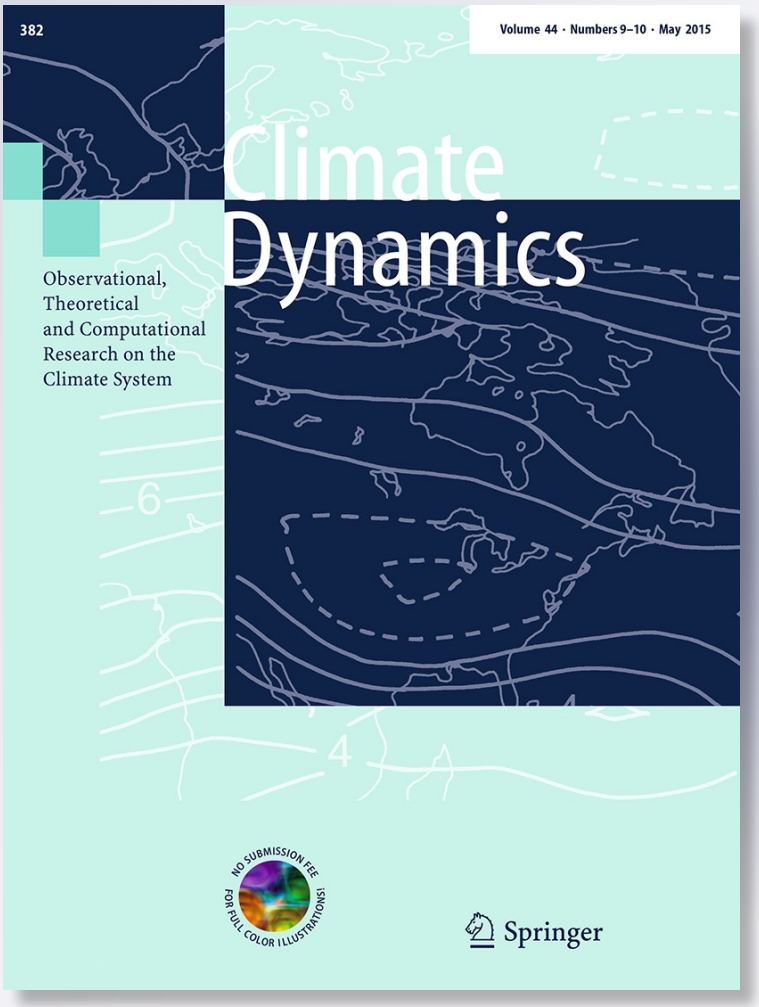
An assessment of a multi-model ensemble of decadal climate predictions

A. Bellucci, R. Haarsma, S. Gualdi, P. J. Athanasiadis, M. Caian, C. Cassou, E. Fernandez, A. Germe, J. Jungclaus, J. Kröger, et al.

Climate Dynamics
Observational, Theoretical and Computational Research on the Climate System

ISSN 0930-7575
Volume 44
Combined 9-10

Clim Dyn (2015) 44:2787-2806
DOI 10.1007/s00382-014-2164-y



Springer

Clim Dyn (2015) 44:2787–2806
DOI 10.1007/s00382-014-2164-y

An assessment of a multi-model ensemble of decadal climate predictions

A. Bellucci · R. Haarsma · S. Gualdi · P. J. Athanasiadis · M. Caian · C. Cassou · E. Fernandez · A. Germe · J. Jungclaus · J. Kröger · D. Matei · W. Müller · H. Pohlmann · D. Salas y Melia · E. Sanchez · D. Smith · L. Terray · K. Wyser · S. Yang

Received: 20 January 2014 / Accepted: 28 April 2014 / Published online: 9 May 2014

Abstract A multi-model ensemble of decadal prediction experiments, performed in the framework of the EU-funded COMBINE (Comprehensive Modelling of the Earth System for Better Climate Prediction and Projection) Project following the 5th Coupled Model Intercomparison Project protocol is examined. The ensemble combines a variety of dynamical models, initialization and perturbation strategies, as well as data assimilation products employed to constrain the initial state of the system. Taking advantage of the multi-model approach, several aspects of decadal climate predictions are assessed, including predictive skill, impact of the initialization strategy and the level of uncertainty characterizing the predicted fluctuations of key climate variables. The present analysis adds to the growing evidence that the current generation of climate models adequately initialized have significant skill in predicting years ahead not only the anthropogenic warming but also part of the internal variability of the climate system. An important finding is that the multi-model ensemble mean does generally outperform the individual forecasts, a well-documented result for seasonal forecasting, supporting the need to extend the multi-model framework to real-time decadal predictions in order to maximize the predictive capabilities of currently available decadal forecast systems. The multi-model perspective did also allow a more robust assessment of the impact of the initialization strategy on the quality of decadal predictions, providing hints of an improved forecast skill under full-value (with respect to anomaly) initialization in the near-term range, over the Indo-Pacific equatorial region. Finally, the consistency across the different model predictions was assessed. Specifically, different systems reveal a general agreement in predicting the near-term evolution of surface temperatures, displaying positive correlations between different decadal hindcasts over most of the global domain.

1 Introduction

Predicting climate evolution over interannual-to-decadal timescales represents a grand challenge for climate scientists, and an unprecedented opportunity for decision-

A. Bellucci (✉) · S. Gualdi · P. J. Athanasiadis
Centro Euro-Mediterraneo sui Cambiamenti Climatici, Viale A. Moro 44, 40127 Bologna, Italy
e-mail: alessio.bellucci@cmcc.it

R. Haarsma
Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

S. Gualdi
Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy

M. Caian · K. Wyser
Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

C. Cassou · E. Fernandez · E. Sanchez · L. Terray
European Centre for Research and Advanced Training in Scientific Computation (CERFACS), Toulouse, France

A. Germe · D. Salas y Melia
CNRM, Météo-France, Toulouse, France

J. Jungclaus · J. Kröger · D. Matei · W. Müller · H. Pohlmann
Max-Planck-Institut für Meteorologie, Hamburg, Germany

D. Smith
Met Office Hadley Centre, Exeter, UK

S. Yang
Danish Meteorological Institute (DMI), Copenhagen, Denmark

makers to calibrate plans and actions over a temporal horizon of a few years. The earliest exploratory investigations on the intrinsic predictive capabilities of a climate model date back to the second half of the 90s (Griffies and Bryan 1997). However, it was only in the late 2000s that successful steps towards valuable decadal predictions initialized with a realistic estimate of the climate state were undertaken (Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2009; Mochizuki et al. 2010). These pioneering single-model efforts highlighted the strong dependency of near-term predictions on specific elements of the adopted prediction system (Hurrell et al. 2009; Meehl et al. 2009). These, essentially, include: (i) a dynamical model, (ii) a strategy for initialization and ensemble generation (i.e., initial state perturbation), and (iii) a set of analyses used to constrain the initial state of the model with a realistic representation of the climatic system. The choice made on each of these aspects introduces, unavoidably, a degree of uncertainty in the predictions. The use of a multi-model ensemble allows for sampling these structural differences among individual prediction systems. A significant fraction of this uncertainty is associated with the varying physical parameterizations and numerical schemes adopted in individual climate models. While the benefits stemming from the use of a multi-model approach have been extensively documented for seasonal predictions (Palmer et al. 2004; Hagedorn et al. 2005) and for long-term climate projections (Lambert and Boer 2001; Tebaldi and Knutti 2007), indications that a similar behaviour holds for decadal hindcasts is supported by a comparatively smaller amount of evidence (van Oldenborgh et al. 2012; García-Serrano and Doblas-Reyes 2012; Kim et al. 2012; Goddard et al. 2013; Meehl et al. 2013). The perspective of an ever-increasing confidence in our knowledge of the ocean state, boosted by the launch of the ARGO observing system, and in the overall performance of coupled ocean–atmosphere general circulation models (CGCMs), has led the climate science community to foster several coordinated efforts aiming at a systematic exploitation of the predictive skill featured by current climate models at the interannual to decadal timescales.

This issue has been tackled in a number of past and ongoing EU-funded projects (ENSEMBLES, THOR, COMBINE, SPECS). The multi-model, decadal prediction experiments performed as part of the forerunner ENSEMBLES exercise (van Oldenborgh et al. 2012; García-Serrano and Doblas-Reyes 2012) in particular, were conducive to the design of a specific set of near-term predictions to be included as core experiments in the 5th Coupled Model Intercomparison Project (CMIP5; Meehl et al. 2009; Doblas-Reyes et al. 2011).

In this article, we analyze the results from the multi-model ensemble of decadal prediction experiments performed as part of the European Framework Program 7 COMBINE (Comprehensive Modelling of the Earth System for Better Climate Prediction and Projection) Project. A primary objective of COMBINE was to improve currently available Earth-system models by including key physical and biogeochemical processes so as to represent more accurately the forcing mechanisms and the feedbacks at work in the climate system. The decadal integrations were part of a broad assessment focusing on the predictive skill featured by a set of European state-of-the-art climate models. The simulations were performed following the CMIP5 protocol for near-term predictions (Taylor et al. 2012), and therefore they contributed to the Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5).

In this study, the COMBINE set of initialized simulations is scrutinized to examine several aspects of decadal predictions within a multi-model framework. The main aims of this study are: to assess the predictive skill, to investigate the impact of the initialization strategy and to assess the uncertainty characterizing the near-term prediction of key climate variables in a multi-model ensemble of CMIP5 decadal integrations.

The paper is structured as follows. The main features of the multi-model ensemble and the experimental design are described in Sect. 2. The metrics and the observational data sets used to verify the forecast skill of the prediction systems are described in Sect. 3. The predictive skill associated with the global-mean surface temperature is shown in Sect. 4. Section 5 is devoted to the analysis of regional forecast skill exhibited both by individual systems and the multi-model ensemble-mean. In Sect. 6, the impact of the initialization method on the quality of decadal predictions is assessed. A quantitative analysis of some aspects of the uncertainty associated with the predictions is presented in Sect. 7. Finally, concluding remarks are provided in Sect. 8.

2 The COMBINE multi-model ensemble

The COMBINE multi-model ensemble (CME) consists of six decadal prediction systems (DPSs; listed in Table 1) blending different dynamical models, initialization and perturbation strategies, and data assimilation products employed to constrain the initial state of the system. The dynamical models used in this study are: CMCC-CM, EC-Earth, CNRM-CM5, MPI-ESM-LR and HadCM3.

Within the CME, two different initialization strategies are used: *full initialization* (CMCC-CM, CNRM-CM5 and EC-Earth in the implementation adopted by KNMI) and *anomaly initialization* (HadCM3, MPI-ESM-LR and EC-Earth in the implementation followed by SMHI and

Table 1 The six decadal prediction systems (DPS) of the ensemble, the corresponding model used, spatial resolution (ocean/atmosphere), initialization method, initialized components, and ensemble size

Institute	Dynamical model	Resolution: AGCM OGCM	Initialization strategy	Initialized components	Ensemble size
Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)—Italy	CMCC-CM	0.7° L31 1.2° L31	Full	Ocean	3
Royal Netherlands Meteorological Institute (KNMI)—Netherlands	EC-Earth	1.1° L62 1° L42	Full	Atmosphere, land Ocean, sea-ice	5
Swedish Meteorological and Hydrological Institute (SMHI)—Sweden Danish Meteorological Institute (DMI)—Denmark	EC-Earth	1.1° L62 1° L42	Anomaly	Atmosphere, land Ocean, sea-ice	7
European Centre lor Research and Advanced Training in Scientific Computation (CERFACS)/ Météo-France/CNRM (MF-CNRM)—France	CNRM-CM5	1.4° L31 0.7° L42	Full	Ocean	10
Max Planck Institute for Meteorology (MPI-M)—Germany	MPI-ESM-LR	1.9° L47 1.5° L40	Anomaly	Ocean	10
Met Office Hadley Centre (MOHC)—UK	HadCM3	3.75° × 2.5° L19 1.25° L20	Anomaly	Atmosphere, land Ocean, sea-ice	10

DMI). In the full initialization strategy, a three-dimensional, observational estimate of the climate state is assigned as the initial condition of the dynamical model at the start of each simulation. In the anomaly initialization strategy, observed anomalies are added to the model's own climatology so as to construct an initial state of the coupled system. Regarding the perturbation techniques represented in the CME, these reflect the best practices currently in use at several climate modeling centres. With each DPS, a set of near-term predictions has been performed in compliance with the CMIP5 experimental design (Taylor et al. 2012). This is summarized as follows. The decadal hindcasts/forecasts consist of 10- or 30-year integrations initialized on the 1st of November of the years 1960–2005 at a 5-year interval, except for CNRM-CM5 and MPI-ESM-LR initialized on the 1st of January of the years 1961–2006 at the same interval, yielding ten hindcasts for each DPS.

