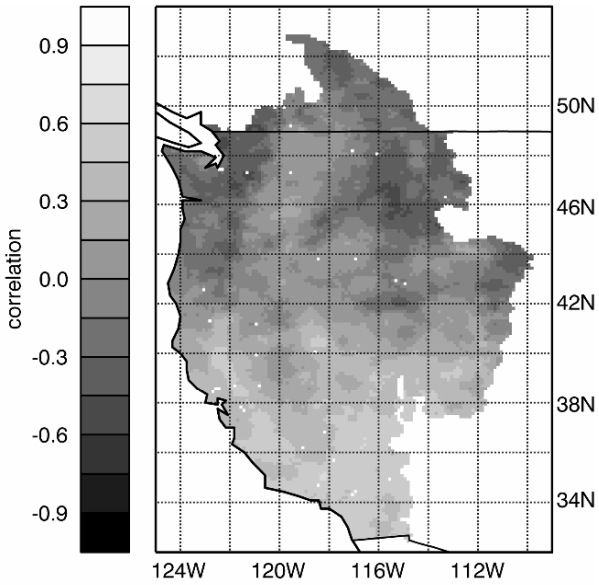

Fig. 8.5 Spearman's correlations between observed January–March seasonal precipitation for 1950–1999 and simultaneous values of the seasonally averaged NIÑO3.4 index

predictability of climate generally is greater for large compared to small area-averages (Gong et al. 2003), performance measures comparing GCM output with commensurate observational data do not necessarily give reliable indications of the accuracy of the models at the spatial scales at which seasonal climate forecasts are to be used. Downscaling is thus required to assess locally specific systematic as well as predictive errors.

If high resolution observational data or data for specific locations are available, detailed spatial corrections can be made to provide forecasts at resolutions that the GCM itself is unable to resolve. To illustrate, the precipitation data for the 50-year period January–March 1950–1999 were used to downscale simulations of precipitation from the ECHAM 4.5 model. A canonical correlation analysis(Chapter 7, Section 7.4.2) was used to downscale the GCM data. Results are shown in Fig. 8.6, which compares the skill of the downscaled predictions with the skill achievable by linearly interpolating the output for surrounding GCM gridpoints. The skill score used (Spearman's correlation) considers only the predictive errors, not any remaining systematic errors.

In Fig. 8.6, results are shown for downscaling the GCM precipitation fields directly, but there have been a number of successful attempts to downscale to station precipitation using other outputs from the GCM. For example, the model's geopotential heights are used frequently, sometimes with more than one level being considered simultaneously. Potential vorticity fields have also been used successfully. However, little attention has so far been given to downscaling multiple fields; if downscaled predictions of precipitation and of temperature are required,

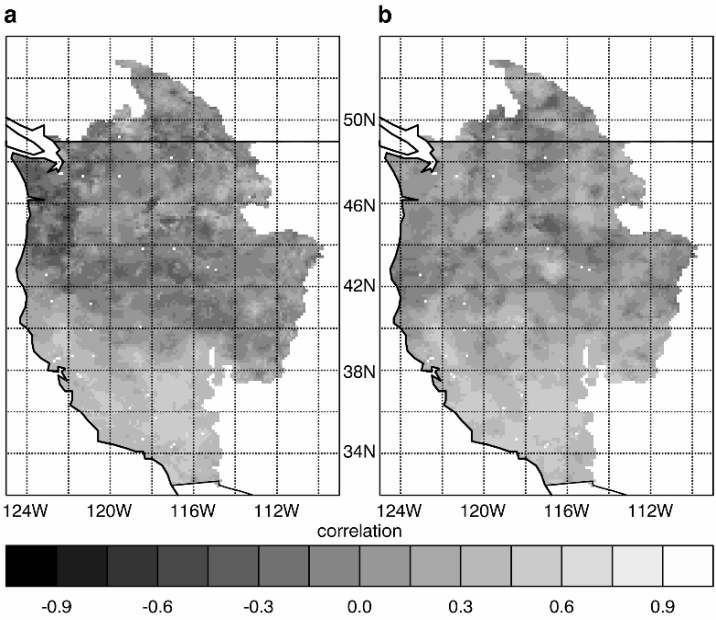

Fig. 8.6 Spearman's correlations between observed and simulated January–March 1950–1999 precipitation. Results are shown for (a) ECHAM 4.5 output linearly interpolated to the 0.125° resolution of the observational data, and (b) ECHAM 4.5 output spatially corrected using canonical correlation analysis. All results are cross-validated using a leave-five-out cross-validation window

for example, these are generally conducted independently, which could result in locally inconsistent results. In contrast, greater attention to correlations between different weather variables has been given in methods of temporal downscaling, and these methods are discussed in the following section.

