Atmos. Chem. Phys. Discuss., 15, 7707–7734, 2015
www.atmos-chem-phys-discuss.net/15/7707/2015/
doi:10.5194/acpd-15-7707-2015

Atmospheric Chemistry and Physics Discussions

Improvement of climate predictions and reduction of their uncertainties using learning algorithms

E. Strobach and G. Bel

Department of Solar Energy and Environmental Physics, Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus, 84990 Israel

Received: 12 February 2015 – Accepted: 26 February 2015 – Published: 12 March 2015

Correspondence to: G. Bel (bel@bgu.ac.il)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Simulated climate dynamics, initialized with observed conditions is expected to be synchronized, for several years, with the actual dynamics. However, the predictions of climate models are not sufficiently accurate. Moreover, there is a large variance between simulations initialized at different times and between different models. One way to improve climate predictions and to reduce the associated uncertainties is to use an ensemble of climate model predictions, weighted according to their past performance. Here, we show that skillful predictions, for a decadal time scale, of the 2 m-temperature can be achieved by applying a sequential learning algorithm to an ensemble of decadal climate model simulations. The predictions generated by the learning algorithm are shown to be better than those of each of the models in the ensemble, the better performing simple average and a reference climatology. In addition, the uncertainties associated with the predictions are shown to be reduced relative to those derived from equally weighted ensemble of bias corrected predictions. The results show that learning algorithms can help to better assess future climate dynamics.

1 Introduction

A group of global climate simulations, referred to as the decadal experiments, was introduced in the Coupled Model Intercomparison Project (CMIP5) multi-model ensemble (Taylor et al., 2012; Meehl et al., 2009). The idea behind these experiments was to investigate the predictability of the climate by atmosphere ocean general circulation models (AOGCMs) in time scales of up to 30 years.

The AOGCMs were initialized with interpolated observations of the ocean, sea ice and atmospheric conditions, together with the atmospheric composition (Taylor and Meehl, 2011). Therefore, they were expected to reproduce the monthly and annual averages of the climate variables and the response of the climate system to changes in the atmospheric composition (Warner, 2011; Collins, 2007; Kim et al., 2012). Indeed,

ACPD
15, 7707–7734, 2015

it was shown (Kim et al., 2012) that in some regions, the CMIP5 simulations have some prediction skill. It was also confirmed (Kim et al., 2012) that the multi-model average provides better predictions than each of the models, similar to what was found for other climate simulations (Doblas-Reyes et al., 2000; Palmer et al., 2004; Hagedorn et al., 2005; Feng et al., 2011). However, the simple multi-model average does not take into account the quality differences between the models; therefore, it is expected that a weighted average, with weights based on the past performances of the models, will provide better predictions than the simple average. As expected, it was shown that the weighted average of climate models can improve predictions when using ensembles of AGCMs (Rajagopalan et al., 2002; Robertson et al., 2004; Yun et al., 2003), AOGCMs (YUN et al., 2005; Pavan and Doblas-Reyes, 2000; Chakraborty and Krishnamurti, 2009) and regional climate models (Feng et al., 2011; Samuels et al., 2013).

The uncertainties in climate predictions can be attributed to three main sources – internal variability of the model, inter-model variability and future forcing scenario uncertainties. The internal variability of the model stems from the sensitivity of the model to the initial conditions, sensitivity to the values of the parameters and the discretization method used. The inter-model variability is the result of different parameterization schemes and modeling approaches adopted in different models. The uncertainties due to different forcing scenarios are mostly related to different scenarios assumed regarding future greenhouse gas emissions. On a decadal time scale, forcing scenario uncertainties and uncertainties due to the internal variability of each model are considerably smaller than the inter-model uncertainties (Meehl et al., 2009; Hawkins and Sutton, 2009) (we also verified that the internal variability of each of the models we used is much smaller than the inter-model variability). Therefore, estimation of the uncertainties from ensemble of climate models is expected to give a meaningful estimation of the total climate prediction uncertainties.

Different methods were used to improve climate predictions using an ensemble of models. A common approach is the simple regression (Krishnamurti et al., 2000; Krishnamurti, 1999). The regression does not assign a weight to each member of the ensem-

ACPD
15, 7707–7734, 2015

ble but rather attempts to find the set of coefficients yielding the minimal square error for a linear combination of the ensemble model predictions. Bayesian methods have also been used for weighting ensembles of climate model projections (Rajagopalan et al., 2002; Robertson et al., 2004; Tebaldi and Knutti, 2007; Smith et al., 2009; Buser et al., 2009, 2010). The weighting scheme of these methods relies on a certain distribution of the errors and other prior assumptions regarding the models; these assumptions are not necessarily valid for climate dynamics and predictions. Many variations of the Bayesian methods were applied to weather forecasting in order to establish the ensemble of models (Kalnay et al., 2006); these methods are less useful for climate predictions in which the variability between different models is larger than the internal variability of each model (Meehl et al., 2009; Hawkins and Sutton, 2009). Here, we use a sequential learning algorithm (SLA) method, which is adopted from the field of game theory (Cesa-Bianchi and Lugosi, 2006; Mallet et al., 2009; Mallet, 2010), to weight ensemble members of global climate models. It is important to note that the SLA method assigns real weights (taking values between zero and one) to the ensemble models rather than to future climate paths (it is straightforward to use the weights of the models to get the probabilities of future climate paths which are common in the Bayesian approaches); this characteristic makes the SLA method appropriate for model evaluation. The SLA method has several advantages compared with other weighting schemes: (i) it makes no assumptions regarding the distribution of the climate variables and the model parameters. Therefore, it can be used for all climate variables and all types of predictions; (ii) there is an upper bound for the deviation of the weighted ensemble average from the best model. For a sufficiently lengthy learning period (the duration of this period depends on the variable, the learning rate (which is described later) and the number of models in the ensemble), the SLA prediction is at least as good as the prediction of the best model in the ensemble; (iii) the weights can be dynamically updated, when new measurements are introduced, with no significant computational cost. The original method (Cesa-Bianchi and Lugosi, 2006) was modified and adjusted to improve decadal climate predictions.

