Multimodels on seasonal and multidecadal time-scales

Andreas P. Weigel

Federal Office of Meteorology and Climatology MeteoSwiss
Zurich, Switzerland
andreas.weigel@meteoswiss.ch

ABSTRACT

Multimodel combination has become an accepted technique to improve the reliability of weather and climate projections on all time-scales. The underlying mechanism is that of a systematic widening of ensemble spread which often leads to a reduction of overconfidence and hence an improvement in prediction skill. If enough training data (e.g. reforecasts) are available, multimodel skill can be further enhanced by applying performance-based model weights. Remaining reliability deficits can, at least in principle, be accounted for by statistical postprocessing. On the long time-scales of climate change, however, the lack of verification data implies that neither the effects of model weighting, nor the realism of the underlying paradigms of ensemble interpretation, can be objectively judged. Any uncertainty estimate obtained for this time-scale is therefore necessarily conditional on the models available as well as on prior assumptions concerning the credibility and statistical properties of the participating single models. These limitations indicate that there is a clear need for more systematic approaches to estimate model uncertainty, particularly on the long time-scale of climate change.

1 Introduction

Multimodel combination is a pragmatic and well-accepted technique to estimate the range of uncertainties induced by model error. The success of multimodels in improving the reliability of weather and climate projections has been demonstrated in numerous studies (e.g. Krishnamurti et al., 1999; Palmer and Co-authors, 2004; Weigel et al., 2008). In the following, some conceptual issues with respect to the construction, interpretation, potential and limitations of multimodel ensembles are discussed. This is first done from the perspective of weather forecasts and seasonal predictions (Section 2), and then from the perspective of long-range climate projections (Section 3). Conclusions are given in Section 4.

2 Multimodels in weather and seasonal forecasting

On the short time-scales of weather and seasonal forecasting, prediction skill of a model can be systematically assessed by verification, i.e. by comparing past forecasts, or reforecasts, with corresponding observations by appropriate skill metrics. Similarly, also the effects of multimodel combination, and the strengths and weaknesses of different combination methods, can be systematically assessed by verification. In this section, the following three questions are discussed: Why do multimodels improve skill (Section 2.1)? What is the conceptual difference between skill gain due to multimodel combination and skill gain due to recalibration (Section 2.2)? And, can the skill of multimodels be further enhanced by assigning skill-based weights to the participating single models (Section 2.3)?

2.1 Why do multimodels improve skill?

In the case of deterministic forecasts, it is straightforward to demonstrate the effect of model averaging on the expected prediction error (e.g. Annan and Hargreaves, 2011): Let m_1, m_2, ..., m_n be (deterministic) forecasts obtained from n models, let $M = \frac{1}{n}\sum_{i=1}^{n} m_i$ be the multimodel mean, and let x be the verifying observation. The expected mean squared error (MSE) of a single model forecast can then be formulated as:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(m_i - x)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(m_i - M) - (x - M)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(m_i - M)^2 - \frac{2}{n}\sum_{i=1}^{n}(m_i - M)\cdot(x - M) + (x - M)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(m_i - M)^2 + (x - M)^2 \quad \geq \quad (x - M)^2 \qquad (1)
\end{aligned}$$

That is, the MSE of the multimodel average, $(x-M)^2$, is always lower than (or equal) the average MSE of the single model forecasts. However, it can also be shown that there is always at least one model that is equal or better than the model average. Yet, in practice it is usually not possible to judge a priori which model that would be at a given time. Indeed, the individual flaws and strengths of models typically vary with forecasting context (location, predictand, initialization time, etc.), so that in the long run, i.e. averaged over a sufficient number of grid-points and forecast realizations, multimodel approaches usually outperform any single-model strategy (Hagedorn et al., 2005).

From a probabilistic perspective, i.e. if several single model ensemble (SME) forecasts are pooled together to a multimodel ensemble (MME) and are verified with a probabilistic skill metric, the picture changes in that situations can be found where even a model that consistently performs better than the other models over the whole range of prediction contexts may be outperformed by a multimodel (e.g. Doblas-Reyes et al., 2005; Weigel et al., 2008). In the following, this apparent paradox is explained and resolved with the help of a simple conceptual model of seasonal predictability (Weigel et al., 2009). Note that, despite the seasonal focus of the following discussion, the same line of argumentation holds for other time-scales, such as weather forecasting.