8.4.2 Temporal Downscaling

Apart from the incompatibility between the spatial resolution of the forecast and that of the observations, other problems with GCM output preclude their application without additional downscaling. An important constraint to the use of GCM output and, more generally, of seasonal climate forecasts, is the temporal resolution of the predictions. As discussed in Chapter 3, seasonal climate is predictable only when the forecast is considered as an aggregate of weather over a period of typically about 3 months; it is not possible to provide accurate predictions of the weather on any given day within the season. However, for many application models, including hydrology and crop models, it is necessary to have forecasts for each day of the season. While the sensitivity of the predictions from such application

models to the precise weather on specific days may be low as long as the seasonal weather statistics are accurate, some means of obtaining atmospheric forecasts at the required temporal resolution is required. In this section, various means of obtaining seasonal forecasts at high temporal resolution are discussed.

Since GCMs are generally run at a temporal resolution of about 20 minutes to generate seasonal forecasts, the simplest solution to the need to obtain weather statistics over the period of the seasonal forecasts would be to use the GCM output. However, there are some severe biases in GCM weather data, which are perhaps best illustrated by considering the frequency distribution of daily precipitation intensities. An example is shown in Fig. 8.7, which compares the frequencies of simulated and observed daily precipitation amounts for San Diego for the 50-year period 01 January 1950–31 December 1999. The model clearly underestimates the frequency of dry days (note that the *y*-axis is logarithmic) and of precipitation intensities exceeding about 4 mm/day. In other words, the model generates too much drizzle, a problem that is characteristic of GCM-based forecasts for all timescales.

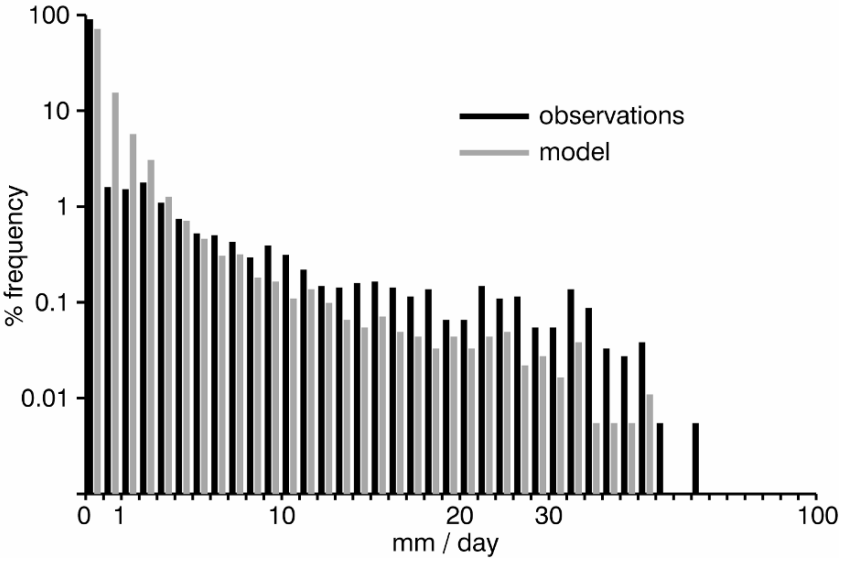

Fig. 8.7 Relative frequencies of observed and ECHAM 4.5-simulated daily precipitation intensities for San Diego for the 50-year period 1951–2000. The simulated precipitation is for the gridpoint nearest to San Diego. Note the logarithmic y-axis, and the uneven intervals on the x-axis

An alternative to using the GCM daily output is to disaggregate the seasonal forecast using statistical methods. Disaggregation involves computing sub-seasonal weather statistics that are consistent with the seasonal forecast. One commonly used statistical procedure is the analogue method, which uses the observed sub-seasonal statistics of seasons that are similar to the forecast for the target season. (See Chapter 7 for more details on statistical forecasting techniques.) Such procedures can be limited severely by sample size, and so are most

commonly used in places such as Australia where datasets are relatively long, and so the number of analogue years large, compared with those for many other countries. An alternative approach is to use simple statistical relationships between seasonal climate and sub-seasonal weather statistics. For example, simple relationships between seasonal rainfall totals and the frequencies of raindays or of heavy raindays can be regressed. Such relationships could then be used to estimate weather statistics contingent upon the forecast for the seasonal aggregate.

