Clim Dyn (2012) 38:1263–1280
DOI 10.1007/s00382-012-1313-4

# Decadal prediction skill in a multi-model ensemble

**Geert Jan van Oldenborgh · Francisco J. Doblas-Reyes · Bert Wouters · Wilco Hazeleger**

Received: 23 September 2010 / Accepted: 7 February 2012 / Published online: 25 February 2012


**Abstract** Decadal climate predictions may have skill due to predictable components in boundary conditions (mainly greenhouse gas concentrations but also tropospheric and stratospheric aerosol distributions) and initial conditions (mainly the ocean state). We investigate the skill of temperature and precipitation hindcasts from a multi-model ensemble of four climate forecast systems based on coupled ocean-atmosphere models. Regional variations in skill with and without trend are compared with similarly analysed uninitialised experiments to separate the trend due to monotonically increasing forcings from fluctuations around the trend due to the ocean initial state and aerosol forcings. In temperature most of the skill in both multi-model ensembles comes from the externally forced trends. The rise of the global mean temperature is represented well in the initialised hindcasts, but variations around the trend show little skill beyond the first year due to the absence of volcanic aerosols in the hindcasts and the unpredictability of ENSO. The models have non-trivial skill in hindcasts of North Atlantic sea surface temperature beyond the trend. This skill is highest in the northern North Atlantic in initialised experiments and in the subtropical North Atlantic in uninitialised simulations. A similar result is found in the Pacific Ocean, although the signal is less clear. The uninitialised simulations have good skill beyond the trend in the western North Pacific. The initialised experiments show some skill in the decadal ENSO region in the eastern Pacific, in agreement with previous studies. However, the results in this study are not statistically significant ($p \approx 0.1$) by themselves. The initialised models also show some skill in forecasting 4-year mean Sahel rainfall at lead times of 1 and 5 years, in agreement with the observed teleconnection from the Atlantic Ocean. Again, the skill is not statistically significant ($p \approx 0.2$). Furthermore, uninitialised simulations that include volcanic aerosols have similar skill. It is therefore still an open question whether initialisation improves predictions of Sahel rainfall. We conclude that the main source of skill in forecasting temperature is the trend forced by rising greenhouse gas concentrations. The ocean initial state contributes to skill in some regions, but variations in boundary forcings such as aerosols are as important in decadal forecasting.

**Keywords** Decadal forecasts · forecast verification · climate change · climate model · Atlantic multidecadal oscillation · ENSO

G. J. van Oldenborgh (✉) · B. Wouters · W. Hazeleger
Royal Netherlands Meteorological Institute (KNMI),
P.O. Box 201, AE 3730 De Bilt, The Netherlands
e-mail: oldenborgh@knmi.nl

F. J. Doblas-Reyes
Institució Catalana de Recerca i Estudis Avançats (ICREA)
and Institut Català de Ciències del Clima (IC3),
Doctor Trueta 203, 08005 Barcelona, Spain

## 1 Introduction

Observations, supported by climate models, indicate that the Earth's climate fluctuates over a wide range of time scales. Several regions, such as the North Atlantic and Pacific Oceans, are characterised by variations on decadal to inter-decadal timescales, which are manifested in substantial changes in sea surface temperature and ocean heat storage. Through coupling with the atmosphere, these low-frequency variations have been linked to changes in precipitation and temperature over land, hurricane activity in the Atlantic Ocean and Indian monsoon intensity (e.g., Zhang and Delworth 2006; Meehl et al. 2006; Knight et al.

2006; Smith et al. 2010). Because of their potentially large socio-economic impact, climate predictions over interannual to decadal time scales have recently gained increased attention (Zhang and Delworth 2006; Räsänen and Ruokolainen 2006; Ruokolainen and Räsänen 2007; Meehl et al. 2009; Keenlyside and Ba 2010). They bridge the gap between seasonal forecasts and century-scale climate projections for the twenty-first century and have the potential to provide valuable information on near-future climate, which ultimately may serve as a base to inform climate change adaptation policy (Cox and Stephenson 2007).

Centennial-scale climate predictions are mainly determined by the prescribed boundary conditions: the scenario chosen to describe the future emissions of aerosols and greenhouse gases to the atmosphere, solar forcing and volcanic activity. On shorter time scales the natural variability is larger than the trend, so that the skill of seasonal forecasts with lead times of a month to a year is mainly due to the initial state. Decadal predictions are intermediate to these two: they are controlled by both the initial and boundary conditions (Hawkins and Sutton 2009).