Consider a set of observations x (e.g. seasonal averages of surface temperature at a given location). Assume that each observation can be formulated as the sum of a model-predictable signal μ and an unpredictable noise term ε_x, that is

$$x = \mu + \varepsilon_x \quad . \qquad (2)$$

μ can be thought of as the expected atmospheric response to slowly varying and predictable boundary conditions such as anomalies in sea-surface temperature, while ε_x represents the chaotic and unpredictable components of the observed dynamical system. x, μ and ε_x are assumed to have zero mean, i.e. anomalies are considered rather than absolute values. Let $\sigma^2_{\varepsilon_x}$ be the unpredictable internal variability, i.e. the variance of the (hypothetical) distribution of possible outcomes, given the predictable signal μ. This situation is illustrated in Figs. 1a and b: the presence of a given predictable signal μ 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 ensemble forecast $\mathbf{f} = (f_1, f_2, ..., f_M)$ with M ensemble members 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

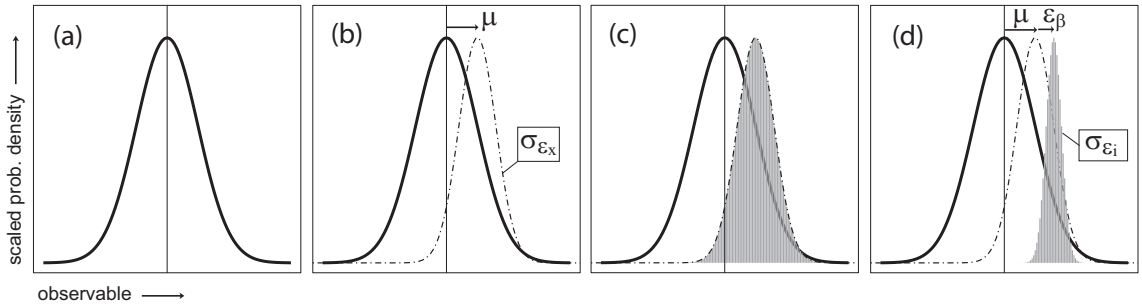

Figure 1: Illustration of reliable and unreliable forecasts (adapted from Weigel et al., 2009). Consider a climatology of observed outcomes (a). Under the influence of anomalies in relevant and predictable boundary conditions (e.g. SST in the context of seasonal forecasting, or a predictable flow regime in weather forecasting), the distribution of possible outcomes is shifted and sharpened w.r.t. climatology (b). The expectation of this constrained distribution is the potentially predictable signal μ, and its standard deviation is σ_{ε_x}. A reliable ensemble (c) would fully sample this distribution of possible outcomes. An unreliable ensemble with ensemble spread $\sigma_{\varepsilon_i} \neq \sigma_{\varepsilon_x}$ does not appropriately sample this distribution (d), and the ensemble mean may differ from μ by an error shift ε_β. Note that the probability densities are scaled differently here for illustrative purposes.

each other. This implies that, for a given predictable signal μ, each forecast member f_i represents an equally likely random sample from the distribution of possible observable states. A reliable ensemble forecast therefore has the following structure:

$$f_i = \mu + \varepsilon_i \tag{3}$$

with $\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon_x}$. This is illustrated in Fig. 1c. The ensemble mean is then an unbiased estimator of the predictable signal μ, and the ensemble spread estimates the uncertainty of the true outcome.

For real ensemble prediction systems, however, the expected ensemble means are not necessarily identical with the predictable signals μ. In fact, ensemble forecasts are often seen to be overconfident, meaning that the ensemble spread is too narrow while being centered at the wrong value. This can be considered in the conceptual model of Eq. 3 by adding an additional scalar error term ε_β - rather like the idea of model error which affects all ensemble members equally:

$$f_i = \mu + \varepsilon_\beta + \varepsilon_i \tag{4}$$

with $\sigma^2_{\varepsilon_i} < \sigma^2_{\varepsilon_x}$. This is illustrated in Fig. 1d. 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 any more an unbiased estimator of the predictable signal, and the forecasts are unreliable. Such overconfidence is penalized by probabilistic skill metrics and implies lower skill scores than if the forecasts were reliable.