ACPD
15, 7707–7734, 2015

2 The sequential learning algorithm

The sequential learning algorithm (SLA) assigns weights to the climate models (*the experts*) in the ensemble based on their past performance. In this work, the output of the models was divided into two periods – a learning period during which the weights were updated and a prediction period during which the weights remained fixed and equal to the weights assigned by the SLA in the last step of the learning process. In order to capture the spatial variability in model performance, the weights were spatially distributed and the weight of each model in each grid cell was determined by the local past performance of the model. For the sake of clarity, the algorithm is described below without spatial indexes although the calculations were done for each grid cell separately. The prediction of the SLA *forecasters* is the weighted average of the ensemble (Cesa-Bianchi and Lugosi, 2006). The weights are assigned to minimize the cumulative regret with respect to each one of the climate models. The cumulative regret of *expert* E is defined as:

$$R_{E,n} \equiv \sum_{t=1}^{n} (l(p_t, y_t) - l(f_{E,t}, y_t)) \equiv L_n - L_{E,n}. \quad (1)$$

t is a discrete time, l denotes some loss function that is a measure of the difference between the predicted (p_t by the *forecaster* and $f_{E,t}$ by *expert* E) and the true (y_t) values. In this work, we defined the loss function to be the square of the difference between the *forecaster* prediction and the "real" value, namely, $l(p_t, y_t) \equiv (p_t - y_t)^2$. $L_n \equiv \sum_{t=1}^{n} l(p_t, y_t)$, $L_{E,n} \equiv \sum_{t=1}^{n} l(f_{E,t}, y_t)$ are the cumulative loss functions of the *forecaster* and *expert* E, respectively. The outcome of the *forecaster*, after $n-1$ steps of learning, is weights assigned to the climate models in the ensemble to be used for forecasting the value at $t = n$. The forecast for $t = n$ is the weighted average of the climate models, that is:

$$p_n \equiv \sum_{E=1}^{N} w_{E,n-1}(R_{E,n-1}) \cdot f_{E,n}. \quad (2)$$

ACPD
15, 7707–7734, 2015

Here, N is the number of models (*experts*) and $w_{E,n-1}$ is the weight of *expert* E, which is determined by the regret up to time $n-1$. We used two *forecasters* (weighting schemes): the Exponentiated Weighted Average (EWA) and the Exponentiated Gradient Average (EGA). The EWA weight is defined as:

$$w_{E,n} \equiv \frac{e^{-\eta \cdot R_{E,n}}}{\sum_{E=1}^{N} e^{-\eta \cdot R_{E,n}}} \tag{3}$$

and its prediction at time n is:

$$p_n = \frac{\sum_{E=1}^{N} e^{-\eta L_{E,n-1}} f_{E,n}}{\sum_{E=1}^{N} e^{-\eta L_{E,n-1}}}. \tag{4}$$

The EGA is similar to the EWA but with the cumulative regret calculated from the summation of the loss gradients. The cumulative regret for the EGA *forecaster* is defined as:

$$R^G_{E,n} \equiv \sum_{t=1}^{n} l'(p_t, y_t) - \sum_{t=1}^{n} l'(f_{E,t}, y_t) \equiv L^G_n - L^G_{E,n} \tag{5}$$

where,

$$l'(p_t, y_t) \equiv \frac{\partial l(p_t, y_t)}{\partial w_{E,t-1}} = 2 \cdot (p_t - y_t) \cdot f_{E,t}. \tag{6}$$

For both *forecasters*, $\eta > 0$ is a parameter representing the learning rate. The deviation between the forecast and the "real" trajectory was quantified using the root mean square error (RMSE). The RMSE of a grid cell with coordinates (i, j), over a period of n time steps (months in our case), is defined as:

$$\mathrm{RMSE}(i,j) \equiv \sqrt{(1/n) \sum_{t=1}^{n} (p_t(i,j) - y_t(i,j))^2}, \tag{7}$$

ACPD
15, 7707–7734, 2015

where $p_t(i,j)$ is the value predicted by the *forecaster* and $y_t(i,j)$ is the "real" value. The global, area-weighted RMSE is defined as:

$$G_{\mathrm{RMSE}} \equiv \left(1/A_{\mathrm{Earth}}\right) \sum_{i,j} A_{i,j} \mathrm{RMSE}(i,j), \tag{8}$$

where A_{Earth} is the earth's surface area and $A_{i,j}$ is the area of the (i,j) grid cell. The learning rate, η, was chosen to minimize the metric $M \equiv \mathrm{RMSE} \cdot (1 + \mathrm{floor}(\max(\Delta w/\Delta t)/(1/N)))$ during the learning period. This metric provides a minimal deviation of the forecast climate trajectory from the observed one and also ensures stable weights of the models (a significant change in the weight of a model was considered the weight a model would be assigned in the absence of learning). We also tested optimization of η using only a fraction of the learning period and found that as long as the optimization period was of the same order of the prediction period, there was no significant change in the outcome. An important difference between the EWA and EGA methods is that after a long enough learning period under ideal conditions (stationary time series), the former converges to the best model while the latter converges to the "real" value assuming that the real value is known. Figure 1 illustrates this difference using a simple case. This difference between the *forecasters* implies that for long enough learning period, using an ensemble that includes one model that perform better throughout the learning period, the weights will be distributed such that the prediction of the EWA will be determined by this best model and the uncertainty will be very small (due to the small weights of the other models). Under the same conditions, the EGA would still assign more significant weights to the other models in order to extract the information they contain regarding the dynamics of the "real" value and this will lead to larger uncertainty (and often better predictions).

ACPD
15, 7707–7734, 2015

3 Improved predictions

We consider an ensemble of eight global climate models for the period of 1981–2011, whose results are part of the CMIP5 decadal experiments (Taylor and Meehl, 2011). Table 1 describes the eight models that we used in this study. These models were first linearly interpolated to the spatial resolution of the NCEP/NCAR reanalysis data using the NCAR command language (NCL) (NCL, 2011). We focus on the model predictions of the 2 m-temperature. The decadal experiments of the CMIP5 project include a set of runs for each of the models, representing different initial conditions. In agreement with the common knowledge (Meehl et al., 2009), we found that on decadal time scales, the internal variability of each model is smaller than the variability between the models. Therefore, we chose, arbitrarily, the first run for each of the ensemble models. The results presented here are based on a learning period of 20 years (1981–2001), followed by predictions for 10 year (2001–2011) validation period.

The learning period served for both learning (i.e., weight assignment) and also to correct the bias of the models. This was simply done by subtracting the average of each of the models during the learning period and adding the average of the NCEP/NCAR reanalysis data (Kalnay et al., 1996) (considered here as reality). This bias correction was applied to each grid cell separately. The bias correction was done to assure that the improvement achieved by the forecasters is beyond the impact of a simple bias correction. In addition, we chose a long enough learning period to ensure that our results are not affected by the drift of the models from the intial condition toward their climate dynamics (Meehl et al., 2009).

