Clim Dyn
DOI 10.1007/s00382-014-2445-5

Can metric-based approaches really improve multi-model climate projections? The case of summer temperature change in France

Julien Boé · Laurent Terray

Received: 1 August 2014 / Accepted: 8 December 2014

Abstract The multi-model ensemble mean is generally used as a default approach to estimate climate change signals, based on the implicit hypothesis that all models provide equally credible projections. As this hypothesis is unlikely to be true, it is in theory possible to obtain more realistic projections by giving more weight to more realistic models according to a relevant metric, if such a metric exists. This alternative approach however raises many methodological issues. In this study, a methodological framework based on a perfect model approach is described. It is intended to provide some useful elements of answer to these methodological issues. The basic idea is to take a random climate model and treat it as if it were the truth (or "synthetic observations"). Then, all the other members from the multi-model ensemble are used to derive thanks to a metric-based approach a posterior estimate of the future change, based on the synthetic observation of the metric. This posterior estimate can be compared to the synthetic observation of future change to evaluate the skill of the approach. This general framework is applied to future summer temperature change in France. A process-based metric, related to cloud-temperature interactions is tested, with different simple statistical methods to combine multiple model results (e.g. weighted average, model selection, regression.) Except in presence of large observational errors in the metric, metric-based methods using the metric related to cloud temperature interactions generally lead to large reductions of errors compared to the ensemble mean, but the sensitivity to methodological choices is important.

Keywords Climate change · France · Metrics · Multi-model · Uncertainties

1 Introduction

Ensemble approaches for climate projections have become ubiquitous. Because of large model-to-model variations and, generally, lack of rationale for the choice of a particular climate model against others, it is widely accepted that future climate change and its impacts should not be estimated based on a single climate model (Wilby et al. 2004). If it is somewhat obvious that an ensemble of models should provide more information than a single one, it is not clear how to use that information optimally in practice.

Most studies to date, as well as the Intergovernmental Panel on Climate Change (IPCC) reports (e.g. Meehl et al. 2007), generally use the multi-model ensemble mean (MMEM) as the best estimate of climate change signals (Tebaldi and Knutti 2007). Yet, the rationale for the use of the MMEM in this context is not as clear as it may first seem (Knutti et al. 2010).

The MMEM approach lies, at least implicitly, on the idea that errors from different models are independent and tend to cancel-out (Tebaldi and Knutti 2007). In that context, the MMEM is supposed to converge to the truth as more models are added (the so-called "truth-centred" or "truth+error" paradigm). This view seems partially supported by the fact that in climate-related fields where verification is possible (e.g. seasonal prediction), the MMEM has been shown to generally outperform each (or at least most of) individual models (Hagedorn et al. 2005). Moreover, when looking at present-day observable features of the climate system (e.g. simulation of climatological features in the historical period), the MMEM tends to have a lower

J. Boé (✉) · L. Terray
URA1875 CNRS/CERFACS, Toulouse, France
e-mail: boe@cerfacs.fr

Published online: 15 January 2015

root mean square error (RMSE) than the vast majority of individual models (Lambert and Boer 2001; Gleckler et al. 2008). However, as put by Annan and Hargreaves (2011), the fact that, in average, the MMEM outperforms individual models can be explained by simple arithmetic: independently of the reference, the mean of the squared errors of individual members is necessarily greater than the squared error of the MMEM. Moreover, as pointed by Sanderson and Knutti (2012), models are generally tuned towards particular observations, which could tend to favor a truth-centred behavior on the instrumental period (Yokohata et al. 2012). None of these points implies that the truth-centred paradigm applies to future climate projections, or that the MMEM necessarily provides the best estimate of future climate changes. Maybe not surprisingly, multimodel ensembles often do not have the statistical properties one would expect if the "truth+error" paradigm were true (e.g. Jun et al. 2008).

As there is no reason to expect that the MMEM always provides the best estimate of climate change signals, non-uniform model weighting could be useful. Even in the present climate context, when the tuning process may favor the "truth+error" behavior (Sanderson and Knutti 2012), weighted averages have been shown to outperform the MMEM (e.g. Weigel et al. 2008). Even if the pool of available observations is the same for all modeling centers, their tuning strategy may differ (e.g. focus on the basic mean state versus long-term trends versus major modes of variability like the El-Nino-Southern-Oscillation; focus on specific regions versus global perspective). Moreover, large portions of code may be shared by different models, even from different groups. As a result, models are not necessarily perfectly randomly distributed about the truth even during the historical period.

Note that another paradigm has been proposed as an alternative to the truth+error one (Annan and Hargreaves 2010). The so-called "statistically indistinguishable ensemble" paradigm, more commonly used in the context of weather forecasting or seasonal prediction (e.g. Toth et al. 2003) stipulates that truth and models are simply drawn from the same distribution. As said in the previously mentioned paper, the truth is then not expected to be at, or even close to, the MMEM. The idea of non-uniform model weighting is theoretically compatible with this paradigm (Annan and Hargreaves 2010).

As noted by Sanderson and Knutti (2012), the truth+error and statistically indistinguishable ensemble paradigms may both be relevant to characterize multi-model ensembles of climate simulations. The tuning to a common target may favor a truth+error behavior in the present climate, while regarding future simulations, as tuning to present-day observations does not necessarily impact the future climate change signal, the indistinguishable paradigm may be more suitable.

If it is clear that moving away from the idea that all climate models are always equally credible, to favor some specific models considered as more realistic, may be useful, in practice major difficulties arise (Knutti et al. 2010). The main one probably lies in the assessment of the relative credibility of future climate projections from different models. Which metric(s) should be used to decide which models can be selected and which models can be rejected in a given context? By metric, it is meant here a quantity calculated in past or present-day climate simulations, and therefore that can be, at least theoretically, estimated with observations (in a loose sense), and that is expected to be informative on the capacity of the models to correctly simulate a given aspect of future climate change. If the metric from a given model is closer to observations, then the future projection by this model is supposed to be more realistic.

A high level of subjectivity exists in the choice of a metric. The most basic metrics are probably based on present-day climatologies. Such metrics have often been used (Giorgi and Mearns 2002; Tebaldi et al. 2005) probably based implicitly on the idea that a better simulation of present-day climatologies should be associated with more realistic simulated future changes. However, there is often little objective evidence that a good representation of present-day climatological conditions is sufficient (or even necessary) to capture correctly the climate response to anthropogenic forcing. A different approach lies in the use of process-based metrics, metrics that are strongly physically and statistically associated with the future changes of the variable of interest. If one finds a physical process responsible for a large part of the spread in the response of climate models to anthropogenic forcing that can be evaluated in the context of present-day climate variability as in Hall and Qu (2006) or Boé and Terray (2008), it may be possible to derive a potentially useful process-based metric (Collins et al. 2012). Once one has identified a potentially useful metric, a statistical method to combine the results of different models taking into account the information on the (supposed) realism of the models given by the metric has to be chosen.

Metric-based approaches to combine multiple models results therefore raise many important questions. How to assess correctly the respective interest of different potential metrics? What is the best statistical method to combine multiple models results based on a given metric? How to be sure in the end that the metric-based estimate of future climate change is not in fact less realistic than the MMEM? Unfortunately, it is not possible to answer those questions in the climate change context without observing future climate change, which is not very satisfying nor useful. In

Table 1 Climate models used in the study

Group	Model	Model	Model
BCC	**bcc-csm1-1**		
CCCma	**CanESM2**		
CNRM-CERFACS	**CNRM-CM5**		
CSIRO-QCCCE	**CSIRO-Mk3-6-0**		
INM	**inmcm4**		
IPSL	IPSL-CM5A-LR	**IPSL-CM5A-MR**	
LASG-CESS	**FGOALS-g2**		
MIROC	MIROC5	**MIROC-ESM**	
MOHC	**HadGEM2-CC**	HadGEM2-ES	
MRI	**MRI-CGCM3**		
NASA-GISS	**GISS-E2-R**		
NCAR	**CCSM4**		
NCC	**NorESM1-M**		
NOAA-GFDL	**GFDL-CM3**	GFDL-ESM2G	GFDL-ESM2M
NSF-DOE-NCAR	CESM1-BGC	**CESM1-CAM5**	CESM1-WACCM

Models in the reduced ensemble (RE) are in bold. The name of the modelling group is given in the first column

this paper, we describe a methodological approach based on a perfect model framework that could provide some interesting and useful elements of answer to all the questions previously mentioned. The basic idea is to take a random climate model in the ensemble and treat it as if it were the truth (we will call data from this model, in both past and future climate "synthetic observations") as done by Räisänen and Palmer (2001) in a somewhat different context, or Knutti et al. (2008), Räisänen et al. (2010) and Räisänen and Ylhäisi (2012). Then, the other members from the multi-model ensemble are used to derive thanks to a metric-based approach a posterior estimate of the climate change signal, based on the synthetic observations of the metrics. Finally, it is possible to compare the posterior estimate with the synthetic observation of future climate change to assess the overall interest of the approach. One cannot consider that the conclusions obtained thanks to this perfect model approach necessarily apply to real world, but as shown in this paper, this framework provides very useful insights.