For each start date, an ensemble of integrations (with a minimum size of three members) is performed. The initial conditions for the ocean are required to be a realistic representation of the observed state, while there are no restrictions regarding the initialization of sea-ice, land surface and the atmosphere. However, as shown in Table 1, some DPS realistically initialize these components, too. Historical radiative forcing conditions, including greenhouse gases (GHGs), aerosols, ozone and solar irradiance variability, are used for the 1960–2005 period, followed by the RCP4.5 scenario settings from 2006 onward. Aerosol forcing for major historical volcanic eruptions is also included in the forcing fields of the DPSs, with the exception of the CMCC-CM system. It is important to emphasize that the inclusion of volcanic emissions among the forcing fields used to constrain a given climate model leads to an overestimation of the real predictive skill, as in a real climate forecast we have no knowledge about future volcanic eruptions.

The following results are based on 10-year hindcasts of annual-mean surface temperature and global precipitation fields. For each individual system, only its own ensemble-mean prediction is considered.

Finally, the multi-model ensemble mean (MME) forecast is computed as a simple, equally-weighted, average of all model forecasts in the COMBINE ensemble.

3 Verification

Following the recommendations by Goddard et al. (2013), here the predictive skill of individual systems and of the multi-model ensemble mean, is evaluated through the anomaly correlation coefficient (ACC; Wilks 2006) between annual mean (Jan–Dec) predictions and observations averaged over the lead-time intervals 2–5 and 6–9 years.

In order to remove the spurious signal associated with the model adjustment following the initialization, a standard bias-correction procedure defined in International CLIVAR Project Office (ICPO 2011) is applied to surface temperature fields. For methodological consistency, the same "full field" approach to bias correction (i.e., with the model climatology defined as the average forecast computed over the collection of available forecasts for the 1960–2010 period; see section 4a in ICPO 2011) is here applied to all systems, therefore including those relying on the anomaly initialization.

A one-tailed test against the null-hypothesis of non-positive correlation is applied to verify the statistical significance of the ACC values. The autocorrelation in the time series, which reduces the effectively independent data, is accounted for through the computation of the effective sample size, using the method outlined in Bretherton et al. (1999).

Predictive skill is evaluated against HadISST (Rayner et al. 2003) and CRUTEM3 (Brohan et al. 2006) data, used as observational references for sea surface temperature (SST) and 2-m air temperature (T2M) over land, respectively, while the GPCC-v4 data set (Schneider et al. 2008) is used for precipitation. Prior to the computation of anomaly correlations, both model and observational fields were linearly interpolated onto a common regular grid at a $1° \times 1°$ spatial resolution.

4 Global mean surface temperatures

Figure 1 displays the evolution of the annual mean, globally averaged surface temperature (middle panel: T2M over land, and bottom panel: SST) for each individual DPS, the multi-model ensemble mean and observations. In the global averaging, all grid points with missing observations have been excluded so that the model results are compared fairly with observations over the same domain. Large areas with missing observations include Antarctica (CRUTEM3) and high-latitude oceans (HadISST).

The hindcasts successfully reproduce the observed long-term surface temperature trend, with the envelope of the multi-model ensemble mostly encompassing the observations, and a good part of the multi-year variability. Additionally, the hindcasts appear to follow the observations in the temperature changes that occurred after major historical volcanic eruptions, whose radiative fingerprint is included as an external forcing in most of the prediction systems (see, in particular, the projected surface temperature drop following the 1991 Mount Pinatubo eruption). On the other hand, significant departures from the observed record occur in correspondence to vigorous El Niño–Southern Oscillation (ENSO) events (most notably, the 1997 El Niño and 2008 La Niña episodes). Seasonal forecasting efforts have shown considerable skill in predicting the evolution of ENSO, yet, the above-mentioned episodes were not correctly predicted. Whether or not this is caused by the wide, 5-year interval between successive initialization dates is an open question due to the limited availability of yearly initialized decadal hindcasts to compare against the 5-year canonical predictions. Doblas-Reyes et al. (2011) compare 1- versus 5-year initialized integrations performed with the UK Met Office DPS (DePreSys) and find no significant differences in terms of 2–5 lead-year skill for surface temperature over the equatorial Pacific (see their Fig. 1b, d). However, the robustness of this result should be tested using a larger set of models.

5 Evaluation of forecast skill

5.1 Multi-model ensemble mean

In the present section, the predictive skill associated with surface temperatures (SST and T2M) and precipitation over land is evaluated with the use of the ACC between annual mean observations and MME anomalies.

Following the approach outlined in van Oldenborgh et al. (2012), we evaluate both the "total" skill (which refers to the total signal including the one associated with the observed warming trend in the initial and boundary conditions) and the "residual" skill associated with climatic fluctuations around the above-mentioned trend. These are obtained by removing a long-term linear fit from both the model output and the observations at every grid point. The multi-model ensemble mean ACC maps for surface temperature (T2M and SST), evaluated over the lead-time periods 2–5 and 6–9 years (with and without the trend removal; Fig. 2) reveal a fairly in-homogenous spatial distribution of skill. Before removing the trend, the ACC is positive and exceeds the threshold value corresponding to the 95 % level of statistical significance over extensive portions of the oceans (particularly, the Indian and North Atlantic oceans) and continents (note that in Fig. 2 stippling is applied to the areas where the ACC is not statistically significant so as to allow a clearer visualization of the features associated with statistically significant high skill). Negative correlations are found in the Pacific and Southern oceans and over parts of the South American continent. This pattern holds for both the lead-time periods examined, but with slightly higher values in the longer (6–9 years) term.

After removing the long-term trends in both the observed and the predicted time series, the local forecast skill undergoes a substantial reduction over most of the global domain, indicating that the warming trend imposed through the initial conditions and the prescribed radiative forcing is the primary cause for the very high forecast skill seen in the non-detrended ACC maps. Significant predictive skill beyond a pure trend is found in the 2–5 year range, over the North Atlantic and over a zonal belt stretching from the Mediterranean basin, to northern Africa and to Eurasia. A similar pattern is found in the 6–9 year range, but with an enhanced skill in the extra-tropical Pacific. The dramatic reduction in skill occurring in the Indian Ocean basin, after trend removal, reflects the fact that the predictable signal in this area is primarily driven by changes in the radiative forcing (Guemas et al. 2012).

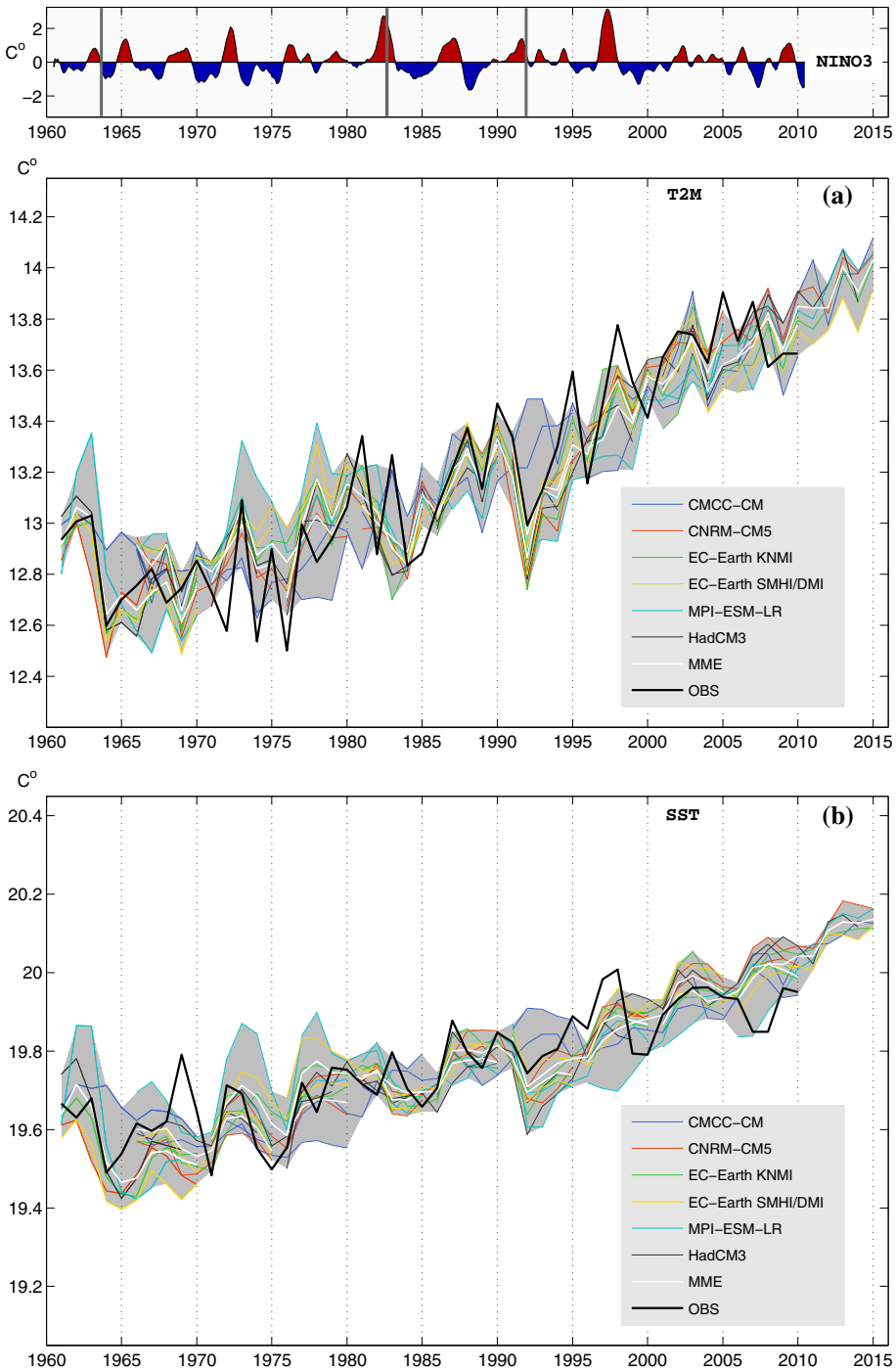


Fig. 1 (*Top*) Time series of observed SST in the NINO3 region (150–90W, 5S–5N). SSTs are taken from HadISST (Rayner et al. 2003). *Vertical grey lines* mark the major volcanic eruptions occurred during the 1960–2010 reference period: Agung (March 1963), El Chichon (March 1982) and Pinatubo (June 1991). (*Middle*) Globally averaged T2M over land. The plot shows annual mean observations and the corresponding time series for all the 10-year long hindcasts, including the multi-model ensemble mean. The *grey envelope* highlights the spread of the ensemble. (*Bottom*) As middle panel, but for SST

The assumption of a linear trend to describe the multi-decadal surface temperature evolution may be questionable, as it is known that the global warming signal exhibits considerable departures from linearity, especially at the regional scale (e.g., Ting et al. 2009 for the North Atlantic). However, given the shortness of the analyzed period, the departures from linearity in our 50-year long time series are arguably small compared to other possible errors, such as the ones arising from the bias-correction. The impact on predictive skill determined by a non-linear trend assumption was assessed by using a polynomial quadratic law instead of the linear fit. The resulting de-trended ACC patterns (not shown) provided similar results to those based on a linear trend assumption, shown in Fig. 2 (see also van Oldenborgh et al. 2012).