The performance of the models was determined by comparing the model predictions to the NCEP/NCAR reanalysis data (Kalnay et al., 1996). We are aware of the spurious variability and trends in the NCEP data and of other reanalysis projects (Uppala et al., 2005; Onogi et al., 2007); however, in order to demonstrate the capability of the SLA to improve global and regional climate predictions, the reanalysis data is the best dataset.

ACPD
15, 7707–7734, 2015

Using the predictions of the climate models only 20 years after they were initialized can debate their ability to generate skillful predictions since it is believed that climate models' skill tends to vanish after that long a period. However, we found that, for most of the models we used, this is not the case. This fact is illustrated in Fig. 2, which shows that the globally averaged RMSE of most of the climate models did not increase considerably during the 30 year-long simulations. Another noticeable and important feature of the climate models of the CMIP5 is the fact that, globally, climatology performs much better than each of the models. In Sect. 5 we show that, despite this fact, the SLA can use the models and the climatology to provide a forecast which is better than the climatology.

Three forecasting methods (*forecasters*) were tested: the Exponentiated Weighted Average (EWA), the Exponentiated Gradient Average (EGA) and a simple average. The simple average represents no learning and is presented to illustrate the superior performance of the SLA. The performance of the *forecasters* is measured by the root mean square error (RMSE), during the validation period, which quantifies the deviation of the predicted climate trajectory from the observed one.

Figure 3 shows the RMSE in the 2 m-temperature monthly average prediction, during the 10 year validation period, for each grid cell. Panels a, b and c correspond to the RMSE of the EWA, EGA and simple average weighting schemes, respectively. Both the EWA and the EGA *forecasters* give better predictions than the simple average. The improvement achieved by the two *forecasters*, compared with the simple average, is more apparent close to the poles and in western South America. In these regions, the models deviate more from each other and the weighting schemes favor those that perform better. Over the oceans and low to mid latitudes the models showed better agreement and therefore the weighting schemes did not yield a large improvement.

The global, area-weighted RMSE can be used to quantify the improvement achieved by the SLA *forecasters*, that is, 1.316 °C for the EWA, 1.297 °C for the EGA and 1.390 °C for the simple average. Since the EWA has the tendency to converge to the best model (if the ensemble includes a model that is always better than the others in certain re-

ACPD
15, 7707–7734, 2015

gions), we also compare the performance of the EWA and EGA *forecasters* with two forecasting methods that predict according to the best model (defined as the model that was assigned the highest weight according to the EWA or the EGA) in each grid cell. The global, area-weighted RMSE was found to be 1.568 °C for the best model based on the EWA and 1.633 °C for the best model based on the EGA. These results show that the SLA *forecasters* outperform the best models in the ensemble. In general, we found that a longer learning period improves the predictions of the *forecasters*. Figure 4 shows that the area-weighted RMSE of the *forecasters* (during the validation period) is reduced when the learning period is extended. By increasing the learning rate we found that shorter learning periods can be selected with no significant increase in error; however, we chose a learning period which is of the order of the prediction period in order to capture the climate dynamics in all the time scales that are relevant to the prediction period.

4 Reduced uncertainties

The weights obtained from the SLA method can be used to better estimate the uncertainties of the predictions. The uncertainties are quantified by the square root of the time average of the weighted variance of the ensemble. This quantity (for a period of n time steps) in the (i,j) grid cell is defined as:

$$\mathrm{STD}(i,j) \equiv \sqrt{(1/n)\sum_{t=1}^{n}\sum_{E=1}^{N} w_E(i,j)(f_{E,t}(i,j)-p_t(i,j))^2}. \quad (9)$$

Here, $f_{E,t}(i,j)$ is the prediction of model E for grid cell (i,j), at time t; $p_t(i,j)$ is the prediction of the *forecaster* for grid cell (i,j), at time t (i.e., the weighted average of all the models); and $w_E(i,j)$ is the weight assigned to model E at grid cell (i,j) (the weights remain constant during the validation period for which the STD is calculated).

ACPD
15, 7707–7734, 2015

The global, area-weighted uncertainty is defined as:

$$G_{\mathrm{STD}} \equiv \left(1/A_{\mathrm{Earth}}\right) \sum_{i,j} A_{i,j} \mathrm{STD}(i,j). \tag{10}$$

Figure 5 shows the uncertainty of the 2 m-temperature during the validation period for the three forecasting methods; panels a, b and c correspond to the EWA, EGA and simple average *forecasters*, respectively. It is important to note that this uncertainty is only due to the different predictions of the ensemble models; other sources of uncertainty are not affected by our forecasting schemes. Both the EWA and EGA *forecasters* yield smaller uncertainties than does the simple average. The improvement is significant in regions where the uncertainties are larger, such as toward the poles and over South America and Africa. The global, area-weighted, uncertainties are: 1.242, 1.381, and 1.593 °C for the EWA, EGA and simple average *forecasters*, respectively. These values show that in addition to improving the predictions, the SLA *forecasters* also reduce the uncertainties of these predictions. Note that the smaller uncertainty of the EWA *forecaster* is simply due to the fact that this *forecaster* converges to the best model in each grid cell (if the ensemble includes a model that is always the best). The uncertainty of the EGA provides a better estimate of the predictions uncertainty because its predictions converges to the observations.

5 Skillful forecast

The skill of a *forecaster* may be defined as its ability to provide better predictions than the reference climatology. In our study the natural choice is the climatology of the learning period, that is:

$$C_m \equiv \frac{1}{L} \sum_{i=1}^{L} y_{i,m}, \tag{11}$$

ACPD
15, 7707–7734, 2015

where, $y_{i,m}$ is the value of the variable (in this study it is the 2 m-temperature as reported in the reanalysis data) in the calendar month m of the year i; the learning period duration is L years; and the climatology, C_m is just the average of that variable during the L years. Prediction that is based on climatology assumes that for each month of the prediction period, the value of the variable will be equal to the climatology of the corresponding calendar month. Therefore, it is reasonable to expect that a skillful model should provide more information on the variability of the climate than the average of previous years (the climatology).