Based on observed relationships between seasonal climate and sub-seasonal weather, fairly sophisticated statistical techniques for simulating weather over a season have been developed. These procedures are based on "weather generators", of which there are a wide range of different designs (Wilks and Wilby 1999). Most weather generators have been constructed to generate series of daily precipitation, and invariably consider the question of precipitation occurrence separately from precipitation amount. Precipitation occurrence is modelled in one of two ways: either as a chain-dependent process or by spell-lengths. As a chain-dependent process, the probability of precipitation is calculated contingent upon the occurrence of precipitation on the previous (day), which is equivalent to modelling precipitation occurrence as a Markov process. For example, the seasonal cycle of probability of precipitation in San Diego given that the previous day was wet is compared for that given that the previous day was dry in Fig. 8.8a and b, respectively. Throughout the year the probability of a wet day is considerably higher given that the previous day was wet compared to when the previous day was dry. These differences in precipitation probability are indicative of the persistence of weather in San Diego, indicating that spells of weather tend to last a few days, rather than weather changing randomly from day to day. Weather generators based on Markov models simulate a series of precipitation occurrence by randomly generating wet and dry days by considering the weather generated on the previous (few) day(s), and should thus generate weather spells with realistic duration. The second approach to modelling precipitation occurrence is to generate a string of

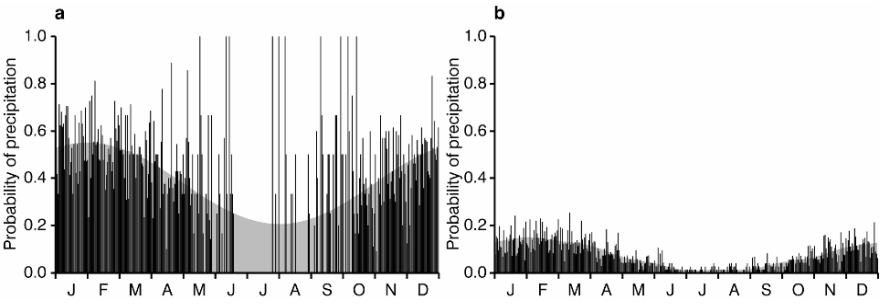

Fig. 8.8 Annual cycle of the probability of precipitation occurrence for San Diego for the 50-year period 1951–2000, given that (a) the previous day was wet, and (b) the previous day was dry. The black vertical bars show the probabilities calculated for each day, while the grey shading indicates smoothed probabilities using the first few harmonics of the annual cycle

alternating wet and dry spell-lengths. The frequency distributions of observed wet and dry spell-lengths are usually modelled using a negative binomial distribution. The spell-length generator operates by randomly drawing alternating random spell-lengths drawn from the corresponding negative binomial distributions.

Whichever way precipitation occurrence is modelled, the generated occurrences of precipitation need to be conditioned somehow on the seasonal forecast. Again there is a range of options for modelling this conditioning (Wilby et al. 2002). For example, if a Markov model is used, the probability of precipitation can be conditioned not only on the generated occurrence of precipitation on the previous day(s), but also on some aspect of the seasonal forecast, such as the predicted rainfall total exceeding some predefined threshold. As a simple example, Fig. 8.9 compares the probabilities during El Niño and La Niña years of daily precipitation during the winter months of January–March in San Diego exceeding various thresholds. Rainfall at all but the highest intensities typically occurs more frequently under El Niño conditions than under La Niña conditions.

Alternatively, the probability of precipitation could be estimated using a statistical model. This regression approach has the advantage of not dividing the degrees of freedom up by the repeated splitting of the dataset when calculating conditioned parameters, but does require the form of the relationship between the conditioning variable and the precipitation probability to be specified.

