CHAPTER

15

Forecast Recalibration and Multimodel Combination

Stefan Siegert, David B. Stephenson

[a]Department of Mathematics, University of Exeter, Devon, United Kingdom

OUTLINE

1 INTRODUCTION

Computational models of the Earth's climate system are based on mathematical abstractions and numerical approximations. Not all physical processes of the real world are included in climate models. The chaotic nature of atmospheric dynamics leads to the forecast's sensitivity to the imprecisely known initial state of the system. Therefore, numerical model forecasts are imperfect representations of the real world. Discrepancies between the model forecast and the real world can be loosely classified into random and systematic. Random forecast errors are unpredictable, whereas systematic errors are (at least to some extent)

https://doi.org/10.1016/B978-0-12-811714-9.00015-2

predictable. The most illustrative example of a systematic forecast error is the mean bias of the forecast (i.e., a constant offset between the time mean of the forecast and the time mean of the real-world predictand). If it is known from past experience that, say, a temperature forecast consistently differs from the real-world temperature by +2 K, it is rational to adjust future forecast downward by 2 K to correct for the bias and thereby improve the forecast. Bias correction is a simple example of forecast recalibration.

There are two distinct uses of the term "calibration" in the literature, both of which are related to, but different from, the technical term "recalibration." Forecast calibration can refer to the *act of calibrating* a forecast by tuning parameters of the numerical model. Forecast calibration can also be used to characterize a forecast as being *reliably calibrated*. We will not be concerned with parameter tuning in this chapter, and use the term "calibration" only in the second sense, to refer to the degree of "reliability" of a forecast model. We will focus on forecast recalibration, which is the process of making a forecast model better calibrated by statistical postprocessing of its output.

It is often the case that there is not only a single forecast model, but also multiple forecast models for the same event. One way of viewing this collection of forecasts is that they are competitors, the best of which should be picked to issue the forecast, thereby discarding the information contained in the output of the other, "suboptimal" models. However, the decision for picking the best model is often ambiguous: Forecast models must be compared by calculating performance measures, such as the correlation between past forecasts and their verifying observations, or proper scoring rules. But these measures are uncertain due to sampling variability, so the forecast model that achieves the best performance measure over a few past cases is not necessarily the best model for future forecasts. Furthermore, there are many different measures of forecast performance, and the ordering of forecasts can depend on the measure used to evaluate them. This ambiguity gives rise to the idea of viewing the various forecasts as complementary sources of information that collectively contain more information about the real world than any one of them individually. When this view is adopted, the challenge changes from picking the best model to combining the various model predictions into a single forecast of the real world.

Fig. 1 provides an illustrative example of seasonal, multimodel ensemble forecast data. The dataset consists of seasonal forecasts of average surface temperatures over the Niño-3.4 region, which is an important indicator for the state of the El Niño-Southern Oscillation (ENSO), and for the occurrence of El Niño and La Niña events. ENSO is a dominant mode of climate predictability on seasonal timescales, and so the correlation coefficients between ensemble mean forecasts and verifying observations vary between 0.81 (CFSv2) and 0.91 (SYST4). Such high predictive skill is rather atypical for seasonal climate predictions. Furthermore, due to the high predictability of ENSO, the ensemble has a rather high signal-to-noise ratio (SNR); that is, the spread of the ensemble is small compared to the variance of the ensemble mean. However, other criteria such as sample size, ensemble size per model, between-model variability, systematic bias, and number of models, this hindcast dataset is representative for forecasts on seasonal timescales. Therefore, the hindcast dataset will be used throughout this chapter to demonstrate various concepts related to forecast recalibration and combination.