The importance of including realistic initial conditions in decadal predictions has been illustrated in a number of recent studies. Smith et al. (2007) found that initialising with observed ocean and atmosphere conditions improves the skill in predicting global temperature and heat content anomalies a decade ahead. Subsequent analysis (Robson 2010) found that the regional patterns of skill presented in Smith et al. (2007) were affected by model drifts. After correcting for this, regional improvements through initialisation were found mainly in the North Atlantic ocean (Robson 2010).

Other studies also found only regional improvement in skill, mainly over the North Atlantic and Pacific Oceans. Keenlyside et al. (2008) showed that including a sea surface temperature (SST) initialisation scheme leads to improved skill in predicting surface temperature in the North Atlantic area, which was attributed to an improved Atlantic meridional overturning circulation (AMOC) in the initialised hindcasts. However, salinity was not constrained in the initial conditions, and it is unclear whether SST alone is sufficient to constrain the AMOC (Dunstone and Smith 2010). The predictability over the ocean was corroborated by Pohlmann et al. (2009) using ocean synthesis fields as initial conditions. Similar results were recently found by Smith et al. (2010); Robson (2010). For the Pacific ocean, Mochizuki et al. (2010); Yasunaka et al. (2011) demonstrated that proper initialisation of their coupled atmosphere-ocean model leads to skilful predictions of upper-ocean temperatures in the regions typically affected by the Pacific Decadal Oscillation (PDO).

These investigations all used a single model, and compared the skill over the uninitialised simulations with the same model. Here, we investigate the hindcast skill of a multi-model ensemble of decadal hindcasts made within the European ENSEMBLES project (van der Linden and Mitchell 2009). In seasonal forecasting, it has been shown that the skill of a multi-model ensemble frequently exceeds the skill of the best contributing model (Hagedorn et al. 2005; Doblas-Reyes et al 2005; Weigel et al. 2008). Unfortunately, uninitialised simulations of identical models are not available. We investigate the total skill of the hindcasts, and separate the skill in a fraction proportional to a non-linear trend, and a fraction that is not described by this simple trend. To aid in the identification of the sources of skill we compare the results to those of the same analysis of the multi-model ensemble from the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl et al. 2007).

Skill in these simulations of past climate comes from the following sources (Hawkins and Sutton 2009).

1. The rising trend of well-mixed greenhouse gases, mainly $CO_2$ (Keeling et al. 1976; IPCC 2007). This trend can be predicted well on the decadal time scale. This is included in all simulations under consideration.
2. Temporal variations in solar activity and stratospheric aerosols due to volcanic eruptions (Robock 2000). These variations cannot be predicted years ahead of time (except in the case of an analysis just after a major tropical eruption). This is included in half the models of the the uninitialised CMIP3 ensemble but not in the initialised ENSEMBLES hindcasts.
3. The temporal and spatial evolution of tropospheric aerosol fields (e.g. Rotstayn and Lohmann 2002; Wild 2009). These can be predicted to some extent on the 10-year time scale based on historical data and scenarios of emissions of aerosols and their precursors. This is in principle included in all simulations, although the effects differ strongly among the models (e.g., Ruckstuhl and Norris 2009).
4. The predictable component of natural climate variability. In principle included in the initialised ENSEMBLES hindcasts, although deficiencies in the model and initial state limit the skill.

We therefore see that the CMIP3 simulations include some information that in a real forecast setting will usually not be available (variations in solar activity and stratospheric aerosols), whereas the ENSEMBLES hindcasts mainly use information that will also be available to real initialised forecasts. From the differences in skill between the initialised and uninitialised ensembles in the trend and beyond the trend we attempt to distinguish between these sources.

## 2 Methods

### 2.1 Models

The ENSEMBLES multi-model for the decadal prediction consists of four forecast systems denoted by ARPEGE4.6, IFS33r1, ECHAM5 and HadGEM2. All models include the main radiative forcings and none have flux adjustments at the ocean surface. ARPEGE4.6 is the atmospheric model employed by CERFACS, it was coupled to the ocean model NEMO (Salas Mélia 2002). The weather forecast model IFS33r1 (Bechtold et al. 2008) was used by the ECMWF at a resolution of TL159/L62 coupled to the HOPE-E ocean model at 1°. The ECHAM5 model (Jungclaus et al. 2006) was used by IFM-GEOMAR coupled to ocean model MPI-OM1. UKMO used HadGEM2-AO, an improved version of the model used for the IPCC AR4 (Johns et al. 2006) with atmospheric resolution N96/L38. Except for the ECMWF model, the forecast systems are the same as those used for the ENSEMBLES seasonal to annual hindcasts (Weisheimer et al. 2009).

