Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better than Multimodels?

ANDREAS P. WEIGEL, MARK A. LINIGER, AND CHRISTOF APPENZELLER

Federal Office of Meteorology and Climatology, MeteoSwiss, Zürich, Switzerland

(Manuscript received 28 August 2008, in final form 13 October 2008)

ABSTRACT

Multimodel ensemble combination (MMEC) has become an accepted technique to improve probabilistic forecasts from short- to long-range time scales. MMEC techniques typically widen ensemble spread, thus improving the dispersion characteristics and the reliability of the forecasts. This raises the question as to whether the same effect could be achieved in a potentially cheaper way by rescaling single model ensemble forecasts a posteriori such that they become reliable. In this study a climate conserving recalibration (CCR) technique is derived and compared with MMEC. With a simple stochastic toy model it is shown that both CCR and MMEC successfully improve forecast reliability. The difference between these two methods is that CCR conserves resolution but inevitably dilutes the potentially predictable signal while MMEC is in the ideal case able to fully retain the predictable signal and to improve resolution. Therefore, MMEC is conceptually to be preferred, particularly since the effect of CCR depends on the length of the data record and on distributional assumptions. In reality, however, multimodels consist only of a finite number of participating single models, and the model errors are often correlated. Under such conditions, and depending on the skill metric applied, CCR-corrected single models can on average have comparable skill as multimodel ensembles, particularly when the potential model predictability is low. Using seasonal near-surface temperature and precipitation forecasts of three models of the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) dataset, it is shown that the conclusions drawn from the toy-model experiments hold equally in a real multimodel ensemble prediction system. All in all, it is not possible to make a general statement on whether CCR or MMEC is the better method. Rather it seems that optimum forecasts can be obtained by a combination of both methods, but only if first MMEC and then CCR is applied. The opposite order—first CCR, then MMEC—is shown to be of only little effect, at least in the context of seasonal forecasts.

1. Introduction

The use of ensemble prediction systems (EPSs) has become a matter of routine in the context of weather and climate risk management, and sophisticated methods of ensemble generation are meanwhile well established. However, while such ensembles successfully quantify the forecast uncertainties arising from the uncertainties in the model initialization, they fail to capture the uncertainties arising from errors and simplifications in the model itself. For example, the uncertainties due to the parameterization of physical processes, the effect of unresolved scales, or imperfect boundary conditions, are not quantified (Buizza et al. 2005; Schwierz et al. 2006; Weigel et al. 2007a). Consequently, ensemble distributions typically underestimate the true forecast uncertainty and tend to be overconfident (or underdispersive), that is, they are too sharp while being centered at the wrong value.

As a pragmatic approach to overcome this problem, it has been suggested to combine several ensemble prediction systems to form a multimodel superensemble (Krishnamurti et al. 1999; Palmer et al. 2004). That way, at least a crude estimate of the range of uncertainties due to model errors can be obtained. The success of this approach has been demonstrated in many studies (e.g., Rajagopalan et al. 2002; Robertson et al. 2004; Hagedorn et al. 2005; Stephenson et al. 2005). Other approaches, such as the introduction of stochastic physics (Buizza et al. 1999) or the perturbed parameter approach (Pellerin et al. 2003) will not be considered here. In essence, multimodel ensemble combination (MMEC) widens the ensemble spread and reduces the root-mean-square error (rmse) of the ensemble means, thus reducing forecast overconfidence and improving the forecast reliability. Indeed, for seasonal forecasts it has been shown that
MMEC is of only little effect if the single model ensembles (SMEs) contributing to the multimodel ensemble (MME) are already reliable (Weigel et al. 2008b; Weigel and Bowler 2009). But does it then really need a multimodel approach to reduce the overconfidence of ensemble forecasts? Could the same effect not be achieved in a cheaper way by simply rescaling unreliable SMEs a posteriori such that they become reliable (i.e., by an appropriate recalibration)?1 Is there a difference at all between MMEs on the one hand and recalibrated SMEs on the other hand with respect to their skill properties?

There are many studies that demonstrate that recalibration, as well as related techniques such as ensemble dressing (Roulston and Smith 2003), do improve the prediction skill significantly (e.g., Atger 2003; Doblas-Reyes et al. 2005; Feddersen and Andersen 2005), but the conceptual differences between a reliability correction by MMEC and a reliability correction by recalibration have only been addressed in very few studies. Doblas-Reyes et al. (2005), for example, conclude from the evaluation of seasonal forecasts that ensemble spread correction does improve the prediction skill, but not beyond the skill of a MME. However, their recalibration procedure not only rescales the ensembles but also corrects for systematic spatial shifts, making it difficult to quantify the mere effect of ensemble spread correction. Moreover, they have not corrected for ensemble size induced biases when comparing the prediction skill of SMEs and MMEs, thus being unfair against the single models (Weigel et al. 2007b,c). Indeed, applying a debiased verification context and a stochastic toy model, the study of Weigel et al. (2008b) indicates that recalibrated SMEs actually can outperform a MME under certain conditions, but at the cost of correlation between the forecasts and the observations.

The present study seeks to close the gaps of Doblas-Reyes et al. (2005) and Weigel et al. (2008b) and seeks to comprehensively answer the following questions: What is the fundamental difference between the skill improvement due to MMEC and the skill improvement due to appropriate recalibration? How do MMEC and recalibration affect the different attributes of prediction skill? And can one of these two techniques be considered more valuable than the other from a user perspective? Or should they be applied in unison? And if yes, in which order?

We will investigate these questions by applying an improved and, for the context of seasonal forecasts, more realistic version of the simple synthetic Gaussian forecast ensemble generator (toy model) used by Weigel et al. (2008b), and by evaluating temperature and precipitation forecasts of a real seasonal MME prediction system.

The paper is structured as follows. Section 2 presents the conceptual background of this study, and the methods of MMEC and recalibration are introduced. Sections 3 and 4 describe the stochastic toy model and the verification context. In section 5, the core of this study, the toy model is systematically applied to investigate and discuss the differing effects of MMEC and recalibration on prediction skill. The findings are substantiated with a real seasonal MME prediction system in section 6, and a generalization of the recalibration method to skewed data is suggested. Concluding remarks are presented in section 7.

2. Methods

a. Multimodel ensemble combination

MMEs are constructed by simply pooling together the participating SMEs with each ensemble member having equal weight (e.g., Hagedorn et al. 2005). More sophisticated approaches in which the participating SMEs are weighted according to their prior performance (e.g., Rajagopalan et al. 2002; Robertson et al. 2004; Stephenson et al. 2005; DelSole 2007; Weigel et al. 2008b; Peña and van den Dool 2008) are not considered here. Note that, when applying MMEC and discussing its effects, we always assume that systematic biases in mean and variance of the model climatologies have been removed prior to model combination as described by Weigel et al. (2008b). The potentially beneficial effect of MMEC on such systematic errors is therefore not considered in this study.