Now assume that an MME is constructed by combining the output of several (overconfident) SMEs stemming from different models. If all models see the same predictable signal μ, and if the model errors ε_β are independent, the combination leads to a widening of the ensemble spread, a reduction in the error of the ensemble mean (the ε_β-terms cancel out), and thus a reduction of overconfidence and an increase of skill. This is illustrated in Fig. 2, where a synthetic generator of forecast-observation pairs based on Eqs. 2-4 (details in Weigel et al., 2009) has been applied to assess the effect of model-combination on skill. As can be seen, multimodel combination reduces overconfidence and improves

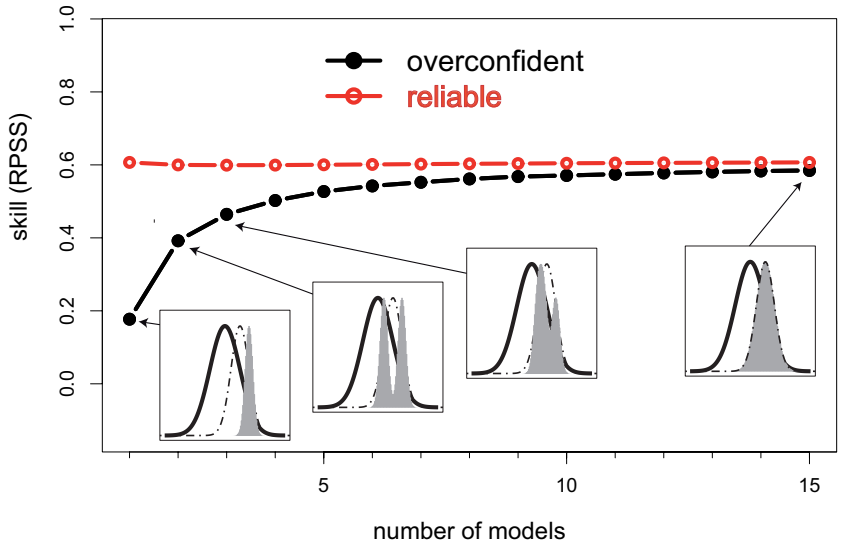

Figure 2: Expected skill of multimodel ensemble forecasts as a function of the number of participating single model ensembles (adapted from Weigel et al., 2008). The red line indicates well-calibrated reliable ensembles and the black line represents highly overconfident ensembles. The ensembles have been generated from synthetic toy model simulations. It can be seen that only in the latter case does model-combination truly enhance prediction skill, because multimodel combination of overconfident single model ensembles widens the spread. The underlying 'mechanics' of multimodel combination is illustrated by the four small panels at the bottom of the plot: The combination of more and more overconfident single model ensembles (shown as grey shading) successively widens the ensemble spread and reduces the ensemble overconfidence until eventually the entire predictable signal is correctly sampled and the forecasts are reliable.

skill beyond the skill of the best participating SME. Under these conditions, even the addition of a poorly performing model can improve the skill of the MME, but only if the poor performance of the SME is due to overconfidence and not lack of predictable signal. This direct link between overconfidence and the success of multimodel combination has also been identified with real seasonal forecasts (Weigel et al., 2008).

2.2 Multimodels versus recalibration

Having seen that the widening of ensemble spread and thus the reduction of overconfidence is a key mechanism to explain the success of multimodel combination, the question arises as to whether a similar effect could be achieved in a cheaper way by a recalibration-based inflation of ensemble spread? In its simplest configuration, such a recalibration strategy could for example consist of multiplying the ensemble mean with a scaling factor r, and the ensemble spread with a scaling factor s (e.g. Doblas-Reyes et al., 2005; Weigel et al., 2009). If applied to the conceptual model of Eq. 4, a recalibrated forecast f_i^{RC} would then be given by:

$$f_i^{RC} = r\left(\mu + \varepsilon_\beta\right) + s\varepsilon_i \quad . \tag{5}$$

Optimum values of r and s can be determined from reforecasts. Note that here it is assumed that systematic biases have already been removed a priori. Weigel et al. (2009) have demonstrated that such an approach can indeed strongly enhance the reliability and thus the skill of (seasonal) ensemble forecasts. However, since r is in most cases smaller than 1 due to the predominating forecast overconfidence (implying that the ensemble mean is shifted towards the climatological mean), such a recalibration scheme

implicitly destroys a part of the predictable signal μ and thus a part of the predictable variance. In contrast to that, MMEs may in the ideal case, that is if all participating single models "see" the same signal μ and have independent model error terms ε_β, become reliable by canceling out the ε_β-terms while retaining the predictable signal μ. This leads to an improvement of the skill attribute of resolution. In principle, multimodel combination therefore has the potential to yield superior results, particularly since the effectiveness of most recalibration schemes strongly depends on the length of the reforecast data record available and on distributional assumptions. In reality, however, only of a finite number of SMEs is usually available ("ensembles of opportunity", see Section 3.1), and the model errors ε_β are often highly correlated and fail to cancel out, so that in many cases the skill of recalibrated SMEs is comparable to or even better than the skill of MMEs (Weigel et al., 2009). Several studies have indicated that optimum results may be obtained by a combination of recalibration and model combination (e.g. Stephenson et al., 2005), i.e. the two techniques should be treated as complementary rather than competing approaches.