The main objective of this paper is to describe and apply the perfect model framework to test different methodological issues associated with non-uniform model weighting and similar metric-based approaches. Is it really possible to outperform the MMEM? Are some statistical approaches better than others in that context? What is the impact of observational errors or internal variability? What is the impact of the ensemble size? The methodology presented here is general, but will be applied to the specific case of summer temperature change over France, for which previous work has suggested a potentially useful metric (Boé and Terray 2014).

Data and the general methodology followed in this study are described in Sect. 2. The different statistical methods used to combine projections from multiple climate models are described in Sect. 3. A first evaluation of the different methods within the perfect model framework is given in Sect. 4. The impact of model similarity on the previous results is assessed in Sect. 5 while the impact of observational errors are assessed in Sect. 6. In Sect 7, it is tested whether a simple climatological metric can provide useful results. The performance of metric-based methods in extreme configurations whereby all models underestimate or overestimate the truth is tested in Sect. 8. Finally, the importance of internal variability in the context of metric-based methods is tested in Sect. 9. Our conclusions are given and discussed in Sect. 10.

2 Data and general methodology

In this paper, climate simulations from 22 global climate models from the CMIP5 (Coupled Model Intercomparison Project Phase 5) archive are analyzed (Table 1). The historical simulations and the projections based on the RCP8.5 scenario (Meinshausen et al. 2011) are used. For some models and simulations, several members are available in the CMIP5 archive. Except where stated otherwise, analyses are done using one single member for each model (the first available member for all the variables studied). Throughout the paper, the present-day variability is characterized using the 1961–2000 period. Future changes correspond to differences between the means of 2080–2099 and 1961–1990 periods.

The potential interest of a process-based metric (met) to constrain the projection of summer temperature change over France is investigated in this study. This metric, described in Boé and Terray (2014) is associated with cloud-temperature interactions. It has been shown to play an important role in the uncertainties of future summer temperature change over Europe. Note that those results have been obtained using the ENSEMBLES regional climate models (RCMs; Déqué et al. 2012). The following analyses will show whether the interest of these metrics is not dependent of the modelling framework and whether they are also useful for the CMIP5 ensemble.

The metric met is the present-day interannual correlation between summer temperature and cloud cover. A large inter-model spread in met in ENSEMBLES RCMs has been found. Models with a strong anti-correlation in the present climate simulate larger decreases in cloud cover over France in the future climate and larger surface temperature changes. It suggests a similar behavior of cloud feedback in summer over France in the context of present-day interannual variability and future climate change.

All the quantities analysed in this paper (e.g. met, change in summer temperature) are scalar quantities. Temperature and cloud cover have been first averaged over France (longitude between −5° and 8°, latitude between 42° and 51° and land area fraction greater than 70 %). In ensemble mean, a warming of 6.3 K is simulated over France in summer, with an intermodel standard deviation of 2 K. More information on climate change in France as simulated by CMIP5 models can be found in Terray and Boé (2013). The intermodel correlation between met and future summer temperature change in France is −0.73. Note that the metric may not simply be a measure of local amplification, as its anti-correlation with global temperature change is also large (−0.65).

In this paper, a perfect model framework is followed to assess the interest of metric-based approaches to constrain multi-model projections. The question we try to answer is the following: given an ensemble of models providing both historical and future climate simulations, are we able to predict the response of a different model in the future climate when only its historical simulation is known? In practice, one model of the grand ensemble is selected to provide synthetic observations. The values of the metric in the present climate, and the future summer temperature change simulated by this model are considered as the observed truth (synthetic observations). The results of all the other models of the ensemble are then combined, based on the different methods described in the next section, including metric-based methods, in order to compute a posterior estimate of France summer temperature change. Finally, the accuracy of this prediction is evaluated by comparing the posterior estimate to the "truth", the synthetic observation of future summer temperature change over France. By considering each of the 22 models successively as the truth, 22 test cases can be built, corresponding to ensemble sizes of 21 models. Skill scores, as the mean absolute bias and the correlation between predicted values and the truths can be computed on those 22 cases.

The impact of the ensemble size N on most results in this paper is tested. In that case, each of the 22 models is successively chosen to provide synthetic observations as previously, then N models are randomly selected among the 21 remaining models (there are $\frac{21!}{N!(21-N)!}$ possible different ensembles with an ensemble size of N). In practice, 300,000 configurations (consisting in 1 model for synthetic observations and N models for prediction) are randomly built for each value of N, to reduce the computing time when N is small. The number 300,000 has been found to be largely sufficient to ensure that the results are robust.

3 Methods to combine multiple models results

The problem that we are interested in is the following: given an ensemble of models that perform more or less well regarding a present-day characteristic of the climate, that is supposed to be indicative of the realism of the models in the future climate, how to combine their results to obtain the best possible posterior estimate of future climate change? In this paper, a few simple approaches, described in this section, are used.

Let's p_i the value projected by the model i (e.g. in this paper, future summer temperature change averaged over France in 2080–2099 compared to the 1961–1990 period), m_i the value of met in the climate model i computed on the 1961–2000 period, M the true value of met (i.e. the value of met in the model that provides the synthetic observations).

Let's P the posterior estimate of future climate change, based on the combination of multiple models results. Finally, d_i is the distance defined a the standardized absolute difference between the value of met in the model i and the "true" value (synthetic observation). N is the total number of models in the predictive ensemble (inferior or equal to 21 throughout the paper).

3.1 Multi-model ensemble mean and random approaches

For the multi-model ensemble mean (MMEM), the posterior estimate is classically calculated as follows:

$$P = \frac{1}{N}\sum_{i=1}^{N} p_i \qquad (1)$$

In the random approach (RANDOM),

$$P = p_i \qquad (2)$$

where i is one model randomly chosen among the N models of the ensemble. It emulates the situation where only one

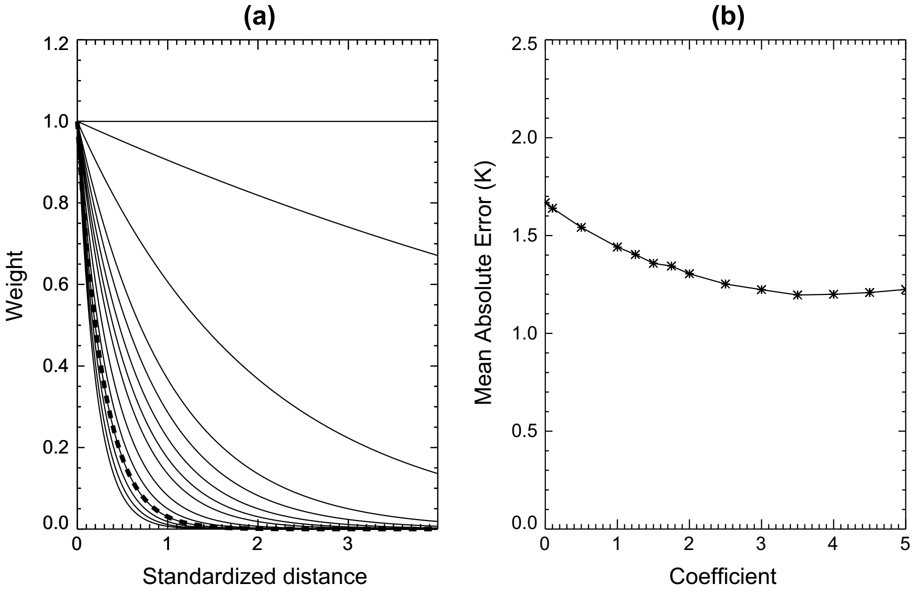


Fig. 1 **a** Weights as a function of the standardized distance for different values of the α parameter. See Sect. 3.2 for details. The value of α tested are [0,0.1,0.5,1,1.25,1.5,1.75,2,2.5 ,3,3.5,4,4.5,5]. The *dashed line* corresponds to $\alpha = 3.5$. **b** Mean absolute error obtained with WAVG in the perfect model framework as a function of the α coefficient for $N = 21$ (22 cases)

model is used to make a projection, without prior knowledge on the realism of this particular model.

