Risks of Model Weighting in Multimodel Climate Projections

ANDREAS P. WEIGEL

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

RETO KNUTTI

Institute for Atmospheric and Climate Science, ETH, Zurich, Switzerland

MARK A. LINIGER AND CHRISTOF APPENZELLER

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

(Manuscript received 23 December 2009, in final form 16 March 2010)

ABSTRACT

Multimodel combination is a pragmatic approach to estimating model uncertainties and to making climate projections more reliable. The simplest way of constructing a multimodel is to give one vote to each model (''equal weighting''), while more sophisticated approaches suggest applying model weights according to some measure of performance (''optimum weighting''). In this study, a simple conceptual model of climate change projections is introduced and applied to discuss the effects of model weighting in more generic terms. The results confirm that equally weighted multimodels on average outperform the single models, and that projection errors can in principle be further reduced by optimum weighting. However, this not only requires accurate knowledge of the single model skill, but the relative contributions of the joint model error and unpredictable noise also need to be known to avoid biased weights. If weights are applied that do not appropriately represent the true underlying uncertainties, weighted multimodels perform on average worse than equally weighted ones, which is a scenario that is not unlikely, given that at present there is no consensus on how skill-based weights can be obtained. Particularly when internal variability is large, more information may be lost by inappropriate weighting than could potentially be gained by optimum weighting. These results indicate that for many applications equal weighting may be the safer and more transparent way to combine models. However, also within the presented framework eliminating models from an ensemble can be justified if they are known to lack key mechanisms that are indispensable for meaningful climate projections.

1. Introduction

Given the reality of a changing climate, the demand for reliable and accurate information on expected trends in temperature, precipitation, and other variables is continuously growing. Stakeholders and decision makers in politics, economics, and other societal entities ask for exact numbers on the climate conditions to be expected at specific locations by the middle or end of this century. This demand is contrasted by the cascade of uncertainties that are still inherent in any projection of future climate, ranging from uncertainties in future anthropogenic emissions of greenhouse gases and aerosols (''emission uncertainties''), to uncertainties in physical process understanding and model formulation [''model uncertainties;'' e.g., Murphy et al. (2004); Stainforth et al. (2007)], and to uncertainties arising from natural fluctuations [''initial condition uncertainty;'' e.g., Lucas-Picher et al. (2008)]. In practice, the quantification of emission uncertainties is typically circumvented by explicitly conditioning climate projections on a range of well-defined emission scenarios (e.g., Nakicenovic and Swart 2000). Initial condition uncertainty is often considered negligible on longer time scales but can, in principle, be sampled by ensemble approaches, as is commonly the case in weather and seasonal forecasting (e.g., Buizza 1997; Kalnay 2003). A pragmatic and well-accepted approach to addressing model uncertainty is given by the concept of multimodel combination (e.g., Tebaldi and Knutti 2007), which is the focus of this paper.

Corresponding author address: Andreas Weigel, MeteoSwiss, Krähbühlstrasse 58, P.O. Box 514, CH-8044 Zürich, Switzerland.
E-mail: andreas.weigel@meteoswiss.ch

DOI: 10.1175/2010JCLI3594.1

So far there is no consensus on what is the best method of combining the output of several climate models. The easiest approach to multimodel combination is to assign one vote to each model (''equal weighting''). Other more sophisticated approaches suggest that assigning different weights to the individual models, with the weights reflecting the respective skill levels of the models, or the confidence we put into them. Proposed metrics as a basis for model weights include the magnitude of observed systematic model biases during the control period (Giorgi and Mearns 2002, 2003; Tebaldi et al. 2005), observed trends (Greene et al. 2006; Hawkins and Sutton 2009; Boé et al. 2009), or composites of a larger number of model performance diagnostics (Murphy et al. 2004).

Given that, in seasonal forecasting, performance-based weighting schemes have been successfully implemented and have been demonstrated to improve the average prediction skill (e.g., Rajagopalan et al. 2002; Robertson et al. 2004; Stephenson et al. 2005; Weigel et al. 2008b), it may appear obvious that model weighting can also improve the projections in a climate change context and reduce the uncertainty range. However, the two projection contexts are not directly comparable. In seasonal forecasting, usually 20–40 yr of hindcasts are available, which mimic real forecasting situations and can thus serve as a data basis for deriving optimum model weights. Even though longer-term climate trends are not appropriately reproduced by seasonal predictions (Liniger et al. 2007), cross-validated verification studies indicate that the climate is nevertheless stationary enough for the time scale considered. Within the context of climate change projections, however, the time scale of the predictand is typically on the order of many decades, rather than a couple of months. This strongly limits the number of verification samples that could be used to directly quantify how good a model is in reproducing the climate response to changes in external forcing and, thus, to deriving appropriate weights. This situation is aggravated by the fact that existing observations have already been used to calibrate the models. Even more problematic, however, is that we do not know if those models that perform best during the control simulations of past or present climate are those that will perform best in the future. Parameterizations that work well now may become inappropriate in a warmer climate regime. Physical processes, such as carbon cycle feedbacks, which are small now, may become highly relevant as the climate changes (e.g., Frame et al. 2007). Given these fundamental problems, it is not surprising that many studies have found only a weak relation between present-day model performance and future projections (Räisänen 2007; Whetton et al. 2007; Jun et al. 2008; Knutti et al. 2010; Scherrer 2010), and only a slight persistence of model skill during the past century (Reifen and Toumi 2009). Finally, not even the question of which model performs best during the control simulations can be readily answered but, rather, depends strongly on the skill metric, variable, and region considered (e.g., Gleckler et al. 2008). In fact, given that all models have essentially zero weight relative to the real world, Stainforth et al. (2007) go a step further and claim that any attempts to assign weights are, by principle, futile. Whatever one's personal stance on the issue of model weighting in a climate change context is, it seems that at present there is no consensus on how model weights should be obtained, nor is it clear that appropriate weights can be obtained at all with the data and methods at hand.

In this study, we want to shed light on the issue of model weighting from a different perspective, namely from the angle of the expected error of the final outcome. Applying a simple conceptual framework, we attempt to answer the following questions in generic terms: 1) How does simple (unweighted) multimodel combination improve the climate projections? 2) How can the climate projections be further improved by appropriate weights, assuming we knew them? 3) What would the consequences be, in terms of the projection error, if weights were applied that were not representative of true skill? Comparing the potential gains by optimum weighting with the potential losses by ''false'' weighting, we ultimately want to arrive at a conclusion as to whether or not the application of model weights can be recommended at all at the moment, given the aforementioned uncertainties.

The paper is structured as follows. Section 2 introduces the basis of our analysis, a conceptual framework of climate projections. In section 3, this framework is applied to analyze the expected errors of both optimally and inappropriately weighted multimodels, taking the skill of unweighted multimodels as a benchmark. The impacts of joint model errors and internal variability are estimated. The results are discussed in section 4, and conclusions are provided in section 5.

2. The conceptual framework

a. Basic assumptions

Our study is based on a conceptual framework of climate change projections, similar to the concept applied by Kharin and Zwiers (2003) and Weigel et al. (2009) for seasonal forecasts. We consider a climate observable x, for example, a 30-yr average of surface temperature over a given region, and assume that x will change by Δx over a specified time period (e.g., the coming 50 yr). We decompose Δx, the predictand, into the sum of a potentially predictable signal, $\Delta\mu$, and an unpredictable

''noise'' term, ν_x: $\Delta x = \Delta\mu + \nu_x$. Thereby, $\Delta\mu$ can be thought of as the *expected* response of the climate to a prescribed change in the external forcing (i.e., the expectation of a hypothetical *perfect* model that is run many times from different initial conditions), while ν_x represents the remaining fluctuations. Now, assume an *imperfect* climate model M is applied to obtain an estimate of Δx. Let Δy_M be this estimate, that is, the climate change signal predicted by M under a prescribed change in external forcing. Assume that there is no scenario uncertainty, that is, that M is subject to the same changes in external forcing as reality. Formally, Δy_M can then be decomposed into the sum of the predictable signal $\Delta\mu$, a random noise term ν_M (often referred to as internal variability), and a residual error term ϵ_M. Thus, we have

$$\begin{aligned} \Delta x &= \Delta\mu + \nu_x \\ \Delta y_M &= \Delta\mu + \nu_M + \epsilon_M. \end{aligned} \tag{1}$$

Henceforth, ϵ_M will be referred to as the model error and can be thought of as a conglomerate of (i) errors due to uncertainties in the model parameters applied to describe unresolvable small-scale physical processes (''parametric uncertainties''), (ii) errors arising from the fact that known processes are missing or inadequately approximated in the model formulation (''structural uncertainty''), and (iii) errors due to our limited understanding of relevant feedbacks and physical processes (''process uncertainties''). A more detailed characterization of these uncertainty terms has been provided by Knutti (2008), among others.

b. Interpretation of the error terms and uncertainties

The quantification of the uncertainties of the error terms ν_x, ν_M, and ϵ_M is a key challenge in the interpretation of climate projections. The uncertainties of ν_x and ν_M stem from the high sensitivity of the short-term evolution of the climate system to small perturbations in the initial state and can, in principle, be sampled by ensemble (Stott et al. 2000) or filtering (Hawkins and Sutton 2009) approaches. For simplicity, we assume that both ν_x and ν_M follow the same (not necessarily Gaussian) distribution with expectation 0 and standard deviation σ_ν, with the understanding that real climate models can reveal considerable differences in their internal variability (Hawkins and Sutton 2009).

Conceptually much more difficult is the quantification of the uncertainty range of the model error ϵ_M. Some aspects of the parameter uncertainty may be quantifiable by creating ensembles with varying settings of model parameters (e.g., Allen and Ingram 2002; Murphy et al. 2004). In addition, some aspects of structural uncertainty may at least in principle be quantifiable by systematic experiments. However, given the enormous dimensionality of the uncertainty space, such experiments can at best provide only a first guess of the uncertainty range. Even more problematic is the quantification of the impacts due to limited physical process understanding, that is, the ''unknown unknowns'' of the climate system.