Figure 6a shows the differences between the 10 year RMSE of the 2 m-temperature monthly mean, of the climatology and of the EGA *forecaster*. Positive values represent locations where the EGA *forecaster* has a smaller RMSE and is, therefore, considered as a skillful *forecaster*. In most regions, the climatology performs better than the EGA *forecaster* (and, obviously, better than the best model!); however, some regions indicate the EGA's advantage, such as eastern North America up to Greenland. We found that the regions in which the EGA *forecaster* performs better are characterized by larger variability (which increases the deviations from the climatology). The global, area-weighted RMSE is 1.188 °C for the climatology and 1.373 °C for the EGA. One could conclude that the EGA *forecaster* is not skillful.

To circumvent this problem, we decided to add the climatology of the learning period as an additional model to our ensemble. In Fig. 6b, we show the difference between the RMSE of the EGA *forecaster*, for the model ensemble that includes the climatology, and the RMSE of the climatology itself. In this figure, one can see that the EGA *forecaster*, for the model ensemble that includes the climatology, provides predictions that are at least as good as the climatology over most of the globe. Adding the climatology to the ensemble reduced the global, area-weighted RMSE of the EGA *forecaster* to 1.156 °C – a small improvement (a reduction of about 2.7 %) over the climatology. The global, area-weighted, uncertainties of the 10 year validation period in this case are: 0.118, 0.953, and 1.552 °C for the EWA, EGA and simple average *forecasters*, respectively. Note, that as we mentioned earlier, the small uncertainty associated with the EWA

ACPD
15, 7707–7734, 2015

forecaster is not representative of the climate prediction uncertainty. In what follows we focus on significance of the results of the EGA *forecaster*.

6 Significance tests

There is more than one test that can be done to demonstrate the significance of the results. We focus on testing whether the EGA *forecaster* improves the predictions beyond climatology (as shown earlier, each of the models performs poorer than the climatology) and whether it reduces the uncertainties below those of equally weighted ensemble. Both tests were done globally and regionally. We start by defining two properties. The first, is the difference between the absolute error of the climatology and the absolute error of the EGA forecaster at a given grid cell and time point, that is $-|(C_t(i,j)-y_t(i,j))|-|(p_t(i,j)-y_t(i,j))|$. The second is the difference between the uncertainties of the equally weighted ensemble and the ensemble weighted according to EGA *forecaster* at a given grid cell and time point, that is $-\sqrt{\frac{1}{N}\sum_{E=1}^{N}(f_{E,t}(i,j)-f_{\cdot,t}(i,j))^2}-\sqrt{\sum_{E=1}^{N}w_E(i,j)\cdot(f_{E,t}(i,j)-p_t(i,j))^2}$ (the dot replacing the E index, represents averaging over that index). For both quantities, positive values represent better performance of the EGA *forecaster*. The 10 year validation period yields for each of these quantities time series with 120 points in each grid cell. The fraction of the time series (number of points out of the total 120) showing positive values can be used to test the significance of the improvement. We define a significant improvement by the EGA *forecaster* to be when the number of successes are above 66 (i.e., when the null hypothesis that the quantities defined above are symmetrically distributed around zero is rejected with $\sim 90\,\%$ confidence).

Figure 7 shows the spatial distributions of the number of positive values (out of the total 120 time points) for the two quantities. The upper panel corresponds to the difference between the absolute error of the climatology and the EGA *forecaster* and

ACPD
15, 7707–7734, 2015

the lower panel corresponds to the difference between the uncertainties of the equally weighted and EGA weighted ensembles.

The upper panel in Fig. 7 shows that there are large regions of improvement which is more apparent over land, close to the poles and to the equator. The lower panel shows that in regions in which the EGA reduces the uncertainty, it does so for almost all time points and vice versa. No correlation between significant improvement of the predictions and significant reduction of the uncertainties was identified.

The global test we performed was done by calculating the area weighted average of the two quantities defined above and to plot the histograms of their time series. These are shown in Fig. 8. The upper panel shows the global average absolute error difference between the climatology and the EGA *forecaster* and the lower panel shows the global average difference between the uncertainties of the equally weighted and EGA weighted ensembles. The x axis is in units of $^{\circ}$C and is zero centered to emphasize the non symmetric distribution of the data. The upper panel shows that there are only 11 negative values out of 120 and a positive peak at around 0.03 $^{\circ}$C. The probability of more than 108 positive values out of 120 in a symmetric distribution with zero mean is practically zero; therefore, we conclude that, globally, the EGA forecaster predicts better than climatology. The uncertainties difference shows that the EGA *forecaster* has lower uncertainty than equally weighted ensemble for all the time points and therefore we can also conclude that the reduction of the globally averaged uncertainties is significant.

7 Summary and discussion

The SLA method does not rely on any assumptions regarding the distributions of the climate variables; therefore, it is robust and can be used for any climate variable. The updating scheme of the weights does not require a considerable computational cost and allows for a fast and easy update of the weights when new measurements become available. In the results presented here, we used the deviation from the trajectory of the

ACPD
15, 7707–7734, 2015

climate variable as the metric for the weighting, but other weighting methods can also be applied. For example, one can use a measure of the statistical distance such as the Kullback–Leibler divergence (Kullback and Leibler, 1951) or the Jensen–Shannon divergence (Manning and Schütze, 1999); a model that yields a probability density function (PDF) which is closer to the measured PDF of a variable will get a higher weight.

One disadvantage of the SLA method (which may also be considered as an advantage for some applications) is the fact that the weights are between zero and one. This means that if the measurements are not spanned by the predictions of the models, the SLA algorithm will not be able to track the measurements but would converge to the best model since by definition the SLA predictions are bounded by the predictions of the models of the ensemble. In this case, other methods, such as the regression that can yield any linear combination of the model predictions, may achieve better predictions than the SLA *forecasters* but will not be able to reduce the ensemble uncertainties.

We showed that climate predictions (on a decadal time scale) of the 2 m-temperature monthly average can be improved and that the associated uncertainties can be reduced using the EGA *forecaster*. The improved predictions and reduced uncertainties considered here are only those arising from the variability between different models. This is because the ensemble used in this study consists of only one run (corresponding to one initial condition) of each of the models. The uncertainties due to the internal variability of each of the models remained unaffected. In principle, the SLA method can be used to quantify the quality of different initialization methods. However, there is no justification for weighting initial conditions generated by the same method at times that are of the same order of magnitude before the prediction period. Therefore, the SLA method cannot reduce uncertainties associated with the internal variability of the models.