Those are the two baseline approaches used in the paper, that have often been used in the literature, and against which the metric-based methods are tested.

3.2 Weighted average

In the weighted average framework (WAVG), the posterior estimate is defined as:

$$P = \frac{1}{\sum_{i=1}^{N} W_i(d_i)} \sum_{i=1}^{N} W_i(d_i) \cdot p_i \qquad (3)$$

The weights $W_i(d_i)$ associated with the metric met in the model i are function of the distance d_i between the value of the metric in the model i and the truth.

Many approaches can be used to define the weight $W_i(d_i)$ (e.g. bayesian model averaging with Bayes factor, Expectation-maximization algorithm, Min and Hense 2006) Simpler approaches, based on pre-defined weight functions are also commonly used (e.g. Lenderink 2010). Here, a parametric function is chosen for the weights to explore different behaviors.

$$W_i = \mathrm{e}^{(-\alpha \cdot d_i)} \qquad (4)$$

As shown in Fig. 1, the parameter α can be varied in order to obtain different shapes. For $\alpha = 0$, all the models have the same weight (i.e. it is equivalent to MMEM). The larger α is, the faster the weights decrease with the distance to the observed metric and for small values of α, the weights decrease almost linearly with the distance.

For different values of α, for $N = 21$, the WAVG approach is applied in our perfect model framework (Fig. 1). WAVG based on met leads to improvements compared to MMEM (i.e. $\alpha = 0$) for all the values of α greater than 0 tested here. The mean absolute error first decreases with increasing α and then very slightly increases. The minimum error is obtained for $\alpha \approx 3.5$. In our application, much more weight has to be given to the models the closest to the metrics.

3.3 The subset approach

A straightforward approach to combine multiple model results is simply to select the models that appear as especially realistic with regard to the metric involved, or to exclude the models that appear as especially unrealistic, and then to compute the ensemble mean for the selected models. This approach is named SUBSET in the following:

$$P = \frac{1}{m} \sum_{j=1}^{m \leq N} p_j \qquad (5)$$

with $m \leq N$, and p_j corresponding to the j–ith model when the models are sorted in increasing order according to their distance to the true value of the metric.

This approach requires to select *a priori* a subset of m models, which is generally quite arbitrary except when models are clearly clustered in two groups around particular value of the metrics. The perfect model framework is interesting here as it allows the impact of the choice of m to be tested, and its optimal value for a given application to be defined for a subsequent real-world application.

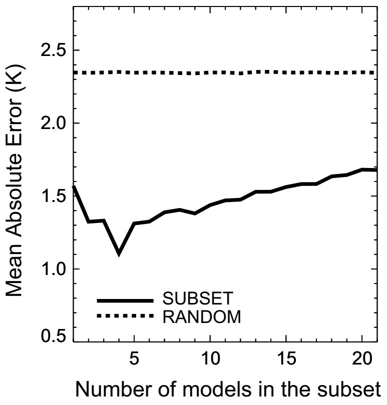


Fig. 2 Mean absolute error obtained for the SUBSET approach as a function of the number of models in the subset. Note that $N = 21$ corresponds to the MMEM. The results of the RANDOM approach in which a model is randomly chosen is also shown

Figure 2 shows the mean absolute error obtained with SUBSET, as a function of the size m of the selected subset. Note that the case $m = 21$ corresponds to the MMEM. In this test, $N = 21$ which corresponds to 22 possible cases, and the results are somewhat noisy. The results obtained with the subset approach are always better than MMEM. Even when only one model is chosen (the closest one relatively to the metric met), SUBSET leads to smaller errors than MMEM. It means that when an objective criterion based on an informative metric is used, even a single model could outperform the MMEM. The reduction of the absolute error compared to RANDOM which also uses a single model is major.

The best results are obtained for $m = 4$. Clearly in our case, it is better to keep only a few high rank models rather than to exclude a few low rank ones. This value of m is chosen for the subsequent analyses.

3.4 Regression

In the regression approach (REG), the linear regression equation between the future changes in the variable of interest and the present-day values of the metric is computed for the models, and then used to estimate future climate change based on the observed value of the metric (e.g. Boé et al. 2009; Bracegirdle and Stephenson 2012; Stegehuis et al. 2013).

The REG approach consists in solving the linear regression equation $p_i = a \cdot m_i + b + \epsilon$ to obtain a and b. Then the posterior estimate is computed as $P = a \cdot M + b$.

3.5 The "close enough" approaches

In the SUBSET approach, the m closest models are always selected whether only a single model or 20 models are in fact really close to the synthetic observation. This approach might not be necessarily optimal.

What is called here the "close enough" (CE) approach tries to overcome this difficulty. It consists in selecting only the models that are judged as sufficiently close to the true value of the metric. If no such model exists, then the value given by MMEM is used as posterior estimate (based on the idea that in that case, there is no reason to give more weight to particular models and all models can be seen as equivalent). The choice to use the MMEM when no close-enough models are found is somewhat arbitrary. The other metric-based approaches described in this section could also be used in that context to exploit the relationship between the metric and temperature change, even if no model close to the truth can be found in the ensemble. The REG approach is therefore also tested in that context (see the CE+REG approach bellow).

A criterion must be chosen to define what a close-enough model is. Even with a real perfect model, the simulated value of a metric would not be expected to be identical to the observed one. The observed value of a metric may indeed suffer from potential observational errors or uncertainties and its estimation is impacted by internal variability. In our perfect model framework, the definition of the acceptability of a model is based on a measure of internal variability. For real case applications, an estimation of the observational error could also be used.

Therefore for CE,

$$P = \frac{1}{n_g} \sum_{i=1}^{n_g \leq N} p_i \tag{6}$$

for the subset n_g of models that satisfy the conditions:

$$|m_i - M| \leq \beta \cdot \sigma(met) \tag{7}$$

If such models are found, the ensemble average on this subset of models is taken. If no model satisfies these conditions, then the MMEM is used.

$\sigma(met)$ is an estimate of the impact of internal variability on met. Because of the limited length of observational records, even for real world application, models generally would have to be used. For each of the models with more than five members on the historical period, we computed the single-model inter-member standard deviation of met and then we took the multi-model average.

The factor β controls here what is judged as "close enough". A large β will lead to the selection of many models, but not necessarily sufficiently similar to the observations. Conversely, a small β makes it particularly difficult to find close enough models. Empirical tests have been made to choose the value of β (Fig. 3). There is no clear optimal value of β, as the results are somewhat noisy, but it is clear that β should not be too large nor too small (Fig. 3b). We chose a value of 0.6 because it is close to the optimum and it leads to the selection of a small number of models n_g, compared to larger values of β (less than four in average,