A more sophisticated approach to conditioning the generator on the seasonal forecast involves modelling the occurrence of precipitation on the basis of the predicted daily sequence of the large-scale atmospheric circulation. Since the daily

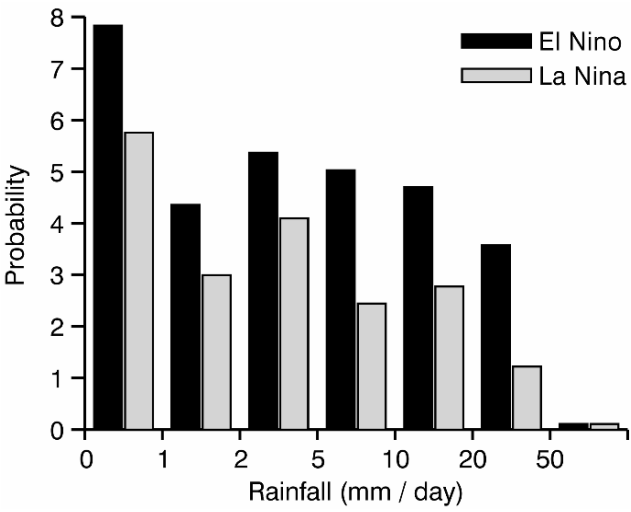

Fig. 8.9 Histograms of wet spells in San Diego commencing any time between 01 January and 31 March for the 50-year period 1951–2000, given that the January–March averaged NIÑO3.4 index was greater than +0.7 (i.e. El Niño conditions prevailed), and less than –0.7 (i.e. La Niña conditions prevailed)

atmospheric circulation has the weather persistence implicitly built in, there is no need to condition the precipitation on the previous days' weather (either by Markov or spell-length modelling). Because the weather persistence is implicit in this approach, these models are called "hidden Markov models" (Robertson et al. 2004). Hidden Markov models condition the precipitation probability by identifying specific weather patterns and then classifying each day into one of the patterns. The daily sequence of the atmospheric circulation over the period of the seasonal forecast would normally be provided by the GCM, and so the procedure is somewhat similar to the spatial downscaling procedures described in Section 8.4.1. Apart from obvious differences in the form of the statistical model used, and in the temporal resolution of the GCM output (daily compared to seasonal average), the procedures have this in common: the large-scale GCM output is statistically corrected to provide an estimate of precipitation (daily occurrence or seasonal total).

Precipitation intensity is modelled in a similar way to spell-lengths: the distribution of non-zero precipitation intensities is represented (frequently by a gamma or mixed exponential distribution), and random intensities are generated for days in which precipitation is specified to occur. Thus intensity is modelled subsequent to occurrence. Again, the intensity of precipitation can be conditioned upon some aspect of the seasonal forecast if there is evidence that seasonal variability is affected by changes in precipitation intensity. Since the inter-annual variability of precipitation can be affected by changes in precipitation frequency and/or intensity, weather generators can be designed to account for both/either effect.

Weather generators can be designed to model a suite of meteorological parameters in such a way that the relationships between the parameters are consistent with the relationship in the real. For example, in many parts of the world there is a relationship between precipitation occurrence and maximum temperature, and some applications of seasonal forecasts it may be important to retain this relationship. Generated temperatures (and other parameters) are conditioned upon the generated precipitation occurrence. In a similar way, it is possible to generate weather sequences at a range of locations so that the generated weather is spatially realistic by accounting for the spatial correlations in the meteorological parameters. This consideration may be important in hydrological modelling, for example, where the spatial distribution of precipitation across a river catchment is important in affecting runoff.

8.5 Using Ensembles

There are two primary motivations for generating an ensemble of predictions (whether from a single model or a set of models). One is that the average of a set of predictions more closely approximates the climate signal than the prediction from any single ensemble member. However, a second motivation is to obtain

some indication of the uncertainty in the prediction.[6] Since the ensemble mean indicates only the central tendency of the predictions, a separate measure is required to indicate the uncertainty. However, it is not obvious how the ensemble members can be used to indicate forecast uncertainty, or even whether they are successful in doing so. In Section 8.5.1 how uncertainty in a forecast can be communicated is discussed. Then some methods for describing the forecast uncertainty using an ensemble are considered (Section 8.5.2). A description of procedures for assessing how well the ensemble can be used for indicating changes in forecast uncertainty is reserved for Chapter 10.

8.5.1 Forecast Uncertainty, Forecast Confidence and Forecast Probabilities

Given the inherent uncertainty in forecasting seasonal climate conditions, the forecaster needs to provide some indication of this uncertainty. A common way of communicating such uncertainty is by indicating the level of confidence to be placed in the forecast. This level of confidence is inversely related to the degree of uncertainty in the forecast: when uncertainty is large a low level of confidence in the forecast is communicated, whereas when uncertainty is reduced confidence increases. The distribution of possible outcomes defines the full extent of the uncertainty in the prediction, but this distribution is unknown and so has to be approximated somehow. Once approximated, the forecaster's confidence can then be defined. The confidence in the forecast can be communicated in a number of ways, and how the ensemble may be used depends on which format is adopted.

