Decadal Climate Predictions Using Sequential Learning Algorithms

EHUD STROBACH AND GOLAN BEL

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

(Manuscript received 11 September 2015, in final form 28 February 2016)

ABSTRACT

Ensembles of climate models are commonly used to improve decadal climate predictions and assess the uncertainties associated with them. Weighting the models according to their performances holds the promise of further improving their predictions. Here, an ensemble of decadal climate predictions is used to demonstrate the ability of sequential learning algorithms (SLAs) to reduce the forecast errors and reduce the uncertainties. Three different SLAs are considered, and their performances are compared with those of an equally weighted ensemble, a linear regression, and the climatology. Predictions of four different variables—the surface temperature, the zonal and meridional wind, and pressure—are considered. The spatial distributions of the performances are presented, and the statistical significance of the improvements achieved by the SLAs is tested. The reliability of the SLAs is also tested, and the advantages and limitations of the different measures of the performance are discussed. It was found that the best performances of the SLAs are achieved when the learning period is comparable to the prediction period. The spatial distribution of the SLAs performance showed that they are skillful and better than the other forecasting methods over large continuous regions. This finding suggests that, despite the fact that each of the ensemble models is not skillful, they were able to capture some physical processes that resulted in deviations from the climatology and that the SLAs enabled the extraction of this additional information.

1. Introduction

Global circulation models are the main tools used to simulate future climate conditions. There are two main practices by which to initialize these models that represent predictions for two different time scales. The first practice corresponds to long-term climate projections. In this type of simulation, the climate models are initialized in the preindustrial era (aka uninitialized runs) and integrated forward in time (usually until 2100). In these simulations, the atmospheric composition in the past is set according to observations, while for the future, several representative concentration pathways (Moss et al. 2008), corresponding to different scenarios of atmospheric composition changes, are used. These climate simulations are expected to provide information about the response of the climate system to different emission scenarios by predicting the changes in the long-term averages (10 yr and more) and the statistics of climate variables under different atmospheric composition scenarios (Collins et al. 2013).

The second practice, which is considered in this work, is near-term (decadal) climate predictions intended to provide information on the dynamics of the climate system in time scales shorter than those of significant changes in the atmospheric concentration and the response time of the climate system to such changes. In this practice, the climate models are initialized with observed conditions close to the prediction period. The expected information from these simulations is the dynamics of the monthly to decadal averages of climate variables (Collins 2007; Meehl et al. 2009, 2014; Warner 2011), which is of great importance for climate services (Cane 2010). Recent studies have demonstrated a potential decadal prediction skill in different regions and for different physical processes (Smith et al. 2007; Keenlyside et al. 2008; Meehl et al. 2009, 2014; Pohlmann et al. 2009).

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-15-0648.s1.

Corresponding author address: Golan Bel, Department of Solar Energy and Environmental Physics, Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus 8499000, Israel.
E-mail: bel@bgu.ac.il

DOI: 10.1175/JCLI-D-15-0648.1

Despite their relatively short term, decadal climate predictions are still accompanied by large uncertainties, and new methods to improve the predictions and reduce the associated uncertainties are of great interest. One of the main approaches to improving climate predictions is to combine the output from an ensemble of climate models. This approach has two known advantages compared with single-model predictions. First, it was shown that the ensemble average generates improved predictions (Doblas-Reyes et al. 2000, 2003; Hagedorn et al. 2005; Palmer et al. 2004, 2000; Kim et al. 2012); second, the distribution of the ensemble member predictions can provide an estimate of the uncertainties. However, the simple average of climate simulations does not account for the quality differences between the ensemble members; therefore, it is expected that weighting the ensemble members based on their past performances will increase the forecast skill.

Uncertainties in climate predictions can be attributed to three main sources. The first is internal variability: that is, uncertainties due to different initial conditions (either different initialization times or different initialization methods) that were used to run a specific model. The second source is model uncertainties due to different predictions of different models. The third source is forcing scenario uncertainties due to different scenarios assumed for the future atmospheric composition (Hawkins and Sutton 2009). The contribution of these sources to the total uncertainty of the climate system varies with the prediction lead time and is also spatially, seasonally, and averaging-period dependent (Strobach and Bel 2015b). It was shown that, for global and regional decadal climate predictions, scenario uncertainties are negligible compared to the first two sources (Hawkins and Sutton 2009; Cox and Stephenson 2007).

There are two contributions to the internal variability: variability due to different starting conditions and variability due to different initialization methods. Uncertainties due to different starting conditions stem from the chaotic nature of the simulated climate dynamics and cannot be reduced using the ensemble approach. However, uncertainties due to different initialization methods and the model variability can be reduced by weighting the members of the ensemble. The total reduction of the uncertainty depends on the relative contribution of these sources to the total uncertainty.

Bayesian inference is one of the methods that have been used in the past to weight an ensemble of climate models. The main part of this method is the calculation of the posterior density, which is proportional to the product of the prior and the likelihood. The Bayesian method optimizes the probability density function (PDF) of the climate variable to the PDF of the data during a learning period and uses it for future predictions. It does not assign weights to the climate models; instead, it gives an estimation for the PDF of the predicted climate variable. Bayesian inference has been used extensively for projections of future climate (Buser et al. 2009, 2010; Smith et al. 2009; Tebaldi et al. 2005; Tebaldi and Knutti 2007; Furrer et al. 2007; Greene et al. 2006; Murphy et al. 2004; Räisänen et al. 2010) and also for near-term climate predictions (Rajagopalan et al. 2002; Robertson et al. 2004). The use of Bayesian inference has reduced the uncertainties of the climate projections and improved their near-term predictions. However, this method relies on many assumptions regarding the distribution of the climate variables that are not always valid, making the Bayesian inference subjective and variable dependent.

A second, and more common, method that has been used to improve climate predictions is linear regression (Feng et al. 2011; Chakraborty and Krishnamurti 2009; Doblas-Reyes et al. 2005; Fraedrich and Smith 1989; Kharin and Zwiers 2002; Krishnamurti 1999; Krishnamurti et al. 2000; Pavan and Doblas-Reyes 2000; Peña and van den Dool 2008; Peng et al. 2002; Yun et al. 2005, 2003). The linear regression method does not assign weights to the ensemble members but rather attempts to find a set of coefficients such that the scalar product of the vector of coefficients and the vector of the model predictions yields the minimal sum of squared errors relative to past observations. The same set of coefficients is then used to produce future predictions. Similarly to the Bayesian method, the regression method also relies on a few inherent assumptions, such as the normal distribution of the prediction errors (therefore, defining the optimal coefficients as those minimizing the sum of squared errors) and the independence of the ensemble member predictions.

Sequential learning algorithms (SLAs, also known as online learning) (Cesa-Bianchi and Lugosi 2006) weight ensemble members based on their past performances. These algorithms were shown to improve long-term climate predictions (Monteleoni et al. 2010, 2011) and seasonal to annual ozone concentration forecasts (Mallet et al. 2009; Mallet 2010). More recently, it was shown that decadal climate predictions of the 2-m temperature can be improved using SLAs and can even become skillful when the climatology is added as a member of the ensemble (Strobach and Bel 2015a). The SLAs have several advantages over the other ensemble methods described above. First, they do not rely on any assumption regarding the models and the distribution of the climate variables. In addition, the weights assigned to the models can be used for model evaluation and the

comparison of different parameterization schemes or initialization methods. Third, the weighted ensemble provides not only predictions but also the associated uncertainties. All these characteristics suggest that the SLAs are suitable for the improvement of various climate variable predictions.

Here, we test the performances of SLAs in predicting the previously investigated 2-m temperature (Strobach and Bel 2015a) and three additional climate variables: namely, the zonal and meridional components of the surface wind and the surface pressure. The performances of the SLAs are compared with those of the regression method. The comparison with the Bayesian method is not straightforward and is not included here. We also study the effects of different learning periods and different bias correction methods on the SLA performances. In addition, we consider a new metric, the reliability, to assess the performances of the forecasters. The SLAs are used here in a nontraditional way. Namely, the weights of the ensemble members are updated during a learning period, and the predictions are made not only for the next outcome but for the whole time series during the validation period. This type of prediction is different from previous climate predictions made using the SLAs (Monteleoni et al. 2010, 2011) and is also beyond the framework in which the SLAs are expected to perform well. The results of phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Taylor et al. 2009) decadal experiments constitute the ensemble, and the NCEP–NCAR reanalysis data (Kalnay et al. 1996) are considered as the observations. This paper is organized as follows. In section 2, we present the data that we used in this study, including the models and the reanalysis data. In addition, we discuss the different bias correction methods that we used. In section 3, we describe the SLAs and the regression forecasting methods as we implemented them. We also provide the details of the climatology that we derived from the reanalysis data. In section 5, we present the predictions of the different forecasting methods. We also evaluate their global and regional performances based on their root-mean-square errors (RMSEs). The global and regional uncertainties and reliabilities of the predictions of the different forecasting methods are presented in sections 6 and 7, respectively. The weights assigned by the SLAs to the different models and to the climatology (all the members of the ensemble) are presented in section 8. The results are discussed and summarized in section 9.

2. Models and data

The decadal experiments were introduced to the Coupled Model Intercomparison Project multimodel ensemble in its fifth phase. The objective of these experiments is to investigate the ability of climate models to produce skillful future climate predictions for a decadal time scale. The climate models in these experiments were initialized with interpolated observation data of the ocean, sea ice, and atmospheric conditions, together with the atmospheric composition (Taylor et al. 2009). The ability of these simulations to produce skillful predictions was not investigated widely, but it was shown that they can generate skillful predictions in specific regions around the world (Kim et al. 2012; Kirtman et al. 2013; Doblas-Reyes et al. 2013; Meehl et al. 2009; Pohlmann et al. 2009; Müller et al. 2012; Meehl et al. 2014; Müller et al. 2014; Kruschke et al. 2014).

The CMIP5 decadal experiments were initialized every 5 yr between 1961 and 2011 for 10-yr simulations, with three exceptional experiments that were extended to 30-yr simulations. One of these 30-yr experiments was initialized in 1981 and simulated the climate dynamics until 2011. The output of four variables from this experiment is tested here: surface temperature, zonal and meridional surface wind components, and surface pressure. In what follows, we analyze the monthly means of these variables.

Table 1 shows the eight climate models included in our ensemble. The decadal experiments of the CMIP5 include a set of runs for each of the models, differing by the starting date and the initialization scheme used. We chose, arbitrarily, the first run of each model. As long as the model variability is the main source of uncertainty, the choice of the realization should not be significant for our analysis. Indeed, it was found that, in the CMIP5 decadal experiments, the model variability is the main source of uncertainty, independent of the prediction lead time, as long as the predictions are not bias corrected. Bias correction reduces mainly the model variability; however, the contribution of the model variability remains important (Strobach and Bel 2015b).

The NCEP–NCAR reanalysis data (Kalnay et al. 1996) were used as the observation data for the learning and for the evaluation of the forecasting methods performances. We are aware of other reanalysis projects (Uppala et al. 2005; Onogi et al. 2007); however, we selected the NCEP–NCAR data based on its wide use (note that the assessment of the quality of the different reanalysis projects is subjective and is beyond the scope of this paper). The effects of using different reanalysis data are left for future research.

Bias correction

The predictions made by the climate models often suffer from inherent systemic errors (Goddard et al. 2013), and it is common to apply bias correction methods to the model outputs before analyzing them.

TABLE 1. Model availability. (Expansions of acronyms are available online at http://www.ametsoc.org/PubsAcronymList.)

Institute	Model name	Modeling center (or group)	Grid points (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 Météorologiques/Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique	128 × 256
LASG–IAP	FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	108 × 128
IPSL[a]	IPSL-CM5A-LR	L'Institut Pierre-Simon Laplace	96 × 96
MIROC	MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo),	128 × 256
	MIROC4h	National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	320 × 640
MRI	MRI-CGCM3	Meteorological Research Institute	160 × 320

[a] Not available for the U and V components of the wind.

For long-term climate projections, this procedure is more straightforward because of the available reference period. Bias correction in decadal climate predictions is not trivial not only because there is no clear reference period but also because some of these experiments are known to have a drift from the initial condition to the model's climatology during the first years of the simulation (Meehl et al. 2009).

Here, two bias correction methods and the original data were considered. The original data without any bias correction is denoted as "no correction." The first bias correction method corresponds to subtracting from each model results their average during the learning period and adding the climatological average (the average of the NCEP–NCAR reanalysis data for the same period). This method is denoted as the "average correction." The second bias correction method corresponds to subtracting from each model and for each calendar month the corresponding average during the learning period and adding the NCEP–NCAR reanalysis average for that calendar month during the same learning period. This method is denoted as the "climatology correction." The two bias correction methods described above do not account for the explicit time dependence of the bias. However, it is reasonable to assume that, for decadal climate predictions, the bias does not change considerably with time.