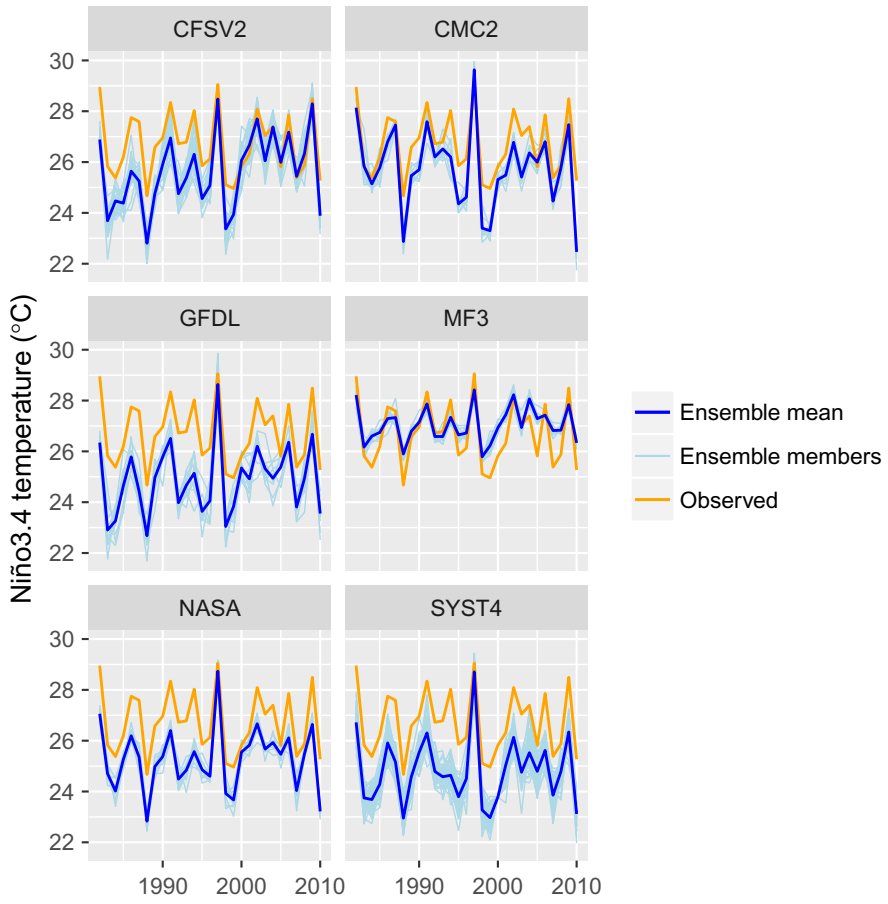

FIG. 1 Hindcast data of a seasonal multimodel ensemble forecasting system: average December temperature in the Niño-3.4 region 1982–2010, with forecasts initialized in August (lead time 5 months). *Light blue lines* represent the ensemble members, *dark blue lines* represent the ensemble mean forecasts, and *orange lines* represent the verifying observations (same in each panel).

Forecasts can be poorly calibrated in a variety of ways. Commonly observed types of forecast uncalibratedness include:

- *Constant bias of the mean*: A difference between forecast mean and observation mean.
- *Dispersion error*: The spread of an ensemble of forecasts does not correctly represent the uncertainty in the observations.
- *Lack of variability*: The year-to-year variability of the forecast is not representative of the variability in the observations.
- *Lack of association*: The correlation between forecasts and observations is low or zero.
- *Error in trends*: Slow average increases or decreases in the observations are not reproduced by the forecasts.

These violations of forecast calibration occur in atmosphere/ocean forecast products on all timescales, from short-term weather prediction to long-term climate projections. Reasons for lack of forecast calibration include initialization errors, structural model errors, model simplifications, numerical truncation errors, missing processes, and simply bugs in the code.

Statistical forecast recalibration is usually necessary for forecast products on all timescales. There are a number of challenges related to forecast recalibration and multimodel combination that are specific to seasonal to sub-seasonal (S2S) climate predictions. The training data to fit statistical recalibration models is often limited, and highly nonstationary. Formulations of the operational forecast model are revised periodically, which can change the statistical behavior of the data and require readjustment of recalibration and combination parameters. The internal variability of the forecasts is high due to the chaotic nature of the atmosphere, which decreases the SNR of the forecast. The correlation between forecast and observations is often low. Finally, multimodel hindcast experiments are not designed with model combinations in mind and are therefore often nonhomogeneous. Strategies to generate hindcast datasets can be loosely characterized as either "on the fly" or "fixed," depending on how changes to the forecast model are accounted for. In the S2S hindcast database (Vitart et al., 2017), for example, most forecasts that are initialized on different days have different hindcast periods and forecast out to different lead times.

By using statistical models to correct forecast errors of dynamical models, forecast calibration bridges the gap between empirical (statistical) and numerical forecasting. To issue reliable forecasts, we need robust statistical methods to issue probabilistic predictions, which take into account the correlation and error structure of multimodel ensemble forecasts.

2 STATISTICAL METHODS FOR FORECAST RECALIBRATION

Forecast calibration is an important diagnostic to differentiate good forecasts from bad forecasts. To characterize forecast calibration, Gneiting et al. (2007) introduced various *modes of calibration* (namely, probabilistic calibration, exceedance calibration, and marginal calibration). All modes of calibration characterize, in different ways, the agreement between the issued forecast distribution and the hypothetical distribution from which the real-world observation is drawn. Forecast recalibration is thus closely related to forecast verification, which is discussed in detail in Chapter 16 of this book.

In a similar spirit, Jolliffe and Stephenson (2012) define forecast calibration in terms of the equality between the forecast and the conditional mean of the observation, given the following forecast:

$$E_Y(Y|X=x)=x. \tag{1}$$

That is, if we collected all instances on which a particular value $X=x$ was forecast, the mean over all verifying observations should be equal to x if the forecast is calibrated. Calibration is thus a joint property of forecasts and observations that can be assessed by comparing several pairs of forecasts and observations. If a forecast is found to be uncalibrated, statistical recalibration methods can be used to correct for the lack of forecast calibration.

Eq. (1) suggests that to recalibrate a poorly calibrated forecast, we could replace the current forecast value x by the conditional mean of the observation, given that forecast value. The exact value of the conditional expectation is not known in general, and it has to be estimated from past forecast and observation data. More generally, one could estimate the conditional distribution of the observation, given the forecast. The conditional mean or distribution can be estimated by collecting all past forecasts that have a given value (or are sufficiently close to a given value) and averaging all past observations corresponding to these forecasts. This *nonparametric* way of recalibrating forecasts is appealing, as it can potentially account for complicated nonlinear relationships between forecasts and observations. However, it requires enough past forecasts that are close to the current forecast value in order to estimate the conditional mean robustly. In data-poor settings, where only a few pairs of past forecasts and observations are available, nonparametric estimation methods will suffer from large estimator variance. In these situations it is often useful (or even necessary) to assume a parametric relationship between forecasts and observations (i.e., to describe the conditional mean of the observation given the forecast by a function of the forecast that is parameterized by a small number of coefficients). Next we discuss two of the most commonly used parametric methods for forecast recalibration—namely, model output statistics (MOS) and nonhomogeneous Gaussian regression (NGR).