2.3 Model weighting

So far, it has been tacitly assumed that the models to be combined "see" the same predictable signal μ, but in fact there may be major differences in how well individual models resolve the physical processes that are relevant for predictability. For instance, a seasonal prediction model which is not able to resolve ENSO events will necessarily fail to exploit the seasonal predictability arising from ENSO. In contrast to the discussions of Section 2.2, the addition of such a model to an ensemble of ENSO-resolving models would reduce skill. One option to avoid such skill degeneration is weighting the participating SMEs according to their prior performance. Many approaches of model weighting have been suggested in literature. They are typically based on a non-linear optimization of past forecasts with respect to a specific skill metrics, Bayesian approaches with climatology as a prior, or regression approaches (e.g. Rajagopalan et al., 2002; Raftery et al., 2005; Coelho et al., 2006; DelSole, 2007; Weigel et al., 2008). All these approaches have in common that they indeed can yield superior skill as compared to equal weighting, but only if enough training data are available to obtain robust weights. If the weights are not robust, more skill may actually be lost than could potentially be gained by model weighting. This is illustrated in Fig. 3. The plot is based on the analysis of seasonal 2-m temperature forecasts stemming from 40 yr of hindcast data of two ensemble prediction systems (details in Weigel et al., 2010). It can be seen that the equally weighted combination of these two models yields on average substantially higher skill than any of the two single models alone, and that skill can be further improved by model weighting. However, if the amount of independent training data is systematically reduced, the weight estimates become more uncertain and the average prediction skill drops. In fact, if the weights are obtained from less than 20 yr of hindcast data, weighted multimodel forecasts are in this example actually outperformed by the equally weighted ones. This issue will be discussed again in the context of long-range climate projections in Section 3.2.

3 Multimodels in long-term climate change projections

As in weather and seasonal climate forecasting, multimodels are also widely used in the context of multidecadal climate change projections to reduce overconfidence and enhance reliability of the projections. In fact, the climate projections and corresponding uncertainty estimates provided in IPCC (2007) heavily rely on multimodels. Fig. 4a, for example, shows the multimodel mean and one standard deviation uncertainty range for global temperature for the historic simulations and projections for three IPCC SRES scenarios. The key challenge associated with such uncertainty estimates is illustrated in Fig. 5: A probability distribution needs to be derived from a finite set of model projections. The uncertainty estimate obtained depends amongst others on three fundamental issues: Has a sufficient

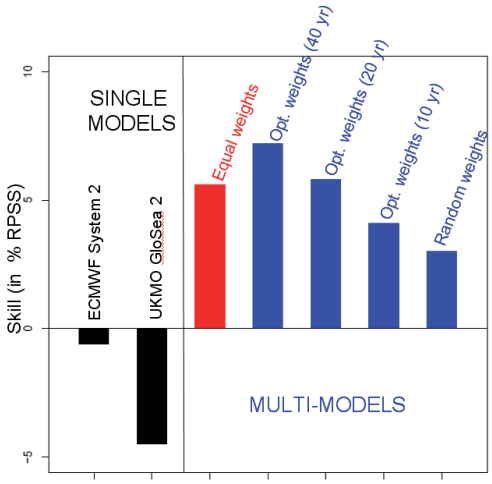

Figure 3: Average global prediction skill (in % RPSS) of seasonal forecasts (JJA, lead-time 1 month) of 2-m temperature, obtained from the DEMETER database and verified against ERA40 data for 1960-2001. Shown if the RPSS for ECMWF's System 2, for the Met Office's GloSea 2, and for multimodels constructed 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; and (iii) with random weights.

number of models been included in the MME to sample all relevant aspects of model uncertainty (Section 3.1)? Is each model equally credible, or should weights be assigned (Section 3.2)? And what is the underlying statistical framework guiding the interpretation of the ensemble members (Section 3.3)? Here only a few brief comments to these questions are provided. For a more in-depth discussion, the reader is referred to Knutti et al. (2010) and Weigel et al. (2010). Note that these aspects are equally relevant for the probabilistic interpretation of multimodel weather and seasonal forecasts. However, due to the existence of verification data on shorter time-scales, the realism and potential benefits of the assumptions made can be systematically assessed and judged, and combination strategies can be adjusted accordingly. Moreover, remaining reliability deficits can at least in principle be corrected a posteriori by statistical post-processing techniques such as recalibration. The key challenge in the context of multidecadal climate change projections arises from the fact that the choices and assumptions made for ensemble combination and interpretation cannot be validated in the sense of a robust verification so that any uncertainty estimate obtained is therefore inherently Bayesian (see also discussion in Section 3.2).