3. Forecasting methods

In this work, we consider three SLAs, introduced below. More thorough descriptions of the SLAs can be found in Cesa-Bianchi and Lugosi (2006) and in Monteleoni and Jaakkola (2003). We also consider the linear regression (REG) (Krishnamurti et al. 2000) method in order to compare the performances of the SLAs to the well-known regression method. The climatology (CLM) is considered here as the threshold for skillful predictions. For clarity, the equations that describe the forecasting methods omit the spatial indices. However, the forecasting schemes were applied to each of the grid cells independently, thereby allowing the spatial distribution of the weights (or the coefficients in the case of the REG) and the reference climatology. The consideration of the effect of geospatial neighborhoods (McQuade and Monteleoni 2012) is beyond the scope of this manuscript. The data that we used consists of time series of monthly means, and the weights were updated in each time step (i.e., every month) during the learning period.

a. The EWA and the EGA

The SLAs use an ensemble of "experts" (climate models), each of which provides a prediction for a future value of a climate variable, to provide a forecast of the climate variable in terms of the weighted average of the ensemble. The process is sequentially repeated, with the weights of the models being updated after each measurement according to their prediction skill. We divide the period of the model simulations into two parts. The first part is the learning (or training) period, the data of which is used to update the model weights in the manner described above, and the second part is used for validating and evaluating the forecaster performance. At the end of the learning period, the learning ends, and the weights generated by the SLA in the last learning step are used to weight the predictions of the climate models during the validation period.

The deviation of the prediction of model E $f_{E,t}$ from the observed value y_t determines the loss function, $l(f_{E,t}, y_t)$, at time t. Similarly, the loss function of the forecaster (the SLA) is determined by the deviation of its prediction p_t from the observed value at time t. The loss function is the metric used to evaluate the models

performances. In our study, we define the loss function as the square of the deviation: namely, $l(f_{E,t}, y_t) \equiv (f_{E,t} - y_t)^2$ for model E and $l(p_t, y_t) \equiv (p_t - y_t)^2$ for the forecaster.

The output of the exponentiated weighted average (EWA), the first SLA described here, at time t is the set of the weights of the models in the ensemble:

$$w_{E,t}^{\mathrm{EWA}} \equiv \frac{1}{Z_t} \times w_{E,t-1}^{\mathrm{EWA}} \times e^{-\eta \times l_{E,t}}, \tag{1}$$

where η is a positive number representing the learning rate of the forecaster, and Z_t is a normalization factor. The EWA prediction at time t is defined below:

$$p_t^{\mathrm{EWA}} \equiv \sum_{E=1}^{N_e} w_{E,t-1}^{\mathrm{EWA}} \times f_{E,t}, \tag{2}$$

where N_e is the number of models in the ensemble.

The second SLA considered here is the exponentiated gradient average (EGA). The EGA assigns the weights according to the following rules:

$$w_{E,t}^{\mathrm{EGA}} \equiv \frac{1}{Z_t} \times w_{E,t-1}^{\mathrm{EGA}} \times e^{-\eta \times l'_{E,t}}, \tag{3}$$

where $l'_{E,t}$ is the gradient of the forecaster loss function with respect to the weight of model E at time $t-1$. The mathematical definition of $l'_{E,t}$ is provided below:

$$\begin{aligned} l'(f_{E,t}, p_t^{\mathrm{EGA}}, y_t) &\equiv \frac{\partial l(p_t^{\mathrm{EGA}}, y_t)}{\partial w_{E,t-1}^{\mathrm{EGA}}} \\ &= 2 \times (p_t^{\mathrm{EGA}} - y_t) \times f_{E,t}, \end{aligned} \tag{4}$$

where the prediction of the EGA, p_t^{EGA}, is defined similarly to the prediction of the EWA:

$$p_t^{\mathrm{EGA}} \equiv \sum_{E=1}^{N_e} w_{E,t-1}^{\mathrm{EGA}} \times f_{E,t}. \tag{5}$$

An important difference between the EWA and the EGA is the fact that, under ideal conditions and stationary time series, the EWA converges to the best model in the ensemble, while the EGA converges to the observations (Strobach and Bel 2015a).

Note that, for the first learning step, one has to assign initial weights to the models. Without any a priori knowledge of the models performances, the natural choice is to assign equal weights to all the models. If the hierarchy of the models is known, it is possible to assign their initial weights accordingly.

The learning rate η was optimized by scanning a wide range of values and using the value that resulted in the minimal RMSE during the learning period. However, we added a restriction that the maximal change in the weight of each of the models, between two learning steps, will be smaller than the weight of each model in an equally weighted ensemble: namely, $1/N_e$. This restriction was added to ensure the stability of the weights. The metric that we used for this optimization is defined below:

$$M \equiv \mathrm{RMSE} \times \left(1 + \Theta\left\{\left[\max_{E=1,..,N_e, t=1,..,n} \frac{\Delta w_{E,t}}{(1/N_e)}\right] - 1\right\}\right), \tag{6}$$

where Θ represents the Heaviside theta function, and RMSE is the root-mean-squared error of the forecaster during the n time steps of the learning period. The RMSE for a grid cell (i, j) is conventionally defined.

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

The value of η that minimizes M was found using a recursive search within a very wide range of values restricted only by the machine precision. The optimization was done for each grid cell separately.

b. *The learn-α algorithm*

The basic form of the EWA was modified to explicitly allow switching between experts. This switching improves the performance of the SLA when dealing with nonstationary time series. The fixed–shared algorithm introduced by Herbster and Warmuth (1998) is defined by the following rules:

$$w_{E,t+1}^{\mathrm{FSA}} = \frac{1}{Z_t} \times \sum_{E*=1}^{N_e} w_{E,t}^{\mathrm{FSA}} \times e^{-\eta \times l_{E*,n}} \times K(E, E*), \tag{8}$$

where

$$\begin{aligned} K(E, E*; \alpha) \equiv{}& (1-\alpha) \times \delta(E, E*) \\ &+ \frac{\alpha}{N_e - 1} \times [1 - \delta(E, E*)]. \end{aligned} \tag{9}$$

Here, $\alpha \in [0, 1]$ is the switching rate parameter, and $\delta(\cdot, \cdot)$ is the Kronecker delta.

Monteleoni and Jaakkola (2003) have extended the fixed-share algorithm by also learning the optimal switching rate parameter α. This modified SLA is known as the learn-α algorithm (LAA). In the LAA, the algorithm scans a range of switching rates, $\alpha_j, j \in 1, \ldots, N_\alpha$, and assigns weights to each value of α_j based on a loss per alpha function, $l_t(\alpha_j) \equiv -\log[\sum_{E=1}^{N_e} w_{E,t}(\alpha_j) e^{-l_{E,t}}]$. The weights are updated sequentially for both the

switching rate and the experts. The updating rule for the weight of a specific value α_j is provided below:

$$W_t(\alpha_j) = \frac{1}{Z_t} W_{t-1}(\alpha_j) e^{-l_t(\alpha_j)}. \quad (10)$$

The updating rule for the weight of expert E, given α_j, is provided below:

$$w^{\mathrm{LAA}}_{E,t}(\alpha_j) = \frac{1}{Z_t(\alpha_j)} \sum_{E^*=1}^{N_e} w^{\mathrm{LAA}}_{E^*,t-1}(\alpha_j) e^{-l_{E^*,t}} K(E, E^*; \alpha_j). \quad (11)$$

The prediction at time t is the weighted average of the experts and the different values of α:

$$p^{\mathrm{LAA}}_t = \sum_{E=1}^{N_e} \sum_{j=1}^{N_\alpha} W_{t-1}(\alpha_j) \times w^{\mathrm{LAA}}_{E,t-1}(\alpha_j) \times f_{E,t}. \quad (12)$$

Here, we adopted a discretization of α to optimize the LAA performance (Monteleoni and Jaakkola 2003).

c. Regression

The linear regression algorithm considered here is described by Krishnamurti et al. (2000). In this algorithm, the forecast is a linear combination of the climate model predictions as described below:

$$p^{\mathrm{REG}}_t = \overline{y} + \sum_{E=1}^{N} a_E (f_{E,t} - \overline{f}_E). \quad (13)$$

Here, $\overline{y} \equiv (1/n)\sum_{t=1}^{n} y_t$ is the temporal mean of the observed values during the learning period (similarly, $\overline{f}_E$ is the temporal mean value of the predicted values by expert E during the learning period), and a_E are the regression coefficients minimizing the sum of squared errors during the learning period G, which is defined below:

$$G \equiv \sum_{t=1}^{n} (p_t - y_t)^2, \quad (14)$$

where n is the number of time steps in the learning period. In a small number of grid cells, some of the models predicted zero (or very close to zero) values throughout the time series. Therefore, we removed these models and applied the regression to the ensemble excluding these models.

d. Climatology

The climatology is defined here as the monthly averages of the observed conditions during the learning period. Namely,

$$C_m = \sum_{t=1}^{n_1} y_{t,m} \quad (15)$$

where $y_{t,m}$ is the observed value in month $m \in [1, 12]$ of year t (t is measured in years from the beginning of the simulations), and n_1 is the duration of the learning period in years (for simplicity, we assume here that both the learning and the validation periods span an integer number of years). The 12 months of the climatology were replicated to match the duration of the validation period; that is,

$$\mathrm{CLM}_{t,m} = C_m, \quad (16)$$

for $t \in [n_1 + 1, n_1 + n_2]$ (n_2 is the duration of the validation period in years). The climatology is often considered as the threshold for a skillful prediction (i.e., a forecaster that outperforms the climatology is considered skillful).

4. Evaluation metrics

Three main evaluation metrics are used here: the average magnitude of the error, quantified by the RMSE of each of the forecasters, the variability of the ensemble predictions, characterized by their standard deviation (STD) and the "reliability" of the predictions, characterized by the deviation of the mean-squared error from the variance (REL). The global averages of the RMSE, the STD, and the REL are calculated by weighting each grid cell by the fraction of Earth's surface it spans. The precise details are provided here for clarity. During the validation period, the RMSE of each forecaster was calculated for each grid cell [because all the climate variables studied here are two-dimensional, each grid cell has two indices (i, j)] from the time series of the forecast and the observations. The RMSE is defined as

$$\mathrm{RMSE}(i,j) = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \mathrm{ERR}^2_t(i,j)}, \quad (17)$$

where

$$\mathrm{ERR}_t(i,j) = p_t(i,j) - y_t(i,j). \quad (18)$$

The global area-weighted average of the RMSE ($\mathrm{RMSE}_{\mathrm{GAW}}$) was calculated as detailed below:

$$\mathrm{RMSE}_{\mathrm{GAW}} \equiv (1/A_{\mathrm{Earth}}) \sum_{i,j} A_{i,j} \mathrm{RMSE}(i,j), \quad (19)$$

where A_{Earth} is the total Earth surface area, and $A_{i,j}$ is the area spanned by the (i, j) grid cell. In what follows, we will present both the spatial distribution of the RMSE and its global average.

Similarly to the RMSE, the variance of the ensemble predictions was calculated for each of the grid cells at each time point:

$$\mathrm{VAR}_t(i,j) \equiv \frac{1}{N} \sum_{E=1}^{N} w_E(i,j)[f_{E,t}(i,j) - p_t(i,j)]^2. \quad (20)$$

The regression does not weight the models but instead optimizes the coefficients in a linear combination of the models. The optimization method can provide not only the optimal values of the coefficients but also the uncertainties in determining these values, and these uncertainties can be used to calculate the variance. Here, we used the method described in Ross (2014). The square root of this temporally averaged (over the validation period) variance is what we define here as the STD of each grid cell. The mathematical definition of the STD is provided below:

$$\mathrm{STD}(i,j) \equiv \sqrt{(1/n)\sum_{t=1}^{n} \mathrm{VAR}_t(i,j)}. \tag{21}$$

The global area-weighted average was then calculated:

$$\mathrm{STD}_{\mathrm{GAW}} \equiv (1/A_{\mathrm{Earth}}) \sum_{i,j} A_{i,j} \mathrm{STD}(i,j). \tag{22}$$

A common method to measure the quality of a forecast is to study its reliability. The reliability attempts to quantify the frequency in which the measured values are within the forecast range of likely values. One of these measures is the reliability score (Leutbecher and Palmer 2008), which is defined as

$$\mathrm{REL}(i,j) \equiv (1/n) \sum_{t=1}^{n} [\mathrm{ERR}_t^2(i,j) - \mathrm{VAR}_t(i,j)]. \tag{23}$$

For a reliable forecaster, REL $\rightarrow$ 0. Overconfidence (a spread smaller than the error) results in positive values, and underconfidence (a spread larger than the error) results in negative values of the REL. It is important to note that the reliability does not favor a forecaster with smaller errors as long as the spread (the range of likely values) grows in accord with the error. Therefore, all three metrics suggested here are necessary for the assessment of the forecasters' performances. The global average of the REL is defined as

$$\mathrm{REL}_{\mathrm{GAW}} \equiv (1/A_{\mathrm{Earth}}) \sum_{i,j} A_{i,j} \mathrm{REL}(i,j). \tag{24}$$

The skill of the forecasters was measured by comparing their RMSE and STD to those of some other reference forecaster. For convenience, we define below the RMSE skill score $R_{\mathrm{ref,fct}}$:

$$R_{\mathrm{ref,fct}} \equiv \frac{\mathrm{RMSE}_{\mathrm{ref}} - \mathrm{RMSE}_{\mathrm{fct}}}{\frac{1}{2}(\mathrm{RMSE}_{\mathrm{ref}} + \mathrm{RMSE}_{\mathrm{fct}})}. \tag{25}$$

The indices ref and fct are used to identify the forecasters whose skills are compared. Similarly, we define below the STD skill score $S_{\mathrm{ref,fct}}$:

$$S_{\mathrm{ref,fct}} \equiv \frac{\mathrm{STD}_{\mathrm{ref}} - \mathrm{STD}_{\mathrm{fct}}}{\frac{1}{2}(\mathrm{STD}_{\mathrm{ref}} + \mathrm{STD}_{\mathrm{fct}})}. \tag{26}$$

For the reliability, we have not defined a skill score, and we present its values explicitly in order to keep both the magnitude and the sign of the REL. Note that the skill scores are defined such that a forecaster with a smaller RMSE than the reference forecaster has a positive $R_{\mathrm{clm,fct}}$ score, and similarly, a forecaster with a smaller STD (i.e., smaller uncertainty) than the reference forecaster has a positive $S_{\mathrm{ref,fct}}$ score. The climatology was used as the reference forecaster for $R_{\mathrm{ref,fct}}$ and the equally weighted ensemble (AVR) as the reference forecaster for $S_{\mathrm{ref,fct}}$.