One of the simplest ways of indicating forecast uncertainty is to specify a range of values within which the observed value is expected to lie with a predefined level of confidence. Usually this level of confidence is kept fixed from forecast to forecast, and the varying uncertainty is reflected by adjusting the width of the interval. Thus, when uncertainty is large (small) the interval is made wide (narrow). For example, forecast A, which states that there is a 90% probability of a seasonal rainfall total being between 100 and 200 mm indicates greater uncertainty than forecast B, which states that there is a 90% probability of the total being between

[6] Here, and elsewhere in this Section, "uncertainty" relates to the range of possible outcomes for a specific target period, and is not the same as, the climatological uncertainty as defined by Murphy (1973a). Murphy's definition is independent of the forecasts themselves, whereas here, as discussed later, uncertainty is represented by the extent to which the forecasts of individual ensemble members for the same target differ. If the forecasts for all the ensemble members are similar, then forecast uncertainty is low, but if they differ substantially then forecast uncertainty is high.

125 and 175 mm.[7] This format is known as a prediction interval (see Chapter 9) and is not widely used in seasonal climate forecasting partly because such intervals are frequently misinterpreted.

An alternative approach to that is more commonly used in seasonal climate forecasting is to fix the interval and to allow the level of confidence to vary. The interval itself can be fixed to meet the user interests, although in practice it is most commonly defined from the terciles of the observed data as measured over a climatological period. The fixed intervals are normally called "categories", along with the unbounded categories either side of the interval. More than one interval can be specified, and in this respect quintiles are being used with increasing frequency. To illustrate: the interval of 100–200 mm used in forecast A above could be used for all forecasts; decreased uncertainty implicit in forecast B would then be communicated by increasing the probability that the seasonal rainfall total will be within this range rather than by narrowing the range. It should be noted that there is no simple relationship between the change in the probability assigned to an interval and the changing level of uncertainty in the forecast, as illustrated in Fig. 8.10. Two forecasts are shown in the figure; forecast A involves less uncertainty

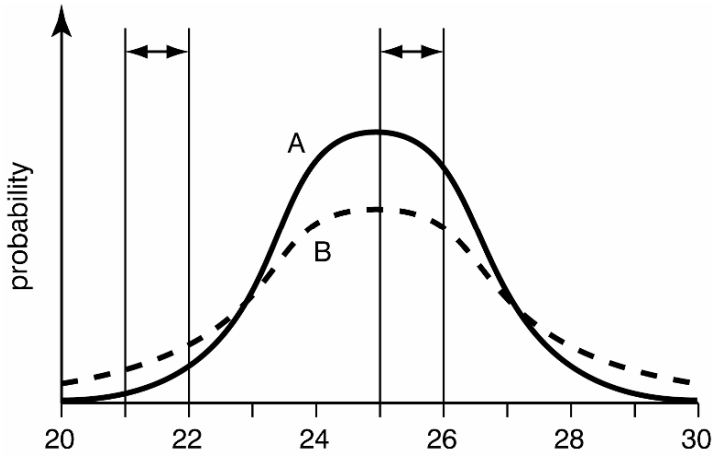

Fig. 8.10 Hypothetical example illustrating the complex relationship between forecast probability and forecast uncertainty. Forecast A (solid line) represents a forecast with relatively low uncertainty, and forecast B (dashed line) represents one with relatively high uncertainty. The narrow vertical lines indicate the limits of intervals for which forecast probabilities are desired. These probabilities are calculated by integrating the areas beneath the lines A and B within the range of the intervals

[7] If the confidence level is α, it would normally be assumed that the probability that the observed value will be less than the lower limit of the interval (125 mm in forecast B) is the same as the probability that the observed value will be greater than the upper limit (175 mm). This probability would be $1–2\alpha$ (5%) However, it is not necessary for the interval to be centred in this way, as long as the corresponding tail probabilities are then specified.

than forecast B. In the interval 21–22°C the probability increases with the more uncertain forecast, but decreases in the interval 25–26°C. This problem in interpretation can be avoided by specifying the probabilities for all categories, and comparing these probabilities to the climatological probabilities for the categories.