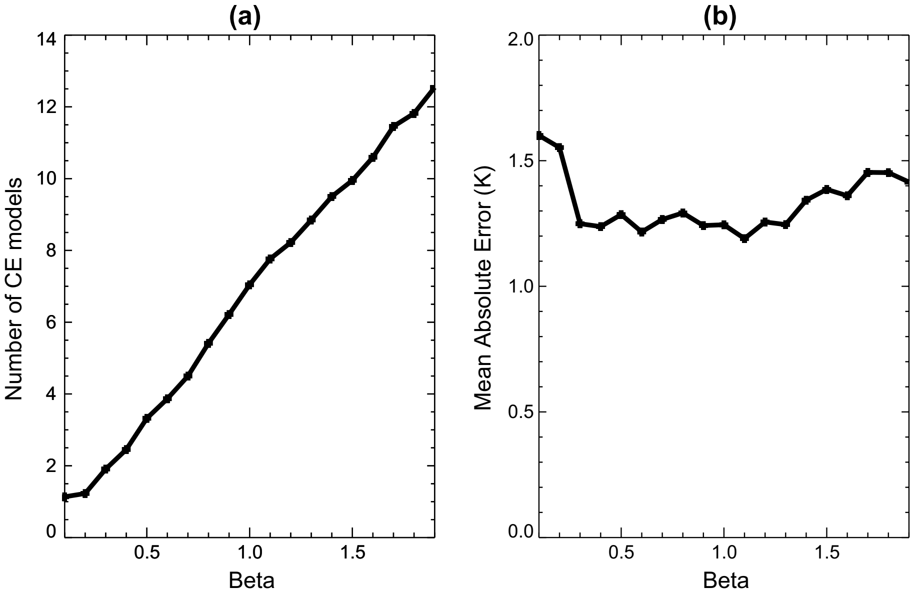


Fig. 3 **a** Mean number of "close enough" models found for different values of the β parameter. See Sect. 3.5 for details. **b** Mean absolute error obtained with CE in the perfect model framework as a function of the β coefficient for $N = 21$ (22 cases)

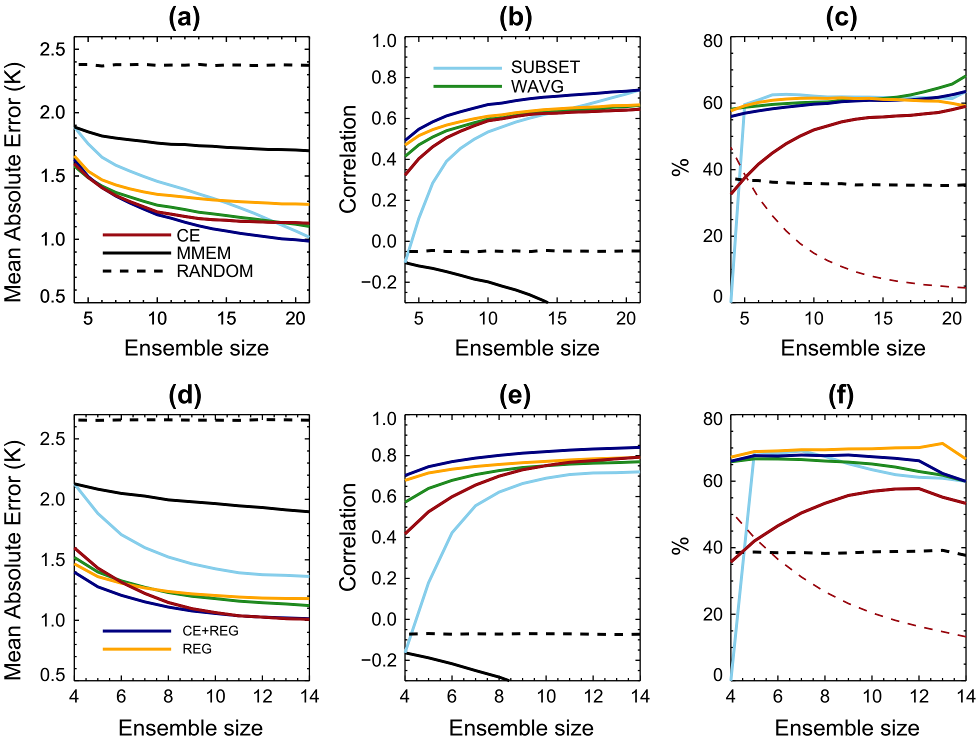


Fig. 4 Skill scores for the different methods to combine multiple model results assessed within the perfect model framework for different ensemble sizes, based on all the models given in Table 1. **a** Mean absolute error, **b** correlation between predicted values and truths, **c** percentage of cases for which the error obtained with the metric-based approach is smaller than the error obtained with MMEM. The *dotted red line* shows the percentage of cases for which the errors obtained with the CE approach and MMEM are identical i.e. the percentage of cases for which no close enough model is found. **d**, **e**, **f** same as (**a**), (**b**), (**c**) for the subset of models RE given in Table 1. The same color code is used in the six panels and is given in (**a**), (**b**) and (**d**)

Fig. 3a). This value of β is therefore expected to be more adapted to small ensemble sizes.

Finally, a variant of the CE approach named CE-REG is introduced. CE-REG is similar to the CE approach except that when no close enough model is found, the value estimated thanks to the REG approach is used rather than the MMEM. The rationale here is that even if no close enough model exists, the REG approach somewhat allows to approximate virtually such a model by exploiting the statistical relationship between the metric and the climate change signal.

4 Evaluation of the different methods and influence of the ensemble size

The skill of the different methods to combine multiple model results described in the previous section is now assessed in the perfect model framework described in Sect. 2, for different ensemble sizes N (Fig. 4a–c). Three skill scores are used: the mean absolute error (MAE) computed on the different test cases, the correlation between the posterior estimates and the true values (synthetic observations), and the percentage of cases for which the absolute error is smaller than the one obtained with MMEM.

First, the MAE for the different approaches and different sample sizes are compared. Not surprisingly, MMEM clearly outperforms RANDOM, with a reduction of 0.7 K of the MAE. Clearly, as widely acknowledged, it is not sound to use the results of a single model rather than MMEM, without a strong rationale. Interestingly, the MAE associated with MMEM is not very sensitive to the ensemble size (for $N > 4$). From $N = 4$ to $N = 21$, the absolute error just decreases by 0.1 K. Therefore, regarding the MAE and for our particular application, large ensemble sizes are not particularly useful when only the ensemble mean is considered.

All the metric-based methods lead to a substantial decrease in the MAE compared to MMEM, for all sample sizes (expect SUBSET for $N = 4$ which is equivalent by construction to MMEM). All the metric-based methods (except SUBSET) for the smallest ensemble size ($N = 4$) even slightly outperforms MMEM for the full ensemble. Generally, the added value of those methods compared to MMEM becomes increasingly clear as the ensemble size increases.

For the largest ensemble size ($N = 21$), the approaches leading to the smaller errors are SUBSET and CE+REG, for which the reduction of MAE compared to MMEM is 0.7 K. The gap in performance between the best metric-based approaches here and MMEM is therefore as large as the gap between MMEM and RANDOM, which clearly shows the potential interest of using metrics to constrain future climate change projections. The relative performances of the different metric-based methods vary with the ensemble size. SUBSET comparatively do not perform very well for small sample sizes. This approach is based on the selection of four models (the closest to the synthetic observations in terms of metric), and the selected models may be in fact far from the observed metric for small sample sizes. The CE approach, also based on the selection of a subset of models, better performs for small sample sizes than SUBSET because the selected models are necessarily close to the observation by construction. WAVG and REG better exploit the information from all the models in the ensemble, which probably explains why they also perform better than SUBSET with small sample sizes. Note that the skill of the different metric-based methods is not much different overall.

There are two different aspects to what is often simply called "improving multi-model climate projections": reducing the error in the posterior estimate or reducing the uncertainty range associated with the posterior estimate. The reduction in the mean absolute error with the metric-based approaches in the perfect model framework indicates improvements related to both aspects. In fact, some approaches described in this paper do not provide a natural "confidence interval" associated with a given posterior estimate (only MMEM, WAVG and REG can readily provide such an uncertainty range). The perfect model framework described in this study could be useful for real world-application to derive an "empirical" estimate of the uncertainty range associated with a given posterior estimate, e.g. the MAE or more classically the standard deviation of the error.

The analysis of the correlations between posterior estimates and synthetic observations of temperature change leads to very similar conclusions (Fig. 4b). Correlations as high as 0.75 are obtained for the best methods (here again CE+REG and SUBSET for large N). Note that the correlations obtained for MMEM are not really meaningful. Normally, a value of 0 would be expected, but in the perfect model approach, the value of P given by MMEM and the truth Pt are not independent, and even perfectly anti-correlated when $N = 21$: the larger the value of Pt is, the smaller is P estimated with MMEM, as it is computed on all the other models in the ensemble.