5. Predictions

a. Global

The simplest measure of the performance of the forecasters is the global average of the root-mean-squared error $\mathrm{RMSE}_{\mathrm{GAW}}$. Figure 1 shows the $\mathrm{RMSE}_{\mathrm{GAW}}$ of the validation period for the six different forecasters—EWA, EGA, LAA, REG, AVR, and CLM—and the different learning periods. The rows of Fig. 1 (from top to bottom) correspond to the surface temperature, zonal wind, meridional wind, and pressure, respectively. The columns of Fig. 1 (from left to right) correspond to no bias correction, average bias correction, and climatology bias correction, respectively. We found that the climatology outperforms all the other forecasters (for an ensemble that does not include the climatology) for all the learning periods and bias correction methods studied here. Therefore, we added the climatology as an expert to the ensemble and its initial weight was 0.5, whereas the initial weight of all the other models was $0.5/(N_e - 1)$ ($N_e - 1$ is the number of the models excluding the climatology). This higher initial weight of the climatology was motivated by its superior performance [as shown in Fig. 1 of the supplementary information and Strobach and Bel (2015a)]. The $\mathrm{RMSE}_{\mathrm{GAW}}$ for the ensemble without the climatology is provided in Fig. 1 and Tables 1–4 of the supplementary information. The decadal climate simulations considered here span a 30-yr period that is split such that the first part is used for learning and the second part is used for the evaluation of the performances; that is, for the 5-yr learning period, the validation period is the next 25 yr, and for the 10-yr learning period, the validation period is the next 20 yr, etc. The $\mathrm{RMSE}_{\mathrm{GAW}}$ s of the individual models are not presented because they are much higher than those of the forecasters. The $\mathrm{RMSE}_{\mathrm{GAW}}$ of the equally weighted ensemble is much lower than those of the models, but for

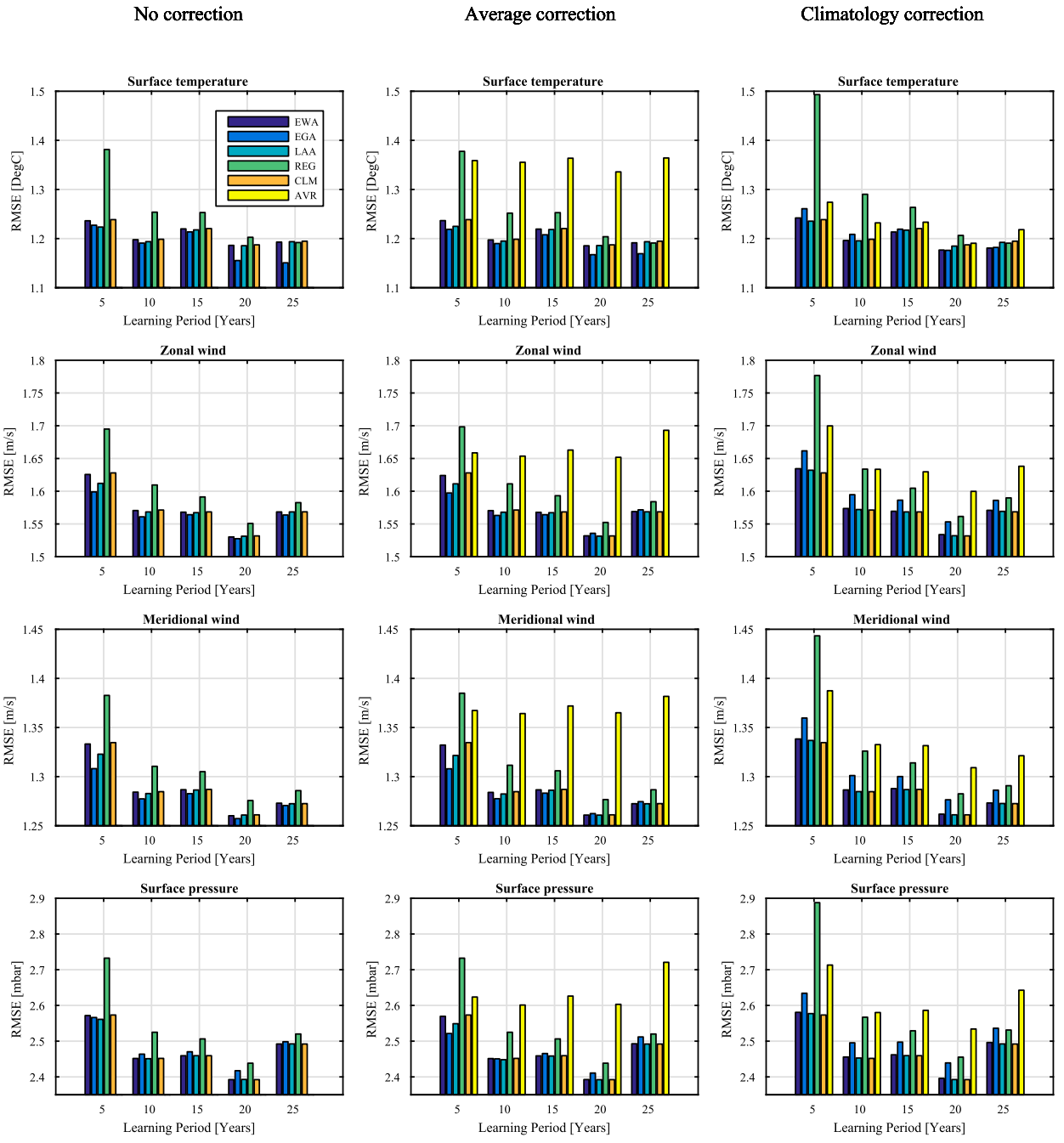

FIG. 1. Globally averaged RMSE with climatology. $RMSE_{GAW}$ for the six forecasting methods [EWA, EGA, LAA, REG, AVR, and CLM (different colors)], learning periods of 5, 10, 15, 20, and 25 yr (x axis) and (top)–(bottom) the four climate variables (surface temperature, two wind components, and pressure). The ensemble used by the forecasters includes the climatology. The columns correspond to (left) no bias correction, (center) average bias correction, and (right) the climatology bias correction. For the no correction case, the RMSEs of the AVR are too large to be included in the panels.

the case of no correction it is too high to be included within the scale shown in Fig. 1 (the relevant data can be found in Tables 5–8 of the supplementary information). The $RMSE_{GAW}$s of the EWA and LAA are less sensitive to the bias correction for the longer learning periods (compared with the EGA and the REG). In addition, we see that, for all the learning periods but the longest (25 yr), the REG is worse than the SLAs (larger $RMSE_{GAW}$). The smallest $RMSE_{GAW}$ was obtained for a learning period of 20 yr and a prediction period of 10 yr.

TABLE 2. The optimal bias correction [no correction (nbias), average correction (bias), and climatology correction (mbias)] for each forecaster and each climate variable: the surface temperature T, zonal wind U, meridional wind V, and pressure P.

	Climate variable			
Forecaster	T	U	V	P
EGA	nbias	nbias	nbias	bias
EWA	mbias	nbias	nbias	nbias
LAA	bias	bias	nbias	bias
REG	nbias	nbias	nbias	nbias
AVG	mbias	mbias	mbias	mbias

The minimal $\mathrm{RMSE}_{\mathrm{GAW}}$ (varying the learning period and the bias correction method) is obtained by the EGA for all the variables but the surface pressure. For the surface pressure, the minimal $\mathrm{RMSE}_{\mathrm{GAW}}$ (in the same sense mentioned above) is obtained by the LAA, but the value is almost equal (deviation of less than a tenth of a percent) to that of the CLM and the EWA. The data that were used to generate Fig. 1 are provided in Tables 5–8 of the supplementary information. It is important to note that the results of Fig. 1 make comparisons between predictions with different learning and prediction periods. However, in each of the experiments, we used all the data of the decadal experiments (viz., all the 30 yr simulated). The fact that the $\mathrm{RMSE}_{\mathrm{GAW}}$ is minimized after 20 yr of learning can be related to two factors: (i) for short learning periods, there is a longer prediction period and, therefore, a larger $\mathrm{RMSE}_{\mathrm{GAW}}$ and (ii) for the 25-yr learning period, the time lead from the initialization to the prediction period is long, and, in addition, the short 5-yr prediction period does not represent the climate variability over a time scale of 25 yr (the duration of the learning period). To better understand the effects of the learning and the prediction periods on the $\mathrm{RMSE}_{\mathrm{GAW}}$, we present in the supplementary information, the results of experiments varying only the learning or the prediction periods. Figure 2 of the supplementary information (the corresponding data are provided in Tables 9–12 of the supplementary information) depicts the results of an experiment in which the prediction period spanned the last 10 yr of the simulation and the learning period varied between 5 and 20 yr. For all the variables but the surface temperature, there is a decrease of the $\mathrm{RMSE}_{\mathrm{GAW}}$ of the SLAs for longer learning periods. The regression does not show a significant change in the $\mathrm{RMSE}_{\mathrm{GAW}}$ for the different learning periods. A possible explanation for the absence of a decrease in the $\mathrm{RMSE}_{\mathrm{GAW}}$ of the surface temperature for learning periods longer than 5 yr might be different trends of this variable during the learning

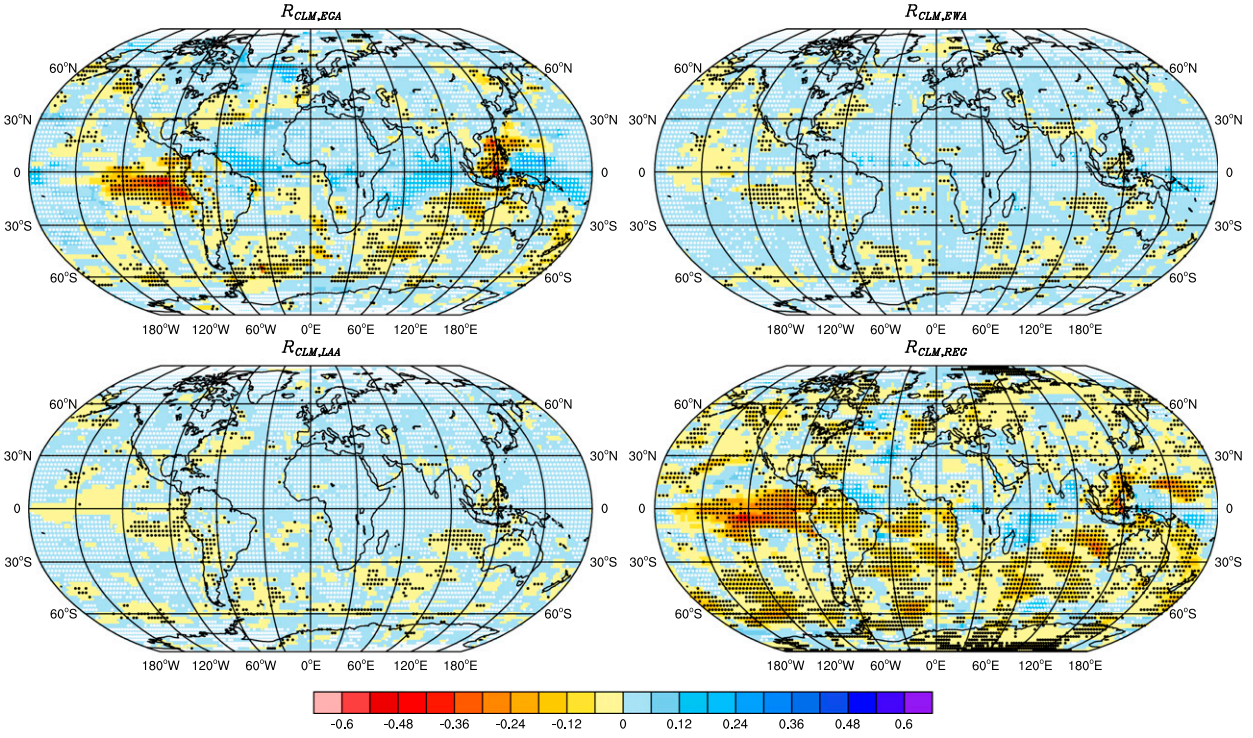

FIG. 2. Surface temperature RMSE skill score: (top left) EGA, (top right) EWA, (bottom left) LAA, and (bottom right) REG. Positive values correspond to a smaller RMSE than the climatology and vice versa. White dots represent significant improvement, and black dots represent a significantly poorer performance.