The SLA method provided better predictions than each one of the models and their simple average. All the models, including the simple average, considered in this study showed no global skill; namely, in averaging over the globe, the climatology provided

ACPD
15, 7707–7734, 2015

a better prediction than each of the models. The SLA *forecasters* do not resolve this issue unless the climatology is added as an additional model to the ensemble. When the model ensemble includes the climatology, the SLA *forecasters* can yield better predictions than the climatology itself by assigning high weight to the climatology in the regions where the models fail and high weight to the best models in regions where they perform better than the climatology (namely, regions where the best models are skillful).

The method and the results presented here provide performance-based, spatially-distributed weights of climate models, which lead to improved climate predictions and reduced uncertainties. These can be relevant for many applications in agriculture and ecology, and for decision makers and other stakeholders. The spatially-distributed weights may also be used for testing new parameterization and physics schemes in global circulation models.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant number [293825].

References

Buser, C. M., Künsch, H. R., Lüthi, D., Wild, M., and Schär, C.: Bayesian multi-model projection of climate: bias assumptions and interannual variability, Clim. Dynam., 33, 849–868, 2009. 7710

Buser, C., Künsch, H., and Schär, C.: Bayesian multi-model projections of climate: generalization and application to ENSEMBLES results, Clim. Res., 44, 227–241, 2010. 7710

Cesa-Bianchi, N. and Lugosi, G.: Prediction, Learning, and Games, Cambridge University Press, Cambridge, UK, 2006. 7710, 7711

Chakraborty, A. and Krishnamurti, T. N.: Improving global model precipitation forecasts over India using downscaling and the FSU superensemble. Part II: Seasonal climate, Mon. Weather Rev., 137, 2736–2757, 2009. 7709

Collins, M.: Ensembles and probabilities: a new era in the prediction of climate change, Philos. T. R. Soc. A, 365, 1957–1970, 2007. 7708

ACPD

15, 7707–7734, 2015

Doblas-Reyes, F. J., Déqué, M., and Piedelievre, J.-P.: Multi-model spread and probabilistic seasonal forecasts in PROVOST, Q. J. Roy. Meteor. Soc., 126, 2069–2087, 2000. 7709
Feng, J., Lee, D.-K., Fu, C., Tang, J., Sato, Y., Kato, H., Mcgregor, J., and Mabuchi, K.: Comparison of four ensemble methods combining regional climate simulations over Asia, Meteorol. Atmos. Phys., 111, 41–53, 2011. 7709
Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, Tellus A, 57, 219–233, 2005. 7709
Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, B. Am. Meteorol. Soc., 90, 1095–1107, 2009. 7709, 7710
Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, B. Am. Meteorol. Soc., 77, 437–471, 1996. 7714
Kalnay, E., Hunt, B., Ott, E., and Szunyogh, I.: Ensemble forecasting and data assimilation: two problems with the same solution?, in: Predictability of Weather and Climate, edited by: Palmer, T. N. and Hagedorn, R., Cambridge University Press, Cambridge, 157–180, 2006. 7710
Kim, H.-M., Webster, P. J., and Curry, J. A.: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, Geophys. Res. Lett., 39, L10701, doi:10.1029/2012GL051644, 2012. 7708, 7709
Krishnamurti, T. N.: Improved weather and seasonal climate forecasts from multimodel superensemble, Science, 285, 1548–1550, 1999. 7709
Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel ensemble forecasts for weather and seasonal climate, J. Climate, 13, 4196–4216, 2000. 7709
Kullback, S. and Leibler, R. A.: On information and sufficiency, Ann. Math. Stat., 22, 79–86, doi:10.1214/aoms/1177729694, 1951. 7721
Mallet, V.: Ensemble forecast of analyses: coupling data assimilation and sequential aggregation, J. Geophys. Res.-Atmos., 115, D24303, doi:10.1029/2010JD014259, 2010. 7710
Mallet, V., Stoltz, G., and Mauricette, B.: Ozone ensemble forecast with machine learning algorithms, J. Geophys. Res.-Atmos., 114, D05307, doi:10.1029/2008JD009978, 2009. 7710

ACPD
15, 7707–7734, 2015

Manning, C. D. and Schütze, H.: Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, 1999. 7721

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T.: Decadal prediction, B. Am. Meteorol. Soc., 90, 1467–1485, 2009. 7708, 7709, 7710, 7714

Onogi, K., TsuTsui, J., Koide, H., Sakamoto, M., Kobayashi, S., Hatsushika, H., Matsumoto, T., Yamazaki, N., Kamahori, H., Takahashi, K., Kadokura, S., Wada, K., Kato, K., Oyama, R., Ose, T., Mannoji, N., and Taira, R.: The JRA-25 reanalysis, J. Meteorol. Soc. Jpn., 85, 369–432, 2007. 7714

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J.-M., Davey, M., Graham, R., Délécluse, P., Lazar, A., Déqué, M., Guérémy, J.-F., Díez, E., Orfila, B., Hoshen, M., Morse, A. P., Keenlyside, N., Latif, M., Maisonnave, E., Rogel, P., Marletto, V., and Thomson, M. C.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), B. Am. Meteorol. Soc., 85, 853–872, 2004. 7709

Pavan, V. and Doblas-Reyes, F. J.: Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features, Clim. Dynam., 16, 611–625, 2000. 7709

Rajagopalan, B., Lall, U., and Zebiak, S. E.: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles, Mon. Weather Rev., 130, 1792–1811, 2002. 7709, 7710

Robertson, A. W., Lall, U., Zebiak, S. E., and Goddard, L.: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction, Mon. Weather Rev., 132, 2732–2744, 2004. 7709, 7710

Samuels, R., Harel, M., and Alpert, P.: A new methodology for weighting high-resolution model simulations to project future rainfall in the Middle East, Clim. Res., 57, 51–60, 2013. 7709

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O.: Bayesian modeling of uncertainty in ensembles of climate models, J. Am. Stat. Assoc., 104, 97–116, 2009. 7710

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: A Summary of the CMIP5 Experiment Design, available at: http://cmip-pcmdi.llnl.gov/cmip5/experiment_design.html (last access: 7 March 2015), 2011. 7708, 7714

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485–498, 2012. 7708

ACPD
15, 7707–7734, 2015

The NCAR Command Language (Version 6.0.0) [Software], doi:10.5065/D6WD3XH5, 2011. 7714

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philos. T. R. Soc. A, 365, 2053–2075, 2007. 7710

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005. 7714

Warner, T. T.: Numerical Weather and Climate Prediction, Cambridge University Press, Cambridge, UK, 2011. 7708

Yun, W. T., Stefanova, L., and Krishnamurti, T. N.: Improvement of the multimodel superensemble technique for seasonal forecasts, J. Climate, 16, 3834–3840, 2003. 7709

Yun, W. T., Stefanova, L., Mitra, A. K., Kumar, T. S. V. V., Dewar, W., and Krishnamurti, T. N.: A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts, Tellus A, 57, 280–289, 2005. 7709

ACPD
15, 7707–7734, 2015

Table 1. Models Availability.