Ten three-member ensemble hindcasts were run for 10 years starting on November 1 of 1960, 1965, ..., 2005. Volcanic aerosol concentrations from eruptions before the analysis date were relaxed to zero with a time scale of one year in ECHAM5, while the other three models did not include any volcanic aerosol effect. In all cases, the effects of eruptions during the hindcasts were not included to reproduce a realistic forecasting context.

Three of the four models (the IFS33r1, HadGEM2 and ARPEGE4.6 models) were used with a full state initialisation strategy similar to the one employed in seasonal forecasting: starting the hindcasts from an ocean analysis that is close to the observations, with perturbations in past wind stress and SST added to sample some of the uncertainties. In contrast, IFM-GEOMAR employed anomaly initialisation, where observed SST anomalies were added to the model climatology and the combined SST restored into the coupled model (Keenlyside et al. 2008). Full details can be found in Doblas-Reyes et al (2010).

The CMIP3 ensemble used consists of 23 models (SST was only available for 22). To cover the period up to 2010, results from the SRES A1b scenario were used to extend the simulations of the twentieth century (20C3M). The temperature change over the last 10 years is not dependent on the scenario chosen (Stott and Kettleborough 2002). All models were weighted equally, i.e., first the different ensemble members of each model were averaged, and next the model means were interpolated to a common 2.5° grid and averaged into a multi-model mean. Half the models (11) include volcanic aerosols, the majority of these also account for variations in solar radiation. This subset is denoted by CMIP3v here. The other half that does not include volcanic aerosols (and often solar variability) is denoted by CMIP3n.

### 2.2 Observations

For the global mean temperature we used the estimate published by the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administration (NOAA) of the US (Smith et al. 2008). The results were checked against other estimates from the Goddard Institute of Space Science (GISTEMP, Hansen et al. 2010) and Hadley Centre/Climatic Research Unit (CRU) (HadCRUT3, Brohan et al. 2006) and no large differences were found (see Fig. 1). The $CO_2$ concentrations were taken from the Mauna Loa series (Keeling et al. 1976) obtained from the Earth System Research Laboratory (ESRL).

Land temperatures were taken from the National Centers for Environmental Prediction (NCEP) GHCN/CAMS dataset (Fan and van den Dool 2008), SST from the NCDC ERSST V3b dataset (Smith et al. 2008). These datasets have little coverage north of 60°N. In this area values from the GISTEMP dataset with 1,200 km decorrelation scale were used (Hansen et al. 2010). The large decorrelation scale is justified for the multi-year averages investigated here. Teleconnections were computed using the longer CRU TS 3.0 analysis (Mitchell and Jones 2005).

Precipitation estimates were taken from the Global Precipitation Climatology Centre (GPCC) v5 (Rudolf et al. 2010).

### 2.3 Verification measures

As the number of verification points is rather low (9 or 10, depending on the lead time) we use simple measures of

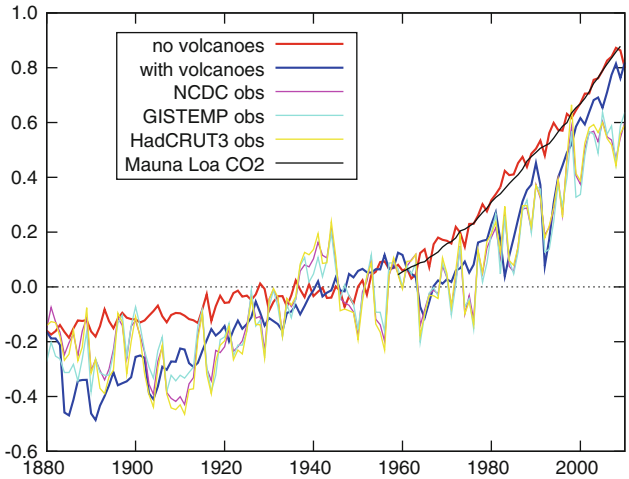


**Fig. 1** Global mean 2 m temperature anomalies (Jan–Dec annual mean relative to 1931–1960) in the CMIP3 20c3m/sresa1b experiments (with/without volcanoes) compared to the NCDC, GISTEMP and HadCRUT3 SST/T2m reconstructions. The model simulations without volcanic aerosols are compared to observed $CO_2$ concentration anomalies scaled by a factor that minimises the RMS difference between the two series

skill: the correlation coefficient $r$ and the root mean square error RMSE. More sophisticated probability-based measures have very large uncertainties for such a small sample (for an example from seasonal forecasting see van Oldenborgh et al. 2008). All forecasts have been bias corrected in the mean, for each model separately, taking into account the evolution of the bias with lead time. Cross-validation was not used, note that correlation coefficients are the same if only a single point would have been left out.