3 REGRESSION METHODS

3.1 Model Output Statistics

The most commonly used parametric methods for forecast recalibration are based on regression techniques. In the meteorological literature, using linear regression to recalibrate a forecast is also referred to as model output statistics (MOS; Glahn and Lowry, 1972; Glahn et al., 2009). In linear regression, the observation y_t at time t is modeled as a linear function of forecasted value (or forecasted values if several forecasts are available) $x_{1,\,t}, \ldots, x_{p,\,t}$, plus an independent, normally distributed error term:

$$y_t = \beta_0 + \beta_1 x_{1,\,t} + \cdots + \beta_p x_{p,t} + \sigma\epsilon_t, \tag{2}$$

where $\beta_0, \ldots, \beta_p$ and σ are unknown parameters and $\epsilon_t \sim \mathcal{N}(0,1)$. Eq. (2) can also be written in vector form as

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma\epsilon_t, \tag{3}$$

using the column vectors $\boldsymbol{x}_t = (1, x_{1,\,t}, \ldots, x_{p,\,t})'$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. A common case is $p = 1$, where there is a single forecast, such as the ensemble mean taken from a single model, using the same variable and location as the predictand y. It is possible that multiple predictors are used (i.e., $p > 1$). These can be output from several forecast models, different variables than the predictand, different ensemble members started from perturbed initial conditions, or variables at different locations that are deemed informative about the predictand. The regression parameters $\boldsymbol{\beta}$ and σ can be estimated from previously observed pairs of forecasts and verifying observations.

It is useful to collect the verifying observations $y_1, \ldots, y_N$ into a column vector $\boldsymbol{y}$ and the row vectors $\boldsymbol{x}_1', \ldots, \boldsymbol{x}_N'$ into the rows of the design matrix $\boldsymbol{X}$. Then we can write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon}$ is assumed to have a multivariate normal distribution with diagonal covariance matrix $\text{var}(\boldsymbol{\epsilon}) = \mathbf{1}$.

Under the assumption that the error term ϵ_t has a standard Gaussian distribution, and ϵ_t and $\epsilon_{t'}$ are uncorrelated for $t \neq t'$, the log-likelihood function of the linear regression model is proportional to

$$\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) \propto -\frac{N}{2}\log(\sigma^2) - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}. \tag{5}$$

The maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are obtained by setting the partial derivatives of ℓ to zero:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{6}$$

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{N}. \tag{7}$$

The commonly used unbiased estimator of σ^2, denoted $\hat{\sigma}_u^2$, is given by subtracting the total number of estimated parameters from N in the denominator of Eq. (7), that is,

$$\hat{\sigma}_u^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{N - p - 1}. \tag{8}$$

After fitting the regression parameters by maximum likelihood, a future observation y^*, given a new forecast $\boldsymbol{x}^*$, is predicted by plugging $\boldsymbol{x}^*$ into Eq. (3), using the maximum likelihood estimators for the regression parameters. By transforming the forecast $\boldsymbol{x}^*$ by the regression relationship (3), some violations of calibration in the raw forecast $\boldsymbol{x}^*$ are corrected, namely constant bias, linear scaling, and ensemble dispersion errors.

It can be shown that the forecast distribution for the new observation y^* based on the new forecast vector $\boldsymbol{x}^*$ is a Student t-distribution:

$$y^* | \boldsymbol{x}^*, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \sim t_{N-p-1}\left[(\boldsymbol{x}^*)'\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2\left(1 + (\boldsymbol{x}^*)'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)\right)\right]. \tag{9}$$

So the forecast mean is at $(\boldsymbol{x}^*)'\hat{\boldsymbol{\beta}}$, and a 95% prediction interval for y^* is given by

$$(\boldsymbol{x}^*)'\hat{\boldsymbol{\beta}} \pm t_{0.975, N-p-1}\hat{\sigma}_u\sqrt{1 + (\boldsymbol{x}^*)'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*}, \tag{10}$$

where $t_{\alpha, n}$ denotes the α-quantile of the Student t-distribution with n degrees of freedom. It is tempting to simply forecast a normal distribution with mean $(\boldsymbol{x}^*)'\hat{\boldsymbol{\beta}}$ and variance σ_u^2. But it has been shown that MOS forecasts issued using the predictive t-distributions are better calibrated than forecasts issued as normal distributions because the t-distribution accounts for the estimation uncertainty of the regression parameters (Siegert et al., 2016a).

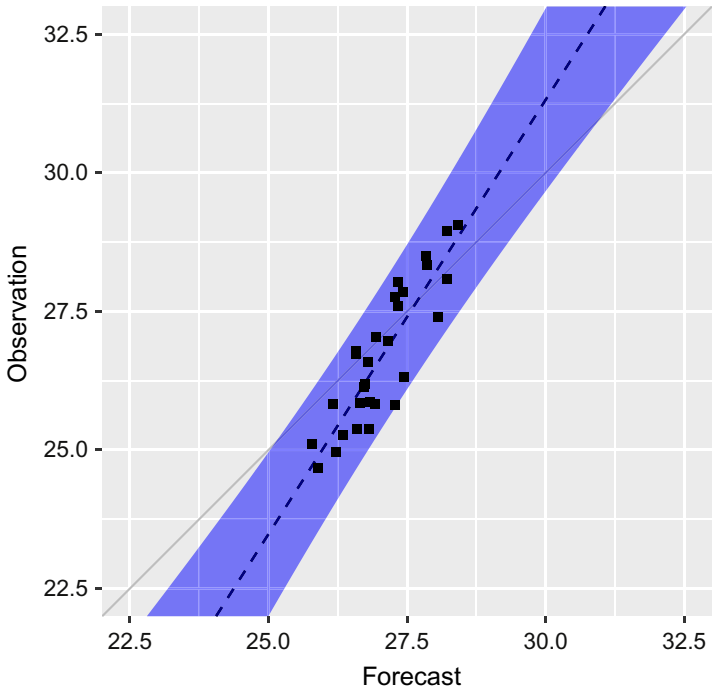

FIG. 2 Recalibration by linear regression applied to MF3 forecasts. The *dashed line* indicates the regression line, and the *blue ribbon* indicates the 95% prediction interval.