Institute ID	Model Name	Modeling Center (or Group)	Grid (lat × lon)
BCC	BCC-CSM1.1	Beijing Climate Center, China Meteorological Administration	64 × 128
CCCma	CanCM4	Canadian Centre for Climate Modelling and Analysis	64 × 128
CNRM-CERFACS	CNRM-CM5	Centre National de Recherches Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	128 × 256
LASG-IAP	FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	108 × 128
IPSL	IPSL-CM5A-LR	Institut Pierre-Simon Laplace	96 × 96
MIROC	MIROC5 MIROC4h	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	128 × 256 320 × 640
MRI	MRI-CGCM3	Meteorological Research Institute	160 × 320

ACPD
15, 7707–7734, 2015

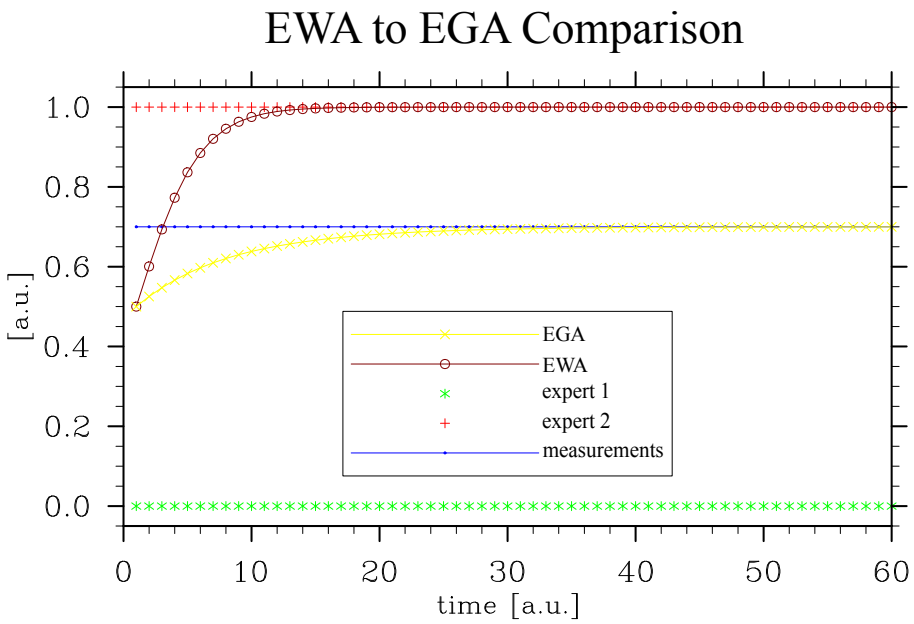

Figure 1. An ideal experiment with two experts. The first always predicts zero and the second always predicts one. The true value is always 0.7. The EWA forecaster converges to the best model (predicting one) while the EGA forecaster converges to the true value.

ACPD
15, 7707–7734, 2015

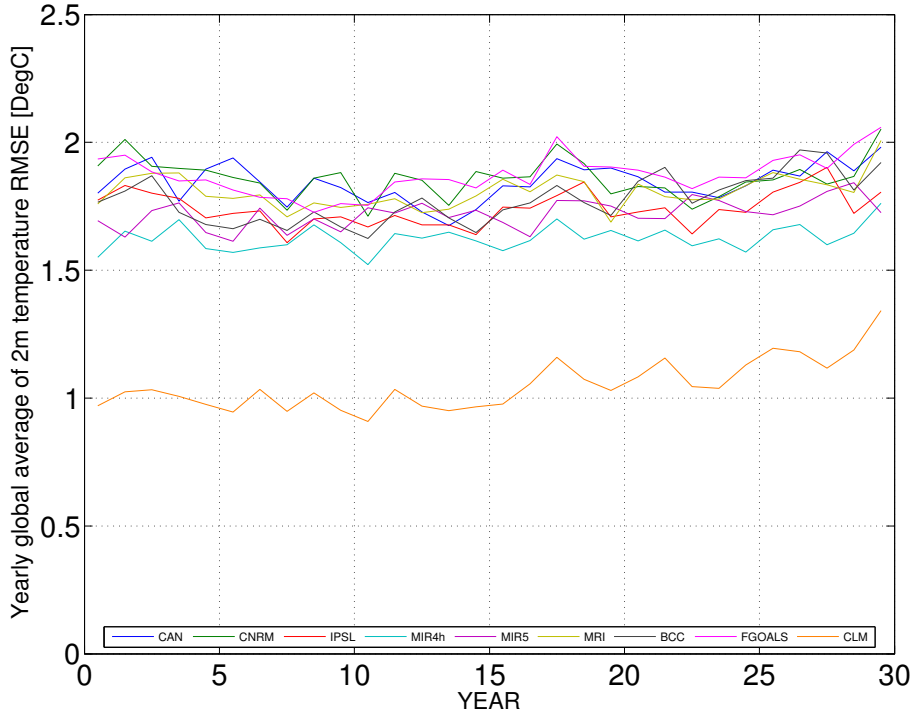

Figure 2. Temporal evolution of the global and annual average of the 2 m-temperature RMSE for the eight climate models (after bias correction) and the climatology. During the 30 years of the simulations, the skill of most of the models did not decline. In fact, a simple linear fit to the models indicates that some of them increased their skill with time.