We consider three lead times: the first year (Nov–Oct) has different characteristics from the other ones due to important initial-condition predictability similar to that found in seasonal forecasting. The rest of the hindcasts is split up in equal-length near-term (years 2–5) and long-term predictions (years 6–9). Assuming normal distributions for predictor and predictand, a one-sided Student's $t$ test appropriate for a skill score shows that the $p$-value $p = 0.1$ is reached for $r = 0.44$ for 10 independent data points (year 1), whereas it has to be $r = 0.47$ to reach this significance with 9 data points (years 2–5 and 6–9) (Press et al. 1992). Serial correlations in the residuals have been taken into account by lowering the effective number of freedom using the lag-1 autocorrelation where this is significantly different from zero.

### 2.4 Trend definition

A large part of the skill in decadal temperature forecasts is due to the trend. To study this trend separately from variability around the trend a good definition of the trend is required. Figure 1 shows the global mean 2 m temperature (T2m) anomalies in the observations and in the uninitialised climate model experiments for the twentieth century and SRES A1b from 2001 onward (20c3m and sresa1b). The simulated global mean temperature rises smoothly but non-linearly in the mean of the 11 models without volcanic aerosols (CMIP3n). The curve can be described very well by the observed $CO_2$ concentrations at Mauna Loa from 1959 onwards, scaled by the regression $(11.6 \pm 0.2)\ 10^{-3}$ K/ppm. The correlation coefficient is $r = 0.994$. This indicates that over this period the global mean effect of other anthropogenic forcings, such as aerosols, are proportional to the $CO_2$ forcing in these climate models. The same result holds for the individual models, with of course a larger contribution from internal variability. The regression coefficients do vary by a factor two, from $(8.0 \pm 0.6)\ 10^{-3}$ K/ppm to $(18.6 \pm 0.5)\ 10^{-3}$ K/ppm, due to the differing climate responses. In all models the global mean temperature changes can be described well by the scaled $CO_2$ concentration changes. Correlations with the observed global mean temperature are less good ($r = 0.90$ for the multi-model ensemble mean). The regression of the modelled global mean temperature on the observations over 1959–2009 is compatible with one (1.05 K/K, with individual models ranging from 0.7 to 1.7 K/K).

The other 12 models in the CMIP3 database do include the effects of large tropical volcanic eruptions (and often variations in the solar constant). The average global mean temperature of this subset is also shown in Fig. 1. Including the effects of large tropical volcanic eruptions brings the modelled temperature anomalies into closer agreement with the observed ones ($r = 0.92$). The multi-model mean of the 12 models that include volcanic aerosols is also compatible with the observed trend 1959–2009 with a regression of 1.11 K/K, individual model results vary from 0.9 to 1.5 K/K.

The decadal hindcasts can be expected to reproduce the warming trend and some of the natural variability around the trend, but not the effects of volcanic eruptions or solar variability after the analysis date. We therefore define the *trend* as the part of the signal proportional to the rising $CO_2$ concentrations as a proxy for the smooth rise of the CMIP3 runs without volcanic aerosols. This part is determined mainly by the boundary conditions of rising greenhouse gas and aerosol concentrations. The residual of the fit gives the variability around the trend. Apart from the effects of solar variability, volcanic aerosols and tropospheric aerosols, this also includes the natural variability of the system, part of which may be predictable from the initial state.

For regional averages, the temperature trend has been attributed to increased greenhouse gases (e.g., Stott 2003). We therefore use the same trend definition on the local scale. This does not imply that we attribute trends to greenhouse gases on the local scale (which is hard given the deficiencies of climate models). This trend definition merely describes a large part of the temperature behaviour over most of the globe (cf. Knutson et al. 2006). This is illustrated in Fig. 2, which compares the observed local long-term temperature trends over 1960–2010 (Fig. 2a, estimated as the long-term regression on the $CO_2$ concentration times the rise in this concentration over 1960–2010) with the standard deviation of running 4-year mean residuals around this trend (Fig. 2b). It is clear that the trend is much larger than the 4-year standard deviation except in areas with low trends (North Pacific, North Atlantic). Fig. 2c shows the part of the 4-year standard deviation that is not explained by uncorrelated annual variability. Similar results based on climate model ensembles were found by Collins (2002); Boer (2004); Pohlmann et al (2004).