b. Climate-conserving recalibration

We now derive the recalibration method applied in this study. The concept itself is not new and has already been applied by Doblas-Reyes et al. (2005). However, a theoretical derivation has not been presented in literature. Since the recalibration algorithm is designed such that it does not introduce systematic biases in mean and variance to the model climatologies, it will henceforth be referred to as climate conserving recalibration (CCR).
We start from a conceptual model of (seasonal) predictability, similar to the one described by Kharin and Zwiers (2003). We consider a set of observations $x$ (e.g., seasonal averages of surface temperature at a given location). Then we assume that each observation can be formulated as the sum of a model-predictable signal $\mu_s$ and an unpredictable noise term $\epsilon_s$, that is $x = \mu_s + \epsilon_s$. Following Kharin and Zwiers (2003), $\mu_s$ can be thought of as the expected atmospheric response to slowly varying and predictable boundary conditions such as anomalies in sea surface temperature, while $\epsilon_s$ represents the chaotic and unpredictable components of the observed dynamical system: $x$, $\mu_s$, and $\epsilon_s$ are assumed to be stochastic Gaussian processes with zero mean (i.e., anomalies are considered rather than absolute values). Let $\sigma^2_s$ and $\sigma^2_{\epsilon_s}$ be the variances of $x$ and $\epsilon_s$ across time. Furthermore, let $\sigma^2_{\epsilon_s}(t)$ be the unpredictable internal variability at time $t$ [i.e., the variance of the (hypothetical) distribution of possible outcomes, given the predictable signal $\mu_s(t)$]. Note that $\sigma^2_{\epsilon_s}$ is time dependent, that is, the level of predictability is allowed to vary from case to case. Under these assumptions, a specific observation at time $t$ can be formulated as

$$x(t) = \mu_s(t) + \epsilon_s(t) \quad (1)$$

with

$$\mu_s(t) \sim \mathcal{N}(0, \sigma^2_{\mu_s}) \quad \text{and} \quad \epsilon_s(t) \sim \mathcal{N}(0, \sigma^2_{\epsilon_s})$$

$\mathcal{N}(\mu, \sigma)$ thereby indicates a random number drawn from a normal distribution with mean $\mu$ and variance $\sigma^2$. This concept is illustrated in Figs. 1a,b: the presence of a given predictable signal $\mu_s$ shifts, and on average also narrows, the distribution of possible outcomes with respect to climatology.

Now assume that prior to each observation $x$ a corresponding $M$-member ensemble forecast $\mathbf{f} = (f_1, f_2, \ldots, f_M)$ has been issued. Assume that these forecasts are issued as anomalies with respect to the mean of the model climatology. If the ensemble forecasts are perfectly reliable, then the observations $x$ and the individual ensemble member forecasts $f_i$ should be statistically indistinguishable from each other for all $i \in [1, \ldots, M]$. This implies (i) that $\sigma^2_{\epsilon}(f_i) = \sigma^2_{\epsilon}$ (where $\sigma^2_{\epsilon}$ is the variance of $\epsilon_s$ across time) and (ii) that, for any given predictable signal $\mu_s(t)$, each forecast member $f_i(t)$ represents an equally likely random sample from the distribution of possible observable states, given the predictable signal $\mu_s(t)$. A reliable ensemble forecast therefore has the following structure:

$$f_i(t) = \mu_s(t) + \epsilon_i(t) \quad (2)$$

with

$$\epsilon_i(t) \sim \mathcal{N}(0, \sigma_{\epsilon_i}(t)).$$

The ensemble mean is then an unbiased estimator of the predictable signal, and the ensemble spread quantifies the uncertainty of the true outcome (illustrated in Fig. 1c).

For real ensemble prediction systems, however, model climatologies tend to be different from the observed climatology (i.e., $\sigma^2_{\epsilon} \neq \sigma^2_s$), and the expected ensemble means $\mu_f$, that is the predicted signals, are not identical to the predictable signals $\mu_s$. In the general case, Eq. (2) must therefore be formulated as follows:

$$f_i(t) = \mu_f(t) + \epsilon_i(t) \quad (3)$$

with

$$\mu_f(t) \sim \mathcal{N}(0, \sigma_{\mu_f}) \quad \text{and} \quad \epsilon_i(t) \sim \mathcal{N}(0, \sigma_{\epsilon_i}(t)).$$

Note that $\sigma_{\epsilon_i}(t)$ quantifies the intraensemble spread at time $t$ and is generally different from $\sigma_{\epsilon_s}(t)$. Also note that the individual member forecasts $f_i$ while still being statistically indistinguishable from each other, are now statistically different from the observations $x$. In such a forecasting system, the ensemble mean is not an unbiased estimator of the predictable signal any more (see Fig. 1d) and the forecasts are unreliable.

To make unreliable forecasts reliable, we employ the following criterion of reliability, which is valid for normally distributed ensemble forecasts (Toth et al. 2003; Palmer et al. 2006): ensembles are reliable if, and only if, the mean square error (MSE) of the ensemble mean forecasts, $\text{MSE}(\mu_f, x)$ is identical to the time-mean intraensemble variance, denoted by $\langle \sigma^2_{\epsilon_i} \rangle$. The basic idea of CCR is to scale the ensemble mean forecasts $\mu_f$ by a factor $r$ and to scale the ensemble spreads by a factor $s$ that is, to construct new forecasts:

$$f_i^{(CCR)} = r \mu_f + s \epsilon_i =: \mu_f^{(CCR)} + \epsilon_i^{(CCR)}, \quad (5)$$

such that (i) the aforementioned reliability criterion is satisfied, and (ii) the forecast climatology is identical to the observation climatology. As shown in appendix A, these conditions are fulfilled if

$$r = \rho(x, \mu_f) \frac{\sigma_x}{\sigma_{\mu_f}} \quad \text{and}$$

$$s = \sqrt{1 - \rho(x, \mu_f)^2} \frac{\sigma_x}{\sqrt{\langle \sigma^2_{\epsilon_i} \rangle}}, \quad (7)$$
where \( \rho(\mu_f, x) \) is the Pearson correlation coefficient between \( x \) and \( \mu_f \). Note that, if the model climatology has a systematic bias in variance (i.e., \( \sigma^2_f \neq \sigma^2 \)), this is automatically corrected for by CCR. Indeed, regardless of whether the model climatology is explicitly calibrated prior to CCR or whether CCR is directly applied, in both cases the same values for \( f_i^{\text{CCR}} \) would be obtained.

3. The stochastic toy model

a. Definition

Motivated from the conceptual model of Kharin and Zwiers (2003), we have developed a synthetic Gaussian generator of forecast—observation pairs. It is designed such that, for a given predictable signal \( \mu_x \), it generates an observation \( x \) and a corresponding \( M \)-member ensemble forecast \( f = (f_1, \ldots, f_M) \) fulfilling preset conditions with respect to forecast skill and ensemble properties. These conditions are controlled by two free parameters, \( \alpha \) and \( \beta \), with \( \alpha \in [0, 1] \) and \( \beta \in [0, \sqrt{1 - \alpha^2}] \).

As will be elucidated further below, \( \alpha \) controls the potential model predictability, while \( \beta \) controls the dispersion characteristics of the forecast ensembles. The toy model has standardized and well-calibrated climatologies (i.e., \( \sigma^2_f = \sigma^2_{\mu_x} = 1 \)) for all \( i \in \{1, \ldots, M\} \).