Unfortunately, the uncertainty characteristics of ϵ_M cannot be simply sampled in the sense of a robust verification. This is for two reasons: (i) the ''sample size problem,'' that is, the fact that the long time scales involved reduce our sample size of independent past observations, and (ii) the ''out of sample problem,'' that is, the fact that any conclusion drawn on the basis of past and present-day observations needs to be extrapolated to so far unexperienced climate conditions. Any uncertainty estimate of ϵ_M is therefore necessarily based on an array of unprovable assumptions and thus is inherently subjective—and volatile. The confidence we put into a climate model reflects our current state of information and belief, but may change as new information become available, or as different experts are in charge of quantifying the uncertainties (Webster 2003). In fact, in a climate change context there is no such thing as ''the'' uncertainty (Rougier 2007), and consequently it is very difficult to give a reproducible, unique, and objective estimate of expected future model performance. On the shorter time scales of weather and seasonal forecasting, model errors exist equally, but their effects can be empirically quantified by sampling the forecast error statistics over a sufficiently large set of independent verification data (e.g., Raftery et al. 2005; Doblas-Reyes et al. 2005; Weigel et al. 2009). In this way, an objective estimate of the forecast uncertainty and thus of model quality is possible; the confidence we put into the accuracy of a model projection is backed up by past measurements of model performance in comparable cases.

Thus, the central conceptual difference between the interpretation of short-range forecasts of weeks and seasons and long-range projections of climate change is in their different definitions of ''uncertainty.'' In the former, uncertainty is defined by long series of repeated and reproducible hindcast experiments and thus follows the relative frequentists' or physical perception of uncertainty, in the sense of a measurable quantity. In the latter, uncertainty is partially subjective and depends on prior assumptions as well as expert opinion, thus following the Bayesian perception of uncertainty. It is for exactly this reason that the concept of model weighting, which requires a robust definition of model uncertainty, is relatively straightforward in short-range forecasting but so controversial on climate change time scales.

In the present study we want to analyze the consequences of ''correct'' and ''false'' weights on the accuracy

of climate projections. However, a weight can only be called correct or false if the underlying uncertainties to be represented by the weights are well defined and uniquely determined. To circumvent this dilemma, we simply assume that enough data were available, or, as Smith (2002) and Stainforth et al. (2007) put it, that we had access to many universes so that the uncertainty range of ϵ_M can be fully sampled and defined in a relative frequentists' sense; that is, we assume that enough information was available such that the relative frequentists' and Bayesian interpretations of model uncertainty converge. This uncertainty, denoted by σ_M, is what we henceforth refer to as the true model uncertainty. We do not know how, or whether at all, the actual value of σ_M can be sampled in practice, but we assume that σ_M *exists* in the sense of a unique physical *propensity* as defined by Popper (1959). While this assumption may appear disputable, it is indispensable for a discussion on the effects of model weighting. Without the existence of a uniquely determined model error uncertainty, the task of defining optimum weights and thus the concept of model weighting in general would be ill-posed by principle.

Finally, we assume that (i) the noise and error terms ν_x, ν_M, and ϵ_M are statistically independent from each other and (ii) that not only ν_x and ν_M, but also ϵ_M, have expectation 0. Both assumptions may be too simplifying. The former assumption implies, among others, that the internal variability of a climate model is not affected by errors in model formulation. The latter assumption implies that, after removing the effects of internal variability, the *expected* mean bias of a model during the scenario period is the same as the *observed* mean bias during the control period (otherwise a nonzero ϵ_M would be expected). This assumption of ''constant biases'' has recently been questioned (e.g., Christensen et al. 2008; Buser et al. 2009). Nevertheless, probably for lack of better alternatives, these assumptions have been applied in most published climate projections (e.g., Solomon et al. 2007), and we will stick to them to keep the discussion as simple and transparent as possible.

c. Definition of skill

As a simple deterministic metric to quantify the expected quality of a climate change projection obtained from a climate model M, we apply the expected mean squared error (MSE) between Δy_M and Δx, henceforth denoted by $\mathcal{S}_M$:

$$\begin{aligned}\mathcal{S}_M &= \langle(\Delta y_M - \Delta x)^2\rangle \\ &= \langle(\nu_M + \epsilon_M + \nu_x)^2\rangle \\ &= 2\sigma_\nu^2 + \sigma_M^2. \end{aligned} \tag{2}$$

The brackets $\langle \ldots \rangle$ thereby denote the expectation. Since σ_ν and σ_M are assumed to be uniquely determined, $\mathcal{S}_M$ is well defined.

3. The effects of model combination and weights

In this section, we apply the conceptual framework of Eq. (1) to analyze how $\mathcal{S}_M$ is affected by the weighted and unweighted combinations of multiple model output. To keep the discussion as transparent as possible, we will restrict ourselves mainly to the combination of only two models. A generalization of the conclusions to more models will not be presented here, but is straightforward by mathematical induction, since the combination of any number of models can be decomposed into a sequence of dual combinations. We start our analysis with the simple and idealized case of fully independent model errors and negligible internal variability $\sigma_\nu = 0$ (section 3a), then we discuss the case when the model errors are not independent (section 3b), and finally we analyze the consequences to be expected if σ_ν is nonnegligible (section 3c).

a. Negligible noise, independent model errors

Assume that the unpredictable noise can be ignored (i.e., $\nu_x = \nu_M = 0$). Under these conditions one has $\Delta\mu = \Delta x$, implying that the true observable climate change signal Δx is in principle fully predictable. Assume that two climate models, $M1$ and $M2$, are applied and yield climate change projections Δy_{M1} and Δy_{M2}. Let ϵ_{M1} and ϵ_{M2} be the corresponding projection errors of $M1$ and $M2$ due to model uncertainty. From Eq. (1) it follows that

$$\begin{aligned}\Delta y_{M1} &= \Delta x + \epsilon_{M1} \\ \Delta y_{M2} &= \Delta x + \epsilon_{M2}. \end{aligned} \tag{3}$$

This situation is illustrated in Fig. 1. Under these assumptions, the expected squared errors of Δy_{M1} and Δy_{M2} are given by $\mathcal{S}_{M1} = \sigma_{M1}^2$ and $\mathcal{S}_{M2} = \sigma_{M2}^2$.

Combining Δy_{M1} and Δy_{M2} with equal weights yields a simple multimodel projection $y_{\mathrm{eq}}^{(2)}$ with

$$\Delta y_{\mathrm{eq}}^{(2)} = \Delta x + \frac{\epsilon_{M1} + \epsilon_{M2}}{2}. \tag{4}$$

The superscript ''(2)'' indicates that two models are combined, and the subscript ''eq'' indicates that they are combined with equal weight. Assuming independence of ϵ_{M1} and ϵ_{M2}, the expected MSE of this multimodel, $\mathcal{S}_{\mathrm{eq}}^{(2)}$, is

$$\mathcal{S}_{\mathrm{eq}}^{(2)} := \langle(\Delta y_{\mathrm{eq}}^{(2)} - \Delta x)^2\rangle = \sigma_{M1}^2\left(\frac{1 + r^2}{4}\right), \tag{5}$$

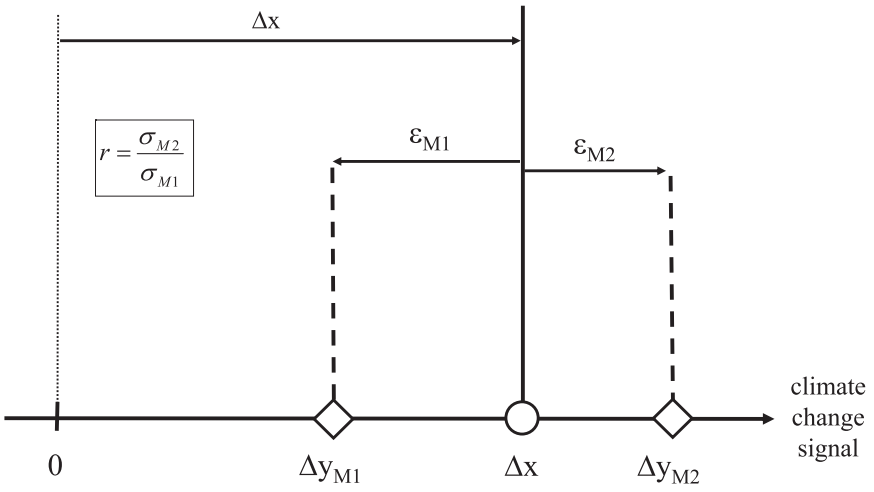

FIG. 1. The conceptual framework of climate change projections for the assumptions applied in section 3a. Here, Δx is the true climate change signal to be observed in response to a prescribed external forcing, and Δy_{M1} and Δy_{M2} are two climate change projections obtained from climate models $M1$ and $M2$ in response to the same forcing. The deviation of Δy_{M1} (Δy_{M2}) from Δx is assumed to be exclusively due to a model error ϵ_{M1} (ϵ_{M2}) with model error uncertainty σ_{M1} (σ_{M2}). The two model error terms are statistically independent from each other. The ratio $r = \sigma_{M2}/\sigma_{M1}$ is referred to as the model error ratio.

with

$$r = \frac{\sigma_{M2}}{\sigma_{M1}}. \tag{6}$$

In the following, r will be referred to as the *model error ratio* between $M2$ and $M1$. It quantifies the relative skill difference between $M1$ and $M2$. If $r = 1$, the errors of both models have the same average magnitude, implying that they have equal skill. As r gets smaller, the expected error magnitude of $M2$ decreases with respect to $M1$, implying that $M2$ has higher skill than $M1$.