Given the size of the trend in comparison with other variability in temperature predictions, we analyse the trend separately from the variability around the trend. The trend is mainly a forced signal but does include climate variability periods of $\mathcal{O}(100)$ years or more (twice the 60-year hindcast period). The variability around the trend includes the effects of initialisation in the ENSEMBLES ensemble,

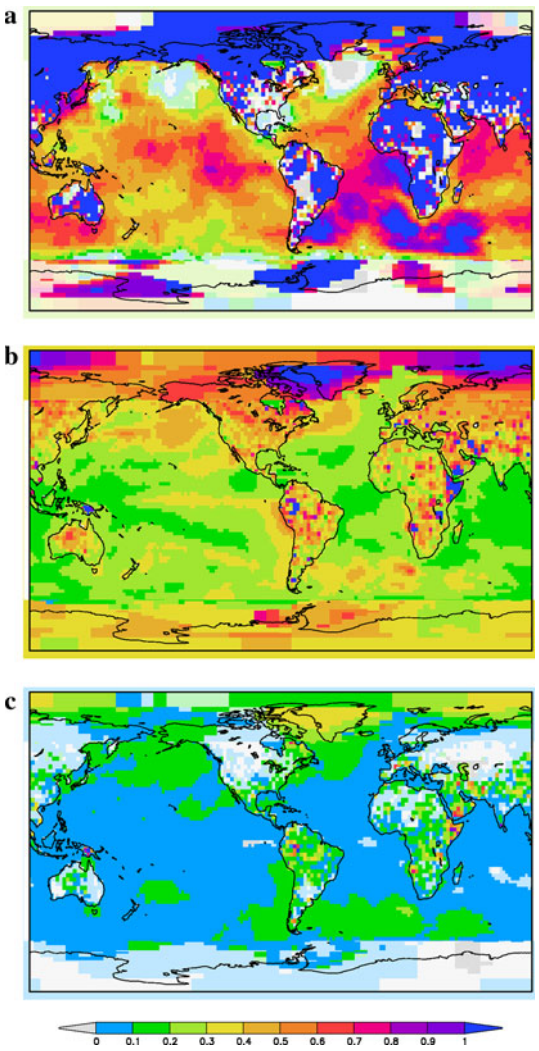


**Fig. 2** **a** Observed temperature trend between 1960 and 2010, computed as the difference in $CO_2$ levels between 1960 and 2010 times the regression of temperature on the $CO_2$ concentration over all data. **b** Standard deviation of 4-year running means of the residual of this regression. **c** As in **b** minus the contribution from uncorrelated interannual variability. *Lighter colours* indicate areas where this is not significantly different from zero ($p > 0.1$)

but also the effects of forcings such as aerosols that are not proportional to the trend. An attempt will be made to distinguish the effects of time- and space-varying forcings from effects of the initialisation by comparing the skill beyond the trend in the initialised ENSEMBLES ensemble by the skill of the uninitialised CMIP3 ensemble using the same trend definition.

In contrast to temperature, for precipitation hindcasts the trend is not larger than natural variability and we only consider the full hindcast skill.

The exact definition of the trend does not affect the results. The fluctuations in the forecasts and observations are so much larger than the non-linearities in the trend that other choices, such as the modelled or observed global mean temperature (used in e.g. van Oldenborgh et al. 2009a) or even a linear trend over this period, give essentially the same results. We prefer the regression on the $CO_2$ concentration on physical grounds and because we expect that it gives better extrapolations into the future than a linear trend definition.

Note that this procedure does not attempt to assess the effect of initialisation on the hindcasts, which would require runs with the same models without initialisation (no-assim) that are not available for the ENSEMBLES experiments. We also do not attempt to separate forced variability from natural variability, which is very hard (Solomon et al. 2011). Finally we avoid the assumption that the trends are correctly modelled by climate models. The trends are strong enough now to identify problems with this assumption (e.g., Stainforth et al. 2005; Knutson et al. 2006; van Oldenborgh et al. 2009a). In a comparison of an initialised run with a no-assim run, a trend bias gives rise to a bias correction that varies as a function of both lead time and analysis time. At lead time zero, it is small as the trend in the analyses is close to the observed trend, but as a function of lead time the gap between the observed trend and modelled trend increases as the model is influenced less by the initial state and more by the forced response. Such a bias structure is very hard to correct for (Robson 2010).