As an example, consider the MF3 Niño-3.4 ensemble mean seasonal forecasts at 5 months lead time, illustrated by a scatterplot in Fig. 2. The points do not lie along the diagonal, which indicates that mean and/or scale of the forecasts do not match the mean and scale of the observations. The forecasts are not well calibrated, and statistical recalibration is therefore necessary. The scatterplot further suggests that linear rescaling of the forecasts might be a good recalibration strategy, which makes MOS a suitable candidate. The maximum likelihood estimator of the coefficient vector $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (-15.70, 1.57)'$ and the (unbiased) maximum likelihood estimator of σ^2 is $\hat{\sigma}_u^2 = 0.39$. For a new forecast value of 27.0°C that lies close to the mean of all previously observed forecast values, the recalibrated prediction equals 26.69°C, and a 95% prediction interval is given by (25.32°C, 27.91°C); that is, the width is 2.59. Likewise, for a new forecast value of 30.0°C, which is large compared to all previously observed forecasts, the prediction is 31.32°C and the 95% prediction interval is given by (29.67°C, 32.96°C) (i.e., the width is 3.30), which is considerably wider than for the intermediate forecast value. Informally, the widening of the prediction interval is caused by the extrapolation beyond previously observed forecast values, which increases uncertainty. Mathematically, the term $(\boldsymbol{x}^*)'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}^*$ in Eq. (9) is responsible for widening of the prediction intervals.

Fig. 3 shows raw forecasts and MOS-recalibrated forecasts of the MF3 model, and their verifying observations. The effect of MOS is to bring the forecast means closer (in a mean-squared-error sense) to the verifying observations, and to increase the forecast variance compared to the ensemble spread. The result is better coverage of the observations by the prediction intervals, and therefore more reliable probability forecasts.

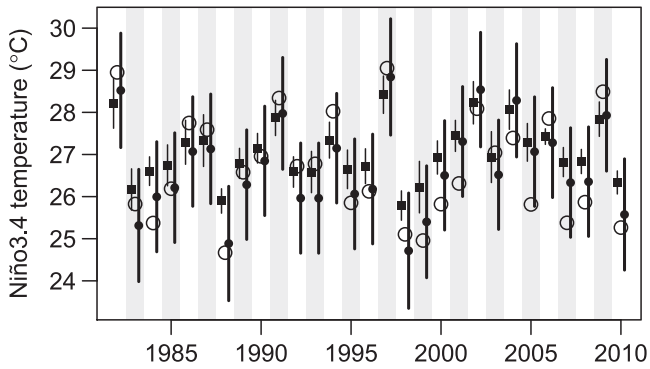

FIG. 3 Illustration of the effect of recalibration by linear regression. *Squares and thin lines* indicate ensemble mean forecasts (generated by the MF3 model) ± two ensemble standard deviations. *Filled circles* and *thick lines* indicate recalibrated ensemble means and 95% prediction intervals. *Open circles* represent observations. The recalibrated forecasts are on average closer to the observations, and the prediction intervals overlap the observations more often than the uncalibrated ensemble spread.

3.2 Nonhomogeneous Gaussian Regression

MOS can be extended to allow use of the ensemble spread information. A good ensemble forecasting system should be able to accurately represent the forecast uncertainty that results from imprecisely known initial conditions and model errors. Therefore, narrow ensembles, corresponding to high confidence in the forecast, should on average incur smaller forecast errors than very wide ensembles, which indicate low confidence in the forecast. Due to model errors and natural variability, the correspondence between ensemble spread and forecast error (the spread-skill relationship) cannot be expected to be perfect, but it is reasonable to assume that a linear relationship exists. A regression framework to recalibrate both ensemble mean and ensemble spread is nonhomogeneous Gaussian regression (NGR; Gneiting et al., 2005). NGR assumes that the observation has a normal distribution whose mean and variance depend linearly on ensemble mean and ensemble variance. In particular, let m_t denote the ensemble mean forecast at time t, and s_t^2 the ensemble sample variance at time t. The conditional distribution of the observations, given the ensemble forecast, is

$$\mathcal{N}(a + bm_t, c + d^2 s_t^2). \tag{11}$$

The recalibration parameters (a, b, c, d) are unknown and have to be estimated from historical forecast and observation data. Unlike linear regression (MOS), the maximum likelihood parameters cannot be determined analytically and therefore have to be estimated by numerical optimization. Given a series of ensemble mean forecasts $m_1, \ldots, m_N$, ensemble variances $s_1^2, \ldots, s_N^2$, and verifying observations $y_1, \ldots, y_N$, the log-likelihood function of the NGR model is proportional to

$$\ell(a,b,c,d;\{m_t, s_t^2, y_t\}_{t=1}^N) \propto -\frac{1}{2}\sum_{t=1}^{N}\left[\log(c + d^2 s_t^2) + \frac{(y_t - a - bm_t)^2}{c + d^2 s_t^2}\right]. \tag{12}$$