Figure 2 shows, as a function of r, the effects of model averaging with equal weights. Without a loss of generality, we only show and discuss $r \leq 1$ (i.e., $\sigma_{M2} \leq \sigma_{M1}$). For the moment, we shall ignore the gray lines. Figure 2 shows the expected MSEs $\mathcal{S}_{M1}$ (thin dotted–dashed line), $\mathcal{S}_{M2}$ (thin dashed line), and $\mathcal{S}_{\text{eq}}^{(2)}$ (heavy black line) in units of $\mathcal{S}_{M1}$. It is easy to see that $\mathcal{S}_{\text{eq}}^{(2)} < 0.5(\mathcal{S}_{M1} + \mathcal{S}_{M2})$ for all r; that is, the expected MSE of the combined projection is always lower than the average of the single model errors, an observation that has also been made in the verification of seasonal multimodel forecasts (e.g., Hagedorn et al. 2005; Palmer et al. 2004; Weigel et al. 2008b). For $r \geq \sqrt{1/3} \approx 0.58$, that is, if σ_{M1} is not too different from σ_{M2}, the multimodel error $\mathcal{S}_{\text{eq}}^{(2)}$ is even lower than that of the better one of the two single models alone. For $r < \sqrt{1/3}$, on the other hand, better skill would be obtained if only $M2$ was considered rather than the multimodel. Thus, the optimum way of combining the available information is obviously a function of r.

We now derive optimum weights to be assigned to $M1$ and $M2$ such that the expected multimodel MSE becomes minimal for a given r. Consider again the two climate projections Δy_{M1} and Δy_{M2}, which are now combined to a weighted average $\Delta y_w^{(2)}$:

$$\begin{aligned}\Delta y_w^{(2)} &= w(\Delta y_{M1}) + (1 - w)(\Delta y_{M2}) \\ &= \Delta x + w\epsilon_{M1} + (1 - w)\epsilon_{M2},\end{aligned} \tag{7}$$

with w being the weight of $M1$, and $(1 - w)$ being the weight of $M2$. The expected MSE of this weighted multimodel, $\mathcal{S}_w^{(2)}$, is then given by

$$\begin{aligned}\mathcal{S}_w^{(2)} &:= \langle(\Delta y_w^{(2)} - \Delta x)^2\rangle \\ &= \sigma_{M1}^2[w^2 + (1 - w)^2 r^2].\end{aligned} \tag{8}$$

Minimizing $\mathcal{S}_w^{(2)}$ on w yields as an optimum weight w_{opt}:

$$w_{\text{opt}} = \frac{r^2}{1 + r^2}. \tag{9}$$

Note that w_{opt} only depends on the error ratio r, but not on the absolute values of σ_{M1} and σ_{M2}. As one would expect, w_{opt} approaches 0.5 as r gets close to 1. For very large (very small) error ratios, on the other hand, w_{opt} approaches 1 (0), implying that all weight is put on $M1$ ($M2$). The values of w_{opt} as a function of r have been added to Fig. 2 on the upper abscissa. Applying w_{opt} in Eq. (8) yields an expression for the optimum expected MSE $\mathcal{S}_{\text{opt}}^{(2)}$:

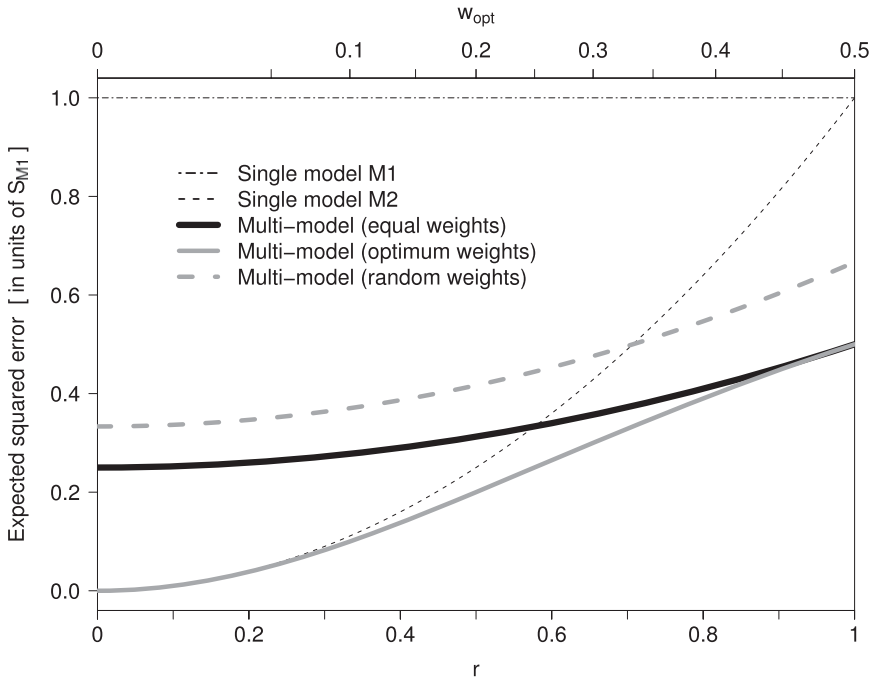

FIG. 2. Expected squared errors of single and multimodels as a function of the model error ratio r for the assumptions applied in section 3a and Fig. 1. Shown are the errors for single model $M1$ ($=\mathcal{S}_{M1}$, thin dotted–dashed line), for single model $M2$ ($=\mathcal{S}_{M2}$, thin dashed line), for the unweighted average of $M1$ and $M2$ ($=\mathcal{S}^{(2)}_{\text{eq}}$, heavy black line), and for the optimally weighted average of $M1$ and $M2$ ($=\mathcal{S}^{(2)}_{\text{opt}}$, solid gray line). The dashed gray line is $\mathcal{S}^{(2)}_{\text{rand}}$, which is the expected squared error if the two models are averaged with random weights. The errors are plotted in units of $\mathcal{S}_{M1}$. The top abscissa shows the corresponding optimum weights w_{opt} of model $M1$ as obtained from Eq. (9).

$$\mathcal{S}^{(2)}_{\text{opt}} = w^2_{\text{opt}}\sigma^2_{M1} + (1 - w_{\text{opt}})^2\sigma^2_{M2} = \sigma^2_{M1}\left(\frac{r^2}{1 + r^2}\right). \tag{10}$$

The curve of $\mathcal{S}^{(2)}_{\text{opt}}$ as a function of r has been included in Fig. 2 (solid gray line), showing that the optimally weighed multimodel clearly outperforms $\mathcal{S}_{M1}$, $\mathcal{S}_{M2}$, and $\mathcal{S}^{(2)}_{\text{eq}}$ for all r. Particularly for small values of r, that is when $M1$ and $M2$ are very different in terms of their expected errors, model weighting can indeed strongly improve the projection quality with respect to the benchmark of equal weighting. However, this requires accurate knowledge of r, which in practice is very difficult if not impossible to obtain (see discussion in section 4). What then happens in terms of the expected MSE if the models are combined with weights w, which may be thought to be optimal, but which in fact do not reflect the true model error ratio? That is, what happens if weights are applied without knowing the true value of r? Assuming that it is equally likely that by chance the optimum weight, the worst possible weight, or any other weight $w \in [0, 1]$ is picked, we introduce $\mathcal{S}^{(2)}_{\text{rand}}$ as a summary measure to quantify the expected MSE of the multimodel for random weights:

$$\mathcal{S}^{(2)}_{\text{rand}} := \int_0^1 \mathcal{S}^{(2)}_w \, dw = \int_0^1 [w^2\sigma^2_{M1} + (1 - w)^2\sigma^2_{M2}] \, dw = \sigma^2_{M1}\left(\frac{1 + r^2}{3}\right). \tag{11}$$

The curve for $\mathcal{S}^{(2)}_{\text{rand}}$ has been added to Fig. 2 as a dashed gray line. It can be seen and shown that $\mathcal{S}^{(2)}_{\text{rand}} \geq \mathcal{S}^{(2)}_{\text{eq}}$ for all r. In other words, the application of weights that are independent of r would on average yield larger errors than if no weights had been applied at all. This conclusion holds for any value of r.

So far, we have assumed that the model errors ϵ_{M1} and ϵ_{M2} are independent of each other, and that the unpredictable noise ν_M and ν_x can be ignored. Under these assumptions, the combination of infinitely many models would eventually cancel out all model errors and yield a perfect climate projection. Indeed, if m models are combined with equal weights, and if $m \to \infty$, the expected multimodel projection $\Delta y^{(m)}_{\text{eq}}$ approaches

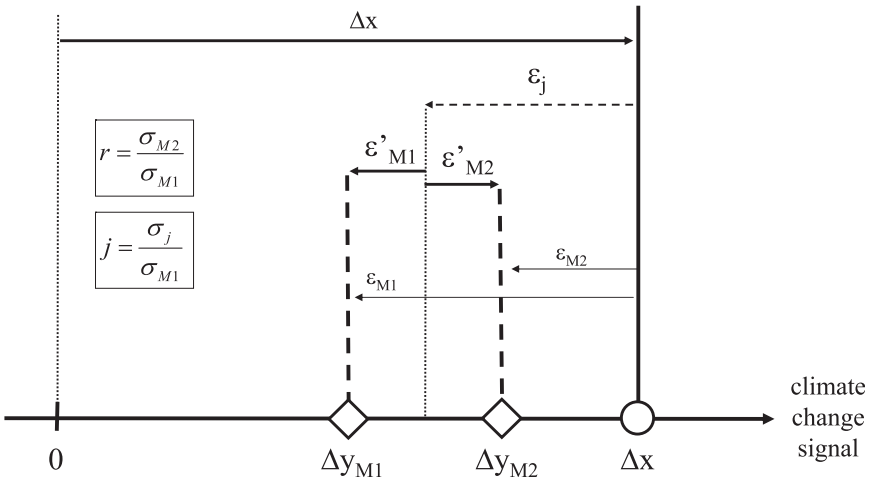

FIG. 3. The conceptual framework for climate change projections for the assumptions applied in section 3b. In contrast to Fig. 1, the deviation of the climate projection Δy_{M1} (Δy_{M2}) from the observation Δx is now thought to be decomposable into two components: (i) an error term ϵ_j (uncertainty σ_j), which is jointly seen by both participating climate models $M1$ and $M2$, and (ii) a residual error term ϵ'_{M1} (respectively ϵ'_{M2}). The residual errors are statistically independent from each other. The ratio $j = \sigma_j/\sigma_{M1}$ is referred to as the joint error fraction.