A skilful simple statistical forecast model would be to extrapolate the non-linear trend up to now given a $CO_2$ concentration scenario. This analysis method addresses the question whether climate models can do better than this baseline forecast.

## 3 Global mean temperature

First we consider hindcasts of the global mean temperature anomalies relative to their respective 1961–1990 climatologies. The model hindcasts of global mean T2m are compared to the NCDC global mean SST/T2m estimate in Fig. 3a–c including the trend. There is good skill in the total global mean temperature at all three lead times with correlations coefficients well above 0.9 for the ensemble mean. The hindcast trend is compatible with the observed trend except in the first year, when it is slightly lower. The skill scores are comparable to those of the CMIP3 ensemble with volcanoes included (Fig. 3d–f). The skill is obviously mainly due to the trend.

Figure 4a–c show the skill after subtracting the trend as defined in Sect. 2.4 in both the ENSEMBLES hindcasts and observations, Fig. 4d–f show the same for the CMIP3 ensembles, both the subset that includes solar variability and volcanic aerosols and the subset that does not include these. The skill in the variations beyond the trend is still sizeable in year 1, $r \approx 0.8$ both in the initialised ensemble and the uninitialised one. In the initialised hindcasts this

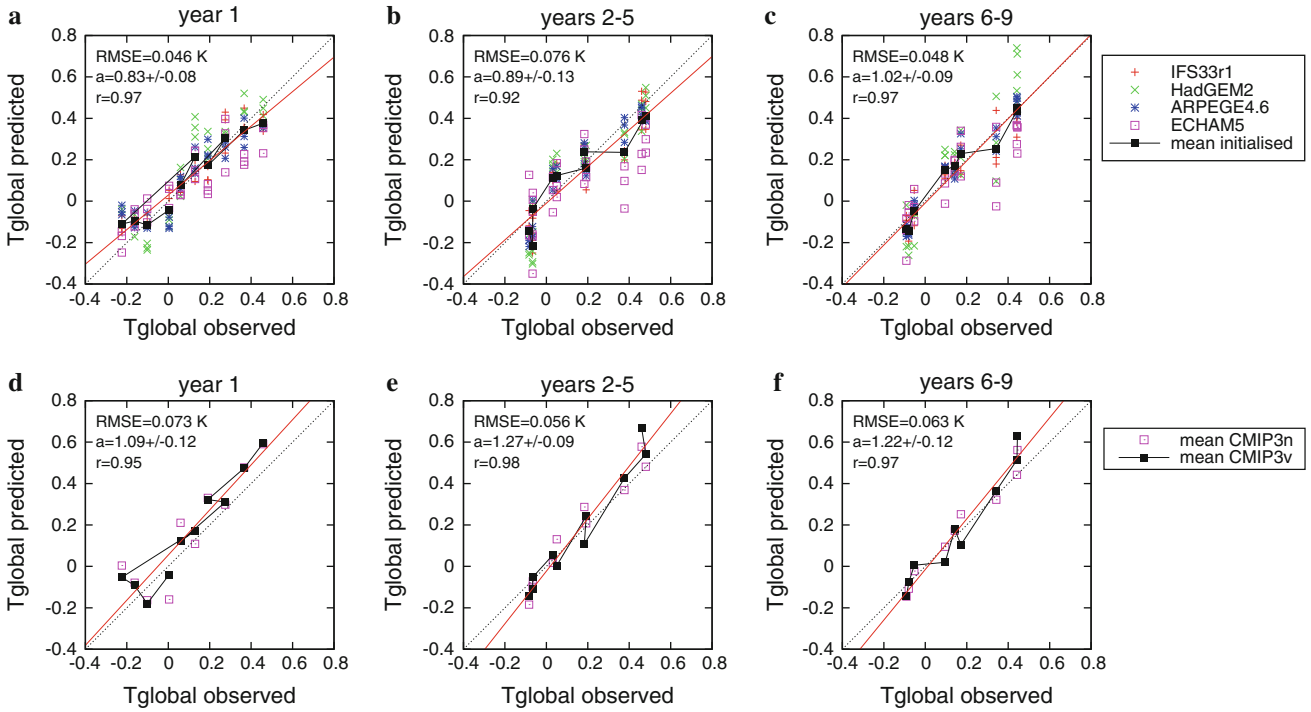


**Fig. 3** Comparison of predicted global mean temperature anomalies (w.r.t. 1961–1990) with observed ones (NCDC) of the ENSEMBLES decadal hindcast experiments (**a–c**) and the CMIP3 ensemble subsets with (CMIP3v) and without (CMIP3n) volcanoes (**d–f**) for year 1 (**a**, **d**), years 2–5 (**b**, **e**) and years 6–9 (**c**, **f**). The *red line* denotes the best fit to the multi-model (**a–c**) and CMIP3v (**d–f**) data, the *dashed line* the ideal 1:1 agreement. The correlation coefficient, RMSE and regression $a$ (with $1\sigma$ error) are given for the multi-model ensemble mean in (**a–c**) and for the CMIP3v mean in (**d–f**). The CMIP3n,v ensembles are sampled at the same years as the ENSEMBLES hindcasts

can be understood as the effects of persistence of the global SST combined with the evolution of ENSO, which can be predicted well for the half year starting