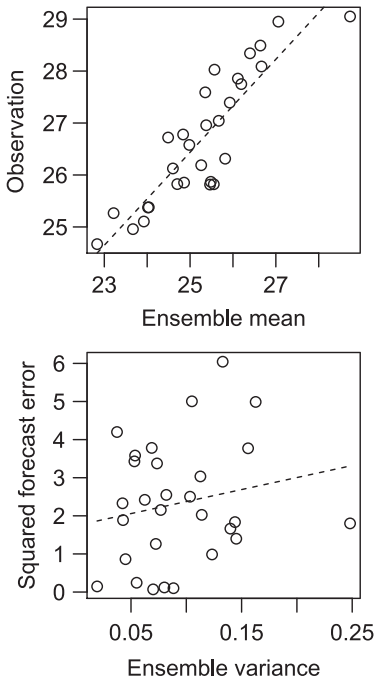

FIG. 4 Scatterplots of observation over ensemble mean and squared forecast error over ensemble variance for the Niño-3.4 ensemble forecasts issued by the NASA model. Least-squares linear fits have been added as guides.

To give a specific example, consider the Niño-34 temperature seasonal forecasts issued by the National Aeronautics and Space Administration (NASA) model at 5 months lead time. Scatterplots of verifying observations over ensemble means and squared forecast errors over ensemble variances are shown in Fig. 4. There is a strong linear relationship between ensemble means and observations (correlation 0.88). There is also a weak positive linear relationship between ensemble variances and squared forecast errors (correlation 0.19). The correlation between variance and error is not statistically significant, but it might still be beneficial for forecast recalibration.

To fit the NGR, we optimize the NGR log-likelihood (Eq. 12) using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, as implemented in the function `stats::optim` of the `R` statistical programming environment (R Core Team, 2017). The estimated values are given in Table 1. The parameter estimates suggest that the forecasts are not well calibrated, and there is scope for improvement by statistical recalibration of mean, scaling, and variance. However, the parameter d is very small, indicating very little relationship between ensemble spread and forecast variance.

It is worth noting that NGR was first proposed, and is mostly applied, in the context of numerical weather prediction (NWP). In NWP, atmospheric prediction is treated as an initial

TABLE 1 NGR Estimates

Parameter	a	b	c	$\lvert d \rvert$
Estimate	4.06	0.89	0.34	2.5×10^{-5}

value problem, so that varying levels of sensitivity to initial conditions can lead to a spread-skill relationship in ensemble forecasts. Seasonal climate forecasting, rather, is a boundary value problem, where long-term predictability is a result of slowly varying drivers of the climate system. A strong spread-skill relationship is therefore unlikely, so it is no surprise that NGR is not beneficial as a recalibration method for the seasonal forecasts shown here. For prediction on sub-seasonal timescales, lying between weather and seasonal climate forecasting; however, a systematic spread-skill relationship might be conceivable.

3.3 Comparing Recalibration Models

Forecast recalibration is a statistical modeling exercise. At any point in time, several recalibration models might be available, and the task of the forecaster is to choose one of them to make a prediction. The task of choosing the "best" among a number of candidate statistical models is called model selection. Here, we give an example to illustrate how to choose between MOS and NGR to recalibrate the NASA model. A good introduction to model selection and statistical modeling in general can be found in Hastie et al. (2009).

A commonly used model selection criterion is the Bayesian Information Criterion (BIC; Schwarz, 1978), defined as

$$\mathrm{BIC} = -2\hat{\ell} + k\log(n), \tag{13}$$

where $\hat{\ell}$ is the log-likelihood function evaluated at the mode (i.e., using the optimized parameter values); k is the number of parameters of the model; and n is the sample size. The model with lowest BIC is to be preferred. A low BIC is achieved by high values of $\hat{\ell}$ and low values of k. Thus, BIC reward models that fit the data well, while at the same time having a small number of free parameters. The BIC is closely related to the Akaike Information Criterion (AIC), which is calculated by replacing $\log(n)$ by 2 in Eq. (13).

We have seen in Table 1 that the optimal value for the parameter d is very small in the NASA ensemble, which suggests that taking the ensemble spread into account in the variance of the forecast distribution might be unnecessary. When d is zero, NGR is equivalent to MOS. The differences between the optimized log-likelihoods of NGR and MOS for this ensemble is on the order of 10^{-10} (i.e., the recalibration by NGR and MOS yields almost identical recalibrated forecasts). But since NGR has four free parameters, where MOS has only three, we get BIC = 64.5 for NGR and BIC = 61.1 for MOS, which suggest that MOS is the preferable recalibration model in this case. In other words, the hypothesized spread-skill relationship in the NASA ensemble cannot be considered useful for forecast recalibration.

Another widely used method for model comparison is cross-validation. In cross-validation the ability of a statistical recalibration model is assessed by evaluating its predictions on unknown data that were not part of the training dataset.

3.4 Further Remarks on Recalibration

Because forecast recalibration is a *statistical modeling* problem, all issues that apply to statistical modeling are relevant to forecast recalibration as well. We have discussed the important areas of parameter estimation and model selection in some detail. Here, we discuss a

number of further problem areas that should be considered and refer the reader to the relevant literature.

If parameters are estimated from a finite number of training data, their estimation uncertainty must be taken into account. Siegert et al. (2016a) have shown that failing to account for parametric uncertainty can lead to degradation of the quality of the recalibrated forecasts. Accounting for estimation uncertainty in the recalibration parameters has the effect of inflating the tails of the forecast distribution, which leads to better calibrated and more skillful forecasts. It is often the case that prior information is available on recalibration parameters, in which case a Bayesian estimation framework is suitable. Siegert et al. (2016b) have shown that prior information on the correlation coefficient of the ensemble mean can improve the performance of recalibrated forecasts compared to standard methods. Furthermore, the Bayesian approach of Siegert et al. (2016b) allows one to address the problems of forecast verification and forecast recalibration in the same coherent statistical framework.