$$\lim_{m\to\infty} \Delta y_{\mathrm{eq}}^{(m)} = \lim_{m\to\infty}\left(\Delta x + \frac{1}{m}\sum_{i=1}^{m} \epsilon_{M,i}\right) = \Delta x, \qquad (12)$$

with $\epsilon_{M,i}$ being the model error of the ith model. Optimally and randomly weighted multimodels can be shown to approach the same limit. The only difference is that the optimally weighted multimodel would converge more quickly than the equally weighted one, while the randomly weighted multimodel would converge more slowly. However, this limit of full error cancellation is not consistent with what has been observed in reality. For example, Knutti et al. (2010) have shown that half of the typical surface temperature biases of climate models would remain, even if an infinite number of models of the same quality were combined. The main reason is probably that different models share similar structural assumptions and in particular share the same unknown unknowns in terms of our physical process understanding, which can lead to correlated errors (e.g., Jun et al. 2008). We are aware that the conclusion of Knutti et al. (2010) refers to an analysis of model mean biases while our discussion focuses on climate projection errors. Nevertheless, their finding illustrates how correlated model errors can influence the effects of model averaging. We therefore now extend our discussion to the situation of joint model errors, that is, model errors which are "seen" by all models, while still ignoring the effects of unpredictable noise.

b. The effect of joint model errors

Assume now that for each climate model M contributing to the multimodel, the model error ϵ_M can be decomposed into a joint error contribution ϵ_j, which is common to all models, and an independent residual error term ϵ'_M; that is, $\epsilon_M = \epsilon_j + \epsilon'_M$. For the combination of two models, $M1$ and $M2$, this implies that the predicted climate change signals Δy_{M1} and Δy_{M2} of Eq. (3) and the weighted multimodel projection $\Delta y_w^{(2)}$ of Eq. (7) become

$$\begin{aligned}
\Delta y_{M1} &= \Delta x + \epsilon_j + \epsilon'_{M1}\\
\Delta y_{M2} &= \Delta x + \epsilon_j + \epsilon'_{M2}\\
\Delta y_w^{(2)} &= w(\Delta y_{M1}) + (1-w)(\Delta y_{M2})\\
&= \Delta x + \epsilon_j + w\epsilon'_{M1} + (1-w)\epsilon'_{M2}. \qquad (13)
\end{aligned}$$

This situation is illustrated in Fig. 3. Note that now the combination of infinitely many models would not converge at Δx as in Eq. (12), but rather at $(\Delta x + \epsilon_j)$, which is more consistent with the observed behavior of real multimodels. Let σ'_{M1}, σ'_{M2} and σ_j be the underlying uncertainties of ϵ'_{M1}, ϵ'_{M2} and ϵ_j. Assuming mutual independence of ϵ'_{M1}, ϵ'_{M2} and ϵ_j, the expected single model squared errors $\mathcal{S}_{M1}$ and $\mathcal{S}_{M2}$ are given by

$$\begin{aligned}
\mathcal{S}_{M1} &= \sigma_j^2 + \sigma_{M1}'^2\\
\mathcal{S}_{M2} &= \sigma_j^2 + \sigma_{M2}'^2, \qquad (14)
\end{aligned}$$

and the expected MSE of the weighted multimodel of Eq. (8) becomes

$$\begin{aligned}
&\mathcal{S}_w^{(2)} = \sigma_{M1}^2[j^2 + w^2(1-j^2) + (1-w)^2(r^2 - j^2)]\\
&\text{with:}\quad j = \frac{\sigma_j}{\sigma_{M1}},\quad r = \frac{\sigma_{M2}}{\sigma_{M1}},\quad j \le r. \qquad (15)
\end{aligned}$$

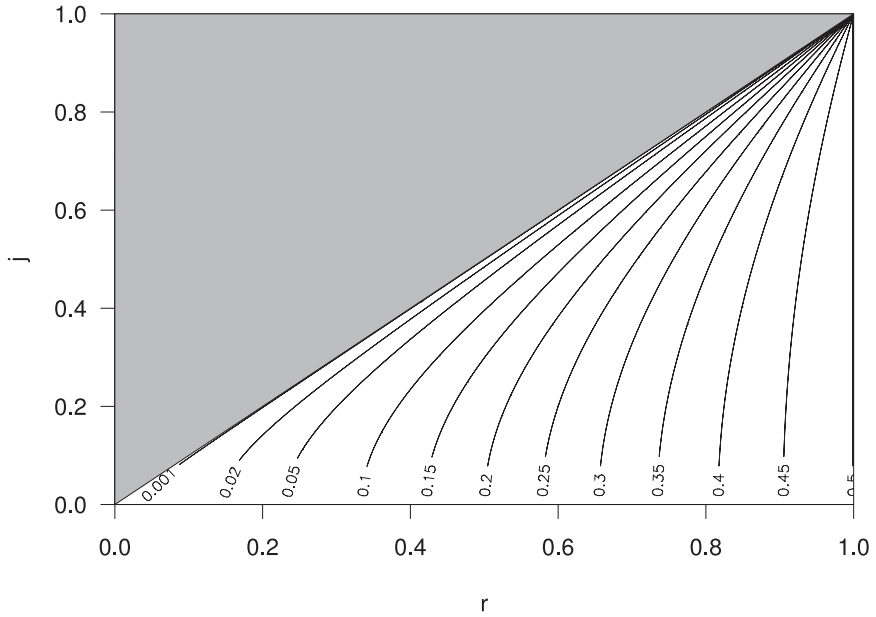

FIG. 4. Optimum weights w_{opt} for the case of dependent model errors and negligible noise as obtained from Eq. (16). The contour lines show w_{opt} as a function of r (model error ratio) and j (joint error fraction). Here, $j = 0$ corresponds to the case of fully independent model errors. Forbidden combinations of r and j are shaded in gray (by construction $j \leq r$ needs to be satisfied).

Henceforth, j will be referred to as the joint error fraction. This term measures the fraction of the root-mean-square error of $M1$, which is equally seen by $M2$. Minimizing Eq. (15) on w yields as an expression for a revised optimum weight

$$w_{\text{opt}} = \frac{r^2 - j^2}{1 + r^2 - 2j^2}. \tag{16}$$

Figure 4 shows these optimum weights w_{opt} as a function of r and j. Note that $j \leq r$ always, since σ_j, the model error uncertainty jointly seen by both $M1$ and $M2$, cannot be larger than σ_{M2}. The contour lines show that, for any r, the optimum weight w_{opt} of $M1$ decreases as j increases. For example, if the model errors ϵ_{M1} and ϵ_{M2} are fully independent ($j = 0$), an error ratio of $r = 0.6$ would correspond to an optimum weight of approximately 0.26. However, w_{opt} would drop to 0.19 if $j = 0.4$, that is if 40% of the root-mean-squared error of ϵ_{M1} contributes to the root-mean-square error of ϵ_{M2}; and w_{opt} would be zero if $j = r = 0.6$. In other words, as j increases, more weight needs to be assigned to the better one of the two models than if the model errors were fully independent. The reason is that the improvement in skill is only possible by minimizing the contributions of the independent error components rather than the total model errors. That is, the error ratio characterizing the *effective* skill difference between $M1$ and $M2$ is no longer given by $(\sigma_{M2}/\sigma_{M1})$, but rather by $(\sigma'_{M2}/\sigma'_{M1})$, which grows as j is increased (for $r \leq 1$). In summary, when the existence of joint model errors is neglected in the formulation of the optimum model weights, then the resulting estimates of w_{opt} would be implicitly biased. Too little weight would be assigned to the better one of the two models, and too much weight to the poorer one.

How does all this then affect the expected MSEs of the multimodel outcome? Figure 5 shows, in analogy to Fig. 2, the expected squared errors $\mathcal{S}_{M1}$, $\mathcal{S}_{M2}$, $\mathcal{S}^{(2)}_{\text{eq}}$, $\mathcal{S}^{(2)}_{\text{opt}}$, and $\mathcal{S}^{(2)}_{\text{rand}}$ for (a) $j = 0.2$, (b) $j = 0.5$, and (c) $j = 0.7$. Here, $\mathcal{S}^{(2)}_{\text{rand}}$ is defined in analogy to Eq. (11). Additionally, Fig. 5 shows (as triangles) the expected MSEs of a weighted multimodel with the weights being determined from Eq. (9) (assuming independent model errors) rather than Eq. (16). This will henceforth be referred to as "simplistic" weights, and the resulting MSE as $\mathcal{S}^{(2)}_{\text{simp}}$. By that, we want to analyze what would happen if r was accurately known and considered, but the existence of the joint model errors was neglected when calculating w_{opt}.