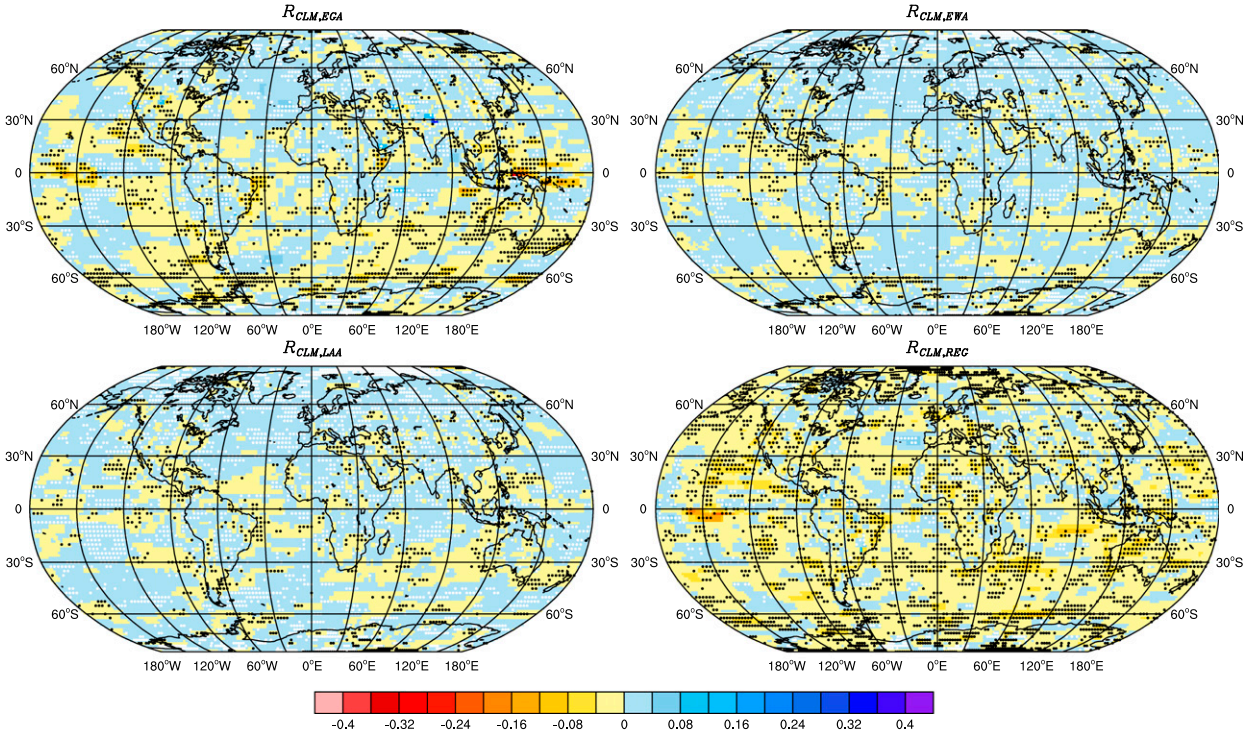

FIG. 3. As in Fig. 2, but for zonal wind.

and prediction periods. Figure 3 of the supplementary information (the corresponding data are provided in Tables 13–16 of the supplementary information) depicts the results of an experiment in which the learning period spanned the first 10 yr of the simulation and the prediction period varied between 5 and 20 yr. It shows that, in general, the longer the prediction period, the larger the $RMSE_{GAW}$ (as expected); there is a large increase in the $RMSE_{GAW}$ when the prediction period is increased from 5 to 10 yr and a smaller increase when the prediction period is increased from 15 to 20 yr. The predictions of the surface pressure show different behaviors (see the supplementary information for more details). Similar results for an ensemble that does not include the climatology are provided in Figs. 4 and 5 and Tables 17–24 of the supplementary information.

In the rest of this section, we focus on the case of 20 yr of learning and 10 yr of prediction. This learning period is chosen because it extends well beyond the drift of the models and it is also long enough to capture the simulated climate dynamics over the time scale of the prediction period. In Table 2, we detail the bias correction that resulted in the smallest $RMSE_{GAW}$ for each forecaster and for each climate variable. We find that all the SLAs have a lower or equal $RMSE_{GAW}$ than the climatology for the surface temperature and wind components. For the surface pressure, only the LAA outperforms the climatology. We also see that, for most climate variables, the $RMSE_{GAW}$s of the EWA and the climatology are almost equal. This is not a coincidence; it reflects the fact that the EWA tracks the best model, which in most grid cells is the climatology. The two other SLAs reduce the $RMSE_{GAW}$ below that of the climatology by extracting information from the other models in the ensemble. The LAA outperforms the EGA for short learning periods (<15 yr) and for all learning periods in the predictions of the surface pressure. This better performance can be attributed to the design of the LAA for the learning of nonstationary data. The poorer performance, relative to the climatology, of most of the forecasters (except for the LAA) in the prediction of the surface pressure is not fully understood. However, we found that, for the surface pressure, the variability between the models is often larger than its seasonal variability, while all the other climate variables considered here show seasonal variabilities that are larger than the variabilities between the models. It is also possible that the model predictions of the monthly mean surface pressure are worse than the predictions of the other climate variables.

b. Regional

The $RMSE_{GAW}$ is convenient because it aims to quantify the performances of the forecasters using only

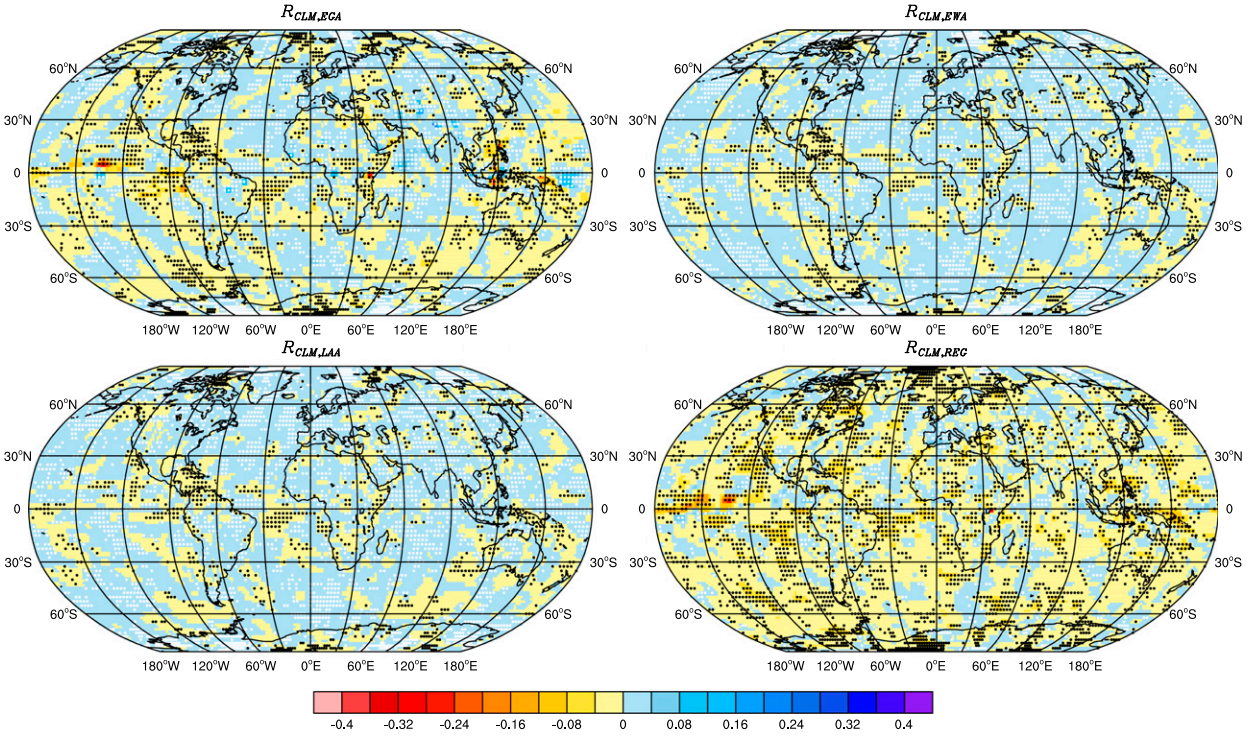

FIG. 4. As in Fig. 2, but for meridional wind.

one number. However, often the more scientifically and practically relevant information is the spatial distributions of the RMSE. We focus on the 20-yr learning period, average bias corrected data and the ensemble that includes the climatology. In this subsection, the spatial distribution of the forecaster performances will be investigated using the $R_{\mathrm{ref,fct}}$ metric defined above. This metric will allow us to compare the performances of the different forecasters and, in particular, to compare their performances to that of the trivial forecaster: the climatology. The statistical significance of the improvement achieved by the forecasters was tested by introducing the null hypothesis that the temporal distribution of $R_{\mathrm{ref,fct}}$ is symmetric around 0. Grid cells in which the hypothesis was rejected with a 90% confidence level in favor of a better forecaster performance are marked with white dots. Similarly, grid cells in which the hypothesis was rejected in favor of a poorer forecaster performance are marked with black dots. Grid cells in which the data do not provide enough evidence to reject the null hypothesis are not marked.

Figure 2 depicts the spatial distributions of $R_{\mathrm{CLM,EGA}}$ (top-left panel), $R_{\mathrm{CLM,EWA}}$ (top-right panel), $R_{\mathrm{CLM,LAA}}$ (bottom-left panel), and $R_{\mathrm{CLM,REG}}$ (bottom-right panel) for the surface temperature. This figure better clarifies the origin of the EGA's superior performance over the other forecasters (as seen from the surface temperature panels, the 20-yr learning period bins of Fig. 1). The largest variability is observed for $R_{\mathrm{CLM,EGA}}$ and the smallest variability for $R_{\mathrm{CLM,LAA}}$. While the LAA shows a positive skill score over large regions, the score is relatively low, reflecting a small improvement in the prediction compared with the climatology. For the EGA, on the other hand, we see that over regions in the North Atlantic, central Africa, the tropics of the Atlantic and Indian Oceans, and Oceania, there is a large improvement relative to the climatology, while in regions in Southeast Asia, west of Australia, and the eastern central Pacific Ocean, there is a much poorer performance compared with the climatology. The regression forecaster shows a poorer performance compared with the climatology (negative skill score) over large regions and, in particular, over the Southern Hemisphere. All the forecasters show a positive skill over regions in North Africa, Asia, and North America, suggesting that the models are capable of capturing deviations from the climatology in these regions.

The spatial distribution of the RMSE skill score for the zonal and meridional wind components is shown in Figs. 3 and 4, respectively. The errors in the predictions of both wind components have similar characteristics. The EGA shows a lower spatial variability in the errors of the wind component predictions compared with the errors of the surface temperature predictions. The EWA

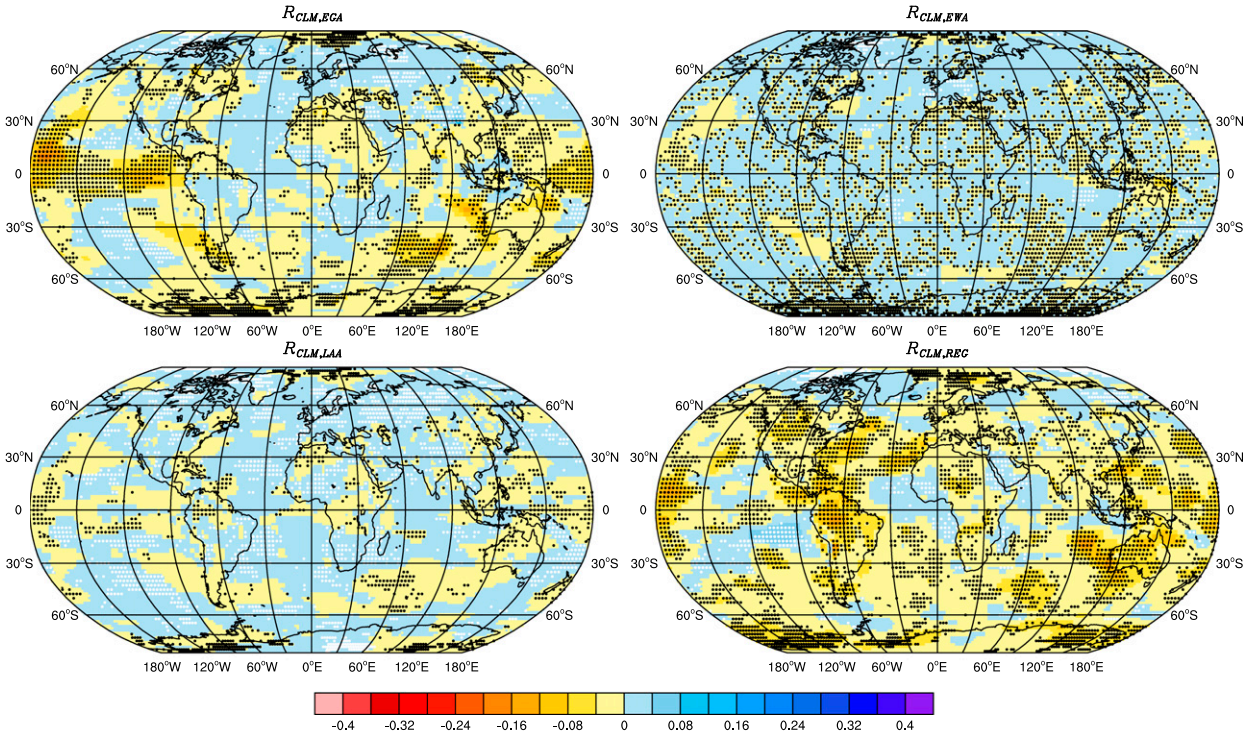

FIG. 5. As in Fig. 2, but for pressure.

and LAA show similar variability to the one found for the surface temperature. All the SLAs show smaller regions of significantly lower errors than the climatology. The REG shows a poorer performance compared with the climatology over most of the globe.