Figure 4c shows how frequently the different approaches outperform MMEM. RANDOM outperforms MMEM in roughly 35 % of the test cases, quite independently of N. For $N = 21$, CE outperforms MMEM roughly 60 % of the cases and gives the same results as MMEM by construction 5 % of the cases. Even with $N = 21$, a close enough model cannot always be found, which suggests that this approach can benefit from larger ensembles.

All the other metric-based methods, quite independently of the sample size, roughly outperform MMEM in 60 % of the cases. Given the large differences in skill noted previously for MAE between MMEM and metric-based methods, such limited differences are at first somewhat surprising. In fact, when the "truth" is close to the ensemble mean, which happens quite often given the distribution of projected summer temperature change in France, MMEM is naturally difficult to outperform. But when there are improvements compared to MMEM, they are often substantial and lead to large differences in MAE (see Sect. 8).

It is clear that the skill of the different methods (except SUBSET) generally becomes less and less impacted by the addition of models in the ensemble, but this effect is more or less pronounced depending on the method. For example visual inspection suggests that no additional model beyond 21 will allow the MAE to be reduced significantly for the REG approach while it seems that the performance of SUBSET could continue to improve as more models are added.

The analyses of this section demonstrate that metric-based methods can potentially lead to major improvements in multi-model climate projections. There are in general as large differences between these methods and MMEM than between MMEM and the use of a single random climate model. CE+REG is generally the best approach, independently of the sample size.

5 Impact of models similarity

The 22 models used for the previous analyses are not necessarily independent. In fact, some of them are very similar. For example, IPSL-CM5A-LR and IPSL-CM5A-MR just differ by the resolution, and therefore it is not expected that these two models lead to very different climate projections (except if some processes or parametrizations are crucially resolution-dependent).

If it is clear that model similarity may have an impact on the results described in the previous section, it is not necessarily easy to anticipate which one. For example, model similarity may make it too easy to find a "close enough" model in the ensemble in the perfect model framework. It may also impact negatively the REG approach, as in that case the errors associated with the regression equation may not be independent. Model similarity is also likely to impact the results obtained with MMEM and RANDOM in the perfect model framework. In any case, model similarity makes the results obtained in the perfect model framework less relevant to the real world and is therefore not desirable in our perfect model framework.

To test the impact of models similarity on the previous results, a subset of models is selected. As models from the same modelling group generally have more in common (Masson and Knutti 2011), only one model by modelling group has been selected. We acknowledge that while this procedure has the advantage to be simple, it is also somewhat arbitrary, as some models from different groups may eventually have similarities as important as two models from the same group (e.g. they may share the same ocean model). It is therefore highly unlikely that even in our reduced ensemble the models are totally independent. Methods using a measure of model similarity based on their results could also have been used (as in Knutti et al. 2013 or Caldwell et al. 2014) to reduce the ensemble. However, as these approaches require the definition of a relevant metric of similarity, they would also involve a certain level of subjectivity. Note that some NSF-DOE-NCAR models are in fact very similar to CCSM4 from NCAR (some of them only differ by the stratosphere or the inclusion of a biogeochemistry component). We choose to use CESM1-CAM5 for that group (and CCSM4 for NCAR), as the atmospheric models are substantially different: CAM5 for the former versus CAM4 for the latter, with substantial differences between the two versions of CAM (Meehl et al. 2013).

The subset of models chosen is given in Table 1. This subset is called the reduced ensemble (RE). The same analyses as in the previous section are now done with the RE subset. The maximum sample size is now 14 as 15 models have been selected.

A degradation of the skill of MMEM and RANDOM are noted in RE compared to the full ensemble (compare Fig. 4d to Fig. 4a). It may be due to the fact that using similar models in the predictive ensemble and as synthetic observation may artificially lead to smaller errors.

Concerning the metric-based methods, in general (except for SUBSET) no degradation of skill is found. In fact, results for the maximum ensemble sizes ($N = 14$ for RE and $N = 21$ for the full ensemble) are often even better or at least similar for RE despite the large difference in ensemble sizes. This is probably simply explained by the fact that if a model B is very similar to a model A, adding the model B to an ensemble that already includes the model A does not provide additional information. For a given sample size, results are better when the different models are less similar and therefore in RE. The relative performances of the different metric-based methods differ between RE and the full ensemble. For example, REG better performs in RE. It might be explained by the issue of error dependency in regression analysis associated with model similarity mentioned previously. Higher correlations and more frequent improvements compared to MMEM are also generally noted for the reduced ensemble compared to the full ensemble.

The results described in this section are consistent with the idea that in the context of multi-model climate projections, what matters ultimately is not the overall size of the ensemble but the effective number of independent models (Annan and Hargreaves 2011). The relevance of the perfect model approach followed in this paper to the real world lies on the hypothesis that any random climate model can be considered as truth. This hypothesis is necessarily less justified when the similarities among models are large. Therefore, it is expected that the results obtained with the RE subset give a better indication of the relative performance of the different methods in real world application and of the potential improvements associated with metric-based methods compared to the ensemble mean.

6 Impact of observational errors

The results described in Sects. 4 and 5 are encouraging for metric-based methods as they suggest these methods can lead to major improvements in the realism of multi-model projections. However, these results are obtained implicitly assuming no error in observations and therefore that the metric can be estimated with perfect accuracy. Obviously, it is seldom the case for real world applications. It is all the more true since in the context of metric-based methods, the term observation is often used in a loose sense: reanalyses data are not observations *stricto sensu* but they can be used when no better estimate is available. It is therefore important to assess to what extent the results of metric-based methods are sensitive to errors in observations, which is possible in the perfect model approach. To do so, an artificial error is added to the synthetic observation of the metric before applying the metric-based methods. The observational errors are expressed as a percentage of the ensemble mean of met (-0.60), and added to the true value of the metric. Errors between 25 % and +25 % are tested (which corresponds to ±0.15).

The results of the analysis are summarized in Fig. 5. The differences in sensitivity to observational errors between the different methods in terms of MAE are important. For relative errors roughly greater than 15 %, the CE approach is outperformed by MMEM, but the REG approach still outperforms MMEM for errors close to 25 %. Moreover the correlations between the posterior estimates and the synthetic observations for REG are very similar independently of the error, while it is not the case for CE. Note that some methods have been somewhat tuned (WAVG, SUBSET, CE, see Sect. 3) assuming no error in observations. It is possible that the tuning is not optimal in presence of observational errors and could be improved (for example the definition of “close enough” models might be relaxed when large errors are expected in observations), but it is not our objective to test this hypothesis here.

It is not surprising that metric-based methods are sensitive to observational errors, and that one must be careful about these issues. In our application even in presence of moderate errors, metric-based methods remain interesting but it may not be always the case. The perfect model approach is very useful before real-world applications in that context. If one can quantify the potential error-range in the observational estimates of the metrics, one can use an analysis such as the one described in this section in a perfect model framework to decide whether or not using a metric-based method remains interesting, and eventually to choose the most skillful statistical method in presence of such errors.

7 Basic climatological metric

In this paper, a process-based metric with significant statistical links with the change in the variable of interest, and which can be understood in terms of physical processes, have been tested. Finding such a metric is not necessarily possible for all applications and in any case it requires some additional work (Boé and Terray 2014). One can therefore wonder whether using such process-based metrics is really useful. In fact, simpler metrics have often been used in practice (Giorgi and Mearns 2002). It is for example sometimes assumed that the realism of the reproduction of climatological means in the present-day climate is somewhat informative about the realism of future climate projections. In this section, it is tested whether climatological temperature over France is a useful metric of future temperature change and how it compares with the process-based metrics previously introduced. Note that summer climatological temperature is weakly correlated to future summer temperature change in France (0.46, $p < 0.05$).