More detailed summaries of the distribution of possible outcomes, and thus of the uncertainty in the forecast, are possible by assuming a distributional form and describing this distribution by its parameter values (for example, the mean and variance of a normal distribution). These distributions can be used to estimate probabilities for any intervals, or intervals for any levels of confidence. Alternatively, some of the percentiles of the distribution (fitted or otherwise) can be specified. These options are discussed in more detail in the following sections.

8.5.2 Forecast Ensembles and Forecast Uncertainty

After correcting for systematic errors in the individual ensemble members (Section 8.3), their distribution is supposed to give an indication of the distribution of possible outcomes. The distribution of the ensemble members should therefore indicate the uncertainty in the forecast: in simple terms, if the various ensemble members are forecasting similar values then uncertainty is low, whereas if the values differ widely then uncertainty is high. However, with a finite ensemble size the distribution of the current forecasts is imperfectly sampled, and so the uncertainty implied by the forecasts has to be estimated.

One of the simplest ways of using the ensemble to indicate forecast uncertainty is to estimate the probabilities for categories by counting the proportions of the ensemble members indicating outcomes within each category. Errors in calculating these probabilities by counting can be derived from the binomial distribution, and can be substantial. The probabilities are more reliably obtained by fitting a distribution to the ensemble members using one of the methods described in Section 8.3.3 and then calculating the probabilities from the fitted distribution (Kharin and Zwiers 2003). Further improvements can sometimes be made by assuming a distributional form for the sampling errors associated with each ensemble member rather than for the ensemble distribution as a whole. Each ensemble member is "dressed" with a fitted distribution (Roulston and Smith 2002). One advantage of this approach is that the prediction errors of the forecasts can be accounted for to some extent. Alternatively, the probabilities could be estimated directly using a statistical model, such as a generalized linear model (Tippett et al. 2007). The statistical model would correct for both the predictive and the systematic errors in the model(s).

The use of a statistical model for estimating probabilities does not necessarily mean that the uncertainty implied by the ensemble distribution provides useful information. The most obvious candidates for predictors in the statistical model are the first few moments of the ensemble distribution, and virtually all of the

usable information is in the ensemble mean. The ensemble mean communicates no information about the uncertainty in the forecast, which, instead, is derived from the error variance of the ensemble mean predictions. Errors in calculating the ensemble variance appear to be too large to derive much useful information in the ensemble spread (Kharin and Zwiers 2003). Alternative measures of spread, such as the inter-quartile range, could be used, but more detailed studies are required to identify how much of the variability in the ensemble spread beyond the sampling variability truly represents variability in forecast uncertainty. There has been minimal investigation into the information content of the shape of the ensemble distribution.

Instead of using the ensemble to estimate probabilities for predefined categories, they could be used to estimate the values associated with specific percentiles of the ensemble distribution. For example, the ensemble median is arguably more informative than the mean since the former is amenable to making a simple probabilistic forecast (there is an estimated 50% probability that the observed value will exceed the ensemble median, but the probability of exceeding the mean is unknown unless some distributional assumptions are made). The percentiles can be estimated either by fitting a distribution to the ensemble or to the individual ensemble members, or by treating the individual ensemble members as percentiles of the distribution. The latter approach is implicit when constructing ranked histograms, as discussed in Chapter 10. Effectively, such procedures are an extension to those used for defining prediction intervals since each end of the interval represents a fixed percentile of the forecast distribution.

8.6 Combining Forecasts

There is ample evidence that combining seasonal climate predictions from a suite of models provides an improved forecast over using even the best of the individual models (Doblas-Reyes et al. 2005; Hagedorn et al. 2005). The improvement is evident not only in forecasts of seasonal averages but also in some of the intra-seasonal statistics such as storm frequencies. Similar conclusions can be drawn for forecasts at medium-range and shorter timescales, and multi-model approaches are being used increasingly in climate change work. At all timescales the improvement in the forecasts results from the improved representation of uncertainty arising from imperfections in model physics. In a single-model ensemble, uncertainty is represented only in terms of the initial conditions, and each ensemble member is subject to the same errors in the model physics, so that clustering of forecasts tends to occur. Alternative ways of accounting for the uncertainties arising from model errors include stochastic parameterization, and perturbed physics approaches, but the use of multi-models is likely to remain popular both in research and operations.