For given values of \( \alpha \) and \( \beta \), the following three steps are undertaken to generate a forecast—observation pair:

Step 1—A predictable signal \( \mu_x \) is sampled:

\[
\mu_x \sim \mathcal{N}(0, \alpha).
\]

Step 2—An observation \( x \) is constructed by sampling an unpredictable noise term \( \epsilon_x \), which is then added to \( \mu_x \):

\[
x = \mu_x + \epsilon_x
\]

with

\[
\epsilon_x \sim \mathcal{N}(0, \sqrt{1 - \alpha^2}).
\]

Note that \( \sigma^2_x = 1 \) for all \( \alpha \in [0, 1] \). Also note that \( \sigma^2_x \) is uniquely determined by \( \alpha \); hence, if \( \alpha \) is kept constant, it does not vary from observation to observation [in contrast to Eq. (1)].

Step 3—A forecast ensemble \( f \) is constructed by imposing a scalar perturbation \( \epsilon_{\beta} \) and an independently sampled vector perturbation \( (\epsilon_1, \ldots, \epsilon_M) \) on the predictable signal \( \mu_x \):

\[
\begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_M
\end{pmatrix}
= \begin{pmatrix}
\mu_x \\
\mu_x + \epsilon_{\beta} + \epsilon_1 \\
\vdots \\
\mu_x + \epsilon_{\beta} + \epsilon_M
\end{pmatrix}
\]

with

\[
\epsilon_{\beta} \sim \mathcal{N}(0, \beta),
\]

\( \epsilon_1, \epsilon_2, \ldots, \epsilon_M \sim \mathcal{N}(0, \sigma_{\text{ens}}) \), and

\[
\sigma_{\text{ens}} = \sqrt{1 - \alpha^2 - \beta^2}.
\]

Note that the forecast signal \( \mu_f = \mu_x + \epsilon_{\beta} \) is generally different from \( \mu_x \), and note that \( \sigma^2_f = 1 \) for all \( \alpha \in [0, 1] \) and all \( \beta \in [0, \sqrt{1 - \alpha^2}] \). Furthermore, note that \( \sigma_{\text{ens}} \) only depends on \( \alpha \) and \( \beta \) and hence, if \( \alpha \) and \( \beta \) are kept constant, does not vary from forecast to forecast [in contrast to Eq. (3)].

If a multimodel consisting of \( N \) SMEs is to be constructed, step 3 is repeated \( N \) times, yielding \( N \) forecast
ensembles $f^{(1)}, \ldots, f^{(N)}$ which are then pooled together to a MME. Note that here it is assumed that all participating SMEs see the same predictable signal $\mu_x$. Transferred to a real prediction context, this implies that all models are assumed to be based on the same sources and processes of predictability, but differ in the way the ensembles represent the remaining uncertainties. At least in the context of seasonal forecasting, this assumption can be justified to some degree (Weigel and Bowler 2009) given that at present, state-of-the-art seasonal prediction systems reveal globally very similar patterns of potential predictability (Yoo and Kang 2005), and given that they perform almost equally well in predicting tropical Pacific SST anomalies (Goddard et al. 2001), which are associated with ENSO, the most dominant signal of global seasonal climate variability.

b. Interpretation of $\alpha$ and $\beta$

How can the two design parameters $\alpha$ and $\beta$ be interpreted? By construction, $\alpha$ controls the variance of the predictable signal $\mu_x$ and thus also the variance of the unpredictable noise $\epsilon_x$. If $\alpha = 0$ the predictable signal $\mu_x$ is zero and the variance of $\epsilon_x$ is 1. As $\alpha$ grows, $\mu_x$ increases in strength with respect to the noise until, for $\alpha = 1$, $\epsilon_x$ is 0. Indeed, the ratio $\sigma^2_{\mu_x}/\sigma^2_{\epsilon_x} = \alpha^2$ is often referred to as the potential predictability of the system (Zwiers 1996; Rowell 1998; Kharin and Zwiers 2003); in this terminology, $\alpha$ therefore controls the potential model predictability$^2$ of the toy model. From Eqs. (9) and (10) one can derive that $\rho(x, f) = \alpha^2$ (i.e., the potential model predictability can be conveniently estimated from the average correlation between the individual ensemble members and the observations; see section 4a).

The second parameter, $\beta$, controls the error term $\epsilon_{\beta}$ and thus the degree to which the predicted signal $\mu_f$ deviates from the predictable signal $\mu_x$—rather like the idea of model error, which affects all ensemble members equally. If $\beta = 0$, $\mu_f$ is identical to $\mu_x$ and the ensemble members truly sample the uncertainties due to the unpredictable noise $\epsilon_x$ (i.e., the forecasts are reliable; see also Weigel and Bowler 2009). As $\beta$ grows, the ensemble spread (controlled by $\sigma_{\text{ens}}$) decreases while the magnitude of $\epsilon_{\beta}$ (i.e., the random error of $\mu_f$) increases. For positive $\beta$, the ensemble forecasts are too sharp while being centered at the wrong location. Thus, $\beta$ controls the degree of ensemble overconfidence (or underdispersion)—a frequently observed characteristic of real ensemble forecasts (e.g., Weigel et al. 2008a,b).

4. Verification

In the following the verification context of this study is discussed. Since forecast quality is a multifaceted term and cannot be summarized by a single skill score (e.g., Murphy 1991), four skill metrics will be applied to characterize the impacts of MMEC and CCR: potential model predictability, reliability, discrimination, and the ranked probability skill score (RPSS).

a. Potential model predictability

In section 3b it has been shown that the Pearson correlation coefficient between the individual ensemble members and the observations is a measure for the potential predictability of the toy model (in the sense as defined by Kharin and Zwiers 2003). We therefore apply

$$\rho_{\text{pot}} = \frac{1}{M} \sum_{i=1}^{M} \rho(f_i, x)$$

(11)

as a measure of potential model predictability, with $\rho(f_i, x)$ being the Pearson correlation coefficient between the $i$th ensemble member and the observations. Here $M$ is the ensemble size.

b. Reliability

Reliability quantifies how consistent the forecast probabilities are with the relative frequencies of the observed outcomes (e.g., Mason and Stephenson 2008). As already mentioned in section 2b, normally distributed ensemble forecasts are reliable if and only if the rmse of ensemble means and observations is identical to the time-mean ensemble spread $\sqrt{\langle \sigma^2_{\text{ens}} \rangle}$. If $\sqrt{\langle \sigma^2_{\text{ens}} \rangle} > \text{rmse}(\mu_f, x)$ the forecasts are underconfident (only rarely observed in real ensemble forecasts), if $\sqrt{\langle \sigma^2_{\text{ens}} \rangle} < \text{rmse}(\mu_f, x)$ the forecasts are overconfident. Based on this fact, we define as a measure of reliability:

$$\text{REL} = \frac{\text{rmse}(\mu_f, x) - \sqrt{\langle \sigma^2_{\text{ens}} \rangle}}{\text{rmse}(\mu_f, x)}.$$

(12)

If REL = 0 (REL > 0; REL < 0) the forecasts are reliable (overconfident; underconfident).

c. Discrimination (resolution)

The forecast attribute of discrimination quantifies the degree to which forecasts differ, given different outcomes. As a measure of discrimination we apply the probability that, given any two observations, the mutual ranking of these two observations can be correctly

---

$^2$ In the literature, the term potential predictability is also often referred to as the skill that is obtained if ensemble members are verified against each other rather than against observations (the perfect model approach; e.g., Müller et al. 2004).
predicted from the corresponding ensemble mean forecasts. This measure is a special case of the generalized discrimination score (also known as the two-alternative forced choice score) \( p_{2\text{AFC}} \) that has been described in detail by Mason and Weigel (2008). In the present context, it is given by

\[
p_{2\text{AFC}} = 0.5[\tau(\mu^*, x) + 1],
\]

\( \tau(\mu^*, x) \) thereby denotes Kendall’s ranked correlation coefficient (Sheshkin 2007) between the ensemble means and the observations.