Delle Monache et al. (2011) and Obled et al. (2002) have used statistical analog techniques to improve forecast recalibration. The underlying idea is to construct the training dataset for parameter estimation by considering only past forecasts that are similar to the present one. A related technique is to use a sliding training window (e.g., Sweeney et al., 2011) to use only the most recent forecast and observation data to construct the training dataset for parameter estimation. A sliding window approach allows the recalibration strategy to adapt to changes in the forecasting system or the climate system.

Forecast data produced by climate models is usually high-dimensional, consisting of multiple climatological variables on a spatial grid and at many points in time for various ensemble members initialized at different times and initial conditions. Various techniques exist for multivariate recalibration, and especially the field of spatial recalibration has undergone rapid development in recent years.

Two important nonparametric methods for spatial recalibration are the *Schaake shuffle* (Clark et al., 2004) and *ensemble copula coupling* (Schefzik et al., 2013). These methods are based on the idea of reordering ensemble forecasts locally so as to better replicate the spatial correlation structure of the predictand (see also Schefzik, 2017; Vrac and Friederichs, 2015; Scheuerer et al., 2017). Parametric approaches for multivariate forecast recalibration have been proposed based on Gaussian random fields (Feldmann et al., 2015) and parametric copulas (Möller et al., 2012; Hemri et al., 2015). It can be noted that multivariate methods such as principal component regression (PCR) and canonical correlation analysis (CCA) have been used to recalibrate seasonal climate forecasts (e.g., Barnston and Tippett, 2017). However, recalibration based on explicit spatiotemporal statistical models is largely unexplored in the field of S2S predictions.

4 FORECAST COMBINATION

The development and maintenance of a climate forecasting system require considerable effort. It is therefore sensible to establish climate modeling centers, where scientists, developers, and administrators provide the necessary expertise and infrastructure. As a consequence, several climate modeling centers exist around the world, each one running its

own forecast system. The multiplicity of modeling centers provides opportunities to share expertise and to compare various modeling strategies. But because each center provides its own climate forecasting products, using slightly various climate models, the user faces a conundrum of choice, having to make an informed decision as to which climate model to use. Better yet, the user might want to benefit from the "wisdom of crowds" and let the multiplicity of climate model forecasts act as a sort of committee that jointly provides a final, combined forecast product.

Various methods have been proposed to optimally combine forecasts from different numerical models. A key reference on forecast combinations in seasonal climate forecasting is DelSole (2007), who presents a unified Bayesian framework that accommodates a number of multimodel combination strategies. Sansom et al. (2013) discuss weighting strategies for multimodel ensembles in a climate change context. Further combination strategies are discussed in Stephenson et al. (2005), Doblas-Reyes et al. (2005), and Rajagopalan et al. (2002). The rest of this section follows the methodologies outlined in DelSole (2007), with particular focus on the hierarchical regression method of Lindley and Smith (1972).

4.1 Hierarchical Linear Regression

Assume, as before, that at times $t = 1, \ldots, N$, we have climate forecasts that were produced by p numerical models, $f_{1,t}, \ldots, f_{p,t}$. Each $f_{i,t}$ is assumed to be scalar, so it could be a spatial, temporal, and ensemble average produced from the output of a single climate model. We assume in this section that the vectors of forecasts $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p$ have been individually standardized to have zero mean and unit variance over time; DelSole (2007) reports that standardization of individual forecasts improved the quality of the combined forecast product. One possible method, motivated by the regression framework discussed in the previous section, is to combine the individual forecasts into a single forecast by a linear combination, and to assume the residual to be independently normally distributed:

$$y_t = \sum_{m=1}^{p} \beta_m f_{m,t} + \sigma \epsilon_t, \tag{14}$$

which can be collected into the matrix equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{15}$$

where $\boldsymbol{X}$ is the $N \times p$ matrix of forecasts, $\boldsymbol{\beta}$ is the vector of combination weights, and $\boldsymbol{\epsilon}$ has a multivariate normal distribution with zero mean and identity covariance matrix. One then can estimate the vector $\boldsymbol{\beta}$ of forecast combination weights, as well as the residual variance σ^2.

There are two possible extremes that we could adopt when estimating the combination weights. On the one hand, we could assume that the combination weights can be completely different and are fully independent, such that we would not be surprised to learn that the weight of one model is orders of magnitude larger, and with a possibly different sign, than the combination weight of another model. On the other hand, we might judge that there should be no difference at all between the combination weights for different models because the individual models are judged to be exchangeable, and we do not expect any performance

differences among them that would warrant upweighting one model forecast in favor of another one.

The framework of DelSole (2007) points out a middle way between those two extremes, using the results of Lindley and Smith (1972) on hierarchical regression. The framework essentially allows the shrinking of the combination weights β_m toward a common, but unknown, value β_0, thus reducing the variability of the combination weights. The underlying idea is that we are usually prepared to assign different weights to different model forecasts, but the weights are not expected to be very different from one another because we would generally not expect large differences between the quality of different forecasts. We will come back to the judgment of similar quality and its implications for forecast combination later in this chapter.