For the non-detrended case, the occurrence of negative ACC values over vast parts of the globe indicates that even for the well-detected twentieth century warming trend,

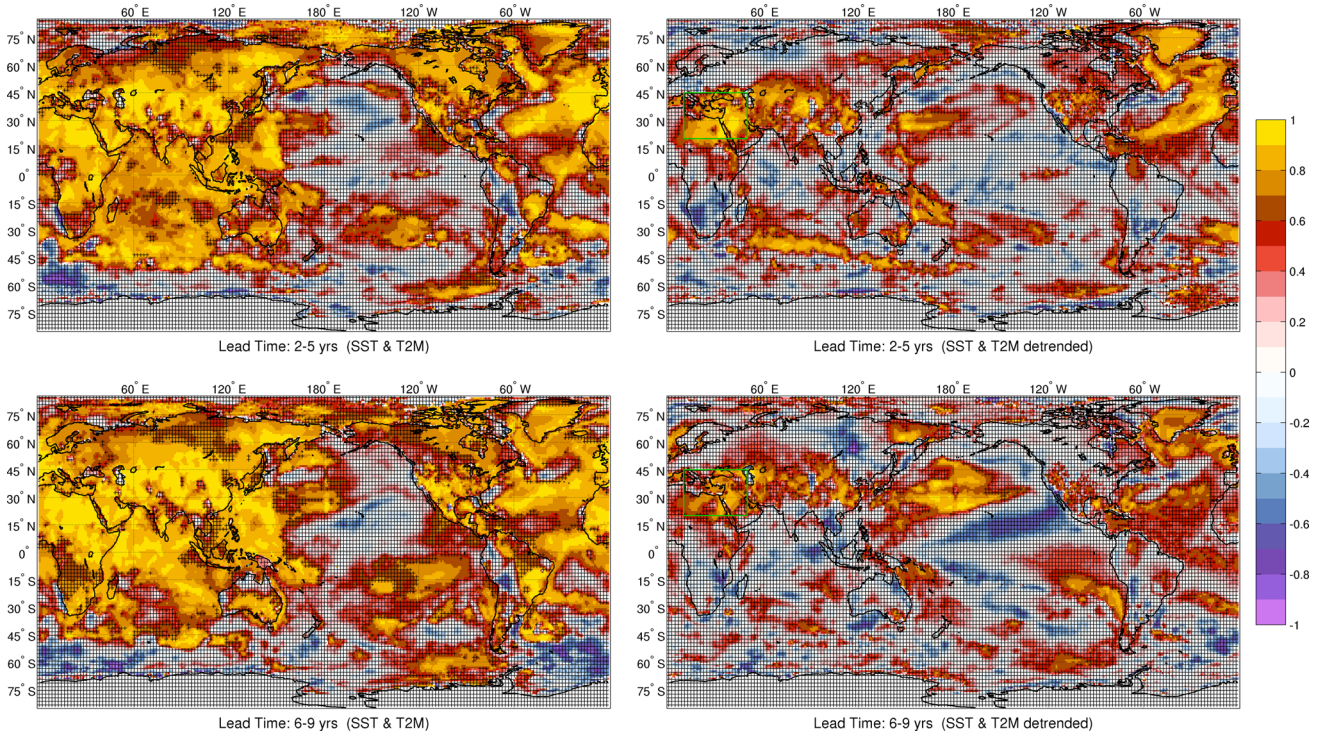


Fig. 2 (*Left*) Multi-model ensemble-mean (MME) anomaly correlation coefficient (ACC) maps for T2M (over land) and SST, for lead-time periods (*top*) 2–5 and (*bottom*) 6–9 years. (*Right*) The corresponding maps after the long-term linear trends have been removed from both the model data and the observations. Stippling is used to indicate points where statistical significance is below the 95 % level, according to a one-tailed *t* test accounting for autocorrelation in the time series. The *green frame* indicates the boundary of the MAME region, used in Fig. 6 (see text for details)

there may be large discrepancies locally between the observed and the predicted surface temperature trends. While strictly speaking negative correlations indicate no skill and are virtually indistinguishable from any other below-significance ACC value, practically their occurrence may help in identifying a specific source of error in models forecast. To better clarify this point, the long-term linear trends are diagnosed from both observations and predictions, for the 1960–2010 period (shown in Fig. 3). While some of the large-scale features in the multi-model trend pattern are consistent with the observations (see, for instance, the pronounced land-to-ocean gradient, with the continents warming at a faster rate than the sea surface), there are extensive areas, mainly over the oceans, where the predicted and the observed rate of change are clearly negatively correlated. These include, most notably, the extra-tropical North Pacific, the western North Atlantic (off the eastern US seaboard) and parts of the western South American continent, in proximity to the Andes. Here the observed trends are weakly negative, in contrast to the warming trends predicted by the models. The close match between these regions and those displaying negative ACC values for non-detrended surface temperature (shown in Fig. 2) suggests that the reason behind the models' poor predictive skill lies in their inability to reproduce correctly the relative fraction of the forced and unforced component in the observed variability signal, with the former dominating over the latter in the above mentioned critical regions.

Figure 4 shows the corresponding ACC maps for precipitation over land. The ACC was computed only at grid points containing at least one rain gauge with a sufficiently long record. Clearly, over large areas of the globe (including all oceans, high latitudes, and deserts) there are no such measurements available. This is reflected in the sparseness of the ACC field, which is also quite noisy spatially indicating poor predictability skill. These characteristics of the ACC field make difficult the visualization of statistical significance through stippling. Therefore, here the statistical significance is assessed by visual inspection of the ACC values. Given the small and more or less uniform autocorrelation characterizing the time series of precipitation, a spatially constant effective sample size can be assumed in the assessment of statistical significance. For $N = 9$ independent values (the last start-date was not considered for consistency with the observed record), corresponding to $N - 2 = 7$ degrees of freedom, the ACC threshold values corresponding to the 90 and 95 % levels of statistical significance are about 0.47 and 0.58, respectively. Given the pioneering stage of decadal forecasts,

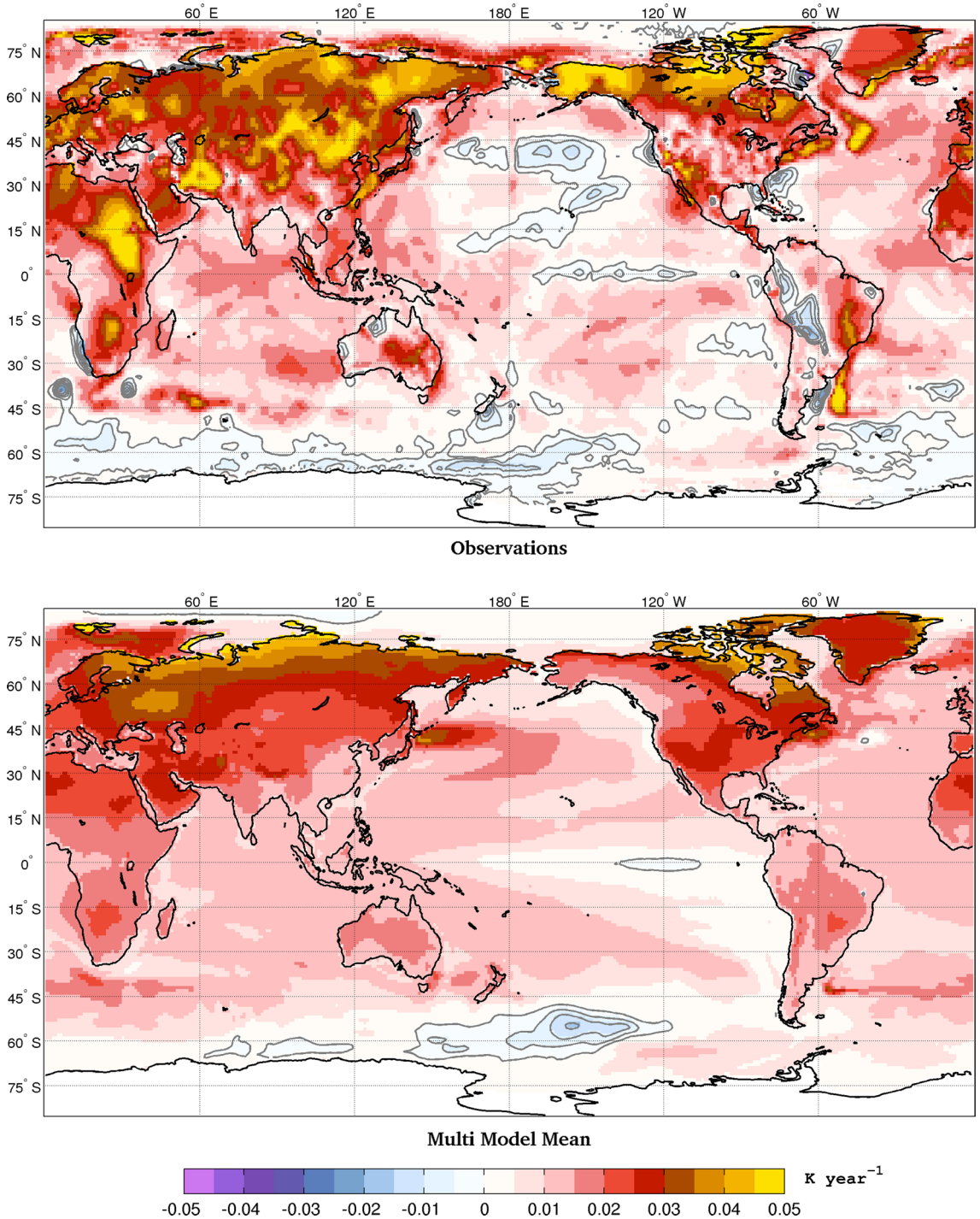


Fig. 3 Patterns of linear surface temperature trends (SST and T2M over land; K year^{-1}) over the period 1961–2010 estimated from (*top*) HadISST/CRUTEM3 data and (*bottom*) multi-model ensemble mean predictions. Trends are computed by a least-square fit of a first order polynomial. For predictions, nine 10-years long hindcasts have been used. The last forecast (2005–2015) is not used for consistency with the observed period (1961–2010). *Contours* are used to highlight negative trends (*countour interval* 0.005 K/year)

particularly when applied to precipitation fields, a 90 % significance level (corresponding to a p value $p = 0.1$) is here considered acceptable.

As for surface temperature (Fig. 2), here ACC maps are presented for the precipitation anomalies before and after the trend removal. However, in contrast to surface temperature, precipitation does not exhibit locally a well-defined trend; consequently the ACC field for precipitation is largely unaffected by the trend removal. In fact, the chances of removing an erroneous trend are quite high, and

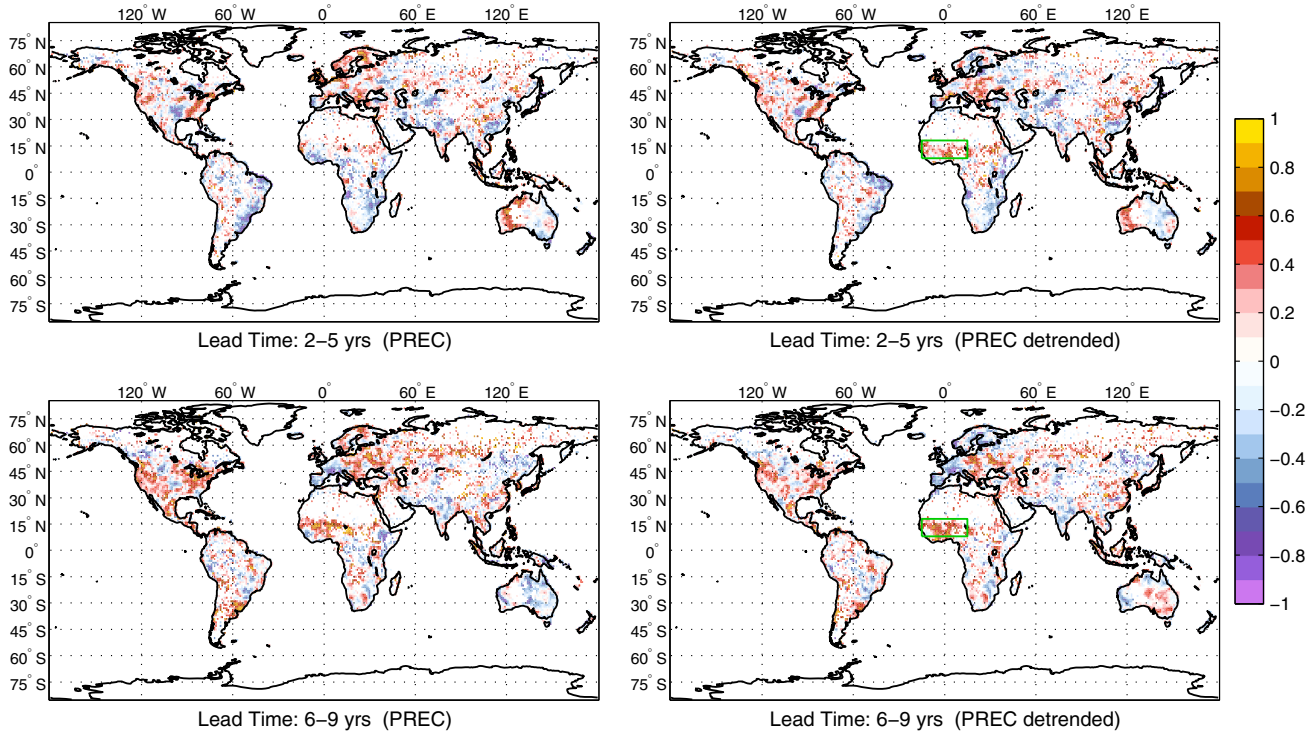


Fig. 4 (*Left*) Multi-model ensemble-mean (MME) anomaly correlation coefficient (ACC) maps for precipitation, for lead-time periods (*top*) 2–5 and (*bottom*) 6–9 years. (*Right*) The corresponding maps after the long-term linear trends have been removed from both the model data and the observations. The *green frame* indicates the boundary of the Sahel region, used in Fig. 6. Correlations are statistically significant at $p < 10$ % for ACC > 0.47 (see also text)

therefore the detrended ACC should be considered with care.