in November ($r = 0.94 \pm 0.04$, see also van Oldenborgh et al. 2005a) and has a large influence on the global mean temperature 3–6 months later ($r \approx 0.7$, see e.g. Thompson et al. 2008). The CMIP3v ensemble profits from the inclusion of volcanic aerosols, knowledge of which is not always available in real forecasts. This is confirmed by the absence of skill beyond the trend in the CMIP3n ensemble.

The multi-model initialised ensemble also does not show any skill in years 2–5 beyond the trend. The positive skill in years 6–9 is not significant at $p < 0.1$. The negative and positive skill scores for years 2–5 and 6–9 can be interpreted as random fluctuations around a low correlation. In contrast the CMIP3v ensemble still shows positive correlations due to the influence of solar variability and volcanic aerosols on the global mean temperature.

These results do not depend on the definition of the trend. Subtracting a linear trend again gives a negative correlation skill score for the initialised hindcasts in years 2–5 and the same skill score for years 6–9.

The low skill scores beyond the first year can be understood from the main causes of the variability of the global mean temperature around the trend. The largest fluctuations in Fig. 1 are due to cooling effects after large volcanic eruptions (Robock 2000). In the 1960–2009 time frame these are the eruptions of Gunung Agung (1963), El Chichón (1982) and Pinatubo (1991). These eruptions cannot be predicted with a lead time of years and are therefore not included in the hindcasts.

Another factor that strongly affects the 4-year averaged global mean temperature in all simulations is the temperature variation in Asia and North America north of 30°N. The 4-year smoothed detrended temperature in this area is strongly correlated with the detrended global mean temperature ($r = 0.7$ over 1960–2010). The low-frequency variability of this temperature is dominated by late winter (January–March). This variability cannot be predicted well by these models beyond the trend (cf. Fig. 5c, d). Variability in these regions is mainly driven by the atmospheric variability described by the Arctic Oscillation, Scandinavia Pattern and Pacific-North America Pattern, which are to a large extent driven by the chaotic nature of the mid-latitude westerly flow.