ACPD
15, 7707–7734, 2015

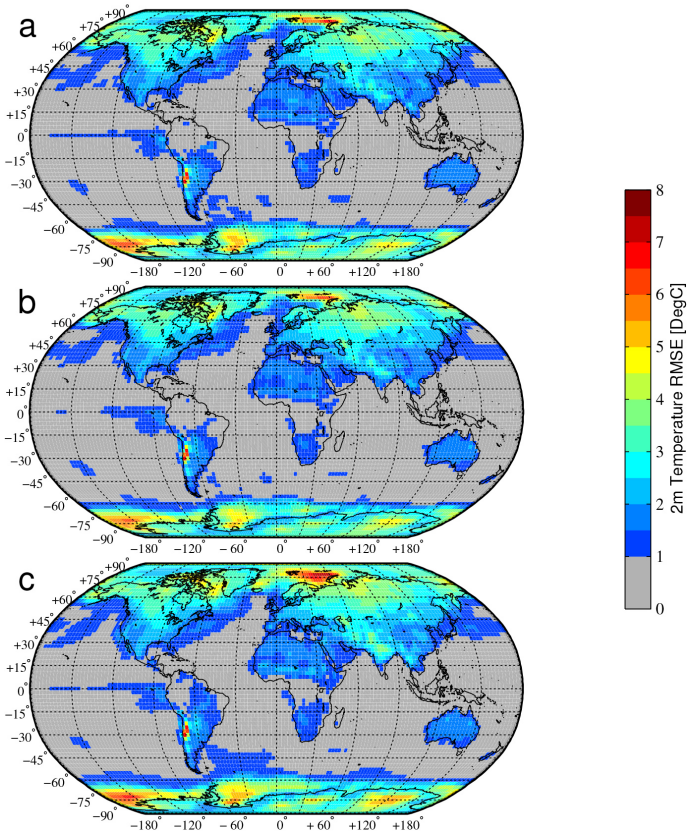

Figure 3. 10 year RMSE of the 2 m-temperature for three forecasting methods: **(a)** EWA, **(b)** EGA, and **(c)** simple average. The colors represent the RMSE of each grid cell. Both SLA forecasters yield a smaller global RMSE than the simple average. The improvements achieved by the EWA and EGA forecasters, compared with the simple average, are more apparent close to the poles and in southwestern America.

ACPD
15, 7707–7734, 2015

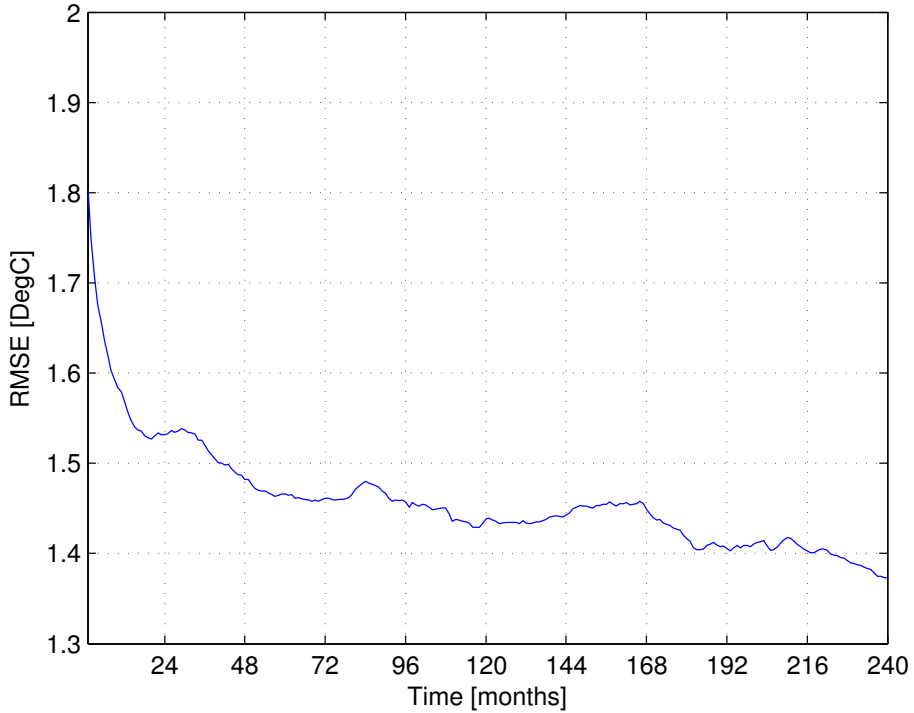

Figure 4. Global, area-weighted RMSE of the 2 m-temperature, during the 10 year validation period, as a function of the learning time. The presented RMSE was calculated for the EGA forecaster; however, a similar trend was obtained for the EWA. In general, a longer learning period improves the *forecaster* predictions.

ACPD
15, 7707–7734, 2015

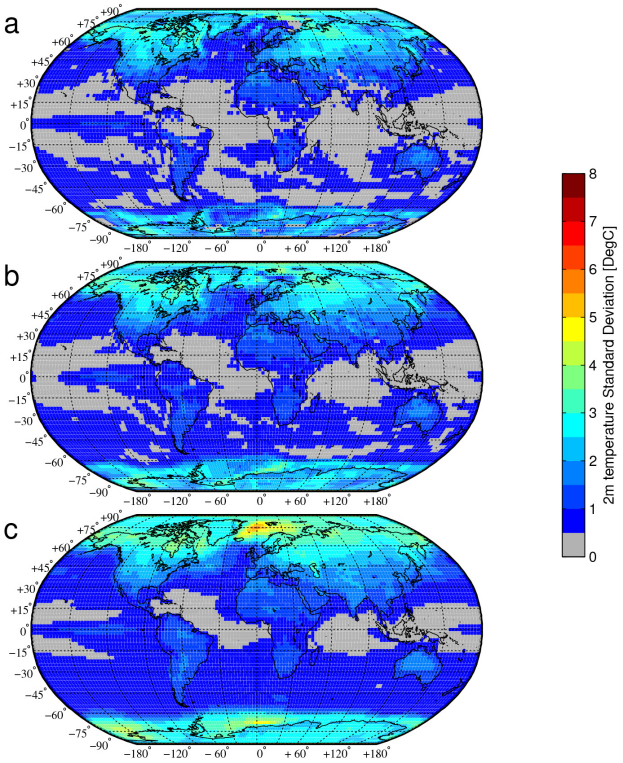

Figure 5. The 2 m-temperature uncertainty during the 10 year validation period for three forecasting methods: **(a)** EWA, **(b)** EGA, and **(c)** simple average. The uncertainties of the EWA are smaller than those of the EGA; however, the predictions of the EGA are better (see the text for a more detailed explanation). Both the EGA and the EWA forecasters yield smaller uncertainties than the simple average. The uncertainties, corresponding to the SLA forecasting schemes, are significantly reduced in regions where the uncertainties are larger, such as toward the poles and over South America and Africa.