Note that a noninformative prediction system has \( p_{2\text{AFC}} = 0.5 \). This can be plausibly interpreted as the probability of getting the relative ranking of any two observations right by simple guessing. Also note that, since \( \tau(\mu^*, x) = \tau(x, \mu^*) \), the \( p_{2\text{AFC}} \) can here equally be interpreted as the probability that the observed outcomes differ, given different forecasts. This is a forecast attribute that is known as resolution (e.g., Mason and Stephenson 2008). Resolution and discrimination will henceforth be used as synonyms.

d. Ranked probability skill score

The RPSS (Epstein 1969; Murphy 1969, 1971) is one of the most widely used summary skill scores and measures both reliability and resolution. It is a squared measure comparing the cumulative probabilities of categorical forecast and observation vectors relative to a climatological forecast strategy. In this study, the RPSS will be applied for three equiprobable categories.

A big caveat of the RPSS is its strong negative bias for small ensemble sizes (e.g., Buizza and Palmer 1998; Richardson 2001; Kumar et al. 2001; Mason 2004). The reason for this bias is the intrinsic unreliability of small ensembles, leading to inconsistencies in the formulation of the RPSS. In the context of the present study this property is problematic, since it implies that the RPSS favors MMEs due to their larger ensemble size. To ensure a fair (i.e., ensemble size independent) comparison between SMEs (ensemble size \( M \)) and MMEs, we randomly sample subensembles of size \( M \) from the multimodel and use these subensembles, rather than the full MMEs, for verification. An alternative strategy that is sometimes applied, namely the use of a correction formula to remove the ensemble size–dependent bias (Weigel et al. 2007b,c; Ferro et al. 2008), cannot be applied here because the underlying assumption of ensemble member exchangeability does not hold for recalibrated ensembles (see appendix B).

5. MMEC and CCR of toy-model forecasts

In this section we apply the toy model of section 3 and the verification context of section 4 to systematically investigate the effects of MMEC and CCR on prediction skill.

a. The effect of MMEC

The effect of MMEC has already been investigated in detail in Weigel et al. (2008b) and is therefore only briefly summarized here. Assume that, for a given potentially predictable signal \( \mu_s \), a total of \( N \) overconfident toy-model SME forecasts have been issued and are to be combined. Furthermore, assume that, without loss of generality, all SMEs are based on the same toy-model parameters \( \alpha \) and \( \beta \), and that the individual model error terms \( \epsilon_p \) are independent of each other. Each of these SMEs then has an ensemble spread of \( \sigma_{\text{ens}} = \sqrt{1 - \alpha^2 - \beta^2} \). The expected ensemble mean of the resulting MME forecast is located at \( \mu_{\text{ens}}^{\text{MME}} = \mu_s + (1/N)\sum_{p} \epsilon_p(N) \), with \( \epsilon_p(n) \) being the \( \epsilon_p \) value sampled for the \( n \)th model. For \( N \to \infty \), the MME mean \( \mu_{\text{ens}}^{\text{MME}} \) converges toward \( \mu_s \), while the expected MME spread widens and approaches a value of \( \sigma_{\text{ens}} = \sqrt{1 - \alpha^2} \), which is the spread of a reliable SME with \( \beta = 0 \). This has been discussed and proved in Weigel et al. (2008b) and is illustrated in Fig. 2 (a corresponding illustration of the effect of CCR is shown in Fig. 3 and will be discussed later in the text). In other words, the combination of independent overconfident models widens the MME spread while reducing the error in the ensemble location. The larger the number of overconfident models contributing to the MME, the more does the MME lose its overconfidence characteristics in favor of the characteristics of well-dispersed ensembles. Such an MME with independent \( \epsilon_p \) and \( N \to \infty \) will henceforth be referred to as ideal MME.

How does this behavior translate into prediction skill? For \( \beta = 0.7 \) and a range of \( \alpha \) values with \( \alpha < \sqrt{1 - \beta^2} \), 100 000 sets of observations, corresponding SME forecasts and ideal (\( N = 100 \)) MME forecasts are generated. Using these data, the expected values of \( \rho_{\text{pot}} \) (Fig. 4), REL (Fig. 5), \( p_{2\text{AFC}} \) (Fig. 6), and RPSS (Fig. 7) are then calculated and plotted as functions of the prescribed potential SME predictability \( \alpha^2 \). For the moment, only consider the black solid lines (overconfident SME forecasts) and the heavy gray solid lines (ideal MME forecasts). The remaining lines will be discussed later in the text. The results can be summarized as follows:

1) Potential model predictability (Fig. 4)—MMEC leaves \( \rho_{\text{pot}} \) unchanged. This is not surprising, given that all contributing SMEs and thus also the MME by construction see the same potentially predictable signal \( \mu_s \sim N(0, \alpha) \). Thus, under the idealizing assumptions made, the potential model predictability is conserved by MMEC (see also Weigel et al. 2008b).
2) **Reliability (Fig. 5)**—The raw SMEs reveal positive REL values over the entire range of $\alpha^2$, implying that the forecasts are overconfident. This is what one would expect, given that $\beta > 0$. The ideal MMEs, on the other hand, have REL = 0 and are therefore perfectly reliable (see also Fig. 2d).

3) **Resolution (Fig. 6)**—The $p_{2AFC}$ score of both the SMEs and the ideal MMEs increases as $\alpha^2$ increases, because higher correlation implies higher discriminative power of the forecasts. The MME thereby consistently scores higher than the SME, because reduced overconfidence not only implies wider ensemble spread, but also a reduction in the random error of the ensemble mean (Weigel et al. 2008b), thus improving the probability to correctly discriminate between the observed outcomes. It is interesting to note that resolution is frequently considered to be a measure of potential predictability, a view that is not supported by the differing behavior of $p_{2AFC}$ and $\rho_{pot}$. Indeed, a contour plot of $p_{2AFC}$ for SMEs as function of $\alpha$ and $\beta$ (Fig. 8) shows that the isolines of $p_{2AFC}$ are inclined and therefore not equivalent with $\alpha$ (i.e., with potential predictability).

4) **RPSS (Fig. 7)**—MMEC strongly improves the RPSS over the entire range of $\alpha^2$ values, which is plausible, given the improvements in reliability and resolution. All in all, the results show that MMEC in the ideal case fully corrects for reliability deficits and improves the forecast resolution, while the potential predictability is conserved. These characteristics become more, respectively less, pronounced as $\beta$ is increased, respectively decreased (not shown). For $\beta = 0$, none of the four skill metrics is modified at all by MMEC, since the participating SMEs are already reliable and all sampled from the same parent distribution as the MME.

**Fig. 2.** Illustration of the effect of multimodel ensemble combination (see also Fig. 12 in Weigel et al. 2008b). The combination of overconfident SMEs [(a) 1 SME, (b) 2 SMEs, (c) 3 SMEs, and (d) 1000 SMEs] successively widens the ensemble spread and reduces the ensemble overconfidence, thus making the forecasts more and more reliable as the number of participating SMEs grows. If many SMEs with independent error terms $\epsilon_i$ (see text for details) are combined, then MMEC eventually adequately samples the full distribution of potential outcomes that are consistent with the predictable signal. Note that the probability densities are scaled differently here for illustrative purposes.