Figure 5 shows the spatial distribution of the surface pressure $R_{\mathrm{CLM,EGA}}$ (top-left panel), $R_{\mathrm{CLM,EWA}}$ (top-right panel), $R_{\mathrm{CLM,LAA}}$ (bottom-left panel), and $R_{\mathrm{CLM,REG}}$ (bottom-right panel). The EGAs performance for the surface pressure is poor compared with its performance for the other variables. Large regions in the Pacific and Indian Oceans show a larger RMSE of the EGA than the climatology, while in some regions in the Atlantic Ocean, northern Eurasia, Greenland, and the South Pacific, the EGA shows a better performance than the climatology. The EWA and LAA assign a very high weight to the climatology and, therefore, show an RMSE skill score close to zero. However, the small improvement achieved by the LAA is statistically significant over large regions. The REG shows a poorer performance than the climatology over most regions, with some exceptions in the central Atlantic and Pacific Oceans and the Arabian Peninsula.

The SLAs show a positive RMSE skill score over most of the globe for the surface temperature and wind components. The LAA shows the highest score (relative to the other forecasters) for the surface pressure. There are several regions (such as the North Atlantic, north Indian Ocean, and northern Eurasia) where all the SLAs seem to provide a smaller RMSE than the climatology. This suggests that at least some of the models capture processes that result in a deviation from the climatology and that the SLAs are capable of tracking these models.

6. Uncertainties

The RMSE is an important measure of the quality of the predictions; however, the uncertainties associated with the predictions of the forecasters are crucial for a meaningful assessment of the predictions' quality. The uncertainties are quantified here using the standard deviation of the ensemble. A natural reference for comparing the variance of the ensemble weighted by the forecasters is the variance of the equally weighted ensemble that represents no learning. It was mentioned earlier that the linear regression does not assign weights to the models in the ensemble but rather attempts to find the linear combination of their predictions that minimizes the sum of squared errors. Therefore, the variance of the regression predictions is based on the uncertainty in determining the regression coefficients. In this section, we will compare the uncertainties of the three SLAs, the regression, and the equally weighted ensemble. Our analysis proceeds similarly to the analysis of the RMSE;

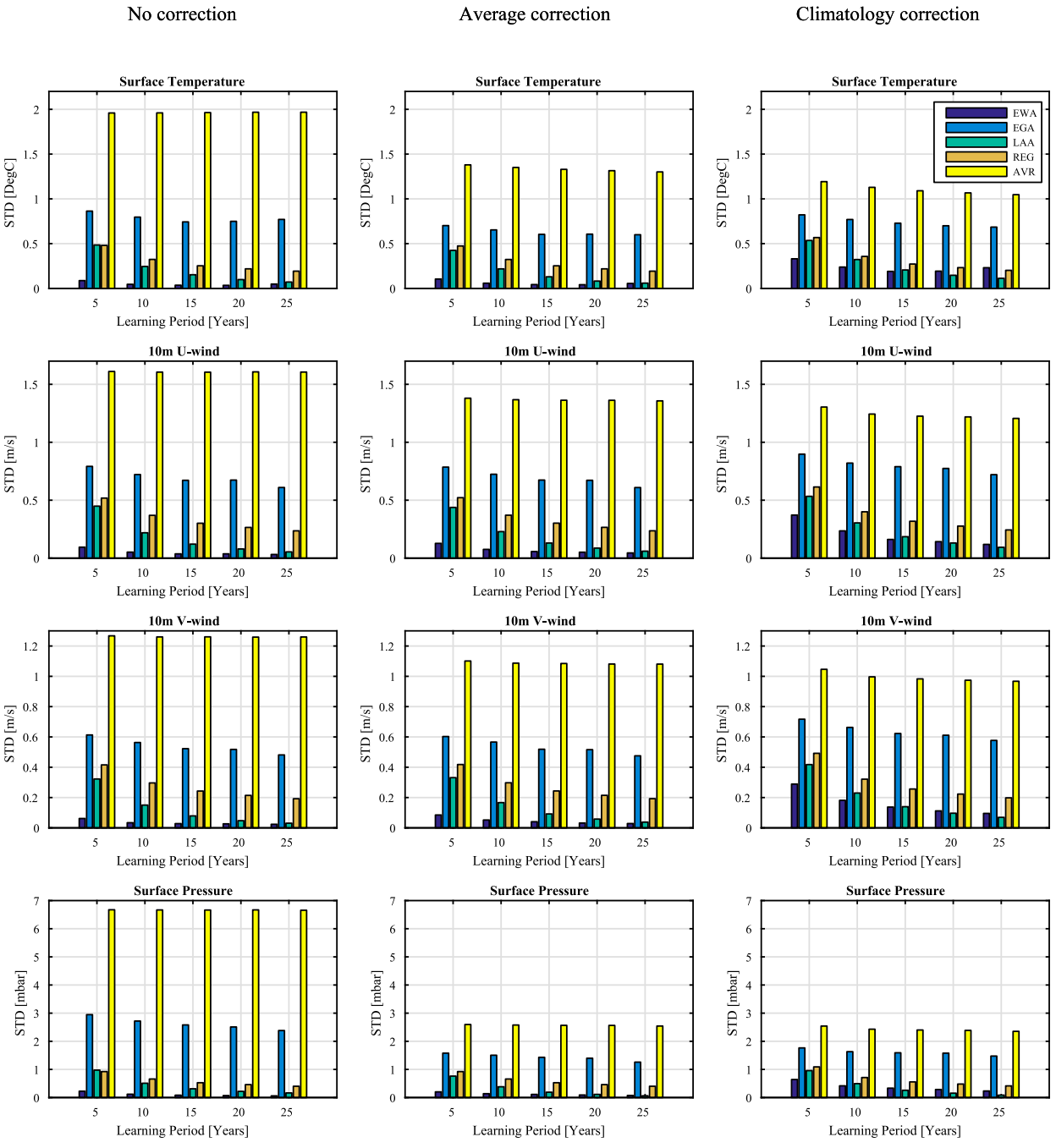

FIG. 6. As in Fig. 1, but for STD with STD_{GAW} for the three SLAs (EWA, EGA, and LAA), for the regression (REG), and for the AVR. The ensemble used by the forecasters (and AVR) includes the climatology. (See section 3a for the details of the different bias correction methods.)

first, we present the globally averaged standard deviation STD_{GAW}, and then we present the spatial distribution of the STD skill score.

a. Global

Figure 6 shows STD_{GAW} of the EGA, EWA, LAA, REG, and AVR for different learning and prediction periods and for the four climate variables considered in this study. The results of Fig. 6 were derived from an ensemble that includes the climatology. The four left panels correspond to no bias correction, the four center panels correspond to average bias correction, and the four right panels correspond to climatology bias correction. The data are provided in Tables 25–28 of the

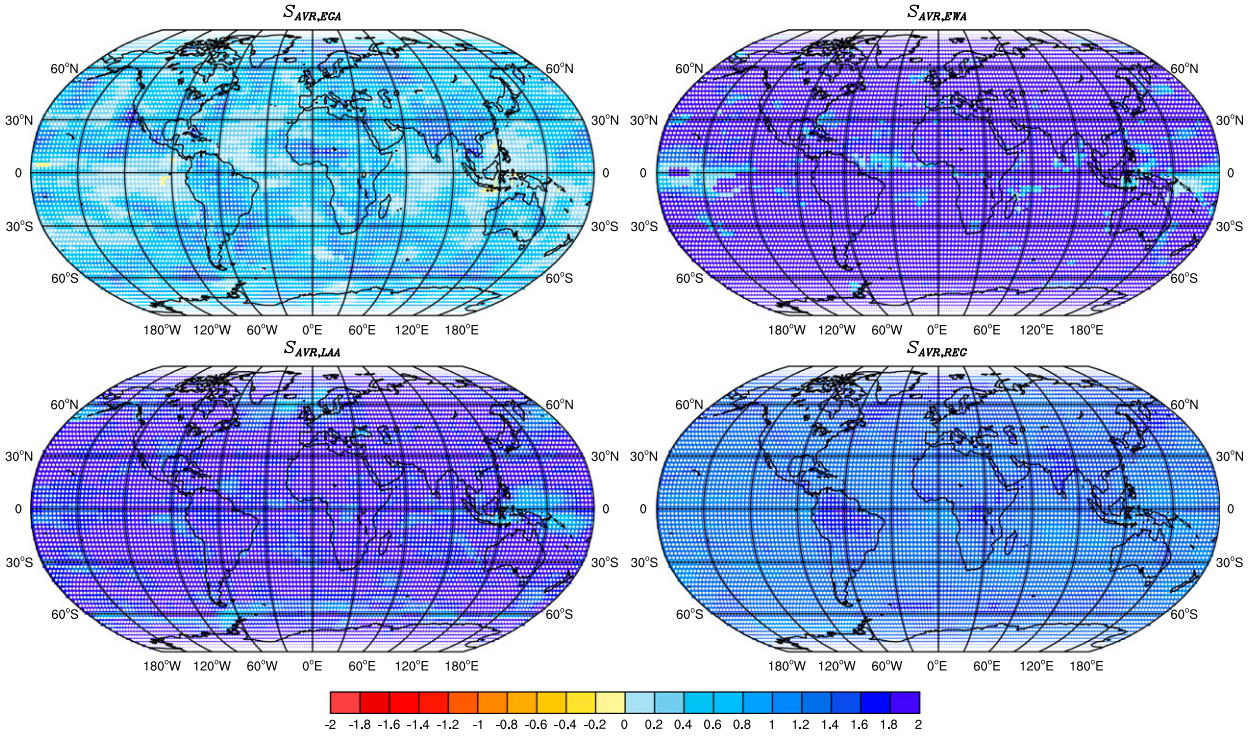

FIG. 7. Spatial distribution of the surface temperature STD skill score: (top left) EGA, (top right) EWA, (bottom left) LAA, and (bottom right) REG. Positive values correspond to a smaller STD than the equally weighted ensemble and vice versa. White dots represent a statistically significant reduction of the STD and black dots represent a statistically significant increase of the STD relative to the STD of the equally weighted ensemble.

supplementary information. The equally weighted ensemble has the largest STD_{GAW}, and the EGA has the second highest STD_{GAW} for all the variables and all bias corrections. The more detailed the bias correction, the smaller the uncertainty of the equally weighted ensemble because it is associated with the anomaly rather than with the actual prediction. For the learning periods longer than 5 yr, the regression has a higher STD_{GAW} than the EWA and LAA for all the variables and all bias corrections. In general, the LAA and EWA have a smaller STD_{GAW} (the smaller uncertainty is associated with the fact that these SLAs are designed to track the best model, if it always performed better than the others). In addition, we notice that, for the bias corrected data, the STD_{GAW} is smaller for longer learning periods, or more precisely, for shorter prediction periods. Figure 6 of the supplementary information (the corresponding data are provided in Tables 29–32 of the supplementary information) depicts the STD_{GAW} for an experiment in which the prediction period was set to the last 10 yr of the simulations and the learning period varied between 5 and 20 yr. It shows similar results to those presented in Fig. 6. Figure 7 of the supplementary information (the corresponding data are provided in Tables 33–36 of the supplementary information) depicts the STD_{GAW} for an experiment in which the learning period was set to the first 10 yr of the simulations and the prediction period varied between 5 and 20 yr. It shows that the duration of the prediction period has almost no effect on the STD_{GAW} of the EGA and the AVR, while for the EWA, LAA, and REG, the longer the prediction period, the larger the STD_{GAW}. Similar results, for an ensemble that does not include the climatology, are provided in Figs. 8–10 and Tables 37–48 of the supplementary information.

b. Regional

The uncertainty has a large spatial variability. We focus on the 20-yr learning period, average bias corrected data and the ensemble that includes the climatology. The STD skill score shows the temporally averaged variability of the ensemble weighted by the forecasters compared with that of the equally weighted ensemble during the validation period. Figure 7 shows the spatial distribution of the surface temperature STD skill score for the three SLAs and the regression. All the SLAs have a positive skill score over almost all the globe, which reflects the fact that they have smaller uncertainties than the equally weighted ensemble. Over