In spite of its noisy character, the ACC fields for precipitation display significant predictive skill over certain regions, particularly at the lead-time period 6–9 years. Specifically, the Sahel, mid-latitude Eurasia and parts of North America feature some patches in which one or both of the above-mentioned threshold values of statistical significance are exceeded. A more in-depth discussion on the potential links between the enhanced forecast skill found over these regions and the Atlantic multi-decadal variability is given in Sect. 5.3. The relative increase of ACC values with lead-time (non-detrended case) indicates that the predictive skill in precipitation stems from the external forcing rather than from the initialization. This specific aspect is further highlighted in the analysis presented in the next section.

5.2 Forecast skill of individual decadal prediction systems

So far, we focused on the point-wise predictive skill of the MME. In the following, the predictive skill associated with each individual prediction system, including the MME, is evaluated through a metric based on the percentage of the area where the corresponding ACC exceeds a 90 % statistical significance threshold. This metric is applied to the previously analysed lead-time periods, before and after the trend removal, for T2M, SST and precipitation. All the results we refer to in this section are illustrated in Fig. 5.

Examining the global skill metric for SST and T2M we see that: (i) for the non-detrended ACC the skill values are slighly higher in the 6–9 range compared to the 2–5 range, and the corresponding ensemble spread is lower, while (ii) for the detrended ACC the skill values are slightly higher in the near-term (2–5 range). These facts may accept the following explanation. In the non-detrended case, most of the predictive skill is dictated by the externally imposed forcings, which are essentially the same across the DPS ensemble. However, the near-term response (2–5 range) is more strongly dependent on the initial conditions, which vary significantly across the DPS ensemble due to the use of different ocean reanalyses, sea-ice initialization, etc. Also, the near-term response is more strongly influenced by the model drift, which differs significantly from model to model (see also Branstator and Teng 2010, 2012; Branstator et al. 2012). These points are consistent with the features mentioned in (i) and (ii), namely: Before detrending the correlations are determined by the models' coherent response to the trends (here the above-mentioned

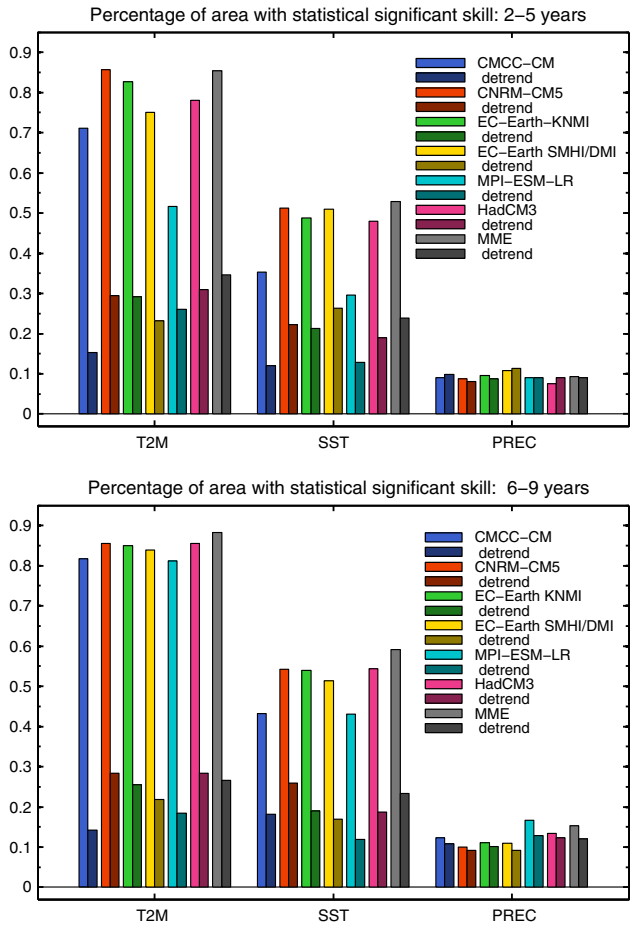


Fig. 5 For each of the three variables (SST, T2M and PREC) the histograms show the percentages of the area where the corresponding ACC is positive and exceeds the 90 % statistical significance threshold (see text). For each variable the ACC is calculated over a different portion of the globe. The lead-year periods 2–5 (*top*) and 6–9 (*bottom*) are used. Each pair of columns refers to before and after removing the trend

differences in the near-term response tend to introduce a degree of discrepancy across the DPS ensemble, leading to lower ACC values and larger ensemble spread). On the other hand, after detrending the correlations (predictive skill) will be mainly due to anomalies determined by the initial conditions, and in this case the near-term response has a better chance to be skillful (e.g. via the memory of the upper ocean heat content).

Regarding precipitation, the global skill is typically low (with or without trends). However, it is interesting to note how, consistently with (i), a few models display slightly higher values in the 6–9 years interval, leading to a consistently higher MME skill over this range.

Another important aspect emerging from this analysis, for T2M and SST, is that the MME exhibits predictive skill that is generally better than (or at least comparable to) the best individual DPS prediction, thus supporting the coordinated efforts that are being made for greater multi-model ensembles in the design of decadal prediction experiments. In the fields of seasonal prediction and weather forecasting this is a well-documented fact (Palmer et al. 2004).

When comparing the global skill scores featured by individual DPSs in Fig. 5, the differences in the corresponding ensemble size must be taken into account. While a robust relationship between the ensemble size and the corresponding predictive skill has been found in the context of single DPS experiments (Chikamoto et al. 2013), such a clear link appears to be elusive in our multi-model analysis.

As shown in Fig. 5, predictions based on a relatively low number of ensemble members do not necessarily yield correspondingly lower quality forecasts, when compared with large-sized ensemble predictions. This suggests that a number of other factors play a more determinant role in setting the overall skill of a DPS, with the inter-model diversity likely blurring the relationship between skill and ensemble size emerging from single-model decadal forecast assessments (Chikamoto et al. 2013).

5.3 Regional assessment

The global ACC fields revealed pronounced spatial variability of predictive skill (Figs. 2, 4). The detection of significant predictability after removing the linear trends suggests that predictable processes exist, related to the internal variability of the system, giving rise to predictive skill beyond the default global warming signal. Based on the global ACC maps for surface temperature and precipitation, a few domains are selected featuring consistently large anomaly correlation values, to perform a more detailed, regional-scale analysis of predictability. The selected areas are: (i) the Atlantic oceanic sector, (ii) the Mediterranean, NE Africa and Middle East region (MAME) [defined over (20–45N, 10–50E]; see box in Fig. 2), and (iii) the Sahel [defined over (8–18N, 15W–15E); see box in Fig. 4), displaying enhanced ACC for SST, T2M and precipitation, respectively.

As a measure of predictive skill at different lead-times, the ACC is computed between the predicted and observed values of specific climate variability indices (Atlantic Multi-decadal Oscillation and Atlantic SST Dipole; see below for a definition of these indices) or the area-weighted average of T2M and precipitation within the selected domains. The skill of individual systems is also tested against a statistical model based on persistence of observed initial conditions. In order to enhance the signal-to-noise ratio, the ACC is calculated using 3-year running averages, namely, for lead-years (1–3, 2–4, etc.). Regarding the regional indices for T2M and precipitation, long-term

trends are removed before the ACC computation, whilst the Atlantic variability indices are not affected by the global warming signature, by definition (see below). Here we define a "predictability limit" as the lead time beyond which the computed ACC falls under a particular threshold value of statistical significance, based on a one-tailed Student's *t* test.

The predictive skill over the Atlantic area is evaluated via two indices characterizing the Atlantic decadal-scale variability: the Atlantic Multi-decadal Oscillation index (AMO) and the Atlantic Dipole index (AD). Following the definition of Trenberth and Shea (2006), the AMO is computed as the area-weighted average SST over the North Atlantic (0–60N, 0–80W) minus the global (60S–60N) mean SST. The AD is defined as the difference of the area-weighted average SST between (40–60N, 60–10W) and (40–60S, 50W–0E) (Latif et al. 2006). Both of these climate variability indices are thought to be strongly dependent on the low-frequency variability of the thermohaline circulation (Knight et al. 2005; Latif et al. 2006).

The AMO (Fig. 6a) appears to be predictable within a time-scale that depends strongly on the particular DPS, ranging from 2–4 up to 8–10 years. MME does generally outperform the individual systems, with ACC values exceeding 0.7 for most of the lead years. The AD index (Fig. 6b) displays a similar behaviour with respect to the AMO index, but with a sizeably higher skill at short lead times (up to 3–5 years) for the MME. A closer inspection of ACC evolution for individual models reveals that this is mostly determined by the specific behaviour of a sub-set of models (in particular, CNRM-CM5, HadCM3 and MPI-ESM-LR) in the short lead-year range, with the other systems showing relatively smaller differences when comparing AD with AMO index predictability. For both Atlantic indices, persistence is systematically beaten by MME, and by most individual systems from lead-years 2–4 onward.

The near-surface air temperature (T2M) over the MAME region (Fig. 6c) features a high degree of predictability, with the MME displaying significant skill up to 10 years. Interestingly, different systems display similar changes of ACC with lead-time (for example, compare EC-Earth and MPI-ESM-LR). Since surface temperature predictability over land is typically weak in CGCMs (Boer and Lambert 2008), the large forecast skill found over the MAME region is suggestive of a possible remote oceanic influence. In particular, considering the influence exerted by the AMO on the adjacent regions (Knight et al. 2005), the long-term predictability found for the AMO index may explain part of the predictive skill found for T2M in the MAME area (Matei et al. 2012a). Regarding the comparison with persistence, similar considerations apply to the MAME region as for the Atlantic indices.

The Sahel is among the few areas displaying (marginally) significant skill for MME precipitation (see Fig. 4). Comparing the predictive skill of different DPSs over this area (Fig. 6d) a considerably large spread is found, with the ACC for individual hindcasts ranging from being marginally significant to systematically below the threshold corresponding to the 90 % level of statistical significance, and often outperformed by persistence, particularly in the near range. This spread is indicative of the large uncertainties associated with the representation of rainfall variability in current climate models. Another feature is the markedly non-monotonic evolution of ACC with lead-time, exhibited by most of the models. This reflects on the MME, which shows the highest predictive skill around years 3–5.

Concerning the noisy behaviour featured by the ACC for most of the regional indices considered in this analysis (largely deviating from the expected monotonic declining evolution with lead-time), this can be ascribed to the low number of points used in the computation of anomaly correlations when using a 5-year spacing between start dates, as also suggested by Doblas-Reyes et al. (2011). This issue is further discussed in Sect. 6.