**Fig. 3.** Illustration of the effect of recalibration. Consider (a) an overconfident ensemble prediction and a potentially predictable signal $\mu_x$. Because of the ensemble overconfidence and the associated uncertainty, a part of $\mu_x$ is perceived as unpredictable noise by the EPS, leading to (b) a reduced effectively predictable signal $\mu_{x_{eff}}$. (c) From the back statistics of past forecasts and observations, recalibration factors can be calculated that (c) rescale the forecast ensembles such that they fully sample the distribution of possible outcomes that are consistent with $\mu_{x_{eff}}$. Note that the probability densities are scaled differently here for illustrative purposes.
However, we want to stress that this conclusion only holds if the model errors are sufficiently independent, and if all participating SMEs are based on the same predictable signal $m_x$ as is the case with the present toy model (see section 3a).

b. The effect of CCR

What is different when CCR is applied on overconfident toy-model forecasts? We start by formulating the CCR factors $r$ and $s$ as functions of $\alpha$ and $\beta$. From Eqs. (8)–(10) $\sigma_x = 1$, $\sigma_{\mu_x} = \sqrt{\alpha^2 + \beta^2}$, $\rho(x, \mu_f) = \frac{\alpha^2}{\sqrt{\alpha^2 + \beta^2}}$, and $\sigma_{\text{ens}} = \sqrt{1 - \alpha^2 - \beta^2}$. Using these identities in Eqs. (6) and (7), expressions for $r$ and $s$ can be formulated as

$$r = \frac{\alpha^2}{\alpha^2 + \beta^2}$$ and

$$(14)$$
s = \sqrt{\frac{\sigma^2(1 - \sigma^2) + \beta^2}{(\sigma^2 + \beta^2)(1 - \sigma^2 - \beta^2)}}.

It is easy to see that \( r \leq 1 \), and it can be shown that \( s \geq 1 \). Thus, for overconfident SMEs, CCR effectively widens (i.e., inflates) the ensemble spread \( (s \geq 1) \) while at the same time the ensemble mean is moved toward the climatological mean \( (r = 1) \). This means that the gain in intraensemble variance due to ensemble inflation is compensated by a reduction of forecast signal variance. This is required to keep the forecast climatology well calibrated [i.e., \( \sigma_j^{(\text{CCR})} = \sigma_j = 1 \)].

As in section 5a, we evaluate how \( \rho_{\text{pot}} \), REL, \( p_{2\text{AFC}} \), and RPSS behave if CCR is applied (displayed as black dashed lines in Figs. 4–7).

1) Potential model predictability (Fig. 4): CCR strongly reduces \( \rho_{\text{pot}} \). Calculation reveals that \( \rho_{\text{pot}} \) is reduced from a value of \( \sigma^2 \) down to a value of \( \rho_{\text{pot}}^{(\text{CCR})} = \sigma^2(1 + \beta^2/\sigma^2)^{-1} \).

2) Reliability (Fig. 5): As for the ideal MMEs, CCR effects a perfect correction of reliability.

3) Resolution (Fig. 6): Resolution is conserved under CCR, because the linear transformation of the ensemble mean forecasts [via \( r \) in Eq. (6)] does not
modify their relative ranking and thus preserves their discriminative power.

4) RPSS (Fig. 7): CCR improves the RPSS, but not as much as MMEC. The reason is that MMEC, in contrast to CCR, not only improves reliability but also improves resolution, which is rewarded by the RPSS.

All in all, the most notable effect of CCR is, apart from the improvement in reliability, the destruction of potential model predictability. Given that the variance of the predictable signal is $s_m^2$ (section 3b), the reduction in $r_{pot}$ due to CCR implies a dilution of the predictable signal. Indeed, SMEs that have been made reliable by CCR do not sample the distribution of possible outcomes that are consistent with $\mu_x$; rather, they sample the wider distribution of outcomes that are consistent with the remaining effectively predictable signal $\mu_{eff} \sim \mathcal{N}(0, \sqrt{r_{pot}^{CCR}})$ (illustrated in Fig. 3). This observed reduction in sharpness is plausible since the CCR-corrected SMEs must additionally account for the uncertainties due to $\epsilon_\beta$.

c. Discussion

From the toy-model simulations and conceptual considerations described above, the following fundamental difference between MMEC and CCR can be crystallized: both methods are successful in making overconfident forecasts reliable; however, MMEC provides a reliability correction with conserved correlation, while CCR provides a reliability correction with conserved resolution. Conserved correlation implies an improvement in resolution, while conserved resolution implies a reduction in correlation, or potential model predictability. This is illustrated in the contour plot of Fig. 8: consider a given overconfident toy-model SME with parameters $\alpha$ and $\beta$, which can be displayed as a point in $\alpha$–$\beta$ space (e.g., point a). Both MMEC and CCR move a down to the $\beta = 0$ line, making the forecasts reliable. MMEC does so without changing $\alpha$, yielding point b, which has improved resolution as compared to point a. CCR, on the other hand, moves a down along the respective isoline of resolution (point c), leading to a reduction in $\alpha$ and thus in potential model predictability.

While these results suggest that MMEC is never inferior to CCR, regardless which skill metric is applied, one must consider that in reality multimodels are not ideal. Usually, the number of participating SMEs is small, and the model errors (i.e., the $\epsilon_\beta$ values in our toy-model context) tend to be correlated (e.g., Yoo and Kang 2005). To address this aspect, Figs. 4–7 additionally show the skill values obtained from a dual model, that is, a MME that consists of only two SMEs: once for independent $\epsilon_\beta$ (thin solid gray line) and once for
dependent $\epsilon_B$ (correlation coefficient 0.5, dashed gray line). For all skill metrics apart from $\rho_{\text{pot}}$, the skill improvement due to MMEC is less pronounced if only two models are combined, and even worse, if the model errors are correlated. In terms of RPSS skill (Fig. 7), it is interesting to note that particularly for forecasts of low potential model predictability (i.e., small $\alpha^2$), the CCR-corrected SMEs are comparable if not better than the realistic MMEs. The reason is that the MMEC advantage of improving resolution (Fig. 6) is comparatively weak for small $\alpha^2$ and is more than outweighed by the better reliability correction of CCR. Having said that, it is essential to note that in a real forecasting context the uncertainties in the CCR parameter estimation also need to be considered (not done here), relativizing the above conclusion. Indeed, the expected reliability and skill improvement due to CCR is reduced if only small training datasets are available, if the ensemble distributions do not satisfy the assumption of normality (further discussed in section 6), or if the system is subject to trends over the training period (Liniger et al. 2007). This discussion shall therefore not be understood as a plea against multimodels, but rather as a plea to combine as many independent models as possible to maximize the beneficial effect of MMEC.

Finally, for the case of imperfect MMEs, we want to discuss whether MMEC and CCR can be combined and used in unison such that the forecasts obtain optimum characteristics w.r.t. reliability and resolution. One could, for example, first recalibrate all participating SMEs and then combine these to a MME. Alternatively, one could first combine the raw SMEs and then recalibrate the resulting MME.

The first option (recalibrate, then combine) is without additional effect beyond the effect of CCR. This is because the CCR-corrected SMEs have been made reliable and see the same predictable signal has been destroyed by recalibration, this loss cannot be recovered by MMEC. Figure 8 provides an illustration of this situation: the combination of several SMEs that have been CCR corrected (point c) is without effect, since point c already is on the $\beta = 0$ line.