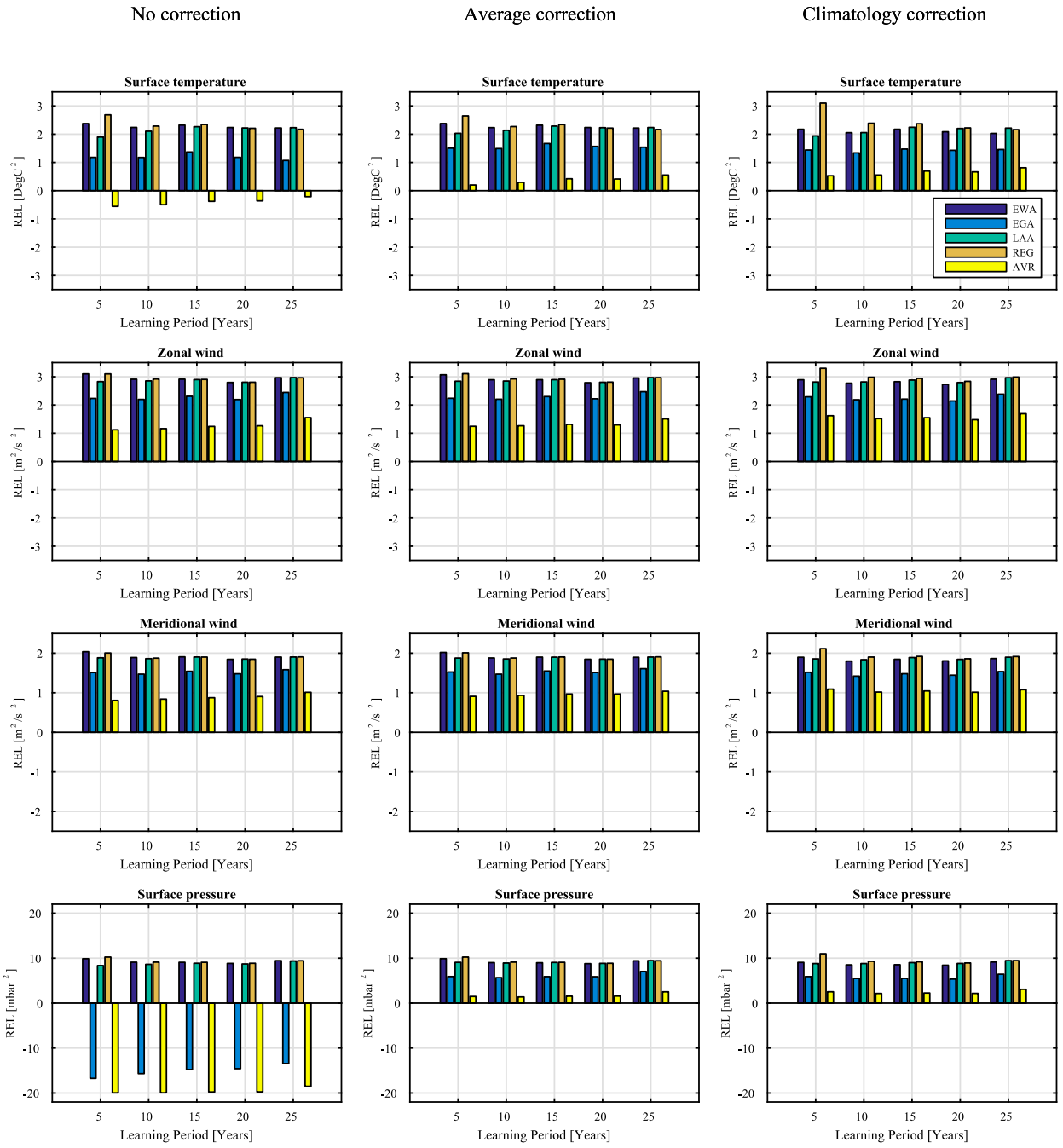

FIG. 8. As in Fig. 6, but for REL.

most of the globe, $S_{AVR,EWA}$ and $S_{AVR,LAA}$ are around 2, which reflects an almost vanishing STD of the EWA and LAA (because they assign a very high weight to the climatology, which is the best member in the ensemble considered here). The regression shows similar STD skill scores to the EWA and LAA but with somewhat lower values, which reflects a less dominant role of the climatology in its predictions. All the forecasters show similar reductions of the uncertainties for the surface wind and pressure. The results are depicted in Figs. 11–13 of the supplementary information.

7. Reliability

An ideal forecaster should have low errors and low uncertainties; however, an uncertainty that is lower than

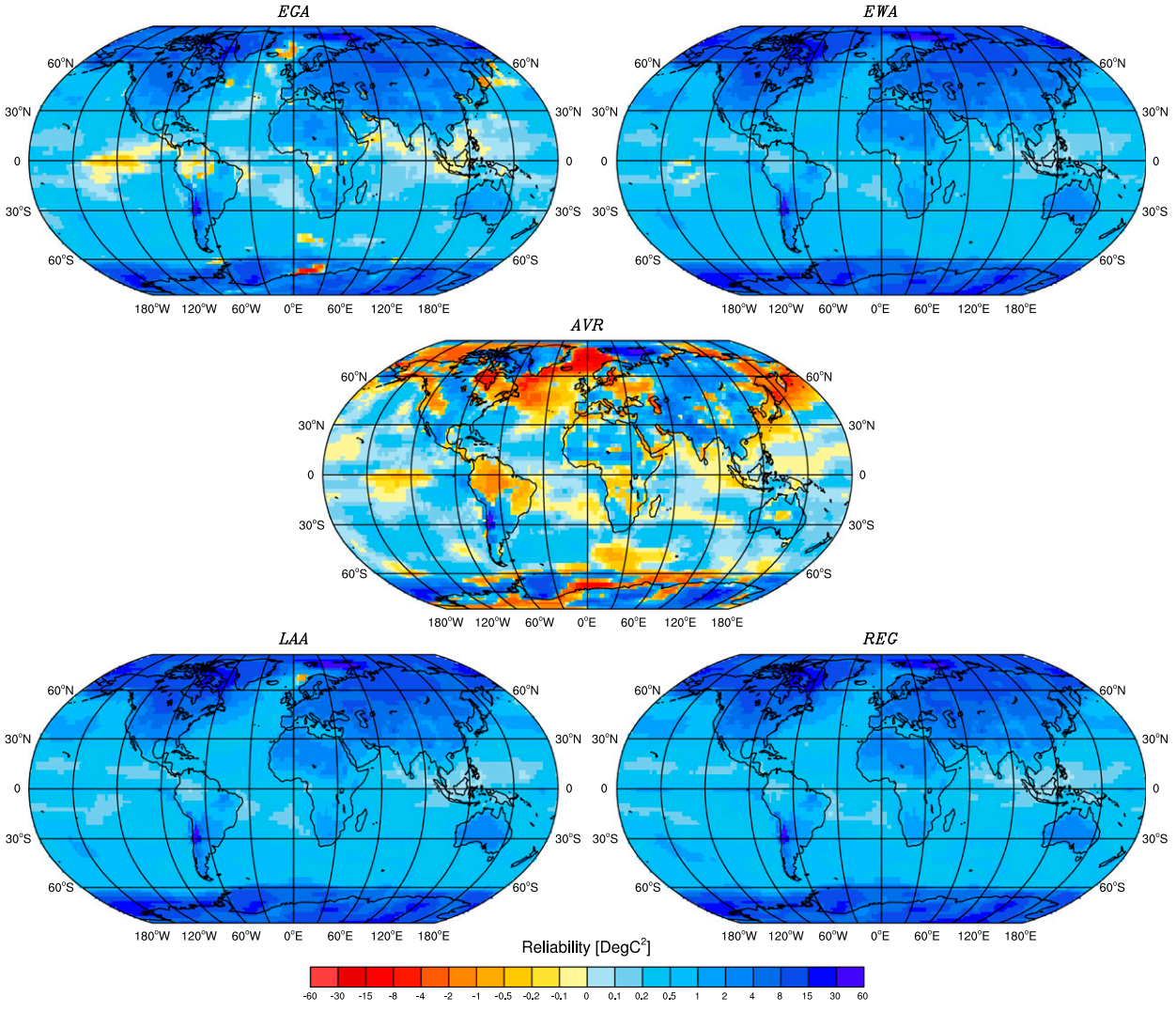

FIG. 9. Surface temperature REL: (top left) EGA, (top right) EWA, (bottom left) LAA, and (bottom right) REG; (middle) The equally weighted ensemble. Positive values indicate overconfidence and vice versa.

the error reflects an overconfident forecaster, while an uncertainty that is larger than the error reflects an underconfident forecaster. The difference between the error and the uncertainty is often used to measure the reliability of the predictions. In this work, we use the reliability score defined in section 4.

a. Global

Figure 8 shows the $\mathrm{REL}_{\mathrm{GAW}}$ of the EGA, EWA, LAA, REG, and AVR for different learning and prediction periods and for the four climate variables considered in this study. The results of Fig. 8 were derived from an ensemble that includes the climatology. The four left panels correspond to no bias correction, the four center panels correspond to average bias correction, and the four right panels correspond to climatology bias correction. The data are provided in Tables 49–52 of the supplementary information. For all the variables, bias correction methods, and prediction periods (except for the surface temperature and pressure with no bias correction), all the forecasters are overconfident; namely, the averaged variance is smaller than the averaged squared error. For the surface temperature with no bias correction, the equally weighted ensemble is underconfident, and for the surface pressure with no bias correction, both the equally weighted ensemble and the EGA are underconfident. In most cases, we find that learning of 10 yr or longer does not affect the $\mathrm{REL}_{\mathrm{GAW}}$ of the SLAs and the REG, while the $\mathrm{REL}_{\mathrm{GAW}}$ of the equally weighted ensemble shows a larger overconfidence for longer learning periods. Figure 14 of the supplementary information (the corresponding data are

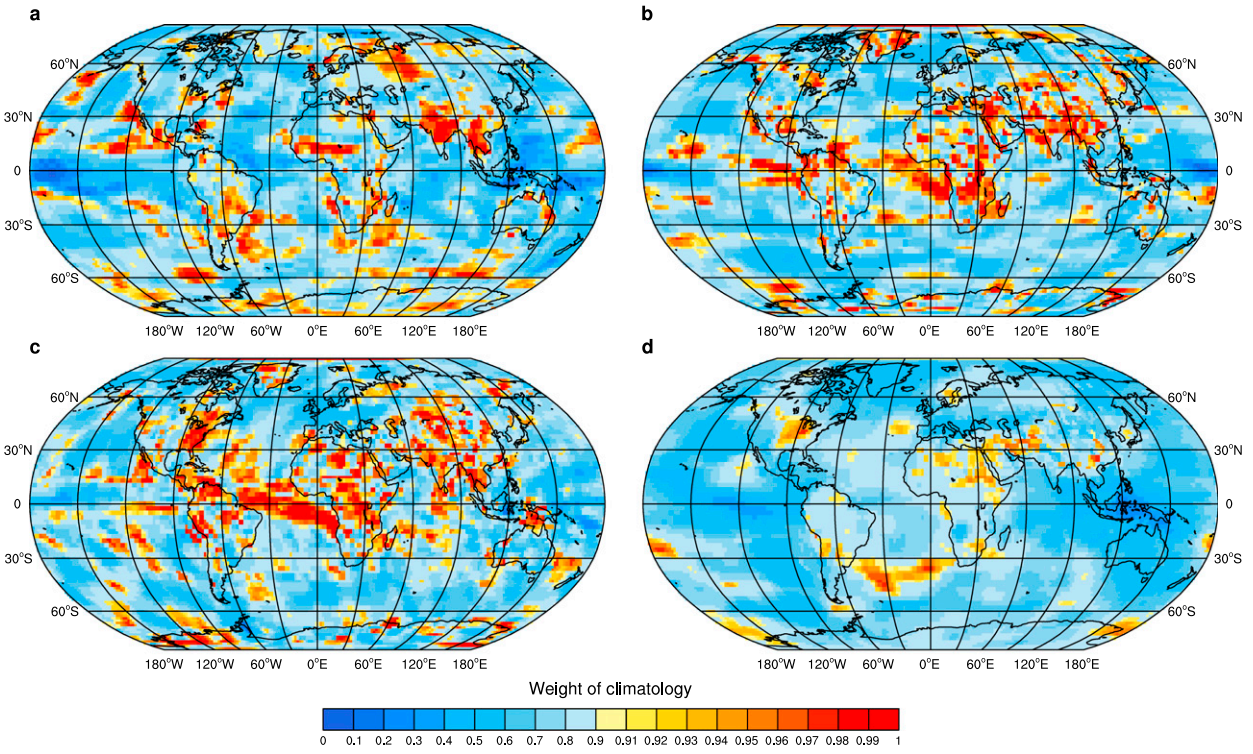

FIG. 10. Spatial distribution of the weight assigned to the climatology by the EGA forecaster for surface (a) temperature, (b) zonal wind, (c) meridional wind, and (d) pressure.

provided in Tables 53–56 of the supplementary information) depicts the REL_{GAW} for an experiment in which the prediction period was set to the last 10 yr of the simulations and the learning period varied between 5 and 20 yr. It shows similar results to those presented in Fig. 8, but the REL_{GAW} of the equally weighted ensemble does not grow for longer learning periods. Figure 15 of the supplementary information (the corresponding data are provided in Tables 57–60 of the supplementary information) depicts the REL_{GAW} for an experiment in which the learning period was set to the first 10 yr of the simulations and the prediction period varied between 5 and 20 yr. It shows similar results to those presented in Fig. 8 and Fig. 14 of the supplementary information. Similar results, for an ensemble that does not include the climatology, are provided in Figs. 16–18 and Tables 61–72 of the supplementary information.

b. Regional

The reliability also varies spatially. We focus on the 20-yr learning period, average bias corrected data, and the ensemble that includes the climatology. The REL shows the temporal average of the reliability of the ensemble weighted by the forecasters during the validation period. Figure 9 shows the spatial distribution of the surface temperature REL for the three SLAs, the regression, and the equally weighted ensemble. The equally weighted ensemble is seen to have higher reliability than the other forecasters over most of the globe. The SLAs and the regression are mostly overconfident; namely, the magnitude of the error is larger than the estimated standard deviation of the results. All the SLAs show higher reliability in the tropics and lower reliability in the midlatitudes and toward the poles. In Figs. 19–21 of the supplementary information, we depict the spatial distribution of the forecasters' reliability for the surface wind and pressure. In all the variables, the equally weighted ensemble shows higher reliability than the other forecasters. For the surface wind components, we find higher reliability over land (except for Antarctica) and lower reliability over the oceans. The surface pressure shows a spatial distribution that resembles the one observed for the surface temperature.