Finally, it is worth noting the differences between the predictive skill of the EC-Earth-KNMI and the EC-Earth-SMHI/DMI systems, as these involve the very same model and only differ in the initialization strategy, thus allowing an assessment of the effect of the latter on the quality of the predictions. It is seen that KNMI (full-value initialization) is more skilful in reproducing the observed variability in the Atlantic and MAME domains, whilst for precipitation over the Sahel, both systems display correlations systematically lower than the 90 % significance level (see Hazeleger et al. 2013 for a detailed analysis of EC-Earth performance under full-value and anomaly initialization). A more thorough analysis of the role of the initialization strategy on the quality of predictions in the COMBINE multi-model ensemble is provided in the next section.

6 The role of initialization strategy on the quality of decadal predictions

The COMBINE ensemble of DPSs is equally partitioned into models employing full-value and anomaly initialization (Table 1). While pros and cons of these two methodologies are known in principle (Meehl et al. 2013) there is no compelling evidence indicating which of the two can be identified as a "best practice" (i.e., yielding higher skill) for decadal predictions. So far, the influence of the initialization strategy on the quality of decadal predictions has been only examined in the context of individual systems

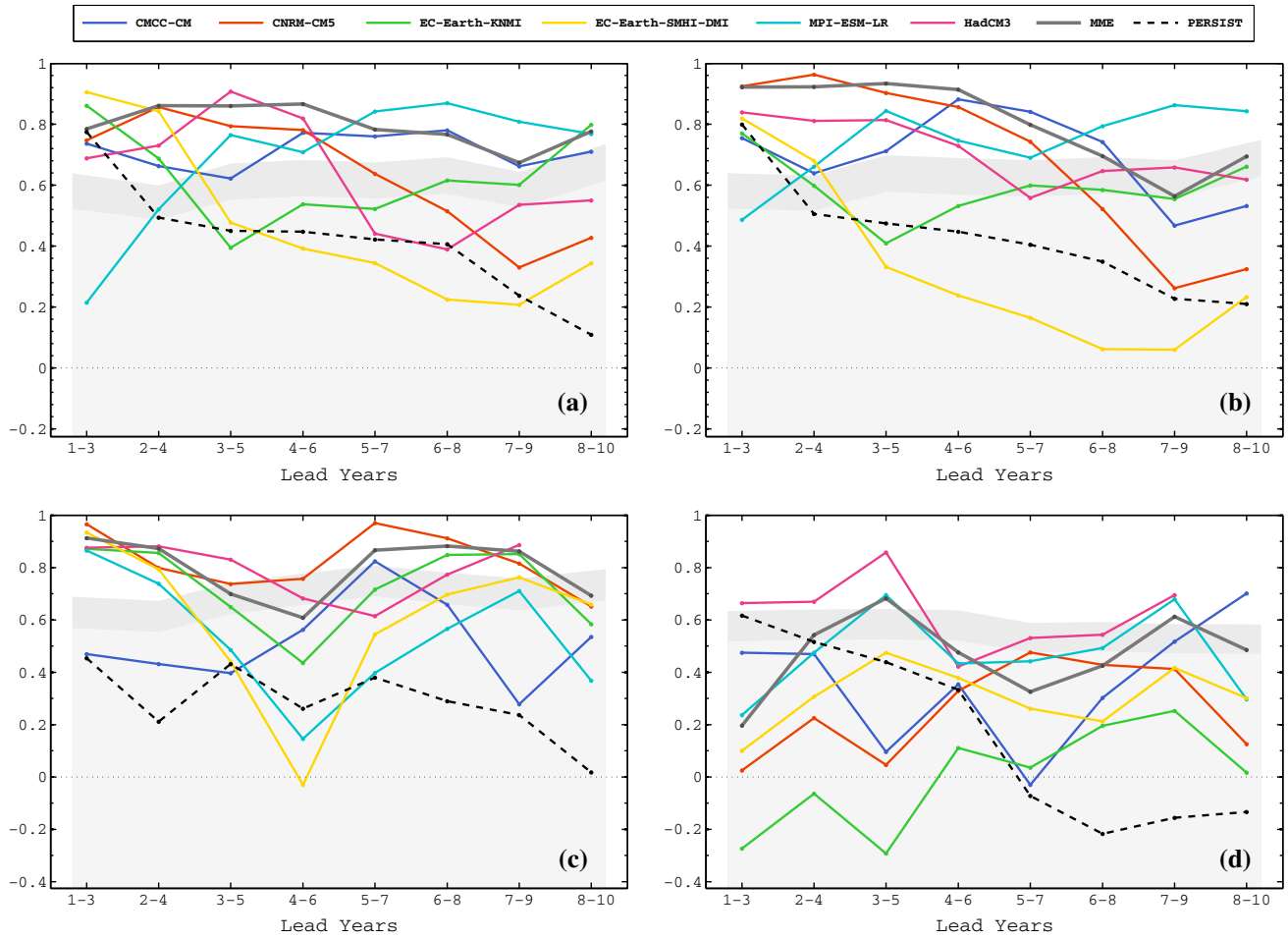


Fig. 6 The ACC as a function of lead time for different climatic indices: **a** AMO, **b** Atlantic SST Dipole, **c** T2M averaged over an area including parts of the Mediterranean, NE Africa and Middle East [20–45N, 10–50E], **d** PREC averaged over the Sahel [8–18N, 15W–15E]. The two thresholds displayed by *grey shading* correspond to the 90 and 95 % levels of statistical significance, accounting for autocorrelation in the time series

(Smith et al. 2013; Hazeleger et al. 2013). The outcomes of these studies were not conducive to any conclusive statement as to whether there is any significant improvement in the predictive capabilities associated with the use of one over the other initialization strategy. Smith et al. (2013) tested the relative merits of the two approaches in the UK Met Office DePreSys system, but over most regions they did not find significant differences in skill, with the exception of some hints of higher predictive skill for full-field predictions at the multiyear timescale. Hazeleger et al. (2013) tackled the same issue in a DPS based on the EC-Earth model, finding similar skill under both initialization methods. Here we make an attempt to assess the relative influence of the initialization method within the framework of a multi-model ensemble of DPSs. We concentrate on surface temperature (SST and T2M) as the predictability associated with these variables is clearly higher than for precipitation.

Different initialization strategies lead to different transient behaviours following the initial state assignment. In particular, full-state initialization leads models to adjust towards their own mean state, which generally differs from the observed climatology (this difference measured by the so-called model bias). In the anomaly initialization case, models are from the start close to their background state, and therefore no strong drift is expected in principle. Before analyzing the predictive skill associated with the two different initialization methods, we first characterize the transient evolution of individual systems following initialization. Specifically, we examine the departure between model (raw hindcast values) and the observed state as a function of lead-time, averaged across all hindcasts. For consistency with the skill analysis we focus on the mean error for the standard 2–5 and 6–9 lead-year ranges. The corresponding error patterns (shown in Fig. 7) reveal pronounced spatial variability, mirroring the

regional structure of the underlying model bias, as well as large model-to-model differences. In particular, the systems relying on anomaly initialization show very little differences between the 2–5 and 6–9 years patterns, suggesting a very rapid adjustment, unlike the systems based on full-state initialization, for which a long-term drift is evident after comparing the bias patterns for the 2–5 with the 6–9 lead-year ranges, particularly in the extra-tropics.

The regional differences in the error rate of change are further analysed for two selected areas: the eastern equatorial Pacific (coinciding with the standard NINO3 region bounded by 90W–150W and 5S–5N) and a box in the subpolar North Atlantic straddling the Gulf Stream extension off the coast of Newfoundland (GS), the latter encompassing a region of cold bias, common to all of the systems being examined (see frames in Fig. 7). Figure 8 shows the evolution of model error with lead time averaged over each of the two selected areas and across all hindcasts. Note that this diagnostic does exactly correspond to the (area-averaged) model drift d_τ term (difference between ensemble mean forecasts and the observations over all cases) which is used to evaluate the bias corrected raw forecast anomalies, following the standard procedure described in the CLIVAR ICPO document (ICPO 2011; see section 4a). The corresponding time series for the average observed surface temperature anomalies (branching off the same start dates as for predictions) are also shown. Comparing GS with NINO3, two distinct behaviours of model error evolution emerge from the full-value initialization subset. A long-term adjustment, possibly extending beyond the 10-year duration of the hindcasts, takes place in the GS area, opposed to a much quicker transient occurring within the first year after initialization (and henceforth not detectable with this diagnostic) in NINO3.

The relatively slow setup of the cold bias in the northern North Atlantic can be ascribed to the advective dynamics associated with the slow adjustment of the large-scale ocean circulation, and specifically the Gulf Stream system, following the initialization. As expected, systems initialized with observed anomalies show no long-term adjustment over the analyzed regions. Interestingly, all systems, regardless of the initialization method, reveal a similar evolution across lead-years in the equatorial Pacific. In particular, the year-to-year fluctuations of different model biases are strongly coherent with each other and with the observed anomalies. A more detailed analysis reveals that the positive peaks in the observed record at lead-years 2 and 7 are generated by an interference of the strong 1982 and 1997 ENSO episodes (when computing the mean observed hindcast, these events show up as lead-year 7 in the decades starting at 1975 and 1990, and as lead-year 2 in the decades starting at 1980 and 1995). The correlated changes in model errors are found to be determined by the apparent lack of coordinated ENSO events in any of the model predictions (see the discussion in Sect. 1). The anomalously warm conditions associated with an El Niño episode lead to an increase (decrease) of the error, if the corresponding model is cold(warm)-biased. Considering the planetary-scale impact of ENSO, the phenomenon just described may also partly explain the non-monotonic evolution of the ACC found for several regional indices (Fig. 5; see also Doblas-Reyes et al. 2011). The noisy character exhibited by the models' drift terms shown in Fig. 8 may in fact reverberate in the computation of the anomaly correlations which depend on the structure of d_τ.

Next, the relative merits and deficiencies of full and anomaly initialization are assessed, by partitioning the CME in two subsets, based on the adopted initialization technique: a full-value and an anomaly initialization ensemble (hereafter, FVI and AI, respectively), with the former including the systems based on CMCC-CM, CNRM-CM5 and the KNMI implementation of EC-Earth, and the latter including the systems based on HadCM3, MPI-ESM-LR and the SMHI/DMI implementation of EC-Earth.

Figure 9 shows patterns of ACC for FVI and AI ensemble mean surface temperature, for lead-years 2–5 and 6–9, and the corresponding difference map. From this comparison, the equatorial Pacific, the western sea-board of US and the northern Indian Ocean, as well as a few isolated spots over the continental domains, including South Africa, northern Australia and the Amazon basin, stand out as the areas featuring the largest differences in predictive skill for the near-term 2–5 year range. On the other hand, differences between the two initialization methodologies are barely distinguishable in the 6–9 year range, except over the southern Pacific. The fact that the major differences emerge in the 2–5 year range is fully consistent with the initialization having a detectable influence in the near-term, with the longer 6–9 year range being mostly affected by the imposed boundary conditions, and therefore less sensitive to the way the initial state is assigned. In light of these results, the overall lack of predictive skill found in the tropical Pacific for the MME mean (Fig. 2), appears to be largely determined by the deteriorating impact of the systems in the AI subset.