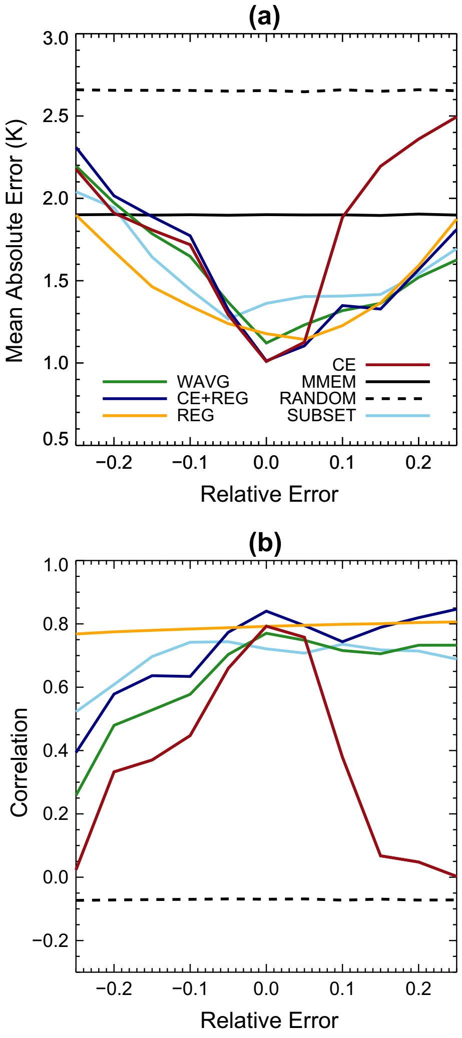


Fig. 5 Skill scores obtained for the subset of models RE described in Table 1 and $N = 14$ for different values of errors added to the synthetic observation of the metric. **a** Mean Absolute Error, **b** correlation between predicted values and truths

For the subset of models RE (see Table 1), we apply the different methods described in Sect. 3, using the

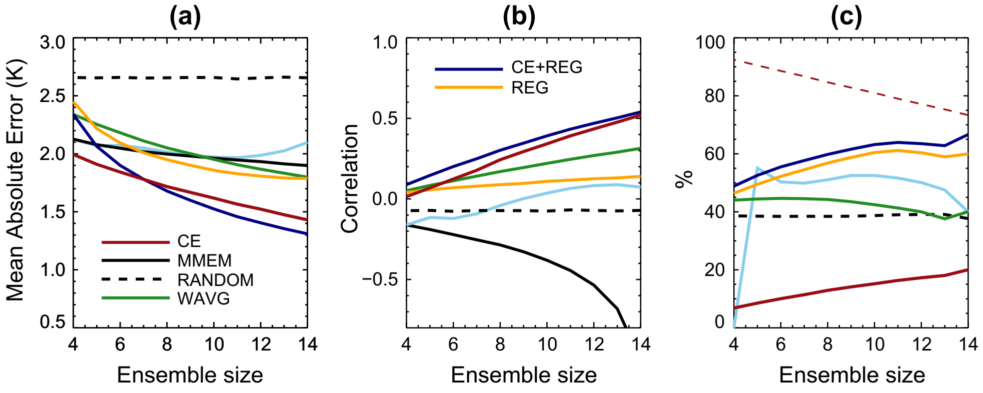


Fig. 6 Same as Fig. 4d–f except that only one metric is used, the summer climatological temperature over France on the 1961–2000 period

climatological temperature over France as a metric and using the same parameters as given in Sect. 3 (Fig. 6).

Climatological temperature as a metric is outperformed by met in virtually all configurations (Fig. 6), which was hoped for. Real improvements with climatological temperature as a metric compared to MMEM are only seen for the CE and CE+REG methods. Note that close-enough models are seldom found (Fig. 6c), but it is sufficient to lead to a rather large reduction of error compared to MMEM.

Using climatological temperature as a metric does lead to improvements compared to MMEM in some configurations when the ensemble size is not too small, but they are smaller that what is obtained with met and much more method-dependent. This result is highly specific to our particular study and cannot be generalized. However, it suggests that one has to be cautious when using metric-based methods. Strong rationale should support the use of a particular metric to obtain potential large reductions of errors. Without the demonstration of a strong link between a metric and the future climate change signal of interest, and/or, better, similar analyses as the ones described in this study in a perfect model framework, it is not necessarily useful, or even, potentially harmful, to use metric-based methods.

8 Extreme cases

When analyzing a multi-model ensemble of climate projections, it is generally hoped that the truth lies within the ensemble spread. Yet, it is not necessarily the case, for example if some important errors are shared by all models.

In our specific case, all the models in the ensemble could ultimately simulate a too large or too small temperature change over France in response to a given GHG scenario. These configurations are called "extreme cases" subsequently. In this section, it is tested to what extent metric-based approaches perform better than MMEM in extreme cases and how skillful they are.

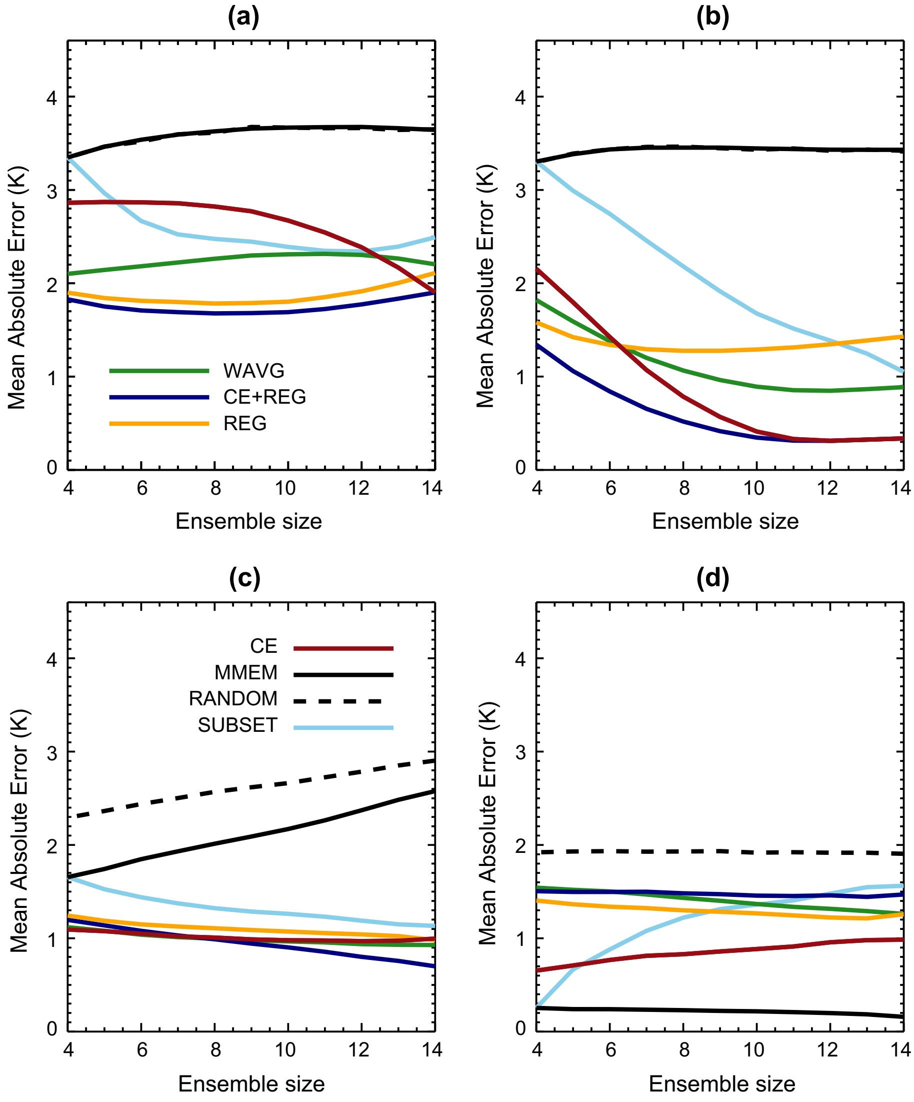
Another question addressed in this section is how the metric-based methods perform in configurations that obey to the truth+error paradigm, i.e. when the truth is equal or very close to the MMEM. By definition, the MMEM provides a virtually perfect prediction in these configurations and it is therefore impossible to outperform it. It remains interesting to quantify the loss of skill of metric-based methods when the truth+error paradigm is respected.