Simple averaging of predictions from different models is usually sufficient to improve the quality of a forecast, but it is tempting to weight the models by their respective skill levels. However, a major difficulty in assigning differing weights arises from the limited availability of hindcasts for which to assess relative model performances robustly. If the skill levels of the models cannot be definitively compared, it is then exceptionally difficult to outperform the simple average of the models' respective predictions (Kharin and Zwiers 2002). Further, as the number of models is increased, problems of over-parameterization arise, and so simple averaging of recalibrated model output again generally proves to be the most effective approach.

To illustrate, monthly predictions of the NIÑO3.4 index from three of the models that participated in the DEMETER experiment were combined using a range of methods. Each of the three models (ECMWF, CNRM, and the Met Office) had nine ensemble members, produced forecasts from four start dates, and generated predictions with lead-times of up to 5 months. Hindcasts were available for the 44-year period 1959–2002, and all results were cross-validated using a 3-year cross-validation window (i.e. 1 year either side of the predicted year was omitted). Results for all lead-times and seasons were pooled.

The details of the various combination schemes are not important, but include two Bayesian model weighting schemes, canonical variate analysis, generalized linear models, multiple regression, and stepwise regression. For all but the Bayesian schemes, the respective model ensemble means were used.[8] Simple model averaging (i.e. equal model weighting) was used as a benchmark level of skill. In all cases the forecasts were expressed as probabilities of the NIÑO3.4 index falling below the lower quartile, above the upper quartile, or within the inter-quartile range. The forecasts were evaluated using the quadratic score, which is a measure of the squared error in the probability assigned to the category that verified (Chapter 10). The score ignores the probabilities assigned to the other categories. Since it is an error score, a perfect set of forecasts would achieve a score of 0.0.

Since each of the schemes can be applied to the models individually to recalibrate the model output, the models can be combined in one of two ways: recalibrate each model individually, and then calculate a simple average of the recalibrated predictions ("recalibration"); apply the schemes to all the models simultaneously ("combination"). For the simple equal weighting, the combination and recalibration schemes will give identical results. The scores for the various schemes are illustrated in Fig. 8.11. Most of the schemes improve the forecasts of individual models, and in most cases the combined forecasts (using either combination approach) improve upon the forecasts from the best single model. The

[8] The use of the ensemble mean only generally gave the best results. Alternatives tried were: to include the ensemble variance (which can provide some marginal improvements in skill), and higher moments; to use all the ensemble members; to use the first few principal components of the ensemble members.

averaging of the recalibrated forecast consistently outperforms the combination method, presumably because of the over-parameterization of the latter.

The results shown in Fig. 8.11 are based on combining forecasts from only three models. Since many of the models used in seasonal climate forecasting are fairly closely related, the predictions from these models are often strongly correlated, potentially creating problems of multicolinearity (see Chapter 7, Section 7.4.1). Some success has been achieved in addressing multicolinearity problems by using procedures equivalent to truncated principal components regression. Such an approach will also help to reduce problems of multiplicity that arise from considering the skill of more than one model (Chapter 7, Section 7.4.1), and which are exacerbated when downscaling approaches are incorporated into combination algorithms. However, it is not clear that principal components regression is appropriate in the context of forecasts of precipitation, for example, where the assumptions of multivariate normality are often violated.

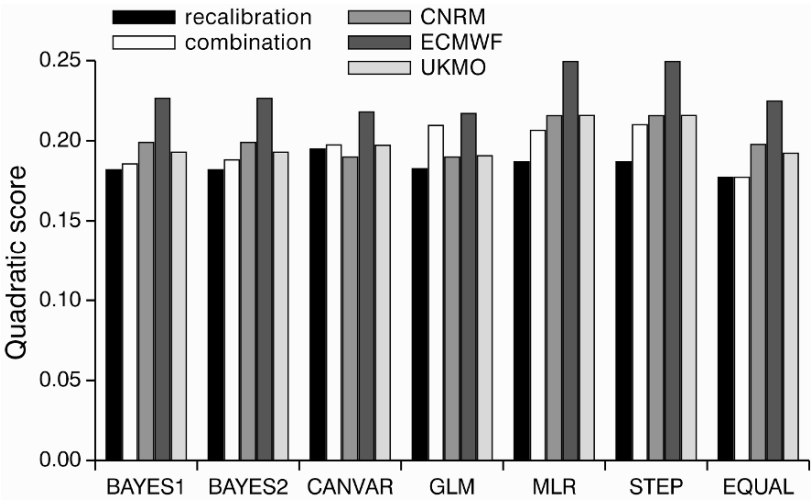

Fig. 8.11 Quadratic scores for monthly predictions of the NIÑO3.4 index, using various forecast combination schemes [Bayesian model weighting schemes (BAYES1 and BAYES2, canonical variate analysis (CANVAR), generalized linear models (GLM), multiple linear regression (MLR), stepwise regression (STEP), and equal weighting (EQUAL)]. The schemes are compared by combining the predictions using two procedures: attempting to account for differences in model skill ("combination"), and by simple averaging of predictions after recalibrating the individual models ("recalibration"). Results for the individual models are shown also