8. Climatology weights

Some of the results above regarding the skill of the forecasters were explained by the weights assigned to the climatology. Because of its superior performance, compared with each of the models in the ensemble, it is expected that the SLAs would assign it a high weight.

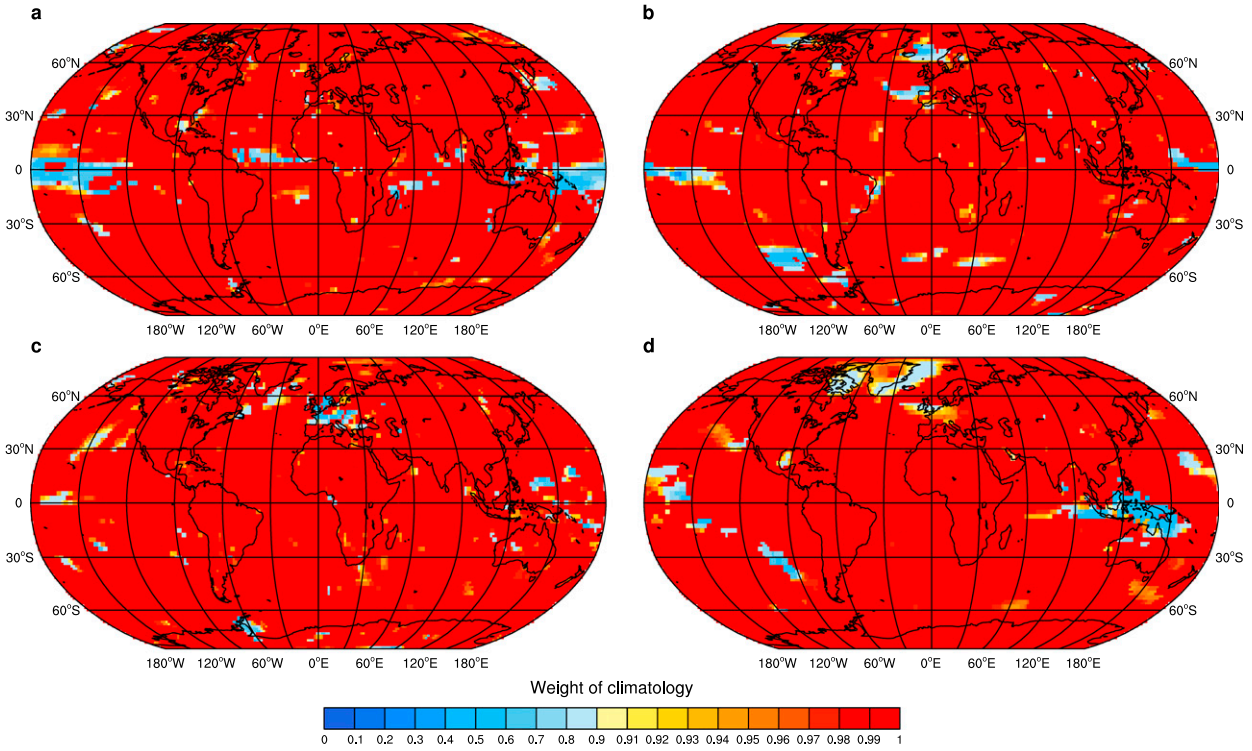

FIG. 11. As in Fig. 10, but for the EWA forecaster.

However, assigning too high a weight to the climatology implies that the forecaster is not capable of capturing deviations from the climatology because of the physical processes captured in the models. Ideally, forecasters should balance between the smaller RMSE of the climatology and the additional information available from the other models.

Figures 10, 11, and 12 show the spatial distribution of the weight assigned to the climatology, for each of the four climate variables, by the EGA, EWA, and LAA, respectively. The weights in these figures correspond to the weights assigned at the end of the 20-yr learning period (i.e., the weights used for the predictions). The average bias correction was applied to the data. The color bar was set to emphasize the differences. The EWA assigns the climatology weights close to 1 over most of the globe for the variables considered here. In the east Pacific tropics regions, the climatology is not the only dominant model in the EWA predictions of the surface temperature, zonal wind, and pressure. The LAA also assigns very high weights to the climatology over most of the globe for all the variables. The exceptions here are the region between the southern westerlies and polar easterlies and the northern Atlantic, in which the weight of the climatology is lower in the predictions of the surface temperature, and the Pacific and Atlantic tropics, in which the weight of the climatology is lower in the LAA predictions of the surface zonal wind. Both the weights assigned by the EWA and those assigned by the LAA stem from the fact that these SLAs are designed to track the best expert, which in our ensemble turns out to be the climatology over most of the globe. Note that the EWA tracks the model that had the lowest cumulative loss during the whole learning period, while the LAA tracks the model that had the smallest cumulative loss during the last part of the learning period. In addition, while the LAA assigns a weight that is almost 1 to the best model, the LAA assigns a lower weight if the values of α that are different from zero obtained a nonzero weight. The EGA assigns lower weights than the EWA and LAA to the climatology over most of the globe for all the climate variables considered here. For the surface temperature, there are still large regions in which the predictions of the EGA are dominated by the climatology. For the surface wind components, the regions dominated by the climatology have some overlap with the regions dominating the EGA's surface temperature predictions. For these variables, the climatology dominates the EGA's predictions over large regions in the tropics and southern Asia. The weight assigned to the climatology by the EGA for the surface pressure is lower in most regions and resulted in a somewhat poorer performance by the EGA in the predictions of this variable. This different performance for

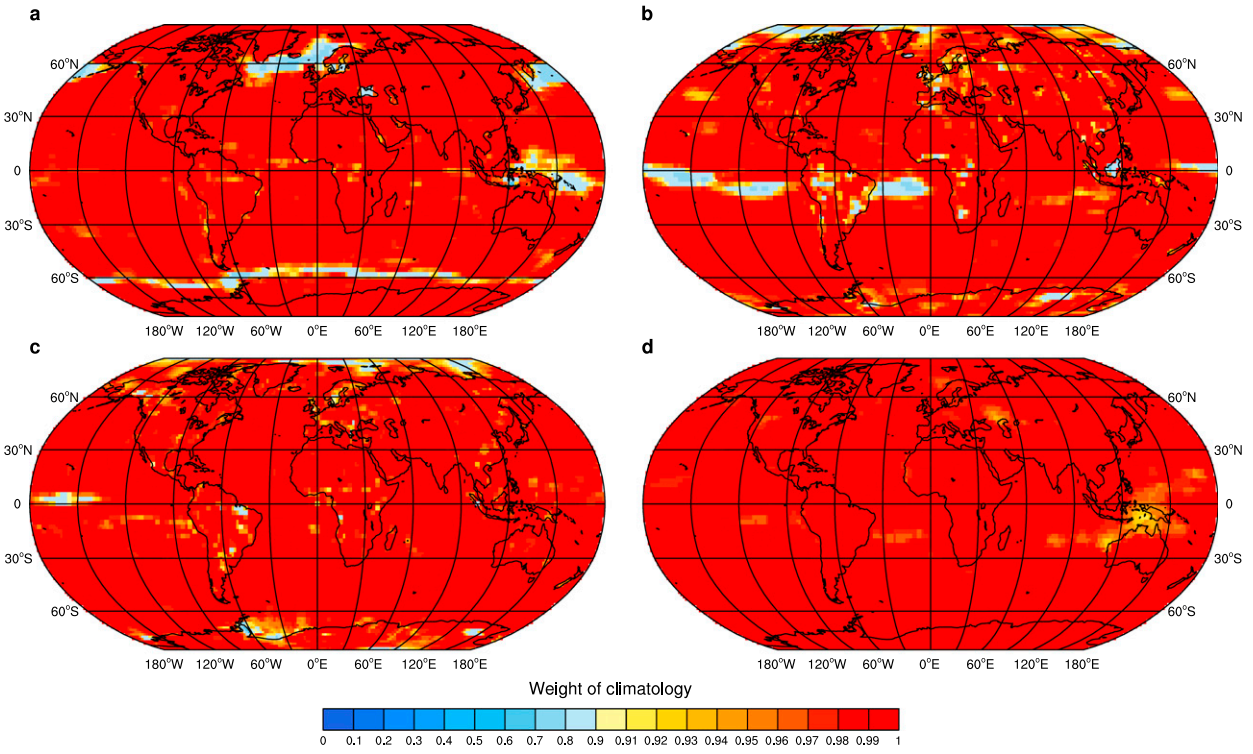

FIG. 12. As in Fig. 10, but for the LAA forecaster.

the surface pressure may be related to the lower quality of the data for this variable. Unlike the EWA and the LAA, the EGA is not designed to track the best expert, but rather to track the measurements. Therefore, the lower weight assigned to the climatology suggests that useful information can be extracted from the models, and their ability to capture some of the processes affecting the climate dynamics in decadal time scales can be quantified by the weight assigned to them by the EGA.

The regression does not assign weights to the models in the ensemble. However, one can try to quantify the significance of the climatology by studying the ratio between the magnitude of the climatology coefficient and the coefficients of the other models. Figure 13 shows the spatial distribution of $\sqrt{a_C^2/\sum_{E=1}^{N} a_E^2}$, where a_E is the coefficient of the Eth member of the ensemble, and a_C is the coefficient of the climatology (see section 3 for more details regarding the regression forecasting method). Similarly to the other forecasters, the climatology dominates the REG predictions over most of the globe, except for some regions in the tropics.

9. Summary and discussion

An ensemble of climate models is known to improve climate predictions and to help better assess the uncertainties associated with them. In this paper, we tested five different methods to combine the results of the decadal predictions of different models: EWA, EGA, LAA, REG, and the equally weighted ensemble. The first three forecasters represent learning algorithms that weight the ensemble models according to their performances during a learning period. The REG attempts to find the linear combination of the model predictions that minimizes the sum of squared errors during the learning period, and the equally weighted ensemble represents no learning.

The learning algorithms were used here to update the weights during the learning period, and the predictions, for the whole time series of the validation period, were made using the weights assigned to the ensemble models at the end of the learning period. This use of the SLAs is different from previous studies and is also beyond the framework in which the SLAs are guaranteed to perform well. Nevertheless, we found that the SLAs performed very well and showed both global and regional skill even in predicting time series that extend long after the learning has ended.