The analyses in Sect. 4 are repeated except that now the different test cases are classified into four categories. The first two categories encompass the cases where the synthetic observation of future temperature change over France is larger or smaller than for all the models in the predictive ensemble. The third category corresponds to cases roughly compatible with the truth+error paradigm: the truth is close to the ensemble mean (in practice, we selected the cases for which the truth is within ±0.2 the inter-model standard deviation). The last category corresponds to all the cases that do not fall into the previous categories. We will call it "normal" as it is the more frequent. Results are depicted in Fig. 7.

By construction, in the truth+error configurations, the MMEM necessarily provides the most accurate, and almost perfect, estimate. In these configurations metric-based methods lead to a degradation of the results, as expected, even if the skill generally remains very good (Fig. 7d). Some substantial differences between methods are found. Note that SUBSET is associated with very small errors for small ensemble sizes as their results are close to the ones given by MMEM by construction in this configuration.

For the normal cases (Fig. 7c) all the methods clearly outperform MMEM for $N > 4$. For large sample sizes, the error is more than two times smaller with the best metric-based methods (CE+REG) compared to the error associated with MMEM (reduction of error close to 1.7 K).

Fig. 7 Same as Fig. 4d except that the different cases are classified into four categories: **a** the synthetic observation of future temperature change over France is smaller than for all the models in the ensemble, **b** the synthetic observation of future temperature change over France is larger than for all the models in the ensemble, **c** the synthetic observation is within the model spread and the models are not truth-centered, **d** the synthetic observation is within the model spread and the models are truth-centered. The models are said truth-centered when the synthetic observations is close to the MMEM, i.e. within ±0.20 intermodel standard deviation of temperature change



Moving to the extreme cases, it is first important to note that the MAE associated with the RANDOM and MMEM approaches are identical by construction in these configurations. If Pt is the truth, the signs of $pi - Pt$ is the same for all the models i for extreme cases, and the sum of $|pi - Pt|$ is therefore identical to the absolute value of the sum of $pi - Pt$.

As the number of extreme configurations is limited, the results for extreme cases are based on small samples, especially when N is large (e.g. just one case for $N = 14$) and therefore intrinsically less robust than for normal cases. The exact tail distribution of temperature changes in the ensemble is more or less important depending on the method in this configuration. A clear dissymmetry in the results between cold and warm extreme cases is noted. Metric-based methods generally perform much better for warm extreme cases. This is probably due to a nonlinearity in the relationship between the metric and future summer temperature change. Very negative present-day inter-annual correlations between temperature and cloud cover are always associated with very strong warming, while the link between temperature change and met is less strong when met is weakly negative (not shown).

All the metric-based methods still lead to more accurate projections compared to MMEM in extreme cases but large inter-method differences exist (Fig. 7a, b). The best methods (CE and CE+REG) lead to a reduction of errors greater than 1.5 K for cold extreme cases and close to 3 K for warm extreme cases for large ensemble sizes.

Note that the proportion of extreme cases depends on the ensemble size: for $N = 14$, the proportion is 13 % and much larger for $N = 4$ (40 %). As the errors tend to be much larger for extreme cases for a given ensemble size as shown in Fig. 7, it partly explains the decrease in error with increased sample sizes noted in Fig. 4.

Table 2 Mean absolute error (K) obtained for different methods with N = 7 using the 8 models for which a least three members are available in present-day and future climate simulations

	CE	WAVG	CE+REG	REG	SUBSET	MMEM	RANDOM
Case 1	1.46	1.37	1.49	1.45	1.46	1.75	2.23
Case 2	1.29	1.18	1.49	1.46	1.39	1.75	2.22
Case 3	0.70	0.74	0.70	0.87	0.99	1.74	2.21

Case 1: a random member is used for each model in the predictive ensemble, and for the synthetic observations. Case 2: the ensemble mean is taken for each model in the predictive and a random member is used for the synthetic observations. Case 3: the ensemble mean is taken for each model in the predictive ensemble and the synthetic observations

The analyses described in this section shows that metric-based methods outperform MMEM in all configurations except when the truth+error paradigm is verified. By construction, it is virtually impossible to outperform MMEM in this configuration. Our analysis shows that metric-based methods lead to a noticeable but generally limited degradation of the results, which varies among the methods, in truth-centered cases.

9 Role of internal variability

In the previous analyses, only one member for each model has been used. Therefore, the estimation of the metrics and future temperature change are impacted to a certain extent by internal climate variability, even if in this study relatively long periods of time are used to compute them.

While to mimic real-world conditions, only one member should be used for the model that provides the synthetic observations, one can wonder whether using across-members averages for each model in the predictive ensemble is useful. In the one hand, doing so limits the impact of internal variability in the estimation of the simulated metrics and future temperature change. In the other hand, as the observed metric is impacted by internal variability in any case, it may not be really useful.

In this section, the eight models for which more than three members are available in present and future simulations are selected and different tests are done. In the first case (case 1), a random member is used for the models and for the synthetic observations, which is similar to what was done previously. In case 2, the ensemble mean is computed for each model in the predictive ensemble and a random member is used for the model that provides the synthetic observations. In case 3, the ensemble mean is taken for each model and for the synthetic observations. Even if it impossible to do so in the real-world, this approach allows the impact of internal variability to be quantified in the results of metric-based methods. The results obtained for the different cases for $N = 7$ are summarized in Table 2.

Using the ensemble means for the predictive models (case 2) sometimes leads to more skillful predictions compared to the use of a single member (case 1), depending on the method, but the improvement remains limited for most methods. These small differences in skill suggest that with metric-based approaches, it is not necessary useful to use multiple members to compute the best posterior estimate, even if it may be helpful. When the ensemble mean is used also for the synthetic observations (case 3), large decreases in MAE are noted with metric-based approaches (in some cases, the error is divided by two). It indicates that a part of the error obtained with metric-based methods is simply due to the impact of internal variability on the observed quantities.

Note that the results presented here are likely to be strongly dependent on the nature of the metrics and periods used to compute them. Our metric is a correlation computed on a 40-year period. While it is impacted by internal variability, the impact is limited. Should the correlations be computed on shorter periods, the impact of internal variability would be greater. Case 3 results suggest that one should try to reduce the impact of internal variability in the observed metric as much as possible, for example by using longer periods. Obviously, data availability limits what is possible to do in real-world applications.

10 Conclusion

The multi-model ensemble mean is not necessarily the best estimator of climate change signals. Alternative approaches based on the idea that more weight should be given to more realistic models are attractive in theory. However, such metric-based approaches raise many theoretical and practical issues that are difficult to tackle.

In this study, a simple approach based on a perfect model framework has been described to test whether and in what conditions metric-based methods could be used to improve future climate change projections. In this framework, each model is successively considered as the truth and its response in the future climate is predicted based only on the knowledge of its present-day climate simulation, and given present-day and future climate simulations from an ensemble of climate models that does not include the first one. It is finally possible to compare the prediction to the truth.

This perfect model framework has been applied to summer temperature change in France. The potential interest of a process-based metric related to cloud-temperature interactions measured by the present-day inter-annual correlation between temperature and cloud cover has been tested. Different statistical methods to combine multiple model results based on the metric have been used. Multiple tests have also been performed to assess the sensitivity of the results to the ensemble size, inter-model similarities, observational errors, and internal variability. Metric-based methods using the metric related to cloud temperature interactions generally lead to major improvements compared to the ensemble mean, leading to much reduced errors in the multi-model estimate. In the absence of observational errors, the improvement in skill between the best metric-based methods and MMEM is as large as the improvement seen when the MMEM is used rather than a single random climate model. Only when large observational errors exist does the MMEM outperform the metric-based methods.

The sensitivity to the statistical method chosen to combine multiple model results based on the metric is generally rather limited for fair ensemble sizes. Overall, the best method for our application is based on the choice of a subset of models (the close-enough and close-enough+regression methods), sufficiently similar to observations in terms of present-day metric. Note that choice of the method could have important implications for impact studies, when climate model projections are used to drive an impact model. The methods that consists in the choice of a small subset of models (SUBSET, CE) are interesting as they can lead to a reduced number of impact simulations compared to the methods that use all the models in the ensemble (WAVG, MMEM). The methods based on the selection of a subset of models also preserve the spatial variability and the inter-variables relationships if several variables are needed by the impact model, which is not necessarily the case for REG.