Problems of multicolinearity in forecast combination algorithms, while not unique to multiple regression, which has been a commonly used method for combining forecasts, are not an issue with some alternative combination methods. Canonical covariate analysis, for example, has some strong similarities to principal component regression, and explicitly addresses the inter-correlations of models (and of individual ensemble members, if used). In addition, the first few principal components of the model predictions could also be used in a wide range of alter-

native statistical schemes, such as generalized linear models (Chapter 7, Section 7.4.3). These approaches deserve further consideration, especially in the context of combining predictions of precipitation amounts, where the standard assumptions of multiple regression (multivariate normality) are sometimes invalid.

8.7 Summary

Producing a seasonal climate forecast from a dynamical model involves a great deal more than simply running the model, and viewing the results. The first problem is to decide which dynamical model(s) should be run given the practical constraints of computing resources. Assuming that dynamical models can represent the underlying physical processes correctly, fully coupled models theoretically should give the best predictions of seasonal climate if they can be initialized accurately, but this initialization can be problematic, and computing resources can be prohibitive. An alternative is to use uncoupled atmospheric models and to prescribe the SST forcing. In the latter case, the SSTs have to be predicted first, and so the uncoupled approach involves a "two-tiered" process.

Once model predictions have been made, they then need to be corrected for systematic errors. These errors result from consistent differences between the model and the observed climatologies, and can be identified by differences in the probability distributions of climate parameters for the model and the observed data. However, since one contribution to the systematic errors in the model is that the geography is distorted, simple gridpoint-by-gridpoint comparisons of model and observed climatologies can be inappropriate. Instead some form of spatial correction to the model output is desirable.

Even after systematic error and spatial correction, the model predictions may require further processing in order to be made relevant for specific locations. All dynamical models produce output that represents an averaged value over a gridded area typically of the order of between 10,000 and 100,000 km^2. Because local climate can vary considerably over fairly short distances, especially in areas of marked terrain, this gridded average may be unrepresentative of specific locations within the grid. The model prediction must therefore be "downscaled". Downscaling can also involve the conversion of a prediction for a gross summary of weather over a season, such as a 3-month rainfall total, to one containing more detailed information about the statistics of weather within the season.

After correcting the model output, the uncertainty in the forecast then needs to be communicated. Apart from the fact that an average of an ensemble of predictions is almost invariably a more accurate forecast than any single prediction, ensembles are a commonly used method of representing the uncertainty in the forecast. (The question of whether the ensemble does in fact provide a reliable indication of forecast uncertainty is deferred until Chapter 10.) If the various predictions from the ensemble are in close agreement then presumably we can place

more confidence in the forecast than when the ensemble members predict widely different outcomes. There are a number of ways of assessing the level of agreement amongst the ensemble members. The most widely used approach is to count the proportion of ensemble members that predict an event of interest. However, more sophisticated procedures are available, and involve fitting a distribution to the predictions, which gives a more reliable indication of the model's forecast distribution, and using statistical models to correct the forecast distribution to account for model skill. Such procedures are discussed in more detail in the following chapter.

Just as an ensemble of predictions from one model provides a more accurate forecast than any single model prediction, so also forecasts obtained by combining predictions from a range of different models are an improvement upon forecasts derived from a single model. There have been numerous attempts recently to combine predictions from different models in ways that account for differences in the skill of the individual models. However, with the typically small sample sizes available for seasonal forecasts, it is difficult to estimate with sufficient accuracy the differences in the performances of the models, and so a simple average of the predictions from the various models is a high standard to beat.

After constructing a forecast, an indication of the reliability of the probabilistic information communicated needs to be performed by conducting a detailed assessment of the quality of a set of historical forecasts produced in a consistent way with the current forecast. The verification of these historical predictions provides an indication of the information content in the forecast, and relevant procedures are discussed in Chapter 10.