A closer look at the individual FVI and AI maps in Fig. 9 reveals that the large ACC differences occurring in the equatorial Pacific in the 2–5 year range are exacerbated by the AI ensemble displaying negative correlations over this region. A rapid loss of predictive skill in the equatorial Pacific (negative ACC meaning essentially no skill) in AI potentially induced by the practice of initializing a dynamical model through observed anomalies which are not consistent with the underlying background state of the model, may be invoked here as a possible explanation. The ENSO-like structure of the ACC difference pattern provides

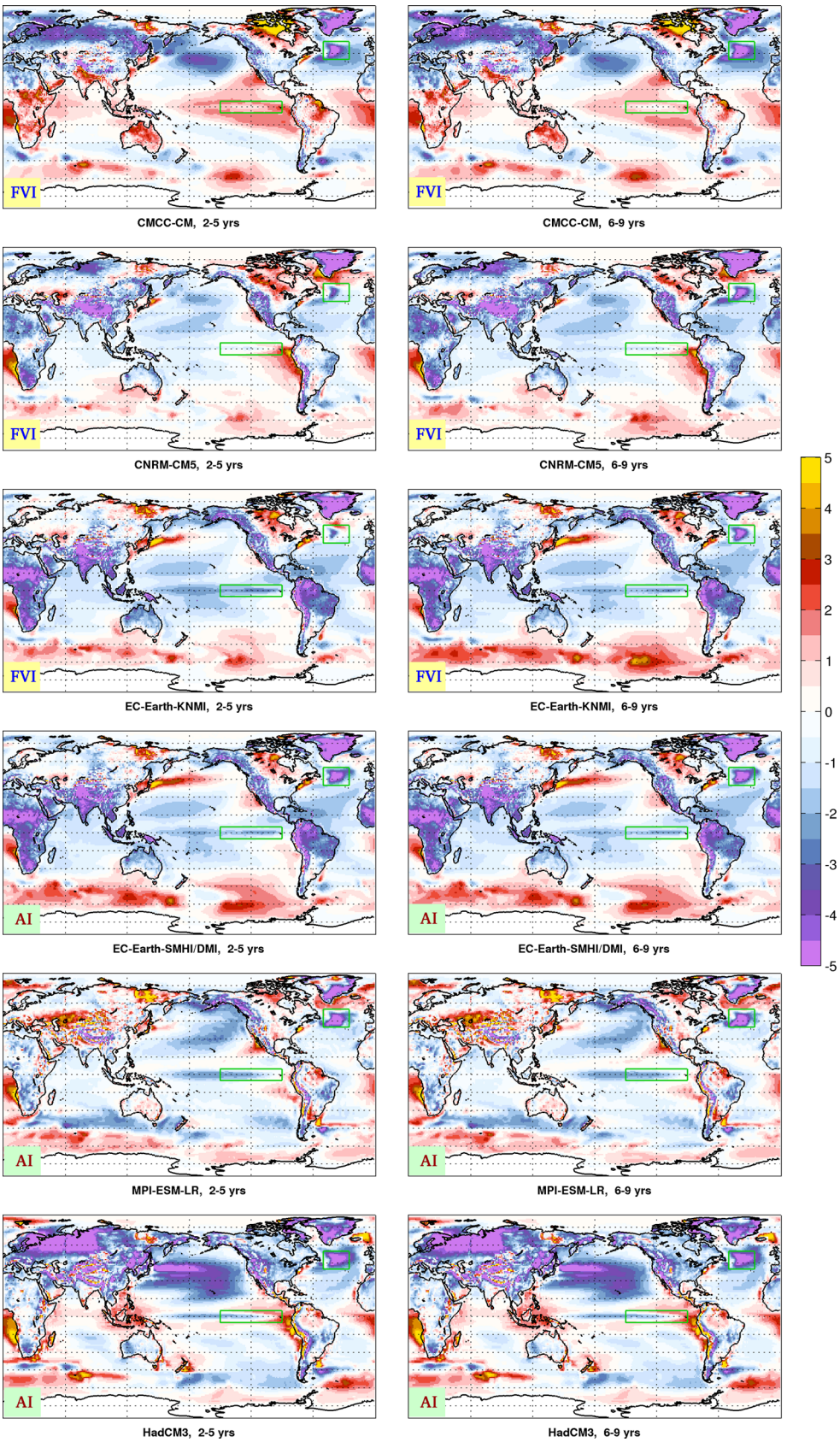


Fig. 7 Surface temperature error patterns (K; computed as model minus observation) averaged over lead-time intervals (*left*) 2–5 and (*right*) 6–9 years and across all hindcasts, for each decadal prediction system of the COMBINE ensemble. The corresponding initialization strategy (full-value/anomaly) is indicated in the *bottom-left corner*. *Green frames* indicate the NINO3 and Gulf Stream extension (see text) areas used in the regional error analysis, illustrated in Fig. 8

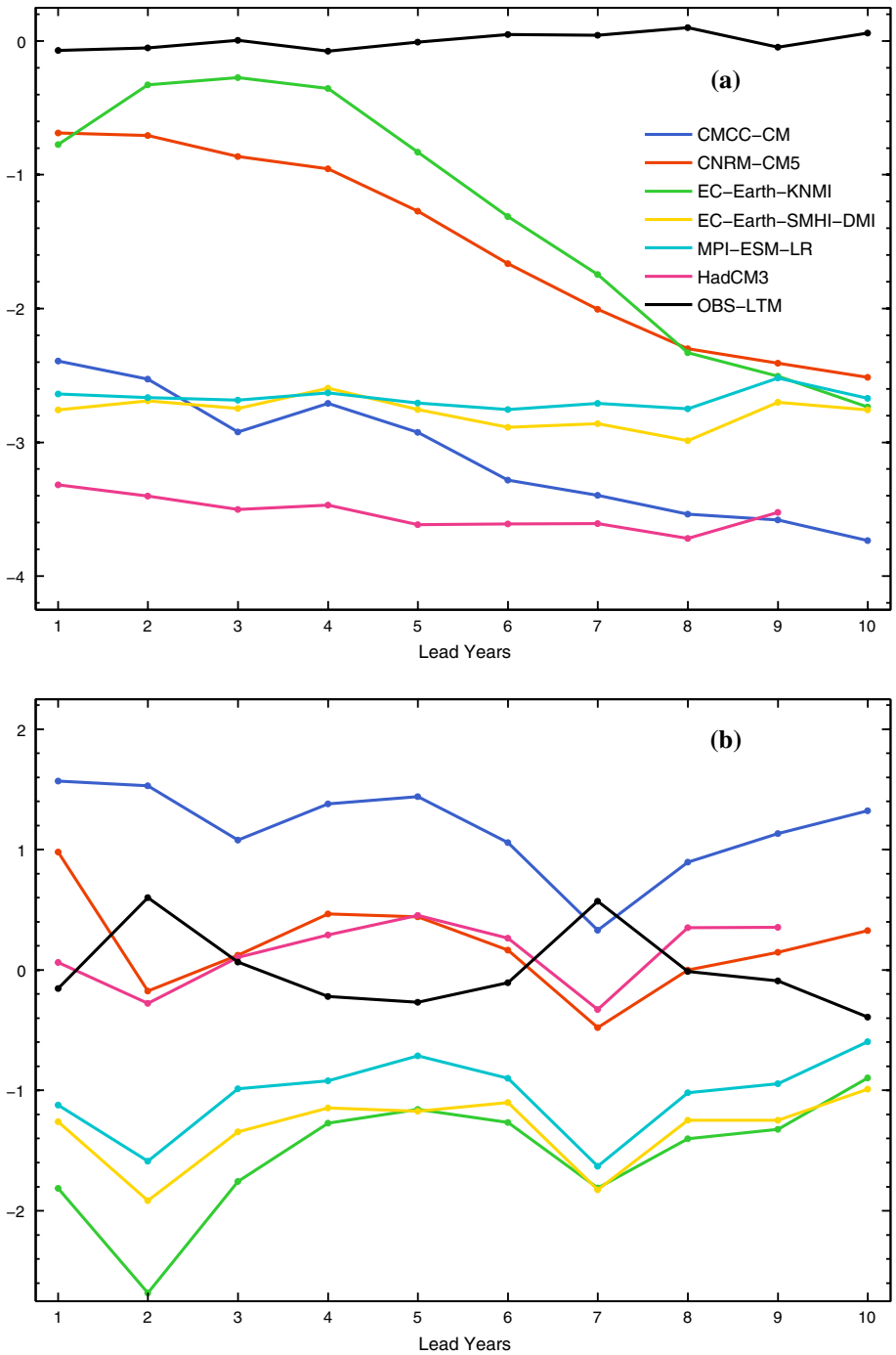


Fig. 8 The model error for surface temperature (in K; model value minus observation) as a function of lead-time, averaged across all hindcasts and over (*top*) the Gulf Stream extension and (*bottom*) the NINO3 areas (see text for details). For the observed temperature (*black*) the long-term mean calculated over all lead-times has been removed

additional hints for the interpretation of the mechanism driving the skill discrepancies between the FVI and the AI ensembles. The influence of the background state on ENSO variability and predictability has been investigated by Magnusson et al. (2012) in a set of coupled integrations performed with the ECMWF forecasting system, where it is shown that correcting the model mean state and seasonal cycle (through flux adjustments) has positive implications for the representation of ENSO, with respect to a twin set of free (i.e., non flux-adjusted) simulations based on anomaly initialization. A consistent mechanism may be at work in the present set of decadal predictions, with the FVI set of simulations being more skilful in reproducing the year-to-year variability (and therefore, displaying higher predictability) in the tropical Pacific, with respect to the AI integrations.

Overall, the present findings are consistent with the results of Smith et al. (2013), providing additional indications of an improved forecast skill in FVI over AI systems in the near-term range, although the differences are only marginally significant and restricted to the Indo-Pacific

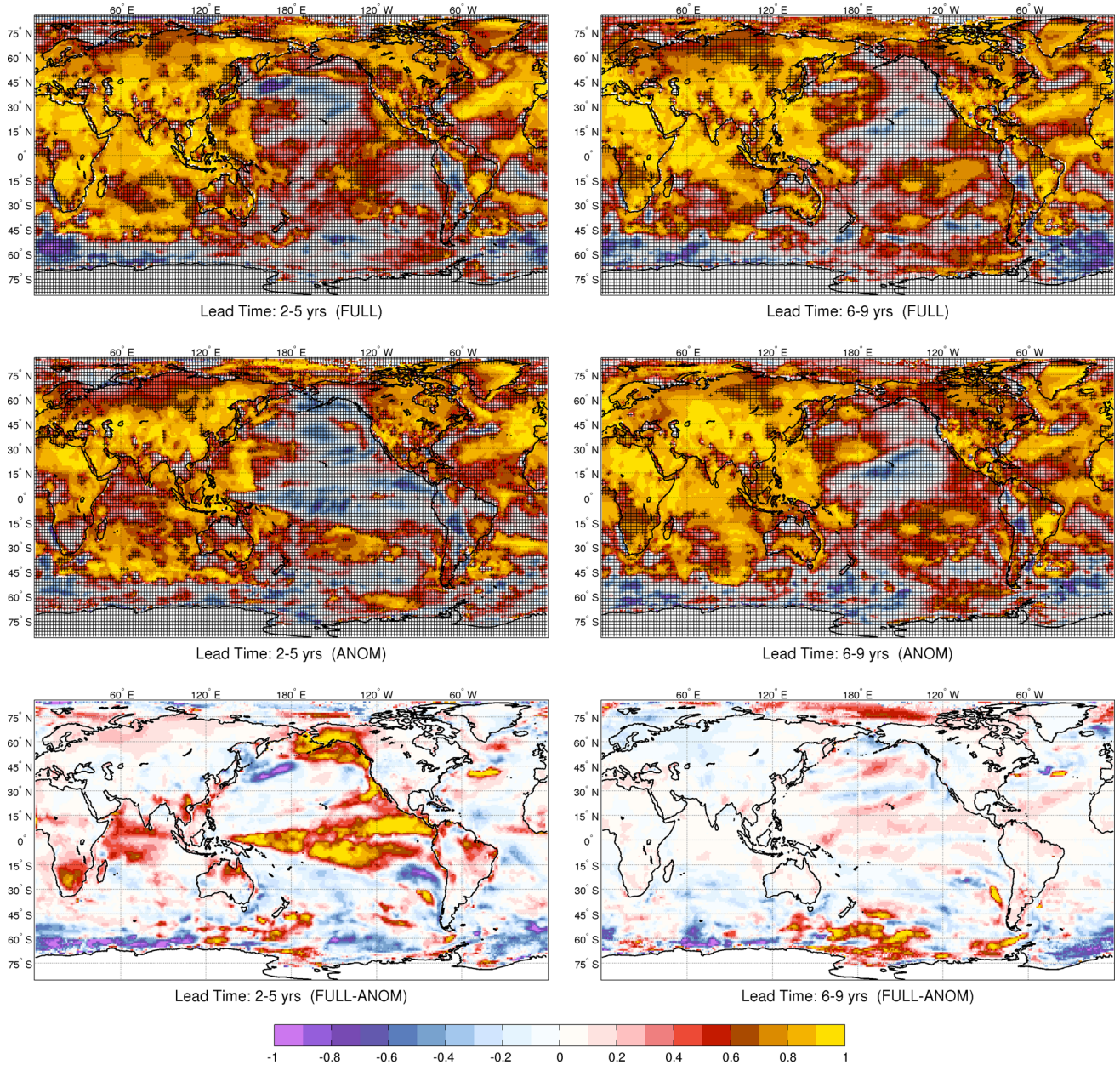


Fig. 9 (*Top*) Anomaly correlation coefficient (ACC) maps for T2M (over land) and SST, for the lead-time periods (*left*) 2–5 and (*right*) 6–9 years for the multi-model mean computed across FVI models. Stippling is used to indicate points where statistical significance is below the 95 % level, according to a one-tailed *t* test accounting for autocorrelation in the time series. (*Middle*) Same as (*top*) but for AI models. (*Bottom*) Difference maps between FVI and AI ACC patterns

equatorial region. The detailed mechanisms determining the lower skill for anomaly initialization deserve further investigation.