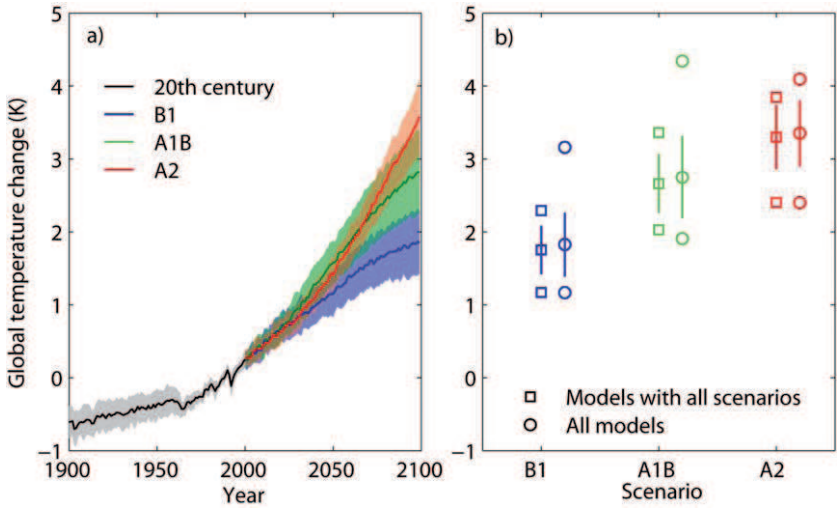

Figure 4: (a) Multi model mean and one standard deviation uncertainty ranges for global temperature (relative to the 1980-1999 average for each model) for the historic simulation and projections for three IPCC SRES scenarios. (b) Mean and one standard deviation ranges (lines) plus minimum maximum ranges (symbols) for the subset of models that have run for all three scenarios (squares) and for all models (circles). The model spread for the scenarios B1 and A1B depends strongly on what models have been included in the ensemble. From Knutti et al. (2010).

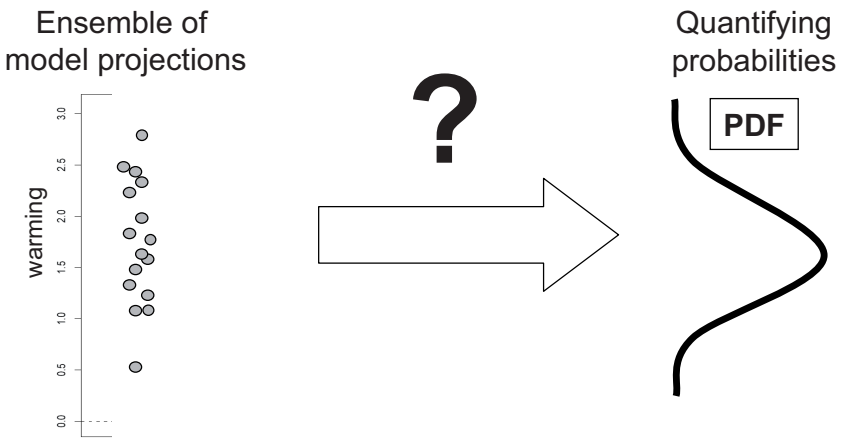

Figure 5: The key challenge in estimating model uncertainty on the basis of multimodel climate projections: A probability distribution needs to be derived from a finite set of model projections. This requires plenty of conceptual decisions and assumptions, such as: Has a sufficient number of models been included in the ensemble to sample all relevant aspects of model uncertainty? Is each model equally credible, or should weights be assigned? If yes, how? And what is the underlying statistical framework guiding the interpretation of the ensemble members?