ACPD
15, 7707–7734, 2015

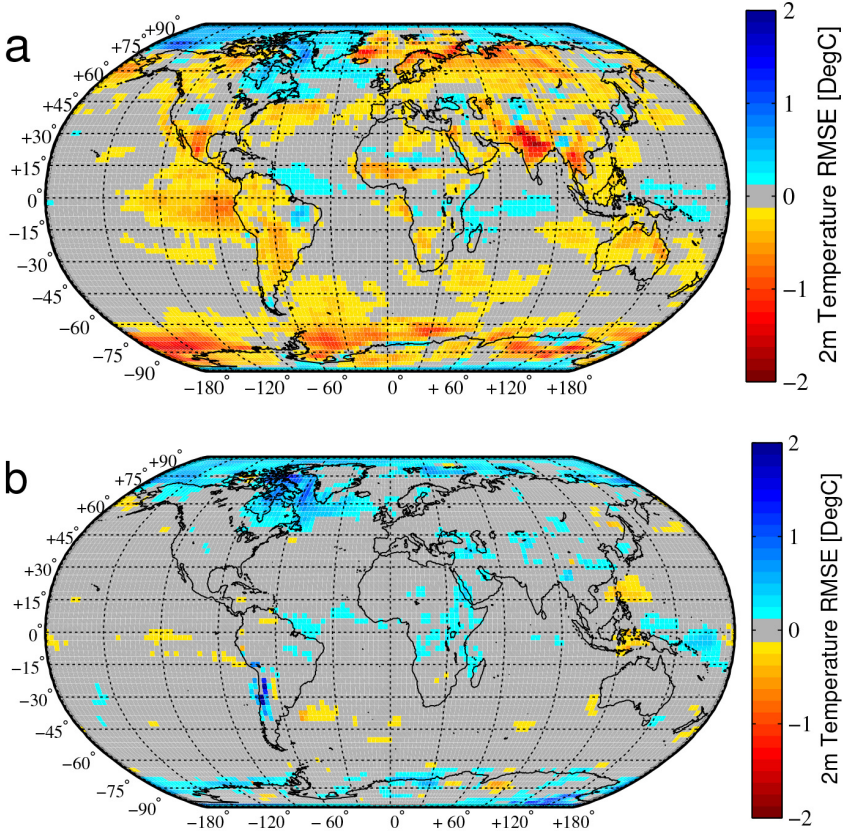

Figure 6. The difference between the 10 year validation period average 2 m-temperature RMSE of the climatology and the EGA forecaster, **(a)** EGA with an ensemble that includes eight models, **(b)** EGA with an ensemble that includes the same eight models and also the climatology of the learning period as an additional model. The results demonstrate that when the ensemble includes the climatology, the EGA forecaster is skillful.

ACPD
15, 7707–7734, 2015

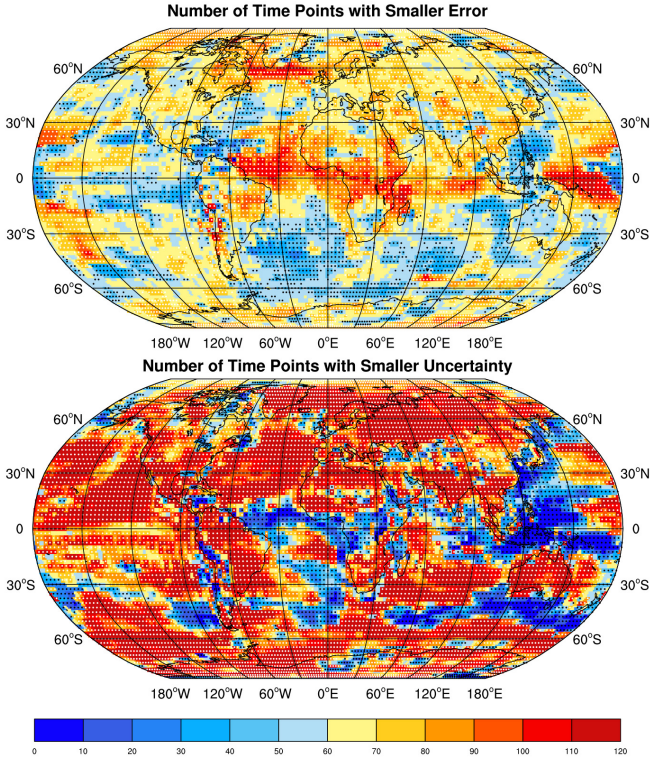

Figure 7. The number of time points in which the EGA *forecaster* performs better. The upper panel shows the spatial distribution of the number of time points in which the absolute error of the EGA *forecaster* is smaller than that of the climatology. The lower panel shows the spatial distribution of the number of time points in which the uncertainty of the EGA weighted ensemble is smaller than that of the equally weighted ensemble. White circles represent significant improvement by the EGA *forecaster* and black circles represent its significantly poorer performance. Both quantities show better performance of the EGA *forecaster* over most of the glob.

ACPD
15, 7707–7734, 2015

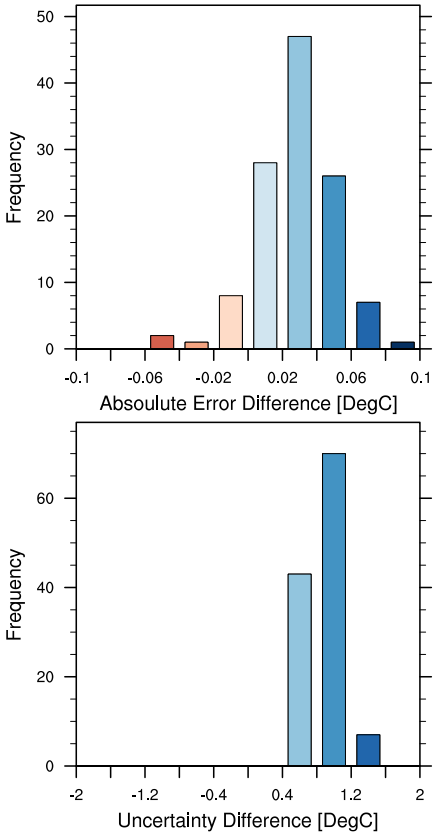

Figure 8. The histograms of the globally averaged differences of absolute error and uncertainty. The upper panel shows the histogram of the globally averaged difference between the absolute error of the climatology and that of the EGA *forecaster*. The lower panel shows the histogram of the difference between the uncertainties of equally weighted and EGA weighted ensembles. Both quantities show significantly improved performance of the EGA *forecaster*.

ACPD
15, 7707–7734, 2015