7 Consensus across model predictions

The purpose of this section is to provide additional insight into the multi-model ensemble predictions by documenting the degree of agreement or disagreement between the different DPS. The final goal is to obtain a spatial mapping of the confidence associated with model forecasts. To achieve this target, we provide a quantitative analysis of the level of consensus between different predictions of surface temperature in the CME. Here, the term "consensus" refers to the level of consistency across predictions performed using different systems, which in turn try to hindcast the same (observed) variability.

While, strictly speaking, this should not be confused with the actual uncertainty in the predictions in respect to

observations (here we only evaluate the cross-model agreement), it can be regarded as an indicator of it, bearing in mind that a large model spread is typically indicative of a consistently high uncertainty.

Generally, the sources of uncertainty in climate predictions can be grouped into three distinct categories (Hawkins and Sutton 2009). These include: (i) uncertainty associated with the initial state of the coupled system (e.g., different ocean reanalyses, all of them providing a plausible estimate of the real system, can be used to constrain the initial state of the model), (ii) the model uncertainty (associated with differences in individual model structure, sensitivity to the external forcings and representation of internal variability) and (iii) uncertainty in future radiative forcing scenarios (e.g., associated to the use of alternative projected changes of GHG emissions). In the present case, since all systems have been forced using a common set of prescribed boundary conditions (historical and RCP4.5 radiative forcing), the leading causes for the ensemble spread (interpreted as uncertainty in the predictions) are only related to points (i) and (ii).

As a measure of consensus between model predictions we use two metrics. Differences in the predictions between the DPSs are quantified through the average root-mean-square error (RMSE) computed from all (15 distinct) DPS pairs.[1] In these calculations we use the predicted annual-mean anomalies after removing the long-term, linear trends. In order to evaluate the temporal coherence between predictions carried out with different DPSs, the average of the ACCs between all prediction-pairs is diagnosed. Namely, in this case the ACC is defined between predictions in the same way that it was previously defined between predictions and observations. In order to distinguish the ACC computed between predictions from the "standard" ACC (i.e., between model and observed values) we will refer to the former as x-ACC. RMSE and x-ACC provide a complementary assessment of the consistency across different members of the CME, in other words, quantifying the DPS-to-DPS discrepancies, and their temporal coherence, respectively. Prior to the computation of RMSE and x-ACC point-by-point, all models have been interpolated onto a common regular grid. The metrics are evaluated separately for the lead-time periods 2–5 and 6–9 years, ultimately yielding a point-wise spatial mapping of the uncertainty characterizing the analysed set of predictions.

In Fig. 10, the RMSE for surface temperature (T2M over land, and SST) is displayed. It is seen that the areas featuring the largest spread correspond to the most intense ocean current systems, such as the western boundary currents (Gulf Stream/North Atlantic Current and Kuroshio) and parts of the Antarctic Circumpolar Current, in the Southern Ocean. Similarly, large departures are found in the Nordic Seas, at the sea-ice edge. Secondary maxima are found in the equatorial Pacific, and in northern Eurasia and North America. The RMSE pattern is qualitatively similar for both lead-time periods, but in the long-term (6–9 years) its amplitude is smaller. This result is consistent with the smaller ensemble spread found in the global mean ACC (Fig. 5) for the 6–9 (compared with the 2–5) year range. Possible causes behind this result have already been discussed in Sect. 5.2. The RMSE patterns were found to match closely the SST interannual standard deviation (SD) (not shown) calculated separately for each DPS using all 10-year hindcasts, and then averaged between the six DPSs. Slow fluctuations in the position of large-scale frontal systems is at the origin of the strong interannual variability associated with western boundary currents and the Antarctic Circumpolar Current. Variability in the Nordic Seas, on the other hand, is likely to be related to processes involving sea-ice formation. The high spatial correlation between these two patterns (RMSE and SD) indicates that the largest spread across predictions occurs at locations where models exhibit strong interannual variability. As expected, a similar result is found after comparing the model-to-model RMSE pattern, shown in Fig. 10, with the standard RMSE between the multi-model mean predictions and the observations, over the same lead-years (not shown). Note also that the RMSE metric bears a strong formal similarity with the standard inter-model ensemble spread, as given by the SD calculated across the multi-model set of climate forecasts. This is not surprising since both diagnostics are designed to quantify the level of dispersion across the multi-model ensemble of decadal forecasts. Their mathematical equivalence is determined by the fact that both metrics provide a measure of the Euclidean distance between different model forecasts. This is confirmed by the high spatial correlation found after comparing the corresponding patterns of RMSE and ensemble spread (not shown).

Next, the x-ACC patterns are shown in Fig. 11. An interesting feature in these maps is that the average correlation between the predictions of DPS pairs is positive over most of the domain for both of the examined lead-time periods, with extensive regions featuring ACC values exceeding 0.6. On the other hand, negative, statistically non-significant, correlations appear to be confined to the Southern Hemisphere. Also, correlations seem to be higher in the 6–9 year range, comparing to the shorter range predictions, a difference that is more pronounced in the Northern Hemisphere. In particular, models display a better

[1] The six available DPSs make 15 distinct pairs. For each pair, the corresponding RMSE is calculated between the two hindcasts. These differences are not "errors" but just deviations. We use the term RMSE because it best describes the calculations involved. In the text, we refer to the average RMSE across all pairs.

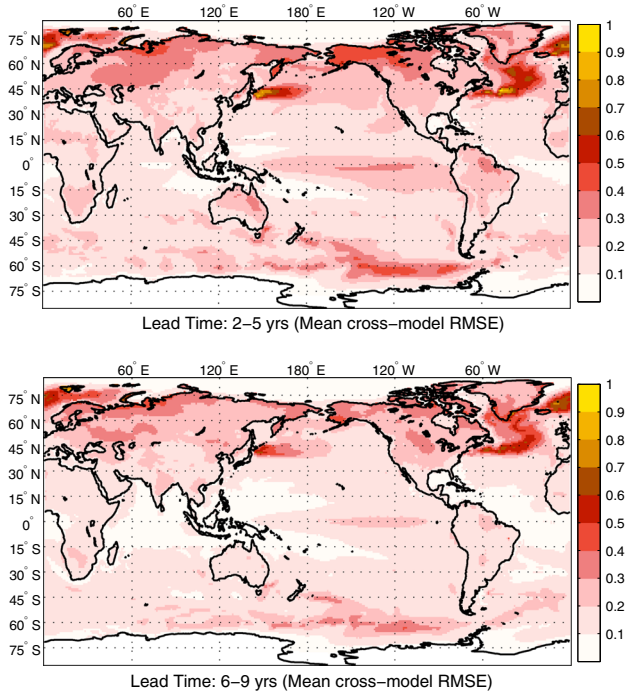


Fig. 10 RMSE of surface temperature (T2M over land and SST, in K) of annual mean predictions for the lead-time ranges 2–5 (*top*) and 6–9 years (*bottom*). The maps display the average RMSE across all 15 distinct DPS pairs; see text for details

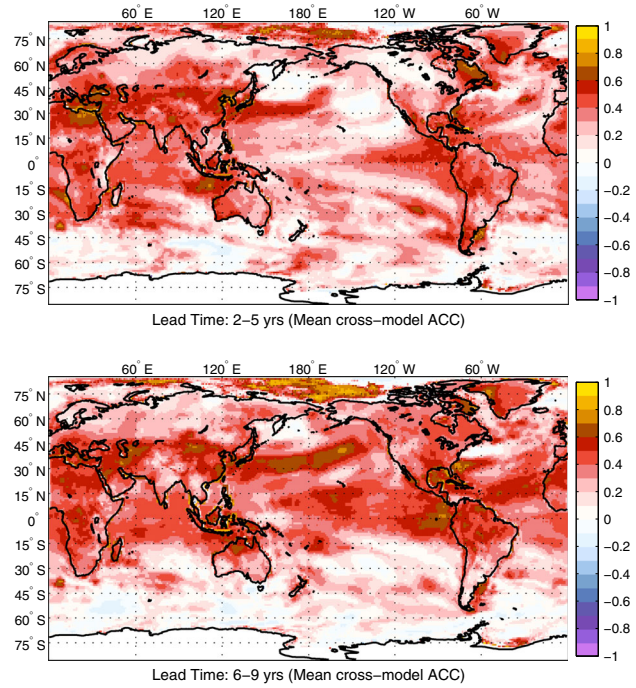


Fig. 11 ACC calculated between annual mean model predictions (x-ACC) for surface temperature (T2M over land and SST) over the lead-time ranges 2–5 (*top*) and 6–9 years (*bottom*). The maps display the average ACC across all 15 distinct DPS pairs; see text for details

agreement on the predicted evolution of tropical SSTs at longer lead times, possibly due to a stronger influence of the predictable signal induced by the radiative forcing (Boer and Lambert 2008). Overall, these results indicate a substantial consistency (in polarity, but not necessarily at the magnitude) between the predicted anomalies of most DPSs. It is worth noting that x-ACC fields display local minima in correspondence to regions featuring the largest RMSE, i.e., at the western boundaries of the North Atlantic and North Pacific ocean basins, as well as over high-latitude Eurasia and North America. A similar orthogonality is found in the equatorial Atlantic and in the northern Indian Ocean, where local maxima in x-ACC correspond to minima in the RMSE. Thus, not surprisingly, the largest intra-model departures are found at the locations where the lowest model-to-model correlations occur.

8 Conclusions

In this study we analyzed a multi-model ensemble of decadal predictions performed using five different CGCMs (combined into six prediction systems) following the CMIP5 protocol, within the framework of the EU COMBINE project. The predictive capabilities of the multi-model ensemble were examined, both at the global and at regional scales, with additional focuses on the influence of the initialization strategy on predictive skill, and on the level of mutual agreement across different model predictions. All of the analyses were conducted on sufficiently well observed variables, including sea surface temperature, near surface air temperature and precipitation over land. Although most of the skill associated with surface temperature fields is dictated by the prescribed boundary conditions, after removing the long-term trends a significant residual predictive skill was found over large oceanic and continental areas. In particular, the multi-decadal variability of SST in the Atlantic basin appears to be skillfully reproduced by individual forecast systems, showing considerable predictive capability up to O(10) years. Similarly, long-term predictability was found for near-surface air temperature over Northern Africa and the adjacent Mediterranean and Middle East. Contrastingly, compared to surface temperatures, precipitation exhibits much lower predictability, except at a few limited areas, including the African Sahel, parts of North America, and Eastern Europe. These results are consistent with a similar analysis performed by Doblas-Reyes et al. (2013), using a different multi-model set. Both the present and Doblas-Reyes et al. analyses suggest a strong connection between the Atlantic multi-decadal variability and the surface temperature and rainfall changes occurring over regions

adjacent to the Atlantic basin. This link may be a signature of the AMO teleconnection pattern, as suggested in a number of previous studies (among the others, Knight et al. 2006; Zhang and Delworth 2006). Goddard et al. (2013) also inspect the forecast skill for precipitation in two DPSs (DePreSys and CanCM4) but over the longer 2–9 lead-years interval, and find significant anomaly correlations over the high latitudes of the northern hemisphere (both sets of hindcasts) and over much of the tropics (only in CanCM4).