The following conclusions can be drawn from Fig. 5. As j increases, the net skill improvement of the multimodels with respect to the single models decreases, regardless of how the multimodel is constructed. This is plausible, since multimodels can only reduce the independent error components, whose magnitude decreases as j is increased. In relative terms, $\mathcal{S}^{(2)}_{\text{eq}}$, $\mathcal{S}^{(2)}_{\text{opt}}$, and $\mathcal{S}^{(2)}_{\text{rand}}$ behave similarly as in section 3a; that is, when taking

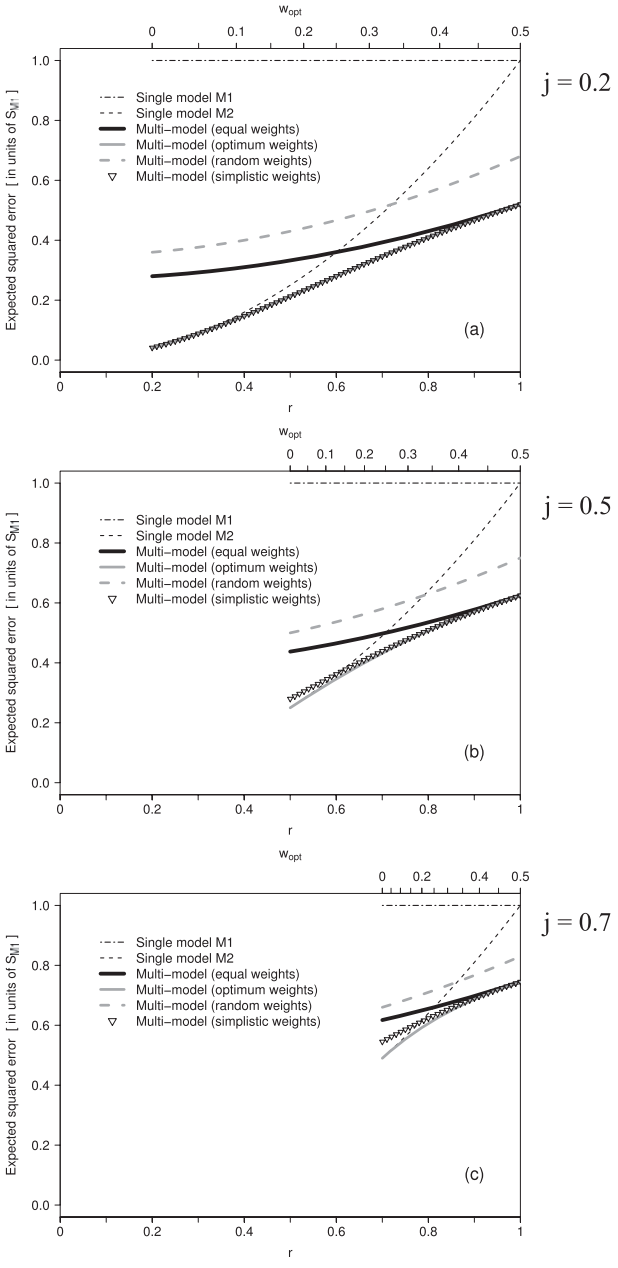

FIG. 5. As in Fig. 2 but for the assumptions applied in section 3b and Fig. 3: j = (a) 0.2, (b) 0.5, and (c) 0.7, with j being the joint error fraction. Additionally, the expected squared errors of a weighted multimodel with the weights being incorrectly determined from Eq. (9) rather than Eq. (16) are shown as triangles (simplistic weights). The top abscissa shows the true optimum weights w_{opt} as obtained from Eq. (16).

the skill of equally weighted multimodels as a benchmark, optimum weighting further reduces the expected MSE, while random weighting significantly deteriorates the error characteristics. Finally, note that the application of "simplistic" weights derived from Eq. (9) rather than Eq. (16) implies squared errors, which are larger than $\mathcal{S}^{(2)}_{\text{opt}}$, but still lower than $\mathcal{S}_{\text{eq}}$. In fact, it is only for values of $j \gtrsim 0.5$ that $\mathcal{S}^{(2)}_{\text{simp}}$ deviates significantly from $\mathcal{S}^{(2)}_{\text{opt}}$. In other words, if one was hypothetically able to determine the value of r accurately but ignored the effects of joint errors, then the results would only be moderately deteriorated with respect to the optimum weights. In summary, correlated model errors have only a minor impact on the results in section 3a concerning the relative performance of weighted versus unweighted multimodels; however, they have a major impact on the absolute multimodel performance in comparison to the single models.

In the last part of this section, we now consider the additional effects arising from unpredictable noise. For simplicity, we return to the assumption of independent model errors; that is, $j = 0$.

c. *The effect of unpredictable noise*

Under the presence of unpredictable noise, all terms in Eqs. (1) and (2) must be considered in the formulation of Δx, Δy_M, and $\mathcal{S}_M$. As described in section 2b, we assume that the noise terms ν_x, ν_{M1}, and ν_{M2} are independent samples from a distribution with expectation 0 and standard deviation σ_ν. The situation is illustrated in Fig. 6. The weighted multimodel projection of Eq. (7) then becomes

$$\Delta y_w = \Delta\mu + w(\epsilon_{M1} + \nu_{\text{M1}}) + (1 - w)(\epsilon_{M2} + \nu_{M2}), \qquad (17)$$

with an expected squared error of

$$\begin{aligned}\mathcal{S}^{(2)}_w &= w^2(\sigma^2_{M1} + \sigma^2_\nu) + (1 - w)^2(\sigma^2_{M2} + \sigma^2_\nu) + \sigma^2_\nu \\ &= \sigma^2_{M1}[w^2(1 + r^2 + 2R^2) - 2w(r^2 + R^2) + r^2 + 2R^2] \\ \text{with} \quad R &= \frac{\sigma_\nu}{\sigma_{M1}} \quad \text{and} \quad r = \frac{\sigma_{M2}}{\sigma_{M1}}. \qquad (18)\end{aligned}$$

Here, R relates the magnitude of the noise to that of the model error of $M1$ and will henceforth be referred to as the relative noise ratio. The values of $R > 1$ imply that the uncertainties due to noise exceed the model uncertainty, while $R = 0$ corresponds to the situation of negligible noise as considered above in sections 3a and 3b. Minimizing Eq. (18) over w yields the following as a revised expression for w_{opt}:

$$w_{\text{opt}} = \frac{r^2 + R^2}{1 + r^2 + 2R^2}. \qquad (19)$$

Figure 7 shows w_{opt} as a function of r and R. The contour lines reveal that, for any r, w_{opt} increases toward 0.5 as R is increased. For instance, if noise is negligible (i.e., $R = 0$), $r = 0.6$ corresponds to $w_{\text{opt}} = 0.26$. However, for $R = 0.5$ one has $w_{\text{opt}} = 0.33$, while for $R = 1$ one has

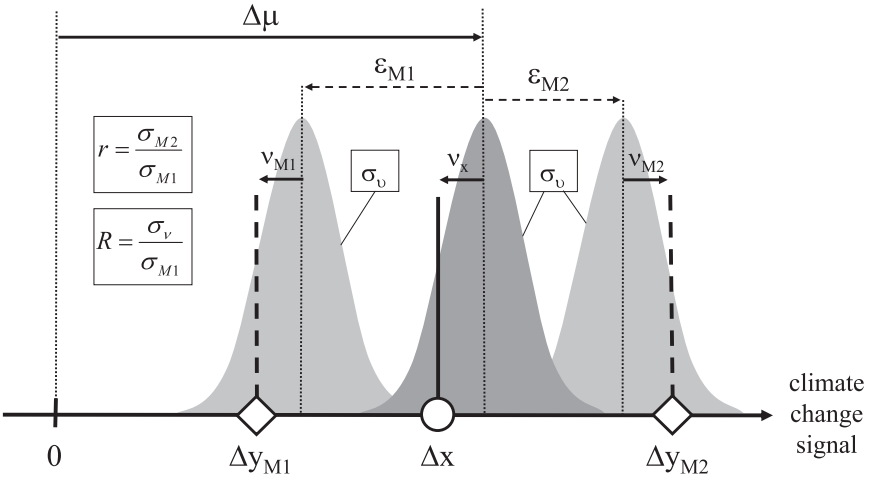

FIG. 6. The conceptual framework for climate change projections for the assumptions applied in section 3c. In contrast to Fig. 1, the observed climate change signal Δx is now thought to be decomposable into a model predictable signal $\Delta\mu$ and an unpredictable "noise" term ν_x. Similarly, the climate change projection Δy_{M1} (Δy_{M2}) is assumed to be decomposable into the predictable signal $\Delta\mu$, a model error term ϵ_{M1} (ϵ_{M2}), and a random noise term ν_{M1} (ν_{M2}). The noise and error terms are statistically independent from each other. All noise terms are assumed to be samples from a distribution with standard deviation σ_ν. The ratio $R = \sigma_\nu/\sigma_{M1}$ is referred to as the relative noise ratio.

$w_{\text{opt}} = 0.40$, and for $R \to \infty$ the optimum weight approaches 0.5 for all r. This behavior is plausible, because multimodel combination not only reduces the model errors but also the errors due to noise. Thus, as R increases, the optimum compensation of noise errors becomes more and more important for the minimization of the total projection error, and under the assumptions made, the noise errors are optimally reduced by equal weighting. In summary, when the effects of noise are neglected in the formulation of optimum model weights, then the resulting estimates of w_{opt} are implicitly biased, with the bias growing quickly as R becomes larger. The bias is such that too much weight would be given to the better one of the two models, and too little weight to the poorer one.

How does the presence of unpredictable noise then affect the quality of the multimodel projections? Figure 8 shows, in analogy to Figs. 2 and 5, the expected squared errors $\mathcal{S}_{M1}$, $\mathcal{S}_{M2}$, $\mathcal{S}^{(2)}_{\text{eq}}$, $\mathcal{S}^{(2)}_{\text{opt}}$, and $\mathcal{S}^{(2)}_{\text{rand}}$ for (a) $R = 0.5$, (b) $R = 1$, and (c) $R = 2$. The definition of $\mathcal{S}^{(2)}_{\text{rand}}$ is analogous to Eq. (11). Additionally, Fig. 8 shows (as triangles) the expected MSE of a weighted multimodel with the weights being determined from Eq. (9) (assuming negligible noise) rather than Eq. (19). As above in section 3b, this will be referred to as simplistic weights, and the resulting MSE as $\mathcal{S}^{(2)}_{\text{simp}}$. By that, we want to analyze what would happen if r was accurately known and considered, but the noise was neglected when calculating w_{opt}.