Conceptually more promising is the second approach (combine, then recalibrate). As discussed above, by combining the available SMEs, reliability and resolution are improved by some amount without reducing the potential predictability (point d in Fig. 8 if we assume a realistic MME). Subsequent CCR on point d could then in principle remove the remaining reliability deficits without changing resolution (i.e., point d would be moved down to point e). That way, a full reliability correction could be achieved under minimum destruction of potential predictability. However, in the context of the present paper, this approach is somewhat hypothetical since realistic MMEs tend to have multimodal distributions (e.g., Figs. 2b,c), thus violating the Gaussian assumptions of CCR. Therefore, other recalibration methods that are beyond the scope of this paper would be required to demonstrate the effect of this approach. Likely, such methods would require even larger training datasets. In the case of well-defined dichotomous events, an approach based on reliability diagrams as applied by Palmer et al. (2008) could be a viable option.

6. MMEC and CCR of real seasonal forecast data

So far, all results have been obtained on the basis of a simple Gaussian-type toy model. It is the aim of this section to investigate whether the conclusions on the effects of CCR and MMEC also hold for real seasonal ensemble predictions.

a. Data

Ensemble forecasts of three operational seasonal prediction systems are evaluated and combined: ECMWF’s system 2 (E; Anderson et al. 2003), the Met Office’s GloSea (U; Gordon et al. 2000), and the coupled ocean–atmosphere model of the Centre National de Recherches Météorologiques of Météo-France (C; Déqué 2001). Hindcast data of these three models are obtained from the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMEME) database (Palmer et al. 2004). Although this database comprises hindcasts of seven different models, we have restricted ourselves to the three models the operational European Multimodel Seasonal to Interannual Prediction System (EUROSIP; Vitart et al. 2007) is based upon.

We consider hindcasts of mean summer near-surface (2 m) temperature and total precipitation, averaged over the months June–August (JJA). All hindcasts have been started from 1 May initial conditions. For temperature, the hindcast period is 1960–2001. The forecast data are CCR corrected and verified gridpointwise against the corresponding observations from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) dataset (Uppala et al. 2005). For precipitation, the observations stem from the Global Precipitation Climatology Program (GPCP, more
Both forecasts and verifying observations are interpolated on a grid with 2.5° × 2.5° resolution. Prior to any recalibration, combination, and verification operations, the model climatology is calibrated gridpointwise [i.e., systematic biases in the mean and variance of the model climatology are removed as described in Weigel et al. (2008b)]. Indeed, when referring to raw SME forecasts, we henceforth assume that the model climatologies have already been calibrated. For the RPSS evaluations, three equiprobable categories are considered, just as in the toy-model experiments above. The terciles separating the three categories are determined from the hindcast and observation data separately.

The temperature forecasts are evaluated in retroactive mode. This means that for each target year to be verified, only data prior to the target year are used as training data for the computation of the observation and model terciles, bias corrections, and the CCR rescaling parameters $r$ and $s$. As target years for verification, we chose 1980–2001. The corresponding training data stem from the 20 yr prior to each target year. Generally, a retroactive evaluation is considered to yield the most realistic approximation of an operational prediction context (Mason and Baddour 2008), particularly in the presence of nonstationarities in the climate system such as trends (Liniger et al. 2007). The estimated $r$ and $s$ values obtained are often substantially different from 1. For example, for model E, typical $r$ values ($s$ values) are on the order of 0.5 (1.2), clearly indicating ensemble overconfidence.

For the precipitation forecasts, a retroactive evaluation is not possible due to the smaller sample size. Instead, a 1-yr-out cross validation (Wilks 2006) is applied, meaning that all years available, apart from the target year, are used as training data. Note that both for the temperature forecasts and the precipitation verification, the length of the training dataset is 20 yr, which is comparable to the hindcast length of real operational state-of-the-art seasonal prediction systems.

### b. Forecasts of near-surface (2 m) temperature

For the evaluation of seasonal forecasts of 2-m temperature we assume that the climatologic and forecast distributions are Gaussian, so that CCR can be applied. The assumption of normality is admittedly a very simplifying one, but can be justified as a first rough estimate for this variable (Wilks 2002, 2006; Weigel et al. 2008b). For each grid point, $p_{pot}$, REL, $p_{2AFC}$, and RPSS are obtained for (i) the raw SMEs, (ii) CCR recalibrated SMEs, (iii) the MME constructed from the raw SMEs, and (iv) the MME constructed from the CCR-corrected SMEs. The results are presented as averages over all high-predictability grid points (HPGs; Fig. 9a) and all low-predictability grid points (LPGs; Fig. 9b). HPGs (LPGs) are thereby defined as those grid points where the average potential model predictability of the three participating SMEs is larger than 0.3 (lower than 0.1). These thresholds have been chosen subjectively to have approximately the same number of HPGs and LPGs. The resulting average skill values are shown in Figs. 10–13. The raw SME forecasts are thereby labeled with E, U, and C, the corresponding CCR-corrected SMEs are labeled with Er, Ur, and Cr; the MME forecasts constructed from the raw (recalibrated) SMEs are labeled with M (Mr). The results can be summarized as follows (for the moment, ignore the columns denoted by Mr):

1) Potential model predictability (Fig. 10)—All participating SMEs have comparable values of potential predictability. At the HPGs, CCR strongly reduces $p_{pot}$ from a value of about 0.45 down to a value on the order of 0.35, while MMEC does not affect $p_{pot}$. At the LPGs, the difference between Er, Ur, and Cr on the one hand and M on the other hand is only marginal because from the beginning there is essentially no potentially predictable signal that could be further reduced by CCR. This is consistent with the toy-model results of Fig. 4.

2) Reliability (Fig. 11)—The SME forecasts have a positive reliability term, implying overconfidence as expected. Both CCR and MMEC clearly improve the reliability, with CCR providing a better, though not perfect, reliability correction, regardless whether HPGs or LPGs are considered. The observation that CCR does not improve REL down to zero differs from the toy-model experiments and is presumably due to the comparatively short record of training data and deviations from Gaussianity, leading to errors in estimating the recalibration parameters $r$ and $s$.

3) Resolution (Fig. 12)—MMEC improves the $p_{2AFC}$ score at HPGs by about 5%, while $p_{2AFC}$ is essentially left unchanged at the LPGs. CCR, on the other hand, destroys resolution, particularly at the LPGs. The latter observation is different from the toy-model results and is, again, presumably due to errors in the recalibration parameter estimates.

4) RPSS (Fig. 13)—At the HPGs (LPGs) the average RPSS of the three SMEs is 0.16 (−0.17). CCR improves this skill score to an average of 0.18 (−0.07), while MMEC yields 0.22 (−0.11). This means that
both CCR and MMEC improve the skill values. However, MMEC yields higher skill scores at the HPGs whereas CCR performs better at the LPGs. In other words, there are conditions (namely low potential predictability), under which an advanced single-model strategy such as CCR can outperform a multimodel approach. This conclusion is in full agreement with the toy-model results in Fig. 7.