The notion of "different, but similar" combination weights can be modeled within a Bayesian statistical framework as follows. The result of a Bayesian computation is a posterior probability distribution of unknown model parameters, given observed data (i.e., $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$), in the present context. The posterior distribution is computed by the Bayes' rule:

$$p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta},\sigma^2), \tag{16}$$

where the likelihood $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ derives from the linear model specification (Eq. 15) and the prior distribution $p(\boldsymbol{\beta}, \sigma^2)$ encodes prior knowledge about the model parameters. In the present context, we will be interested only in the maximum a posteriori (MAP) estimators of the model parameters (i.e., the values that maximize $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$), and so the proportionality constant in Eq. (16) is unimportant.

The hierarchical regression framework developed by Lindley and Smith (1972) allows us to encode the notion that the combination weights β_m are different but similar in the prior distribution $p(\boldsymbol{\beta}, \sigma^2)$. The elements of $\boldsymbol{\beta}$ are assumed to be independently normally distributed around a common (unknown) mean β_0 and variance σ_β^2:

$$\beta_i \sim N(\beta_0, \sigma_\beta^2). \tag{17}$$

The normal distribution allows for the β_m to be different, but a small variance σ_β^2 will constrain them to be close to one another. We have to make further assumptions about β_0 and σ_β^2 to close the calculations. Either the values of β_0 and σ_β^2 must be specified, or if this is not possible, vague assumptions must be encoded as probability distributions over β_0 and σ_β^2. The following choices seem justified in the specific context of climate forecast combination and also will lead to a convenient and tractable method of estimating the MAP values of the combination weights. Users will probably not have strong prior beliefs about β_0, and therefore a very wide (uninformative) prior distribution for the central value β_0 is appropriate. A convenient choice of the prior for β_0 is therefore a normal distribution with zero mean and diverging variance. On the other hand, a user who wants to encode the idea of "not too different" combination weights will usually have an idea about what "too different" means quantitatively. For example, if we think that combination weights for our forecasts are unlikely to differ from their common value by more than 0.2, this can be encoded by specifying the variance $\sigma_\beta^2 = 0.1^2$. Finally, to complete the prior specifications, we choose an uninformative prior distribution for the residual variance σ^2—namely, an inverse χ^2 distribution with degrees of freedom $\nu = 0$, such that $p(\sigma^2) \propto 1/\sigma^2$.

Lindley and Smith (1972) show that under these prior assumptions, the MAP estimators of the combination parameters $\boldsymbol{\beta}$ and σ^2 can be obtained by solving the following system of equations:

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{X} + \frac{s^2}{\sigma_{\beta}^2}(\mathbf{1}_p - p^{-1}\boldsymbol{J}_p)\right]^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{18}$$

$$s^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n+2}, \tag{19}$$

where $\mathbf{1}_p$ is the $p \times p$ identity matrix, and $\boldsymbol{J}_p$ is a $p \times p$ matrix with each element equal to 1. The equations cannot be solved analytically, but an approximate solution can easily be found iteratively by solving the two equations in turn, each time substituting the solution of one equation into the other. We found that this algorithm leads to convergence within a few (<10) iterations, with little dependence on the choice of initial values.

Note that the two extreme cases mentioned here (equal weighting and fully flexible unequal weighting) correspond to particular choices of the prior variance parameter σ_{β}^2: By setting $\sigma_{\beta}^2 \to \infty$, the additional term in the brackets in Eq. (18) vanishes and the estimate of $\boldsymbol{\beta}$ reduces to the least-squares estimator $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Imposing no constraints on the elements of $\boldsymbol{\beta}$ by setting $\sigma_{\beta}^2 \to \infty$ thus amounts to ordinary multiple linear regression (MLR). On the other hand, by setting $\sigma_{\beta}^2 = 0$ (i.e., assuming that all combination weights are equal to β_0), the MAP estimate converges to the same estimate of β that would be obtained if we fitted a simple linear regression (SLR) to the multimodel ensemble mean; see Appendix of DelSole (2007) for a proof.

Fig. 5 shows the results of estimated combination weights for the six seasonal Niño-3.4 temperature forecasts shown in Fig. 1. The ensemble mean forecasts of all models were standardized before estimating the combination weights. Three distinct values of σ_{β}^2 were chosen:

- $\sigma_{\beta}^2 \to \infty$, corresponding to unconstrained MLR
- $\sigma_{\beta}^2 = 0.1^2$, corresponding to HLR
- $\sigma_{\beta}^2 = 0$, corresponding to SLR on the multimodel ensemble mean

Combination with equal weights yields $\beta_i = \beta = 0.16$. Using MLR with no constraints on the variability of the parameters, the combination weights vary wildly, between -0.16 and 0.66.

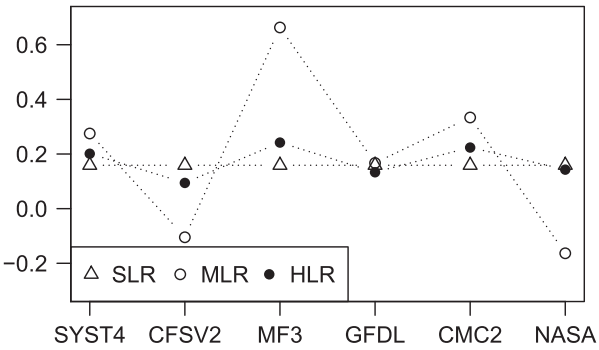

FIG. 5 Niño-3.4 combination weights assigned to numerical models by different methods: SLR = simple linear regression ($\sigma_{\beta}^2 = 0$); MLR = multiple linear regression ($\sigma_{\beta}^2 \to \infty$); HLR = hierarchical linear regression ($\sigma_{\beta}^2 = 0.1^2$).