The results can be summarized as follows. As R increases, the difference between $\mathcal{S}_{M1}$ and $\mathcal{S}_{M2}$ decreases and the two models become more similar in terms of their net skill, because the individual model error terms ϵ_{M1} and ϵ_{M2} lose are diminished in relative importance with respect to the unpredictable noise. At the same time, the range of r values for which the equally weighted multimodel outperforms $M2$ (i.e., the better one of the two single models) grows. Indeed, in section 3a it has been noted that, under the absence of unpredictable noise, $\mathcal{S}^{(2)}_{\text{eq}} \le \mathcal{S}_{M2}$ only if $r \ge \sqrt{1/3} \approx 0.58$. However, if $R = 0.5$, then $\mathcal{S}^{(2)}_{\text{eq}} \le \mathcal{S}_{M2}$ for all $r \ge \sqrt{1/6} \approx 0.40$; and if $R \ge \sqrt{0.5} \approx 0.71$, then the equally weighted multimodel outperforms both single models for any $r \in [0, 1]$. Taking $\mathcal{S}^{(2)}_{\text{eq}}$ as a benchmark, the additional error reduction by optimum weighting decreases as R becomes larger. This is simply because w_{opt} approaches 0.5 for large R, and thus $\mathcal{S}^{(2)}_{\text{eq}}$ approaches $\mathcal{S}^{(2)}_{\text{opt}}$. The application of random weights, on the other hand, still strongly diminishes the expected skill for all r and R. Finally, note that the application of simplistic weights derived from Eq. (9) rather than Eq. (19) leads to a massive increase of the MSE with respect to $\mathcal{S}^{(2)}_{\text{eq}}$, if r is small and R is on the order of 1 or larger. This illustrates how essential it is that the effects of unpredictable noise be quantified and considered when determining optimum weights. The implications and relevance of these findings will be further discussed in section 4. We finish this section with four remarks.

REMARK 1: Note that here we have made the simplifying assumption that the noise terms ν_x, ν_{M1}, and ν_{M2} are samples from the same distribution with variance σ^2_ν.

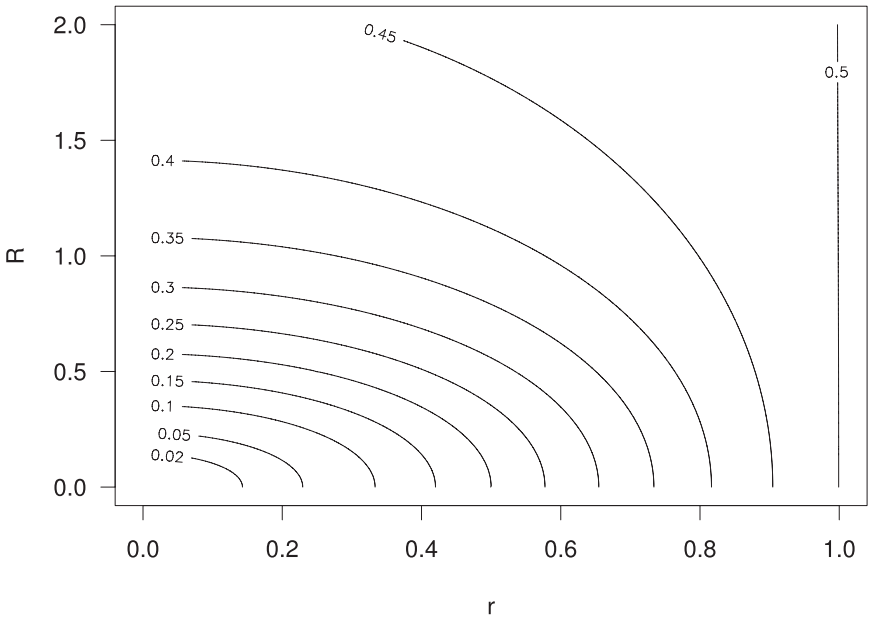

FIG. 7. Optimum weights w_{opt} under the influence of internal variability (''noise'') as obtained from Eq. (19). The contour lines show w_{opt} as a function of r (model error ratio) and R (relative noise ratio). Here, $R = 0$ corresponds to the case of negligible noise.

However, our conceptual framework can be easily generalized to differing internal variabilities $\sigma^2_{\nu_x}$, $\sigma^2_{\nu_{M1}}$, and $\sigma^2_{\nu_{M2}}$. Under these conditions, w_{opt} of Eq. (19) generalizes to

$$w_{\text{opt}} = \frac{r^2 + R_2^2}{1 + r^2 + R_1^2 + R_2^2}$$

$$\text{with} \quad R_1 = \frac{\sigma_{\nu_{M1}}}{\sigma_{M1}} \quad \text{and} \quad R_2 = \frac{\sigma_{\nu_{M2}}}{\sigma_{M1}}. \tag{20}$$

For large internal variabilities, w_{opt} then approaches $R_2^2/(R_1^2 + R_2^2)$ rather than 0.5 as above.

REMARK 2: What is the impact of noise if an infinite number of models are combined as in Eq. (12)? If m models are combined with equal weights, and if $m \to \infty$, the expected multimodel projection $\Delta y^{(m)}_{\text{eq}}$ and the expected MSE $\mathcal{S}^{(m)}_{\text{eq}}$ approach

$$\lim_{m\to\infty} \Delta y^{(m)}_{\text{eq}} = \lim_{m\to\infty}\left[\Delta\mu + \frac{1}{m}\sum_{i=1}^{m}(\epsilon_{M,i} + \nu_{M,i})\right] = \Delta\mu \tag{21}$$

and

$$\lim_{m\to\infty} \mathcal{S}^{(m)}_{\text{eq}} = \lim_{m\to\infty}\left[\sigma^2_\nu + \frac{1}{m^2}\sum_{i=1}^{m}(\sigma^2_{M,i} + \sigma^2_\nu)\right] = \sigma^2_\nu. \tag{22}$$

A multimodel can thus at best provide an unbiased estimate of the predictable signal $\Delta\mu$, but not the actual outcome Δx. This is plausible, because model combination can only cancel out the noise terms ν_M stemming from internal model variability; the unpredictable noise of the observations, ν_x, remains. Optimally and randomly weighted multimodels can be shown to approach the same limit. The only difference is that the optimally weighted multimodel converges more quickly than the equally weighted one, while the randomly weighted multimodel converges more slowly.

REMARK 3: What happens if two models have been run with several ensemble members stemming from different initial conditions? Let N_{M1} and N_{M2} be the ensemble sizes of $M1$ and $M2$; that is, N_{M1} (N_{M2}) independent samples of ν_{M1} (ν_{M2}) are available. Averaging the ensemble members of each model prior to model combination yields the following expected MSEs:

$$\mathcal{S}_{M1} = \sigma^2_\nu + \frac{\sigma^2_\nu}{N_{M1}} + \sigma^2_{M1}$$

$$\mathcal{S}_{M2} = \sigma^2_\nu + \frac{\sigma^2_\nu}{N_{M2}} + \sigma^2_{M2}. \tag{23}$$

Thus, in comparison to Eq. (2) the contribution of noise to the total projection uncertainty is strongly reduced, but the contribution of model error remains. This has implications on w_{opt}, which is now given by

$$w_{\text{opt}} = \frac{r^2 + \dfrac{R^2}{N_2}}{1 + r^2 + \left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)R^2}. \tag{24}$$

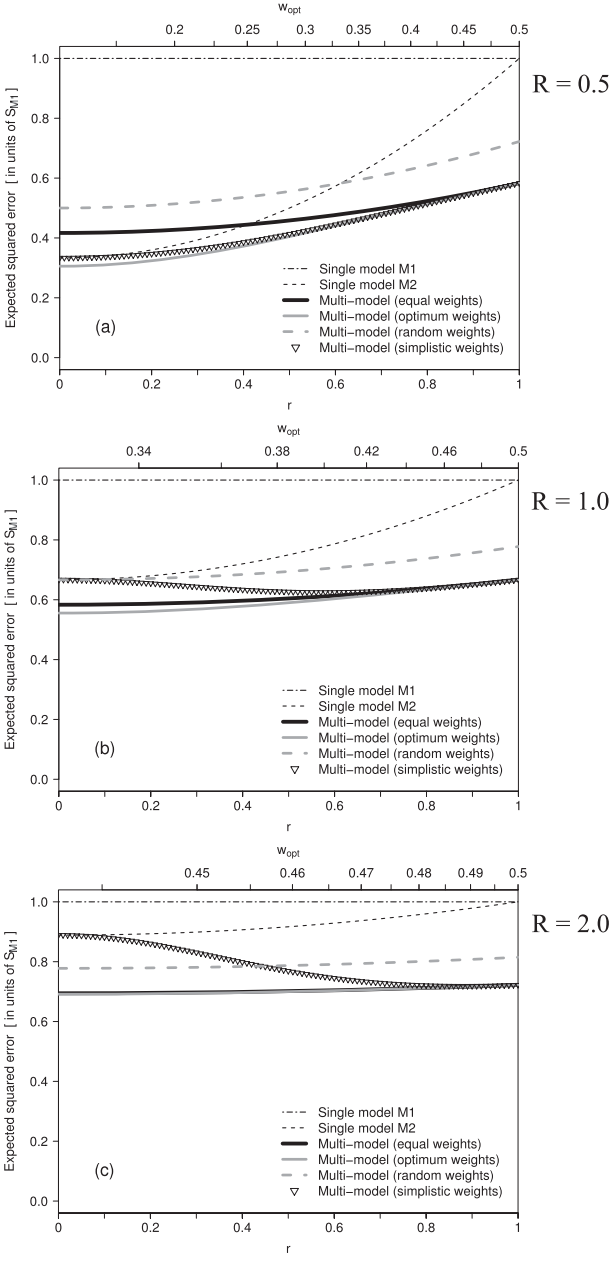

FIG. 8. As in Fig. 2, but for the assumptions applied in section 3c and Fig. 6: R = (a) 0.5, (b) 1, and (c) 2, with R being the relative noise ratio. Additionally, the expected squared errors of a weighted multimodel with the weights being wrongly determined from Eq. (9) rather than Eq. (19) are shown as triangles (simplistic weights). The top abscissa shows the true optimum weights w_{opt} as obtained from Eq. (19).