5) Now consider MMEs constructed from CCR-correction SMEs (Mr). Figures 10–13 show that, regardless which skill metric is considered, Mr is by and large of the same order of magnitude as Er, Ur, and Cr (i.e., the combination does not induce much added value beyond the effect of CCR alone). In particular, observed losses in resolution and correlation due to CCR cannot be regained by subsequent MMEC.

All in all, this evaluation shows that, despite the comparatively short verification record available, and despite the very simplifying assumptions concerning the Gaussian behavior of observations and forecasts, the key conclusions drawn from the toy-model experiments are reproduced astonishingly well (apart from the conservation of resolution by CCR). Most notably, the real forecasts indicate that MMEC not only outperforms the skill values of raw SMEs, but also of recalibrated SMEs.

![Fig. 9. Grid points (in gray) of (a) high seasonal predictability and (b) low seasonal predictability, evaluated for JJA averages of 2-m temperature with a lead time of 1 month. Predictability is considered high (low) if the average correlation of the forecasts of the E, U, and C models with the observations is larger than 0.3 (lower than 0.1).](image)

![Fig. 10. Potential model predictability ($\rho_{pot}$) for real seasonal forecasts of JJA-averages of 2-m temperature, with a lead time of 1 month, obtained from the DEMETER database for the period 1980–2001. Values of $\rho_{pot}$ are determined gridpointwise and averaged over (a) all HPGs and (b) all LPGs as shown in Fig. 9. The evaluations are carried out for the raw single model forecasts E, U, and C; for the CCR-corrected single model forecasts Er, Ur, and Cr; for the multimodel M that is constructed from the raw forecasts E, U, and C; and for the multimodel Mr that is constructed from the recalibrated forecasts Er, Ur, and Cr. The recalibration parameters are estimated in retroactive mode from the 20 yr prior to each target year.](image)
however, only if a pronounced potential model predictability is present. For situations with low predictability, similar if not better skill scores can be achieved by recalibration.

c. Generalization to skewed distributions: Forecasts of precipitation

It is a major limitation of the applicability of CCR that it requires normally distributed forecasts and climatologies. Here we suggest a generalization of this method such that it can also be applied to skewed distributions such as precipitation. In essence, we follow the approach of Tippett et al. (2007) and apply so-called Box–Cox transformations (see appendix C), which only depend on a parameter \( \lambda \) and make the data approximately Gaussian. More specifically, the following steps are applied to recalibrate the precipitation forecasts. First, both for the observations and the forecasts, optimum Box–Cox-transformation parameters \( \lambda_{\text{obs}} \) and \( \lambda_{\text{fcst}} \) are estimated from a maximum likelihood approach (appendix C) and applied to make the data normal. CCR as defined in Eqs. (5)–(7) is then applied on the transformed data. The resulting recalibrated forecast data are finally transformed back into observation space, applying an inverse Box–Cox transform with parameter \( \lambda_{\text{obs}} \).

As above, skill is evaluated both for the three raw SMEs (i.e., E, U, C), the corresponding recalibrated
SMEs (i.e., Er, Ur, Cr), and the MMEs (i.e., M and Mr). Again, the results are stratified on HPGs and LPGs as shown in Fig. 14. Note that the number of HPGs is much lower than for the temperature forecasts in Fig. 9. Here we only consider the RPSS skill score since the ensemble-mean based metrics of reliability (REL) and resolution $p_{2AFC}$ as introduced in section 4 are problematic to interpret if applied on skewed data. Figure 15 shows that, similarly to the discussion above, both CCR and MMEC improve the skill values, with MMEC being more effective at HPGs and CCR being more effective at LPGs. Again, the skill value of the MME constructed from recalibrated SMEs (i.e., Mr) is comparable to the skill values of the recalibrated SMEs alone. However, note that at the HPGs the gain in prediction skill due to CCR is less pronounced than for temperature forecasts. This is probably due to additional uncertainties arising from the small number of HPGs and the estimation of the Box–Cox parameters, whose accuracy sensitively depends on the length of the training record. Nevertheless, these results imply that the application of suitable transformations can indeed be a viable option to generalize Gaussian recalibration methods to skewed data such as precipitation.

7. Conclusions

Multimodel ensemble combination (MMEC) is a well-established technique to improve the prediction skill of ensemble forecasts. However, given that MMEC essentially aims at improving the forecast reliability, we have raised and discussed the question as to whether the same effect could be achieved by an appropriate recalibration. For that purpose, an easy-to-implement climate-conserving recalibration (CCR) technique has been derived and applied. While this CCR technique is based on the
assumption of Gaussian forecast distributions, it can be made applicable to skewed distributions such as precipitation by applying an appropriate transformation.

Our discussion has been largely based on a stochastic generator of synthetic and Gaussian forecast–observation pairs. This toy model has two free parameters controlling two essential statistical properties of forecast ensembles: the underlying potential model predictability of the forecasting system and the reliability of the ensemble distributions. The toy model has been used to systematically generate forecast ensembles of varying characteristics. These forecasts have then been corrected by CCR or combined to a multimodel. It is thereby assumed that all single model ensembles (SMEs) contributing to a multimodel ensemble (MME) see the same predictable signal, an assumption that can mostly be justified in the context of seasonal forecasting. Four skill metrics have been applied to assess the impacts of CCR and MMEC: potential model predictability, reliability, resolution, and the ranked probability skill score (RPSS).

The central conclusion of this study is that both MMEC and CCR improve the forecast reliability. However, while MMEC simultaneously improves resolution, resolution is in principle conserved by CCR. Potential predictability, on the other hand, is conserved by MMEC but reduced by CCR. These findings suggest that MMEC is superior from a principle point of view compared to CCR. However, this statement only holds if ideal multimodels are considered (i.e., MMEs consisting of infinitely many SMEs with independent model error terms). In a real forecasting context, the success of MMEC strongly decreases if only a few SMEs contribute to the MME, or if the individual model errors are correlated. Under such conditions, CCR-corrected SMEs can be much more reliable than a MME and consequently yield higher RPSS skill values, at least in regions of low potential predictability when the dilution of predictable signal essentially does not matter while overconfidence does. Having said that, the effect of CCR can be strongly deteriorated if the estimation of the recalibration parameter is not robust (e.g., due to short data records or wrong distributional assumptions). All these conclusions have been confirmed by an evaluation of real seasonal ensemble forecasts of near-surface temperature and precipitation.

Many forecast providers and users may now ask the question: “Which method is better then?” In short, our evaluations have shown that this question cannot be easily answered in such generic terms, since it depends on many aspects, including the multifaceted nature of prediction skill, economic considerations, and the potential predictability of the system itself. Indeed, the value of MMEC depends on questions such as the following: How many models are available for a multimodel? How independent are these models from each other in terms of structure and model errors? How expensive is it to run several models to obtain model data from different weather and climate prediction centers? Can the systematic biases of the SMEs be identified and removed prior to combination? Does the user want forecasts with optimum sharpness and resolution, rather than optimum reliability? The value of CCR, on the other hand, depends on questions such as the following: Is a sufficiently long record of hindcast and observation data available so that robust estimates of the CCR parameters can be obtained? How expensive are these hindcast data? How well are the distributional
assumptions satisfied? And does the user put a higher priority on the reliability of the forecasts rather than on optimum resolution?

All in all, and given the principle superiority of MMEC, we encourage the combination of as many models as possible as a first choice to maximize the prediction skill. CCR, on the other hand, is suggested as a reasonable alternative to obtain reliable forecasts if a good multimodel is not available or too expensive.