3.1 Ensembles of opportunity

In practice, MMEs are usually not designed according to certain criteria (e.g. criteria ensuring that structural and parameter uncertainty are optimally sampled and that all models satisfy similar performance criteria), but rather are simply constructed on the basis of the model runs available. That is, it is more the number of climate modeling centers, their budgets, priorities and their modeling experience that determine the composition of a MME than physical reasoning. Such ensembles are therefore often referred to "ensembles of opportunity". The addition or removal of a model from an MME is often seen to have major consequences on the uncertainty estimates obtained, implying that the model uncertainty space is very likely to be undersampled. This is for example evident in Fig. 4b, which shows the distributions of the MMEs Fig. 4a is based upon, once only considering the subset of those models that have been run for all three emission scenarios (squares), and once considering all models available for each scenario (circles). Particularly for the scenarios B1 and A1B, the ensemble spread depends sensitively on which models have been included in the ensemble. This dependency of the results on the "arbitrariness" of the number of models available, together with a common lack of system in ensemble design, imposes severe challenges in the interpretation of multimodel climate projections and thus represents a major limitation of the multimodel approach.

3.2 Model weighting

The second important question for the interpretation of an MME of climate projections is whether each participating model should be equally weighted ("one model one vote"), or whether they should be weighted according to some criteria of prior performance. Given that, in weather and seasonal forecasting, performance-based weighting schemes have been successfully implemented and have been demonstrated to improve the average prediction skill, it may appear obvious that model weighting can also improve the projections in a climate change context and reduce the uncertainty range. However, as mentioned above, the two projection contexts are not directly comparable. In seasonal forecasting, for instance, usually 20 to 40 yr of hindcasts are available, which mimic real forecasting situations and can thus serve as a data basis for deriving optimum model weights. 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 derive 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, 2011), 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). Evidence from several studies suggests that the task of finding robust and representative weights for climate models is certainly a difficult problem. This is mainly due to (i) the inconveniently long time scales considered, which strongly limit the number of available verification samples; and (ii) nonstationarities of model skill under a changing climate. 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 (Weigel et al., 2010). This is illustrated in Fig. 6 which shows results obtained with a simple conceptual toy model of climate change projections described in Weigel et al. (2010). This toy model has been used to assess the effects of equal, optimum and inappropriate weighting in generic terms by controlled combination experiments of two models as a function of their relative skill. Note that this does not imply that the derivation of performance-based weights is impossible by principle. However, it does imply that a decision to weight climate models should be taken with great 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 long-range climate models, equal weighing may for many applications well be the safer and more transparent way to go.

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. This is for example evident in Fig.7, which has been published in Masson and Knutti (2011) and shows the results of a hierarchical cluster analysis of the performance characteristics of the CMIP3 models during the control period. Models stemming from the same institution or sharing versions of the same atmospheric model are in most cases grouped into the same cluster, indicating that they are more similar to each other than to the other models. 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. ENSEMBLES, 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.

3.3 Statistical interpretation

The third, and probably most fundamental aspect for obtaining reliable uncertainty estimates from multimodels is the underlying statistical framework that guides the probabilistic ensemble interpretation. Many approaches have been suggested in literature, and most of them can be assigned to one of two interpretational paradigms. The first paradigm ("truth plus error", e.g. Tebaldi et al., 2005; Buser et al.,

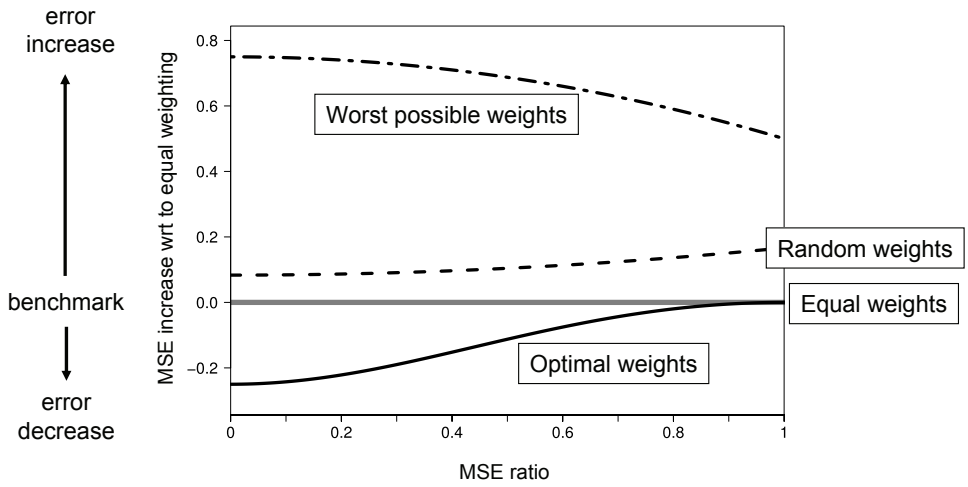

Figure 6: Increase/decrease of the expected mean squared error (MSE) of weighted averages of two single models (solid black: optimum weights; dot-dashed: worst possible weights; dashed: random weights) with respect to the benchmark of equal weighting. The results are plotted as a function of the MSE ratio of the two single models to be combined. The combination experiments are based on the conceptual model of Weigel et al. (2010).