This variability is somewhat damped using the hierarchical regression estimators with $\sigma_\beta^2 = 0.1^2$, which restricts the combination weights to a much more reasonable range of 0.09–0.24.

A relevant question to ask is which of the three combination method performs best. If we addressed this question simply by looking at the sum of squared residuals after fitting the regression models, we would find that $\sigma_\beta^2 \to \infty$ performs best. But this is, at least partly, due to the great flexibility of the unconstrained MLR model, which allows parameters to adapt to random variations in the data by setting one regression coefficient to a very large positive value and another coefficient to a very low negative value. But the sum of squared residuals is a measure of in-sample goodness of fit, which is not really relevant in practice. In practice, we would like to estimate how well the methods perform out of sample, on as-yet-unseen data that was not part of the training dataset.

To estimate out-of-sample performance, we conduct a leave-one-out cross-validation. We leave out 1 year of the N years in the hindcast archive and fit the combination weights using the $N - 1$ remaining pairs of forecasts and observations. We then use the fitted combination weights on the left-out forecasts to predict the left-out observation. This process is repeated N times, each time leaving out a different year, which results in N out-of-sample predictions whose squared prediction errors can be used to assess out-of-sample performance. The equally weighted forecast combination ($\sigma_\beta^2 = 0$) obtains a leave-one-out mean-squared prediction error of 0.281. The forecast combination obtained by unconstrained multiple regression ($\sigma_\beta^2 \to \infty$) has a much larger mean-squared prediction error of 0.330. The constrained unequal weighting approach with $\sigma_\beta^2 = 0.1^2$ achieves a leave-one-out mean squared error of 0.277, which is a large improvement over multiple regression and a minor improvement over simple regression on the multimodel mean.

The choice of the prior parameter σ_β^2 can be guided by different principles. DelSole (2007) suggests using a nested cross-validation approach to estimate the optimal value σ_β^2. Lindley and Smith (1972) show how $\boldsymbol{\beta}$ and σ^2 can be estimated when an informative prior distribution (in the form of a scaled inverse-χ^2 distribution) is specified for σ_β^2, rather than setting a fixed value as we did earlier in this chapter. It also should be noted that the choice of the prior variance of the β_m should depend on the number of models. We might be more willing to accept larger differences between the weights of 2 models than between the weights of 10 models. It is also possible to specify a different prior distribution than a normal distribution for $\boldsymbol{\beta}$. In particular, a Laplace prior distribution, which leads to the so-called Lasso regression (Tibshirani, 1996), might be beneficial. The Laplace distribution has more probability mass close to the mode and more probability mass in the tails than the normal distribution. Therefore, it would set some of the weights to exactly equal values, while giving significantly higher or lower weight to only a few models. However, no closed-form solutions are available for the Lasso estimates of $\boldsymbol{\beta}$, and thus computationally more expensive numerical optimization methods would be required.

4.2 Why Is It So Hard to Beat the Recalibrated Multimodel Mean?

It is interesting to note the tiny difference between the leave-one-out prediction errors of constrained unequal weighting (0.277) and equal weighting (0.281). The improvement from unequal weighting compared to equal weighting is so tiny that it could well be simply

due to chance, and even if it were a genuine improvement, its practical utility would be limited at best. Based on this data, we have no reason to believe that unequal weighting offers any considerable improvement over equal weighting. As a matter of fact, there are a number of reasons why we should not expect a large improvement of unequal weighting over equal weighting in the first place. All climate models in the multimodel ensemble have roughly the same complexity—they all run on supercomputers and are maintained and developed by government agencies. Obviously, all models contain the same basic physics—namely, a discretized and simplified version of the Navier-Stokes equations, thermodynamic closure relations, and parameterizations of unresolved processes. The models were not developed independently, but rather rely on the same body of knowledge about the practicalities of numerical climate modeling. Furthermore, from a statistical point of view, the small sample size of $N = 31$ years naturally limits the precision with which the combination weights can be estimated. The ensuing estimation variance of the combination weights will degrade the quality of out-of-sample predictions. Some authors (e.g., Weigel et al., 2010) have explicitly warned against using unequal weighting at all and recommend treating the different models as exchangeable, even though small differences are conceivable in principle. It should be noted, however, that there are cases where a single model is superior to all other models, and therefore, forecast combination with less skillful models is always detrimental (e.g., Vitart, 2017). We have shown that unconstrained MLR with a small training dataset can indeed degrade the performance of the combined forecast compared to equal weighting. But a suitable shrinkage strategy that limits the variability of the combination weights can reduce this problem and has the potential to gain slight improvements over equal weighting. However, for the reasons stated in this chapter, we should not expect the improvement to be large, even if we knew the "true" optimal combination weights.

5 CONCLUDING REMARKS

Forecasts of physical-dynamical models can suffer from various forecast biases that can be corrected by statistical methodology. Furthermore, the availability of several forecast models for the same predictand calls for statistical methods to optimally combine multiple forecasts into a single forecast. In this chapter, we have outlined various regression approaches and discussed relevant statistical concepts such as model selection, in-sample versus out-of-sample performance, and the incorporation of prior knowledge. The methods discussed are based on developments from short-term weather forecasting to longer-term seasonal climate forecasting, and thus they are fully applicable at the sub-seasonal scale.

Acknowledgments

The authors thank Caio Coelho for providing the seasonal Niño-3.4 hindcast dataset, and Thordis Thorarinsdottir for helpful input on multivariate recalibration. Andrew Robertson and Frédéric Vitart provided helpful feedback and comments on earlier drafts.