We tried different learning periods and found that a learning period that is at least as long as the prediction period yields better results. In our experiments, learning periods longer than 10 yr ensure that the learning exceeds well beyond the drift of the models. The globally averaged root-mean-squared error, $\mathrm{RMSE}_{\mathrm{GAW}}$, is

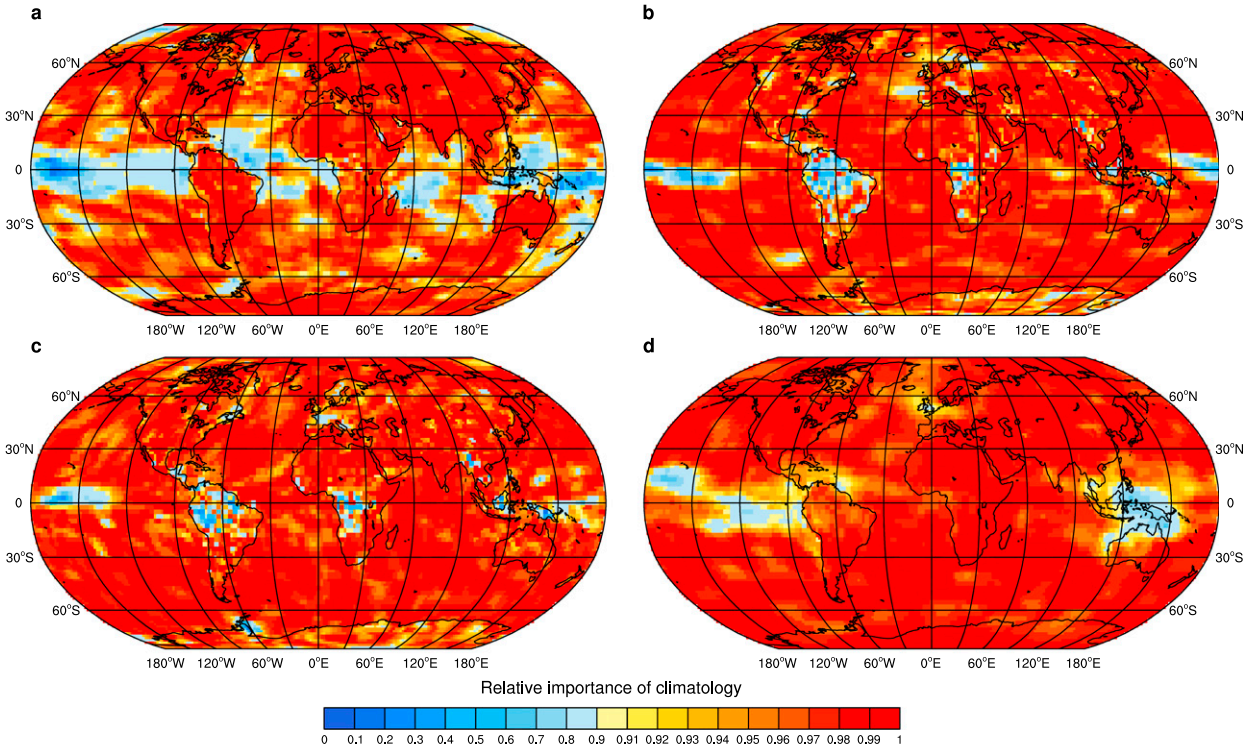

FIG. 13. As in Fig. 10, but for the relative significance of the climatology in the predictions by the REG forecaster. See section 8 for the exact definition of the significance.

smaller for 10–20 yr of learning, which ensures a long enough learning period and a not too short prediction period. The globally averaged spread of the weighted ensemble predictions, STD_{GAW}, is smaller for shorter prediction periods, as expected. The globally averaged reliability is less sensitive to the duration of the learning and prediction periods. The predictions of the surface temperature, wind, and pressure were studied, and their qualities were assessed.

The simple average was shown to have larger errors and larger uncertainties than the forecasters that used a learning period to weight–combine the model predictions. Over most of the globe, the SLAs performed better than the regression in terms of the metrics we used to quantify the forecasters' performances. This poorer performance of the regression is associated with the basic assumptions of the linear regression and its oversimplified method to linearly combine the model predictions. The SLAs do not rely on these assumptions and use more advanced methods to weight the models, resulting in smaller errors. The EWA and the LAA were found to be more appropriate in cases in which tracking of the best model is of interest. The climatology outperformed all the other models in the ensemble; therefore, the EWA and the LAA converged to it over most of the globe and for all the four climate variables. The equally weighted ensemble was shown to be underconfident in most cases, while all the other forecasters were found to be overconfident. However, the measure used for the reliability does not favor forecasters with smaller errors and uncertainties but only forecasters with uncertainties that are close to their errors. Therefore, we believe that a more appropriate reliability score should be defined. However, it is beyond the scope of this work.

Although the globally averaged RMSE of the SLAs is only a few percentage points smaller than that of the climatology, it was shown to be statistically significant. In addition, we found that, in many regions, the improvement is larger. The spatial distribution of the SLAs' performance showed that they are skillful over large continuous regions. This finding suggests that the models were able to capture some physical processes that resulted in deviations from the climatology and that the SLAs enabled the extraction of this additional information. Similarly, the large regions over which the climatology outperforms the forecasters may suggest that physical processes, associated with the climate dynamics affecting these regions, are not well captured by the models. The SLAs' performances were much poorer for the surface pressure than for the other variables. This poorer performance might be related to the quality of the models output or to the large fluctuations of this

variable. The better predictions of the EWA and LAA (relative to the EGA) for the surface pressure result from their tracking of the climatology; therefore, it is difficult to extract from their predictions new information regarding the physics of the climate system. The reduction of the uncertainties, relative to the equally weighted ensemble, is much more substantial than the reduction of the errors and can reach to about 60%–70%, globally. The uncertainties considered here are only those associated with the model variability within the ensemble. The internal uncertainties, scenario uncertainties, and other sources of uncertainty were not studied here.

The results presented here are in agreement with previous results [see Meehl et al. (2009) and references therein]. However, in this work, monthly means were considered, whereas in previous works, the averages of longer periods, which have smaller fluctuations, were considered. Smaller errors than the climatology (i.e., skillful predictions) of the SLAs can be observed in the North Atlantic, north Indian Ocean, northern Eurasia, and some regions in the Pacific Ocean. In addition, the SLAs showed predictive skill for the surface temperature over many land areas, such as northern Eurasia, Greenland, and, to some extent, also the Americas. The results suggest that learning algorithms can be used to improve climate predictions and to reduce the uncertainties associated with them.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant [293825]. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for the CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For the CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. E.S. wishes to acknowledge a fellowship from the Israel Water Authority.

REFERENCES

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

——, ——, and C. Schär, 2010: Bayesian multi-model projections of climate: Generalization and application to ENSEMBLES results. *Climate Res.*, **44**, 227–241, doi:10.3354/cr00895.

Cane, M. A., 2010: Climate science: Decadal predictions in demand. *Nat. Geosci.*, **3**, 231–232, doi:10.1038/ngeo823.

Cesa-Bianchi, N., and G. Lugosi, 2006: *Prediction, Learning, and Games.* Cambridge University Press, 408 pp.

Chakraborty, A., and T. N. Krishnamurti, 2009: Improving global model precipitation forecasts over India using downscaling and the FSU superensemble. Part II: Seasonal climate. *Mon. Wea. Rev.*, **137**, 2736–2757, doi:10.1175/2009MWR2736.1.

Collins, M., 2007: Ensembles and probabilities: A new era in the prediction of climate change. *Philos. Trans. Roy. Soc. London*, **A365**, 1957–1970, doi:10.1098/rsta.2007.2068.

——, and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136, doi:10.1017/CBO9781107415324.024.

Cox, P., and D. Stephenson, 2007: A changing climate for prediction. *Science*, **317**, 207–208, doi:10.1126/science.1145956.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelievre, 2000: Multimodel spread and probabilistic seasonal forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.*, **126**, 2069–2087, doi:10.1256/smsqj.56704.

——, V. Pavan, and D. B. Stephenson, 2003: The skill of multimodel seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dyn.*, **21**, 501–514, doi:10.1007/s00382-003-0350-4.

——, R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

——, and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, doi:10.1038/ncomms2704.

Feng, J., D.-K. Lee, C. Fu, J. Tang, Y. Sato, H. Kato, J. L. Mcgregor, and K. Mabuchi, 2011: Comparison of four ensemble methods combining regional climate simulations over Asia. *Meteor. Atmos. Phys.*, **111**, 41–53, doi:10.1007/s00703-010-0115-7.

Fraedrich, K., and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting. *J. Climate*, **2**, 291–294, doi:10.1175/1520-0442(1989)002<0291:CPSILR>2.0.CO;2.

Furrer, R., S. R. Sain, D. Nychka, and G. A. Meehl, 2007: Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environ. Ecol. Stat.*, **14**, 249–266, doi:10.1007/s10651-007-0018-z.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, doi:10.1007/s00382-012-1481-2.

Greene, A. M., L. Goddard, and U. Lall, 2006: Probabilistic multimodel regional temperature change projections. *J. Climate*, **19**, 4326–4343, doi:10.1175/JCLI3864.1.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1107, doi:10.1175/2009BAMS2607.1.

Herbster, M., and M. K. Warmuth, 1998: Tracking the best expert. *Mach. Learn.*, **32**, 151–178, doi:10.1023/A:1007424614876.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84–88, doi:10.1038/nature06921.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799, doi:10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **39**, L10701, doi:10.1029/2012GL051644.

Kirtman, B., and Coauthors, 2013: Near-term climate change: Projections and predictability. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 953–1028, doi:10.1017/CBO9781107415324.023.

Krishnamurti, T. N., 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi:10.1126/science.285.5433.1548.

——, C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Kruschke, T., H. Rust, C. Kadow, G. Leckebusch, and U. Ulbrich, 2014: Evaluating decadal predictions of northern hemispheric cyclone frequencies. *Tellus*, **66A**, 22830, doi:10.3402/tellusa.v66.22830.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Mallet, V., 2010: Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. *J. Geophys. Res.*, **115**, D24303, doi:10.1029/2010JD014259.

——, G. Stoltz, and B. Mauricette, 2009: Ozone ensemble forecast with machine learning algorithms. *J. Geophys. Res.*, **114**, D050307, doi:10.1029/2008JD009978.

McQuade, S., and C. Monteleoni, 2012: Global climate model tracking using geospatial neighborhoods. *Proc. 26th AAAI Conf. on Artificial Intelligence*, Toronto, Canada, Association for the Advancement of Artificial Intelligence, 335–341.

Meehl, G. A., and Coauthors, 2009: Decadal prediction. *Bull. Amer. Meteor. Soc.*, **90**, 1467–1485, doi:10.1175/2009BAMS2778.1.

——, and Coauthors, 2014: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243–267, doi:10.1175/BAMS-D-12-00241.1.

Monteleoni, C., and T. Jaakkola, 2003: Online learning of non-stationary sequences. *Adv. Neural Inf. Process. Syst.*, **16**, 1093–1100.

——, G. A. Schmidt, and S. Saroha, 2010: Tracking climate models. *Proc. NASA Conf. on Intelligent Data Understanding*, Mountain View, CA, NASA, 1–15.

——, ——, ——, and E. Asplund, 2011: Tracking climate models. *Stat. Anal. Data Min.*, **4**, 372–392, doi:10.1002/sam.10126.

Moss, R., and Coauthors, 2008: Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. IPCC Expert Meeting Rep., 25 pp. [Available online at https://www.ipcc.ch/pdf/supporting-material/expert-meeting-ts-scenarios.pdf.]

Müller, W. A., and Coauthors, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.*, **39**, L22707, doi:10.1029/2012GL053326.

——, H. Pohlmann, F. Sienz, and D. Smith, 2014: Decadal climate predictions for the period 1901–2010 with a coupled climate model. *Geophys. Res. Lett.*, **41**, 2100–2107, doi:10.1002/2014GL059259.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.

Onogi, K., and Coauthors, 2007: The JRA-25 Reanalysis. *J. Meteor. Soc. Japan*, **85**, 369–432, doi:10.2151/jmsj.85.369.

Palmer, T. N., Č. Branković, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.*, **126**, 2013–2033, doi:10.1002/qj.49712656703.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, doi:10.1175/BAMS-85-6-853.

Pavan, V., and F. J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over the Euro-Atlantic: Skill scores and dynamic features. *Climate Dyn.*, **16**, 611–625, doi:10.1007/s003820000063.

Peña, M., and H. van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate*, **21**, 6521–6538, doi:10.1175/2008JCLI2226.1.

Peng, P., A. Kumar, H. van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, **107**, 4710, doi:10.1029/2002JD002712.

Pohlmann, H., J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926–3938, doi:10.1175/2009JCLI2535.1.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi, 2010: Weighting of model results for improving best estimates of climate change. *Climate Dyn.*, **35**, 407–422, doi:10.1007/s00382-009-0659-8.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.*, **130**, 1792–1811, doi:10.1175/1520-0493(2002)130<1792:CCFTRA>2.0.CO;2.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744, doi:10.1175/MWR2818.1.

Ross, S. M., 2014: Regression. *Introduction to Probability and Statistics for Engineers and Scientists*, 5th ed. S. M. Ross, Ed., Academic Press, 357–444, doi:10.1016/B978-0-12-394811-3.50009-5.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799, doi:10.1126/science.1139540.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *J. Amer. Stat. Assoc.*, **104**, 97–116, doi:10.1198/jasa.2009.0007.

Strobach, E., and G. Bel, 2015a: Improvement of climate predictions and reduction of their uncertainties using learning algorithms. *Atmos. Chem. Phys.*, **15**, 8631–8641, doi:10.5194/acp-15-8631-2015.

——, and ——, 2015b: The contribution of internal and model variabilities to the uncertainty in CMIP5 decadal climate predictions. 35 pp. [Available online at http://arxiv.org/pdf/1508.01609v1.pdf.]

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the CMIP5 experiment design, 33 pp. [Available online at http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.]

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **A365**, 2053–2075, doi:10.1098/rsta.2007.2076.

——, R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate*, **18**, 1524–1540, doi:10.1175/JCLI3363.1.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, doi:10.1256/qj.04.176.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction.* Cambridge University Press, 550 pp.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate*, **16**, 3834–3840, doi:10.1175/1520-0442(2003)016<3834:IOTMST>2.0.CO;2.

——, ——, A. K. Mitra, T. S. V. V. Kumar, W. Dewar, and T. N. Krishnamurti, 2005: A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. *Tellus*, **57A**, 280–289, doi:10.1111/j.1600-0870.2005.00131.x.