If N_{M1} and N_{M2} become very large, Eq. (24) approaches Eq. (9), that is, the value of w_{opt} for negligible noise. In other words, the availability of many ensemble members increases (reduces) the weight to be put on the better (weaker) of the two models—a pattern of behavior that has already been observed and discussed within the context of seasonal forecasting by Weigel et al. (2007).

REMARK 4: In this section we have assumed that the model errors are independent (i.e., that $j = 0$). However, also under the presence of noise, it is straightforward to generalize the projection context to the situation of $j > 0$, as in section 3b. In this case, the optimum multimodel mean would converge to $(\Delta\mu + \epsilon_j)$ rather than $\Delta\mu$, and the limit of the MSE would be $(\sigma_j^2 + \sigma_\nu^2)$ rather than σ_ν^2. However, even more than in section 3b, the presence of joint errors has only minor implications on the relative performance of the weighted versus unweighted multimodels and will, therefore, not be further discussed here.

4. Discussion

As all results presented above are based on a simple conceptual framework, they are as such only valid to the degree that the underlying assumptions hold. Most likely, our most unrealistic assumption is that the emission uncertainty has been entirely ignored. In principle, emission uncertainty could be conceptually included in Eq. (1) by adding an emission scenario error term s to Δy_M, such that $\Delta y_M = \Delta\mu + \epsilon_M + \nu_M + s$. However, in a multimodel ensemble, all contributing single models are typically subject to the same emission scenario assumptions and thus the same scenario error s. Therefore, the impacts of emission uncertainty on the *relative* performance of single models versus multimodels are probably very small. Rather, it is that the *absolute* projection accuracy would be heavily affected, in that both single-model and multimodel MSEs would be systematically offset by s^2 with respect to the errors discussed in section 3. This of course has severe consequences for our interpretation of climate projections in general, but does not affect our discussion on model weights. Apart from the issue of emission uncertainty, the conceptual framework involves many more simplifying assumptions, such as the omission of interaction terms between the different uncertainty sources, as mentioned by Déqué et al. (2007). However, we believe that by having explicitly considered the effects of skill difference (via r), model error dependence (via j), and noise (via R), the conceptual framework, despite its simplicity, is realistic enough to allow some generally valid conclusions.

The least surprising conclusion to be drawn is probably that equally weighted multimodel combination *on average* improves the reliability of climate projections—a conclusion that is fully consistent with what is known from many verification studies in weather and seasonal forecasting (e.g., Hagedorn et al. 2005; Palmer et al. 2004; Weigel et al. 2008b). Regardless of which values for r, j, and R are chosen, the expected MSE of the

multimodel is lower than the average MSE of the participating single models. Moreover, and again consistent with experience from shorter time scales, it has been shown that in principle model weighting can optimize the skill, if properly done. However, this requires an accurate knowledge of r—the key problem in the context of climate change.

Any estimate of r is to some degree necessarily based on the assessment of past and present model performance, and it needs to be extrapolated into the future to be of use for model weighting. In essence, the assumption must be made that r is stationary under a changing climate, which is problematic since other physical processes may become more relevant and dominant in the future than they are now (Knutti et al. 2010). This apprehension is backed by recent analyses of Christensen et al. (2008) and Buser et al. (2009), who have shown that systematic model errors are likely to change in a warming climate. However, even if r was stationary under a changing climate, we would still be confronted with the problem of how to determine a robust estimate of r on the basis of the available data. In contrast to, say, seasonal forecasting, the multidecadal time scale of the predictand strongly limits the number of independent verification samples that could be used to quantify r. This problem is aggravated by the fact that over the larger part of the past century, the anthropogenic climate change signal was relatively weak in comparison to the internal variability. Indeed, Kumar (2009) has shown that for small signal-to-noise ratios (on the order of 0.5) even 25 independent verification samples, a sample size which would actually be very large on multidecadal time scales, is hardly enough to obtain statistically robust skill estimates. Attempts have been made to try and circumvent this sampling issue by estimating model error uncertainties on the basis of other variables that can be verified more easily, such as systematic model biases (e.g., Giorgi and Mearns 2002). However, this leaves the question as to whether such ''alternative'' variables are representative for a model's ability to quantify multidecadal climate change signals. Whetton et al. (2007), Jun et al. (2008), Knutti et al. (2010), and other studies, for example, show that the correlations between present-day model performance (in terms of such alternative variables) and future changes are in fact weak, and within the context of monthly forecasting, Weigel et al. (2008a) have shown that those areas with the best bias characteristics are not necessarily those areas with the highest monthly prediction skill.

Given all of these fundamental problems in quantifying r, it seems that at the moment there is no consensus on how robust model weights can be derived in the sense of Eq. (9)—apart from one exception: If we *know* a priori that a given model $M1$ *cannot* provide a meaningful estimate of future climate while another model $M2$ *can* (e.g., because $M1$ is known to lack important key mechanisms that are indispensable to providing correct climate projections, while $M2$ has them included), then it may be justifiable to assume that $\sigma_{M2} \ll \sigma_{M1}$ and thus $r = 0$. For small R, this would then correspond to removing $M1$ entirely from the multimodel ensemble. In fact, some studies have found more consistent projections when eliminating poor models (e.g., Walsh et al. 2008; Perkins and Pitman 2009; Scherrer 2010). In the general sense, however, model weights bear a high risk of not being representative of the underlying uncertainties. In fact, we believe that the possibility of inadvertently assigning nearly random weights as analyzed in section 3 is not just an academic play of thoughts, but rather a realistic scenario.

Under such conditions, the weighted multimodel yields on average larger errors than if the models had been combined in an equally weighted fashion. In fact, unless r and R are very small, the potential loss in projection accuracy by applying unrepresentative weights is on average even larger than the potential gain in accuracy by optimum weighting. Also this aspect finds its equivalent in the context of seasonal forecasting. In an analysis of 2-m temperature forecasts stemming from 40 yr of hindcast data of two seasonal prediction systems, Weigel et al. (2008b) have shown that the equally weighted combination of these two models yields on average higher skill than any of the two single models alone, and that the skill can be further improved if optimum weights are applied (the optimum weights have thereby been defined grid-point wise). However, if the amount of independent training data is systematically reduced, the weight estimates become more uncertain and the average prediction skill drops (see Table 1 for skill values). In fact, if the weights are obtained from less than 20 yr of hindcast data, weighted multimodel forecasts are outperformed by the equally weighted ones. Particularly low skill is obtained for random weights, as can be seen in Table 1. However, note that even the randomly weighted multimodel still outperforms both single models.

In summary, our results suggest that, within the context of climate change, model combination with equal rather than performance-based weights may well be the safer and more transparent strategy to obtain optimum results. These arguments are further strengthened if the magnitude of the noise becomes comparable to or even larger than the model error uncertainty; that is, if $R \gtrsim 1$. Under these conditions, the optimum weights have been shown to approach 0.5. This means, for large R, equal weighting essentially *is* the optimum way to weight the models (see Figs. 8b and 8c), at least if the models to be

TABLE 1. Average global prediction skill of seasonal forecasts (June–August) of 2-m temperature with a lead time of 1 month, obtained from the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) database (Palmer et al. 2004) and verified against 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) data (Uppala et al. 2005) for the period 1960–2001. Skill is measured by the positively oriented ranked probability skill score (RPSS; Epstein 1969). The verification context is described in detail in Weigel et al. (2008b). Shown is the RPSS for ECMWF's "System 2" (*M*1), for the Met Office's "GloSea" (*M*2), and for multimodels (*MM*) constructed from *M*1 and *M*2 with (i) equal weights; (ii) with optimum weights obtained grid-point wise from 40, 20, and 10 yr of hindcast data by optimizing the ignorance score of Roulston and Smith (2002); and (iii) with random weights. Skill values are given in percent.

*M*1	*M*2	*MM* Equal *w*	*MM* Optimum *w* (40 yr)	*MM* Optimum *w* (20 yr)	*MM* Optimum *w* (10 yr)	*MM* Random *w*
−0.6	−4.5	5.6	7.2*	5.8	4.1	3.0

* Weigel et al. (2008b) obtain a higher value (9.4) because they include climatology (i.e., information from observation data) in the weighting process.

combined have comparable internal variability. Table 2 provides some rough estimates of R obtained from other studies found in the literature. While these studies are based on different methods and projection contexts, which can lead to considerably different estimates of R, they all show that R can indeed be large enough so that the application of model weights would only be of moderate use, even if the model error ratios were accurately known. This is particularly relevant if variables with low signal-to-noise ratios are considered (e.g., precipitation rather than temperature), if relatively small spatial and temporal aggregations are evaluated (e.g., a 10-yr average over central Europe rather than a 30-yr global average), if the lead times are comparatively short (e.g., 20 yr rather than 100 yr), and if no ensembles are available to sample the uncertainty in the initial conditions.

5. Conclusions

Multimodel combination is a pragmatic and well-accepted technique to estimate the range of uncertainties induced by model error and to improve the climate projections. The simplest way to construct a multimodel is to give one vote to each model, that is, to combine the models with equal weights. Since models differ in their quality and prediction skill, weighting the participating models according to their prior performance has been suggested, which is an approach that has been proven to be successful in weather and seasonal forecasting. In the present study, we have analyzed the prospects and risks of model weighting within the context of multidecadal climate change projections. It has been our aim to arrive at a conclusion as to whether or not the application of model weights can be recommended.