The overall emerging structure of the COMBINE multi-model ensemble predictive skill for surface temperature fields is largely consistent with results from analogous assessments based on both multi-model (Doblas-Reyes et al. 2011, 2013; García-Serrano and Doblas-Reyes 2012; van Oldenborgh et al. 2012; Kim et al. 2012; Goddard et al. 2013; Chikamoto et al. 2013) and single-model (Pohlmann et al. 2009; Bellucci et al. 2013; García-Serrano et al. 2012; Matei et al. 2012b) decadal hindcasts. Common features include, in particular: (i) a strong predictive skill in the Atlantic sector, (ii) a pronounced asymmetry in the predictive skill between the Atlantic and Pacific oceanic basins, and (iii) negative ACC over parts of South America. The consensus in predictive skill featured by sensibly different DPSs corroborates the view that current CGCMs share common predictability features, as well as common deficiencies. In particular, the high predictive skill found in the Atlantic region supports the widely recognized role of the slowly evolving changes in the strength of the thermohaline circulation and meridional heat transport as primary drivers of the low-frequency variability in the Atlantic. The generally low skill found in the Pacific basin for SSTs, is a similarly robust feature, also emerging in Kim et al. (2012; their Fig. 3) and Doblas-Reyes et al. (2011; their Fig. 1). Consistent with this finding, Kim et al. (2012), Branstator and Teng (2012) and Bellucci et al. (2013) show a rapid decay with lead-time of the forecast skill associated with SST in the extra-tropical North Pacific in a number of near-term predictions performed with CMIP5 models [however, hints of improved predictability in the North Pacific sub-surface temperatures are found by Mochizuki et al. (2010), Chikamoto et al. (2013)]. In light of the surface temperature trend analysis (Fig. 3), the poor skill found in the extra-tropical North Pacific seems to reflect the inability of the models to correctly reproduce the observed ratio between forced and unforced variability in this region, where the warming trend explains a small fraction of the total variability.

An important finding of the predictive skill analysis is that the multi-model ensemble mean does generally outperform the individual forecasts (when this is not the case, it is at least comparable to the best of them). This finding has been well-documented for seasonal forecasting, but here is found to hold also at the decadal range, supporting the need for large multi-model ensembles to provide valuable decadal predictions (see also Kim et al. 2012).

One of the open questions for the decadal prediction praxis relates to the identification of an optimal (i.e., skill-maximizing) initialization strategy. By clustering the six analysed DPSs in two groups according to the initialization method (full-value and anomaly initialization), it was possible to frame this issue (so far only examined for individual systems; Smith et al. 2013; Hazeleger et al. 2013) in a multi-model perspective. From the comparison between the predictive skill patterns associated with these two groups, the equatorial Pacific emerged as an area particularly sensitive to the details of the initialization method. In particular, the anomaly initialization seems to exert a deteriorating impact on the skill in the equatorial Pacific, while a sensible improvement is obtained in the full-state initialized systems. Should this result be confirmed (possibly by extending the same analysis to a larger set of systems), there would be implications for the design of multiple model ensembles for operational decadal climate predictions (Smith et al. 2012), as the balance between anomaly and full-value initialized systems may strongly affect the regional skill of the multi-model mean.

Finally, the consistency across the different model predictions of surface temperature was assessed. The model-to-model RMSE pattern indicated that the largest departures between different DPSs occur at the same locations where the largest year-to-year variability is found (in particular, over the areas dominated by intense western boundary currents and their open ocean extensions). The pattern of temporal coherence across different systems reveals a general agreement between models in predicting the near-term evolution of surface temperature fields (referring to the sign of the corresponding anomalies), displaying positive correlations between different decadal predictions over most of the global domain.

The present analysis adds to the growing evidence that the current generation of climate models adequately initialized have significant skill in predicting years ahead not only the anthropogenic warming but also part of the internal variability of the climate system. The high skill detected in the Atlantic sector, and extending over the potentially linked surrounding areas (including the Mediterranean and Sahel climatic hot-spots), emerges as a particularly robust feature of CMIP5 models. This finding discloses a promising future for the new-born decadal predictions field, envisaging the possibility of valuable assessments of climatic fluctuations at the regional scale over a multi-year horizon. This calls for additional efforts aiming at the improvement of the existing Earth observing network, so as to better constrain future climate predictions. This step needs to involve not only the oceanic sub-

system, but should also be extended to those less-observed components (land-surface, cryosphere, stratosphere, aerosols) which may introduce additional memory (and thus predictability) in the climatic system, beyond the seasonal scale.

Finally, in the perspective of providing trustworthy decadal climate forecasts to inform end-users, a step-change in the amount of dedicated computational resources is required. The statistical robustness of the forecast skill featured by most of the decadal predictions made available by the modeling groups worldwide through the CMIP5 effort, is severely hampered by the low number (10, according to the CMIP5 protocol for near-term predictions; Taylor et al. 2012) of the canonical initialization start-dates typically used to perform the decadal integrations. The lessons we are learning in this pioneering pre-operational stage will help to design the next generation of coordinated decadal climate forecast experiments, out to CMIP6 and beyond.

Acknowledgments The authors gratefully acknowledge the support from the EU FP7 COMBINE Project (Grant Agreement Number 226520). A.B., S.G. and P.J.A. did also receive support from the Italian Ministry of Education, University and Research and Ministry for Environment, Land and Sea through the Project GEMINA. We also wish to thank Dr. G. J. van Oldenborgh for providing some of the data used in this assessment by means of the KNMI Climate Explorer utility. Finally, the insightful comments from two anonymous reviewers are thankfully acknowledged.

References

Bellucci A, Gualdi S, Masina S, Storto A, Scoccimarro E, Cagnazzo C, Fogli P, Manzini E, Navarra A (2013) Decadal climate predictions with a coupled OAGCM initialized with oceanic reanalyses. Clim Dyn 40:1483–1497. doi:10.1007/s00382-012-1468-z

Boer GJ, Lambert SJ (2008) Multi-model decadal potential predictability of precipitation and temperature. Geophys Res Lett 35:L05706

Branstator G, Teng H (2010) Two limits of initial-value decadal predictability in a CGCM. J Clim 23:6292–6310

Branstator G, Teng H (2012) Potential impacts of initialization on CMIP5 decadal predictions. Geophys Res Lett. doi:10.1029/2012GL051974

Branstator G, Teng H, Meehl G, Kimoto M, Knight J, Latif M, Rosati A (2012) Systematic estimates of initial value decadal predictability for six AOGCMs. J Clim 25:1827–1846

Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Blad I (1999) The effective number of spatial degrees of freedom of a time-varying field. J Clim 12:1990–2009

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

Chikamoto Y, Kimoto M, Ishii M, Mochizuki T, Sakamoto TT, Tatebe H, Komuro Y, Watanabe M, Nozawa T, Shiogama H, Mori M, Yasunaka S, Imada Y (2013) An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. Clim Dyn 40:1201–1222. doi:10.1007/s00382-012-1351-y

Doblas-Reyes FJ, van Oldenborgh GJ, García-Serrano J, Pohlmann H, Scaife AA, Smith D (2011) CMIP5 near-term climate prediction. CLIVAR Exch 16(2):8–11

Doblas-Reyes FJ, Andreu-Burillo I, Chikamoto Y, García-Serrano J, Guemas V, Kimoto M, Mochizuki T, Rodrigues LRL, van Oldenborgh GJ (2013) Initialized near-term regional climate change prediction. Nat Commun. doi:10.1038/ncomms2704

García-Serrano J, Doblas-Reyes FJ (2012) On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. Clim Dyn 39. doi:10.1007/s00382-012-1413-1

García-Serrano J, Doblas-Reyes F-J, Coelho CAS (2012) Understanding Atlantic multi-decadal variability prediction skill. Geophys Res Lett 39:L18708. doi:10.1029/2012GL053283

Goddard L et al (2013) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn 40:245–272. doi:10.1007/s00382-012-1481-2

Griffies SM, Bryan K (1997) A predictability study of simulated North Atlantic multidecadal variability. Clim Dyn 13:459–487

Guemas V, Corti S, García-Serrano J, Doblas-Reyes F, Balmaseda M, Magnusson L (2012) The Indian Ocean: the region of highest skill worldwide in decadal climate prediction. J Clim. doi:10.1175/JCLI-D-12-00049.1

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

Hawkins E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. Bull Am Meteorol Soc 90:1095–1107

Hazeleger W, Guemas V, Wouters B, Corti S, Andreu-Burillo I, Doblas-Reyes FJ, Wyser K, Caian M (2013) Multiyear climate predictions using two initialization strategies. Geophys Res Lett 40:17941798. doi:10.1002/grl.50355

Hurrell JW et al (2009) Decadal climate predictions: opportunities and challenges. OceanObs'09, Community White Paper, pp 1–21

International CLIVAR Project Office (ICPO) (2011) Data and bias correction for decadal climate predictions. International CLIVAR Project Office, CLIVAR publication series no 150, p 6

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Röckner E (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453:84–88

Kim HM, Webster PJ, Curry JA (2012) Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. Geophys Res Lett 39:L10701. doi:10.1029/2012GL051644

Knight JR, Allan R, Folland CK, Vellinga M, Mann ME (2005) A signature of persistent natural thermohaline circulation cycle in observed climate. Geophys Res Lett 32:L20708

Knight JR, Folland CK, Scaife AA (2006) Climate impacts of the Atlantic multidecadal oscillation. Geophys Res Lett 33:L17706

Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of climate models. Clim Dyn 17:83–106. doi:10.1007/PL00013736

Latif M, Böning C, Willebrand J, Biastoch A, Dengg J, Keenlyside N, Schweckendiek U, Madec G (2006) Is the thermohaline circulation changing? J Clim 19:4631–4636

Magnusson L, Alonso-Balmaseda M, Molteni F (2012) On the dependence of ENSO simulation on the coupled model mean state. Clim Dyn. doi:10.1007/s00382-012-1574-y

Matei D, Pohlmann H, Jungclaus J, Müller W, Haak H, Marotzke J (2012a) Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. J Clim. doi:10.1175/JCLI-D-11-00633.1

Matei D, Baehr J, Jungclaus J, Haak H, Müller WA, Marotzke J (2012b) Multiyear prediction of monthly mean Atlantic meridional overturning circulation at 26.5N. Science 335:76–79

Meehl GA, Goddard L, Murphy J, Stouffer RJ, Boer G, Danabasoglu G, Dixon K, Giorgetta MA, Greene AM, Hawkins E, Hegerl G, Karoly D, Keenlyside N, Kimoto M, Kirtman B, Navarra A, Pulwarty R, Smith D, Stammer D, Stockdale T (2009) Decadal prediction: can it be skillful? Bull Am Meteorol Soc 90:1467–1485

Meehl GA et al (2013) Decadal climate prediction: an update from the trenches. Bull Am Meteorol Soc. doi:10.1175/BAMS-D-12-00241.1

Mochizuki T et al (2010) Pacific decadal oscillation hindcasts relevant to near-term climate prediction. PNAS 5:1833–1837. doi:10.1073/pnas.0906531107

Palmer TN et al (2004) Development of a European multi-model ensemble system for seasonal to inter-annual prediction. Bull Am Meteorol Soc 85:853–872

Pohlmann H, Jungclaus JH, Köhl A, Stammer D, Marotzke J (2009) Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic. J Clim 22:3926–3938

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14):4407. doi:10.1029/2002JD002670

Schneider U, Fuchs T, Meyer-Christoffer A, Rudolf B (2008) Global precipitation analysis products of the GPCC. Technical report, Global Precipitation Climatology Centre (GPCC), Deutscher Wetterdienst, Offenbach

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317:796–799

Smith DM et al (2012) Real-time multi-model decadal climate predictions. Clim Dyn. doi:10.1007/s00382-012-1600-0

Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. Clim Dyn. doi:10.1007/s00382-013-1683-2

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. doi:10.1175/BAMS-D-11-00094.1

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

Ting M, Kushnir Y, Seager R, Li C (2009) Forced and internal twentieth-century SST trends in the North Atlantic. J Clim 22:1469–1481

Trenberth KE, Shea DJ (2006) Atlantic hurricanes and natural variability in 2005. Geophys Res Lett 33:L12704

van Oldenborgh GJ, Doblas-Reyes FJ, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Clim Dyn 38:1263–1280. doi:10.1007/s00382-012-1313-4

Wilks DS (2006) Statistical methods in the atmospheric sciences. International geophysics series, vol 91, 2nd edn. Academic Press, London

Zhang R, Delworth TL (2006) Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. Geophys Res Lett 33:L17712. doi:10.1029/2006GL026267