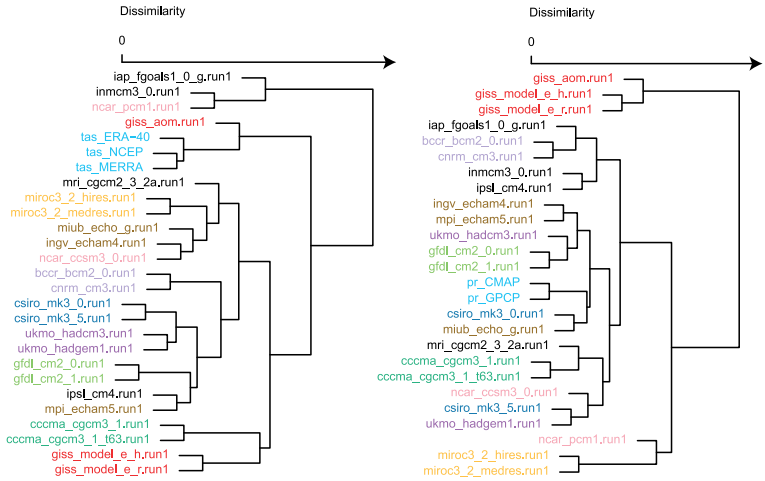

Figure 7: Hierarchical clustering of the CMIP3 models for (left) surface temperature and (right) precipitation in the model control state. Models from the same institution and models sharing versions of the same atmospheric model are shown in the same color. Observations also are marked by the same color. Models without obvious relationships are shown in black. From Masson and Knutti (2011).

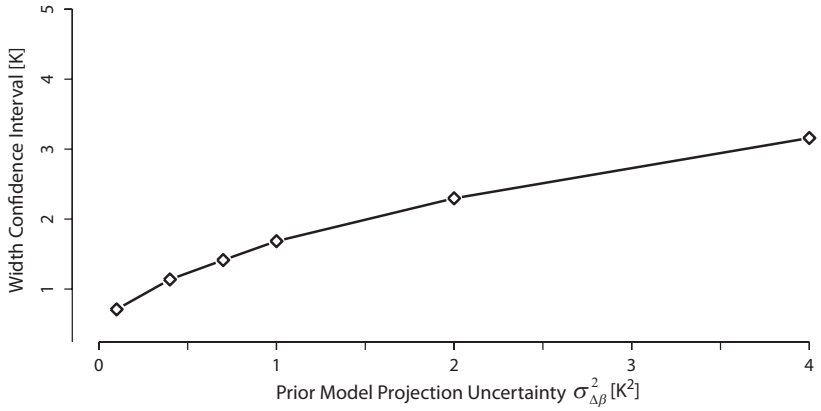

Figure 8: Uncertainty (95% confidence interval width) in the posterior estimate of the change signal of summer temperature in northeastern Switzerland. The uncertainty estimates have been obtained with the Bayesian algorithm of Buser et al. (2009) and are shown as a function of the prior choice of 'tolerable' model projection uncertainty. 6 GCM-RCM-model chains of the ENSEMBLES-project (ENSEMBLES, 2009) have been used. Scenario period is 2020-2049, reference period is 1980-2009. The figure has been adapted from Fischer et al. (2011).

2009) is based on the assumption that each ensemble member is sampled from a distribution centered around the truth. The other paradigm assumes that each one of the ensemble members considered is 'exchangeable' with the other members as well as with the real system (e.g. Murphy et al., 2007; Annan and Hargeaves, 2010). Again, due to the long time-scale considered, it is difficult to judge which interpretation is more appropriate in a climate change context. This problem is further complicated by the fact that any probabilistic framework applied relies on an array of more or less subjective prior assumptions (discussed for example in Fischer et al., 2011). For instance, the Bayesian 'truth plus error' algorithm of Buser et al. (2009) requires that a prior is specified on the 'tolerable' range of projection errors. The choices made for this prior largely determine the posterior estimates of model uncertainty (see Fig. 8). This high dependency of model uncertainty estimates on the underlying statistical framework and prior assumptions raises the question as to whether it is possible at all at present to formulate reliable probabilistic climate change projections on the basis of a multimodel ensemble of opportunity.