On shorter time scales, such an assessment can be carried out in the form of a statistically robust verification of the predictand of interest. For climate change projections, however, this is hardly possible due to the long time scales involved. Therefore, our study has been based on an idealized framework of climate change projections. This framework has been designed such that it allows us to assess, in generic terms, the effects of multimodel combination independently of the model error magnitudes, the degree of model error correlation, and the amount of unpredictable noise (internal variability). The key results, many of which are consistent with experience from seasonal forecasting, can be summarized as follows:

1) Equally weighted multimodels yield, on average, more accurate projections than do the participating single models alone, at least if the skill difference between the single models is not too large.
2) The projection errors can be further reduced by model weighting, at least in principle. The optimum weights are thereby not only a function of the single model error uncertainties, but also depend on the degree of model error correlation and the relative magnitude of the unpredictable noise. Neglecting the latter two aspects can lead to severely biased estimates of optimum weights. If model error correlation is neglected, the skill difference between the two models is underestimated; if internal variability is neglected, the skill difference is overestimated.
3) Evidence from several studies suggests that the task of finding robust and representative weights for climate models is certainly a difficult problem. This is due to (i) the inconveniently long time scales considered, which strongly limit the number of available verification samples; (ii) nonstationarities of model skill under a changing climate; and (iii) the lack of convincing alternative ways to accurately determine skill.
4) If model weights are applied that do not reflect the true model error uncertainties, then the weighted multimodel may have much lower skill than the unweighted one. In many cases, more information may actually be lost by inappropriate weighting than can potentially be gained by optimum weighting.
5) This asymmetry between potential loss due to inappropriate weights and potential gain due to optimum weights grows under the influence of unpredictable noise. In fact, if the noise is of comparable or even

TABLE 2. Selection of relative noise ratio values (R) as estimated from the literature. Note that different methodologies have been applied in the studies cited.

Reference	Evaluated region	Lead time/ projection range	Averaging period (yr)	Parameter	R
Hawkins and Sutton (2009)	Global	90 yr	20	T	0.1
Hawkins and Sutton (2009)	BI	50 yr	10	T	0.5
Hawkins and Sutton (2010)	Europe	50 yr	10	P	1.0
Cox and Stephenson (2007)	Global	90 yr	10	T	0.6
Cox and Stephenson (2007)	Global	50 yr	10	T	1.4
Solomon et al. (2007), Fig. 10.27	Global	100 yr	20	T	0.2
Solomon et al. (2007), Fig. 10.27	Global	100 yr	20	P	0.2
Murphy et al. (2004)	NE	Double CO_2	20	T	0.6
Murphy et al. (2004)	NE	Double CO_2	20	P	0.8

BI = British Isles, NE = northern extratropics, T = surface air temperature, P = precipitation.

larger magnitude than the model errors, then equal weighting essentially becomes the optimum way to construct a multimodel, at least if the models to be combined have similar internal variability. In practice, this is particularly relevant if variables with low signal-to-noise ratios are considered (e.g., precipitation rather than temperature), if high spatial and temporal detail is required, if the lead times are short, and if no ensemble members are available to sample the uncertainty of the initial conditions.

These results do not imply that the derivation of performance-based weights is impossible by principle. In fact, near-term (decadal) climate predictions, such as those planned for the Intergovernmental Panel on Climate Change's (IPCC) fifth assessment report (Meehl et al. 2009), may contribute significantly to this objective in that they can serve as a valuable test bed for assessing projection uncertainties and characterizing model performance. Moreover, also within the presented framework eliminating models from an ensemble can be justified if they are known to lack key mechanisms that are indispensable for meaningful climate projections. However, our results do imply that a decision to weight the climate models should be made with the greatest care. Unless there is a clear relation between what we observe and what we predict, the risk of reducing the projection accuracy by inappropriate weights appears to be higher than the prospect of improving it by optimum weights. Given the current difficulties in determining reliable weights, for many applications equal weighing may well be the safer and more transparent way to proceed.

Having said that, the construction of equally weighted multimodels is not trivial, either. In fact, many climate models share basic structural assumptions, process uncertainties, numerical schemes, and data sources, implying that with a simple ''each model one vote'' strategy truly equal weights cannot be accomplished. An even higher level of complexity is reached when climate projections are combined that stem from multiple GCM-driven regional climate models (RCMs). Very often in such a downscaled scenario context, some of the available RCMs have been driven by the same GCM, while others have been driven by different GCMs (e.g., Van der Linden and Mitchell 2009). Assigning one vote to each model chain may then result in some of the GCMs receiving more weight than others, depending on how many RCMs have been driven by the same GCM.

Given these problems and challenges, model combination with equal weights cannot be considered to be a final solution, either, but rather a starting point for further discussion and research.

Acknowledgments. This study was supported by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) Climate and by the ENSEMBLES project (EU FP6, Contract GOCE-CT-2003-505539). Helpful comments of Andreas Fischer are acknowledged.

REFERENCES

Allen, M. R., and W. J. Ingram, 2002: Constraints on future changes in climate and the hydrological cycle. *Nature,* **419,** 224–232.

Boé, J., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat. Geosci.,* **2,** 341–343, doi:10.1038/NGEO467.

Buizza, R., 1997: Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.,* **125,** 99–119.

Buser, C. M., H. R. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multimodel projection of climate: Bias assumptions and interannual variability. *Climate Dyn.,* **33,** 849–868, doi:10.1007/s00382-009-0588-6.

Christensen, J. H., F. Boberg, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change

projections of temperature and precipitation. *Geophys. Res. Lett.,* **35,** L20709, doi:10.1029/2008GL035694.

Cox, P., and D. Stephenson, 2007: A changing climate for prediction. *Science,* **317,** 207–208.

Déqué, M., and Coauthors, 2007: An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections. *Climatic Change,* **81,** 53–70.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multimodel ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus,* **57A,** 234–252.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Frame, D. J., N. E. Faull, M. M. Joshi, and M. R. Allen, 2007: Probabilistic climate forecasts and inductive problems. *Phil. Trans. Roy. Soc.,* **365A,** 1971–1992.

Giorgi, F., and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the ''reliability ensemble averaging'' (REA) method. *J. Climate,* **15,** 1141–1158.

——, and ——, 2003: Probability of regional climate change based on the reliability ensemble averaging (REA) method. *Geophys. Res. Lett.,* **30,** 1629, doi:10.1029/2003GL017130.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.,* **113,** D06104, doi:10.1029/2007JD008972.

Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Climate,* **19,** 4326–4343.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multimodel ensembles in seasonal forecasting. Part I: Basic concept. *Tellus,* **57A,** 219–233.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.,* **90,** 1095–1107.

——, and ——, 2010: The potential to narrow uncertainty of regional precipitation change. *Climate Dyn.,* in press, doi:10.1007/s00382-010-0810-6.

Jun, M., R. Knutti, and D. W. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.,* **103,** 934–947.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate,* **16,** 1684–1701.

Knutti, R., 2008: Should we believe model predictions of future climate change? *Philos. Trans. Roy. Soc.,* **366A,** 4647–4664.

——, R. Furrer, C. Tebaldi, and J. Cermak, 2010: Challenges in combining projections from multiple climate models. *J. Climate,* **23,** 2739–2758.

Kumar, A., 2009: Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Mon. Wea. Rev.,* **137,** 2622–2631.

Liniger, M. A., H. Mathis, C. Appenzeller, and F. J. Doblas-Reyes, 2007: Realistic greenhouse gas forcing and seasonal forecasts. *Geophys. Res. Lett.,* **34,** L04705, doi:10.1029/2006GL028335.

Lucas-Picher, P., D. Caya, R. de Elía, and R. Laprise, 2008: Investigation of regional climate models' internal variability with a ten-member ensemble of 10-year simulations over a large domain. *Climate Dyn.,* **31,** 927–940.

Meehl, G. A., and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Nakicenovic, N., and R. Swart, Eds., 2000: *Special Report on Emissions Scenarios. A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 599 pp.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Perkins, S. E., and A. J. Pitman, 2009: Do weak AR4 model bias projections of future climate change over Australia? *Climatic Change,* **93,** 527–558.

Popper, K. R., 1959: The propensity interpretation of probability. *Brit. J. Philos. Sci.,* **10,** 25–42.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Räisänen, J., 2007: How reliable are climate models? *Tellus,* **59A,** 2–29.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.,* **130,** 1792–1811.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.,* **36,** L13704, doi:10.1029/2009GL038082.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.,* **132,** 2732–2744.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change,* **81,** 247–264.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Scherrer, S. C., 2010: Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming. *Int. J. Climatol.,* in press, doi:10.1002/joc.2170.

Smith, L. A., 2002: What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA,* **99,** 2487–2492.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Sciences Basis*. Cambridge University Press, 996 pp.

Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. Roy. Soc. London,* **365A,** 2145–2161.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: A unified framework for the combination of multimodel weather and climate predictions. *Tellus,* **57A,** 253–264.

Stott, P. A., S. F. B. Tett, G. S. Jones, M. R. Allen, J. F. B. Mitchell, and G. J. Jenkins, 2000: External control of 20th century temperature by natural and anthropogenic forcings. *Science,* **290,** 2133–2137.

Tebaldi, C., and R. Knutti, 2007: The use of the multimodel ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.,* **365A,** 2053–2075.

——, R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Van der Linden, P., and J. F. B. Mitchell, Eds., 2009: ENSEMBLES: Climate change and its impacts at seasonal, decadal and centennial timescales. Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, 160 pp. [Available from Met Office Hadley Centre, FitzRoy Road, Exeter AU3 EX1 3PB, United Kingdom.]

Walsh, J. E., W. L. Chapman, V. Romanovsky, J. H. Christensen, and M. Stendel, 2008: Global climate model performance over Alaska and Greenland. *J. Climate,* **21,** 6156–6174.

Webster, M. D., 2003: Communicating climate change uncertainty to policy-makers and the public. *Climatic Change,* **61,** 1–8.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: Generalization of the discrete Brier and ranked probability skill scores for weighted multimodel ensemble forecasts. *Mon. Wea. Rev.,* **135,** 2778–2785.

——, D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller, 2008a: Probabilistic verification of monthly temperature forecasts. *Mon. Wea. Rev.,* **136,** 5162–5182.

——, M. A. Liniger, and C. Appenzeller, 2008b: Can multimodel combination really enhance the prediction skill of ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **134,** 241–260.

——, ——, and ——, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Mon. Wea. Rev.,* **137,** 1460–1479.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady, 2007: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophys. Res. Lett.,* **34,** L14701, doi:10.1029/2007GL030025.