Finally, note that the joint application of both MMEC and CCR could be a promising approach to further optimize the forecasts. However, this requires that CCR and MMEC are used in the correct order: The multimodel combination of CCR-corrected SMEs is only of little effect, since the participating SMEs already are reliable. If, on the other hand, the raw SMEs are first combined, thus improving resolution, and then recalibrated, the forecasts can at least in principle be made reliable under minimum dilution of a potentially predictable signal. However, a more sophisticated recalibration scheme than the one presented in this study is required for this task. Such a recalibration scheme must be able to deal with multimodal ensemble distributions, which are typical for (nonideal) MMEs.

Acknowledgments. This study was supported by the Swiss National Science Foundation through the National Center for Competence in Research (NCCR) Climate, and by the ENSEMBLES project (EU FP 6 Contract GOCE-CT-2003–505539). Helpful discussions with Francisco Doblas-Reyes, Michael Tippett, Simon Mason, and Chris Ferro on different aspects of this manuscript are gratefully acknowledged. We thank two anonymous reviewers for their constructive comments on an earlier version of this paper.

APPENDIX A

Derivation of the CCR Parameters

In this appendix Eqs. (6) and (7) are derived. Let \( \langle \ldots \rangle_t \) denote averaging over time \( t \) and let \( \langle \ldots \rangle_i \) denote averaging over the ensemble members \( i \). Similarly, let \( \text{var}_r(\ldots) \) denote a variance evaluated across \( t \), and \( \text{var}_i(\ldots) \) a variance evaluated across \( i \). Furthermore, assume that the individual ensemble members \( i \) are statistically indistinguishable, and that the number of samples and ensemble members is sufficiently large that removing one sample or ensemble member does not substantially affect the results. We start from Eq. (5):

\[
 f_i^{(\text{CCR})} = r\mu_f + s\varepsilon_i. 
\]

As explained in section 2b, \( r \) and \( s \) are chosen such that the following two conditions are satisfied:

Condition 1—The climatology of any ensemble member \( i \) is identical to the observation climatology:

\[
 \sigma_x^2 = \text{var}_i[f_i^{(\text{CCR})}] 
\]

\[
 = \text{var}_i(r\mu_f + s\varepsilon_i) 
\]

\[
 = r^2\sigma_{\mu_f}^2 + s^2\text{var}_i(\varepsilon_i). \quad (A1) 
\]

Given that the ensemble members are statistically indistinguishable from each other, one has for all \( i \in [1, \ldots, M] \):

\[
 \text{var}_i(\varepsilon_i) = \langle \text{var}_i(\varepsilon_i) \rangle_i, 
\]

\[
 = \langle \langle \varepsilon_i^2 \rangle_i \rangle_t, 
\]

\[
 = \langle \langle \varepsilon_i^2 \rangle_i \rangle_t, 
\]

\[
 = \langle \sigma_{\varepsilon_i}^2 \rangle_t. \quad (A2) 
\]

Using Eq. (A2) in Eq. (A1) yields

\[
 \sigma_x^2 = r^2\sigma_{\mu_f}^2 + s^2\langle \sigma_{\varepsilon_i}^2 \rangle_t. \quad (A3) 
\]

Condition 2—The MSE of the ensemble means is identical to the time-mean intraensemble variance:

\[
 s^2\langle \sigma_{\varepsilon_i}^2 \rangle_t = \text{MSE}(\mu_f, x) 
\]

\[
 = \text{var}_r(r\mu_f - x) 
\]

\[
 = \text{var}_r(r\mu_f) + \text{var}_r(x) - 2\text{cov}(r\mu_f, x) 
\]

\[
 = r^2\sigma_{\mu_f}^2 + \sigma_x^2 - 2rp(\mu_f, x)\sigma_{\mu_f}\sigma_x. \quad (A4) 
\]

Solving Eqs. (A3) and (A4) for \( r \) and \( s \) yields Eqs. (6) and (7). Note that a second solution is given by \( r = 0 \) and \( s = \sigma_x/\sqrt{\langle \sigma_{\varepsilon_i}^2 \rangle_t} \), which corresponds to random sampling from the climatology.

APPENDIX B

Nonexchangeability of Recalibrated Ensemble Members

As mentioned in the text, the RPSS is negatively biased for small ensemble sizes. Ferro et al. (2008) have derived a formula for an unbiased estimator of the RPSS that would be obtained was the ensemble size infinite. However, as will be shown in the following, the key assumption of ensemble member exchangeability is violated once ensembles have been recalibrated, thus forbidding the application of such a bias correction formula. Exchangeability implies that
(i) the correlation between any two ensemble members does not depend on which ensemble members are chosen [i.e., \( \rho(f_i, f_j) = \rho \) for all \( i \neq j \) with \( i, j \in \{1, \ldots, M\} \)];

(ii) \( \rho \) is independent of the ensemble size \( M \) (i.e., new ensemble members can be hypothetically added without changing the statistical properties of the ensemble members).

Without loss of generality, consider an unskilled \( M \)-member ensemble prediction system with \( \rho(x, \mu) = 0 \). Applying CCR on such an ensemble yields \( r = 0 \) [Eq. (6); i.e., the ensemble mean is shifted to 0]. From this it follows that

\[
f_M^{(\text{CCR})} = -\sum_{i=1}^{M-1} f_i^{(\text{CCR})}. \tag{B1}
\]

Were the recalibrated SME members exchangeable, condition (i) would require that

\[
\rho = \rho[f_1^{(\text{CCR})}, f_M^{(\text{CCR})}] = -\rho \left[ f_1^{(\text{CCR})}, \sum_{i=1}^{M-1} f_i^{(\text{CCR})} \right] = -[1 + (M - 2) \rho],
\]

implying that

\[
\rho = -\frac{1}{M - 1}. \tag{B2}
\]

From this it follows that condition (ii) is not fulfilled and that the recalibrated forecast ensemble members are not exchangeable. Note that this result also forbids the bias correction formulas of Weigel et al. (2007b,c), which are based on the even stricter assumption of independent ensemble members.

**APPENDIX C**

**Box–Cox Transformations**

To apply CCR on skewed precipitation data, we use a suitable power transformation to transform the original data such that their distribution becomes normal.

Box and Cox (1964) have proposed a useful family of parametric power transformations, which are often referred to as Box–Cox transformations. These transformations map a set of \( n \) data values \( y = (y_1, \ldots, y_n) \) to another set of transformed data values \( y^{(\lambda)} = (y_1^{(\lambda)}, \ldots, y_n^{(\lambda)}) \), with the parameter \( \lambda \) defining a particular transformation. This family of transformations is given by

\[
y^{(\lambda)} = \begin{cases} \frac{y - 1}{\lambda} (\lambda \neq 0), \\
\log y (\lambda = 0), \end{cases}
\]

An optimum value for \( \lambda \) is commonly obtained by maximizing the logarithm of the likelihood function \( L \) (for details see Box and Cox 1964), which is given by

\[
\log [L(y, \lambda)] = -\frac{n}{2} \log \left\{ \frac{\sum_{i=1}^{n} [y_i^{(\lambda)} - \bar{y}^{(\lambda)}]^2}{n} \right\} + (\lambda - 1) \sum_{i=1}^{n} \log y_i
\]

with

\[
\bar{y}^{(\lambda)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(\lambda)}.
\]

**REFERENCES**