Even if the impact of inter-model similarities is not as large for all the statistical methods, there is generally no rationale to prefer a large ensemble with important similarities among models compared to a more limited number of quasi-independent models.

These results are encouraging and suggest that a better estimation of future summer temperature change in France is achievable. However, as our most important metric is based on cloud cover, a variable for which the length of observations with a good spatial coverage is limited and the observational uncertainties are large, a careful preliminary assessment of cloud products is necessary. Specific tests in the perfect model framework could be used to assess whether given the length of observed series (which directly influences to extent to which the metrics are impacted by internal variability) and their uncertainties (e.g. using different observed data-sets to estimate the metric), metric-based methods remain interesting in practice.

Note the results described in this paper are likely to be very specific to our application. Indeed, a metric that explains an important part of the inter-model differences and that can be understood in terms of physical processes exists. This metric is also robust across ensembles, as it is relevant both in ENSEMBLES regional climate models (Boé and Terray 2014) and in CMIP5 models. The literature suggests that finding a metric with such properties for a given application is seldom possible. It is important, as our results suggest that a strong metric is needed to be really useful. The most important part of a study aiming to constrain multiple models results based on a metric is therefore likely the search for a good metric itself, rather than the other methodological aspects, discussed in our study. In our case, the metric has been found in Boé and Terray (2014) by trying to understand the physical mechanisms responsible for the inter-model spread in the changes in our variable of interest, and how those mechanisms can also impact present-day climate properties. Note that this effort may also be useful to improve climate models in a way that directly impacts future climate projections, by highlighting some crucial mechanisms that deserve particular attention in the development and evaluation of the models.

The perfect model framework described in this study could be used to assess the theoretical interest of metrics that cannot be computed in the real-world because of lack of data, too poor spatial and temporal coverage etc. It could be useful to highlight which observed variables would be needed, as well as their minimum length and the acceptable level of uncertainties to reach a given improvement compared to MMEM. This knowledge could therefore be used to support the development of observation systems, or data-rescue efforts in order to improve climate change projections.

From a general perspective, this study has illustrated the interest of a perfect model framework to test metric-based methodologies before real world applications. But it is important to remember that it is not sufficient that a given approach performs well in the perfect model framework for it to succeed in real world applications. Good performance in the perfect model framework can mainly be seen as a quasi-necessary condition but not as a sufficient one. If all the models share the same systematic deficiency (for example, if an important feature of the real climate system is missing in all models), the link between the metric and the change in the variable of interest might be largely spurious and the perfect model framework will have little relevance to the real world. Note that if model deficiencies make the perfect model framework irrelevant for real world applications, the hypothesis that the models are centred on the truth, necessary to support the use of the MMEM, will also be wrong.

This study has illustrated the great potential interest of metric-based approaches for multi-model climate projections. However, as some recent studies (Masson and Knutti 2013; Weigel et al. 2010), it has also highlighted some pitfalls associated with such approaches. In particular, a weakly informative metric may not always lead to no improvement compared to the MMEM: it could lead to a degradation of the results in some configurations. The use of metrics in a particular application should therefore always be supported by strong rationale, based on physical arguments.

Acknowledgments This work has been supported by the French National Research Agency (ANR) in the framework of its JCJC program (ECHO, Decision No. ANR 2011 JS56 014 01).

References

Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geophys Res Lett 37:L02703

Annan JD, Hargreaves JC (2011) Understanding the CMIP3 multi-model ensemble. J Clim 24:4529–4538

Boé J, Terray L (2008) Uncertainties in summer evapotranspiration changes over Europe and implications for regional climate change. Geophys Res Lett 35:L05702

Boé J, Terray L (2014) Land-sea contrast, soil–atmosphere interactions and cloud–temperature interactions: interplays and roles in future summer European climate change. Clim Dyn 42(3–4):683–699

Boé J, Hall A, Qu X (2009) September sea-ice cover in the Arctic Ocean projected to vanish by 2100. Nat Geosci 2:341–343

Bracegirdle TJ, Stephenson DB (2012) More precise predictions of future polar winter warming estimated by multi-model ensemble regression. Clim Dyn 39:2805–2821

Caldwell PM, Bretherton CS, Zelinka MD, Klein SA, Santer BD, Sanderson BM (2014) Statistical significance of climate sensitivity predictors obtained by data mining. Geophys Res Lett 41:1803–1808

Collins M, Chandler RE, Cox PM, Huthnance JM, Rougier J, Stephenson DB (2012) Quantifying future climate change. Nat Clim Change 2:403–409

Déqué M, Somot S, Sanchez-Gomez E, Goodess CM, Jacob D, Lenderink G, Christensen OB (2012) The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability. Clim Dyn 38:951–964

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the Reliability Ensemble Averaging (REA) method. J Clim 15:1141–1158

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 13:D06104

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—part I: basic concept. Tellus A 57:219–233

Hall A, Qu X (2006) Using the current seasonal cycle to constrain snow albedo feedback in future climate change. Geophys Res Lett 33:L03502

Jun M, Knutti R, Nychka D (2008) Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? J Am Stat Assoc 103(483):934–947

Knutti R, Meehl GA, Allen MR, Stainforth DA (2008) Constraining climate sensitivity from the seasonal cycle in surface temperature. J Clim 19:4224–4233

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Climate 23:2739–2758

Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. Geophys Res Lett 40:1194–1199

Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of coupled climate models. Clim Dyn 17:83–106

Lenderink G (2010) Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. Clim Res 44:151–166

Masson D, Knutti R (2011) Climate model genealogy. Geophys Res Lett 38:L08703

Masson D, Knutti R (2013) Predictor screening, calibration, and observational constraints in climate model ensembles: an illustration using climate sensitivity. J Clim 26:887–898

Meehl GA et al (2007) Climate Change 2007: the physical science basis. Contribution of Working Group I to the fourth assessment report of the intergovernmental panel on climate change. In: Solomon S et al (eds) Climate Change 2007: the physical science basis. Cambridge University Press, Cambridge

Meehl GA et al (2013) Climate change projections in CESM1(CAM5) compared to CCSM4. J Clim 26:6287–6308

Meinshausen M, Smith SJ, Calvin KV, Daniel JS, Kainuma M, Lamarque JF, Matsumoto K, Montzka SA, Raper SCB, Riahi K, Thomson AM, Velders GJM, van Vuuren D (2011) The RCP greenhouse gas concentrations and their extension from 1765 to 2300. Clim Change. doi:10.1007/s10584-011-0156-z

Min S-K, Hense A (2006) A bayesian assessment of climate change using multimodel ensembles. Part I: global mean surface temperature. J Clim 19:3237–3256

Räisänen J, Palmer T (2001) A probability and decision-model analysis of a multimodel ensemble of climate change simulations. J Clim 14(15):3212–3226

Räisänen J, Ruokolainen L, Ylhäisi JS (2010) Weighting of model results for improving best estimates of climate change. Clim Dyn 35:407–422

Räisänen J, Ylhäisi JS (2012) Can model weighting improve probabilistic projections of climate change? Clim Dyn 39:1981–1998

Sanderson BM, Knutti R (2012) On the interpretation of constrained climate model ensembles. Geophys Res Lett 39:L16708

Stegehuis AI, Teuling AJ, Ciais P, Vautard R, Jung M (2013) Future European temperature change uncertainties reduced by using land heat flux observations. Geophys Res Lett 40:2242–2245

Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. J Clim 18:1524–1540

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc Ser A 365(1857):2053–2075

Terray L, Boé J (2013) Quantifying 21st-century France climate change and related uncertainties. Comptes Rendus Geosci 345(3):136–149

Toth Z, Talagrand O, Candille G, Zhu Y (2003) Probability and ensemble forecasts. In: Jolliffe IT, Stephenson DB (eds) forecast verification: a practitioners guide in atmospheric science. Wiley, Chichester, pp 137–163

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260

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

Wilby RL, Charles SP, Zorita E, Timbal B, Whetton P, Mearns LO (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis (TGICA). http://www.ipccdata.org/guidelines/dgm_no2_v1__09_2004

Yokohata T, Annan JD, Collins M, Jackson CS, Tobis M, Webb MJ, Hargreaves JC (2012) Reliability of multi-model and structurally different single-model ensembles. Clim Dyn 39:599–616