4 Conclusions

Plenty of studies have shown that multimodels improve the skill of weather and climate predictions, both in a deterministic and a probabilistic context. Multimodels represent an effective ad-hoc method to obtain first-guess estimates of model uncertainty and to make the forecasts more reliable. In contrast to perturbed parameter or stochastic approaches, multimodels not only sample parameter and physical uncertainty, but also structural uncertainty and numerical uncertainty of the dynamical cores. Moreover, multimodels are "politically attractive" in that the information provided by different modeling centers can be jointly considered. The improvement of prediction skill by multimodels is relatively simple to understand, regardless which time-scale is considered: Multimodel combination usually widens ensemble spread, thus reducing overconfidence and enhancing reliability. This also explains why multimodels are often seen to even outperform the best participating single model. On the short time-scales of weather and seasonal predictions, forecasts often come along with a set of past forecasts or reforecasts that can be used for a systematic verification. With this, it is relatively straightforward to judge the success of multimodels, to optimize the combination method applied, to assign meaningful model weights, and to correct for remaining reliability deficits by statistical post-processing approaches. On

longer time-scales, however, the choices and assumptions made cannot be assessed by a robust verification. Estimates of model uncertainty thus become increasingly Bayesian, i.e. increasingly conditional on more or less subjective prior assumptions. At the moment, there are no convincing concepts to derive probabilistically meaningful model weights for long-range climate models, nor is there a consensus on how quantitative estimates of projection uncertainty should be derived. Particularly the last aspect is aggravated by the fact that multimodels typically represent ensembles of opportunity, i.e. they are not constructed in a systematic way with a clear underlying probabilistic concept but are put together on the basis of what is available. This highlights the need for more systematic approaches to estimate model uncertainty, approaches which should be based on realistic assumptions and principles.

Acknowledgments

This work has been supported by Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) Climate.

References

Annan, J. D. and J. C. Hargeaves, 2010: Reliability of the CMIP3 ensemble. *Gephys. Res. Let.*, **37**, L02703, doi:10.1029/2009GL041994.

Annan, J. D. and J. C. Hargreaves, 2011: Understanding the CMIP3 multimodel ensemble. *J. Clim. In press.*.

Buser, C. M., H. R. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Clim. Dyn.*, doi:10.1007/s00382-009-0588-6.

Coelho, C. A. S., D. B. Stephenson, F. J. Doblas-Reyes, M. Balmaseda, A. Guetter, and G. J. van Oldenborgh, 2006: A bayesian approach for multi-model downscaling: seasonal forecasting of regional rainfall and river flows in south america. *Met. App.*, **13**, 73–82.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Clim.*, **20**, 2810–2826.

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 A*, **57**, 234–252.

ENSEMBLES: 2009, *ENSEMBLES : Climate change and its impacts at seasonal, decadal and centennial timescales. Summary of research and results from the ENSEMBLES project.*, P. van der Linden and J. F. B. Mitchell (eds.). Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. 160 pp.

Fischer, A. M., A. P. Weigel, C. M. Buser, R. Knutti, H. R. Künsch, M. A. Liniger, C. Schär, and C. Appenzeller, 2011: Climate change projections for switzerland based on a bayesian multi-model approach. *Int. J. Clim.* accepted.

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

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

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 A*, **57**, 219–233.

IPCC: 2007, *Climate change 2007: the physical science basis*, S. Salomon and D. Qin and M. Manning and M. Marquis and K. Averyt and M. M. B. Tignor and H. L. Miller Jr. and Z. Chen, eds. Cambridge University Press.

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. Am. Stat. Assoc.*, **103**, 934–947.

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

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550.

Masson, D. and R. Knutti, 2011: Climate model geneology. *Geophys. Res. Let.*, **38**, L08703, doi:10.1029/2011GL046864,.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Phil. Trans. Royal. Soc. A*, **365**, 1993–2028.

Palmer, T. N. and Co-authors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872.

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. Let.*, **36**, L13704, doi:10.1029/2009GL038082.

Scherrer, S. C., 2011: Present-day interannual variability of surface climate in CMIP3 models and its relation to the amplitude of future warming. *Int. J. Clim.*, **31**, 1518–1529.

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

Tebaldi, C., 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. Clim.*, **18**, 1524–1540.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Clim.*, **23**, 4175–4191.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of ensemble forecasts? *Quart. J. Roy. Met. Soc.*, **134**, 241–260.

— 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. Let.*, **34**, L14701, doi:10.1029/2007GL030025.