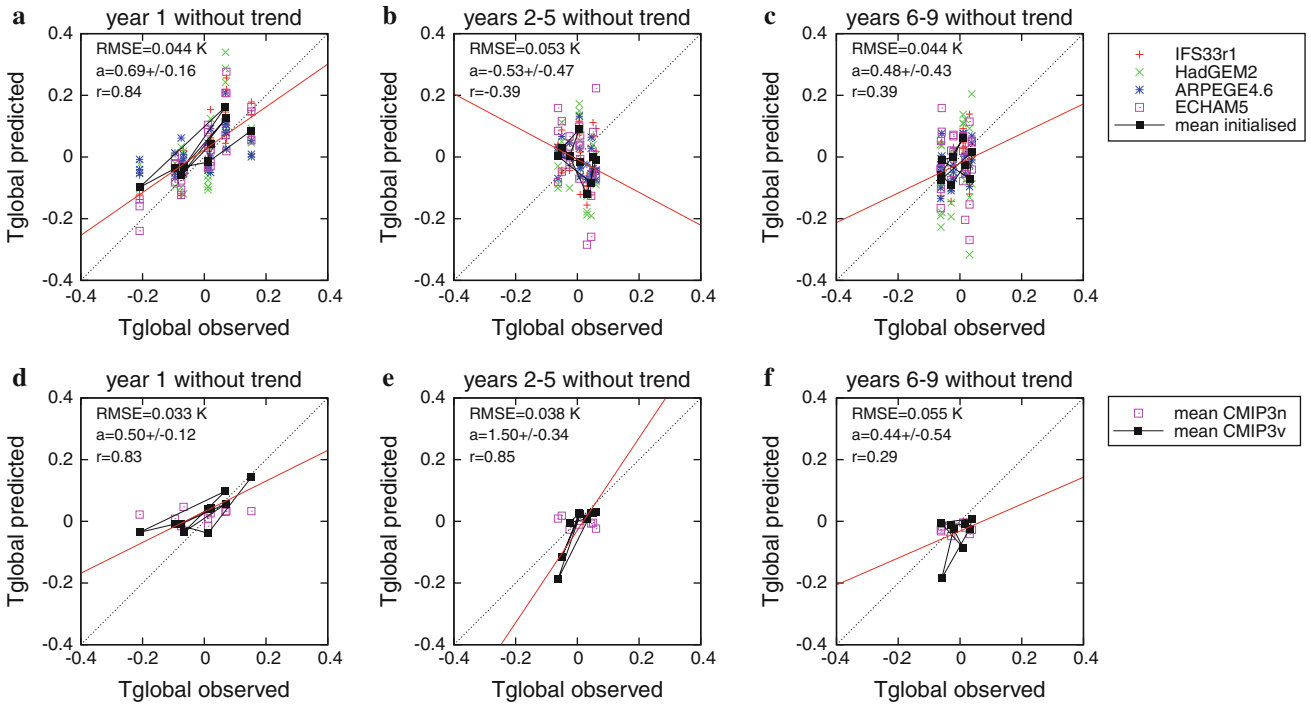


**Fig. 4** As Fig. 3 but with the trend subtracted

## 4 Local temperature forecast skill

Having established that the ENSEMBLES multi-model ensemble shows limited skill in the global mean temperature beyond the first year we next consider the spatial distribution of forecast skill. For sea points we verify SST against ERSST v3b, land point T2m is verified against the GHCN/CAMS dataset and polar regions (south of 60°S, north of 60°N) against the GISTEMP 1,200 km T2m dataset. In Fig. 5 we show the correlation skill in the total temperature forecasts and the skill after subtracting the local trends in the observations and models. (Using the model T2m fields over sea instead of SST does not make a noticeable difference.)

The skill of the T2m/SST forecasts including trends is shown in Fig. 5a, b. The correlation coefficients have values of 0.5–0.8 over most of the globe. These values are statistically significant at $p < 0.1$. Exceptions are SST in the North Pacific and Southern Oceans and T2m in parts of the Andes where other datasets have no data. These are all regions with low trends in the observational datasets used.

The next question is how much of the skill is due to factors beyond the trend. We subtract local trends (i.e., the local regressions against $CO_2$ concentration) from both the hindcasts and the observations, and recompute the skill scores. (Note that the trends are not necessarily the same in the model and the observations, the trends are compared in Sect. 5.) The correlation coefficients are much lower without trends, see Fig. 5c,d.

Statistically, the correlation in these maps is on average positive. We computed the field significance of this signal to be $p \approx 0.1$ using the method of Sterl et al. (2007), which entails estimating the number of degrees of freedom from the autocorrelation of the maps of local $p$-values under the assumption that this autocorrelation is the same over the whole globe. The number of degrees of freedom is then $4\pi/(\pi a^2)$ with $a$ the decorrelation scale in radian. A Monte Carlo test showed that this procedure gives comparable results to the method of (Livezey and Chen 1983) that requires a time-dependent field. Our method here results in an estimate of $\mathcal{O}(200)$ degrees of freedom, which together with the mean and standard deviation give $p \approx 0.1$ with a one-sided $t$ test.

However, the signal to noise ratio for each individual region is so large that one cannot identify regions of skill based on statistics alone. There are almost as many regions with negative correlation coefficients as there are regions with positive ones. Instead, we focus on two well-known areas of low-frequency variability that have also been identified in long climate model runs (Collins 2002; Boer 2004; Pohlmann et al. 2004). There is positive skill over the North Atlantic (significantly in years 2–5) and the eastern subtropical Pacific in years 6–9. Both these signals are stronger than persistence (Fig. 5e, f), which has been computed from the detrended observational datasets over the same time period 1960–2009. Other areas of positive skill can at this stage not be distinguished from random fluctuations.

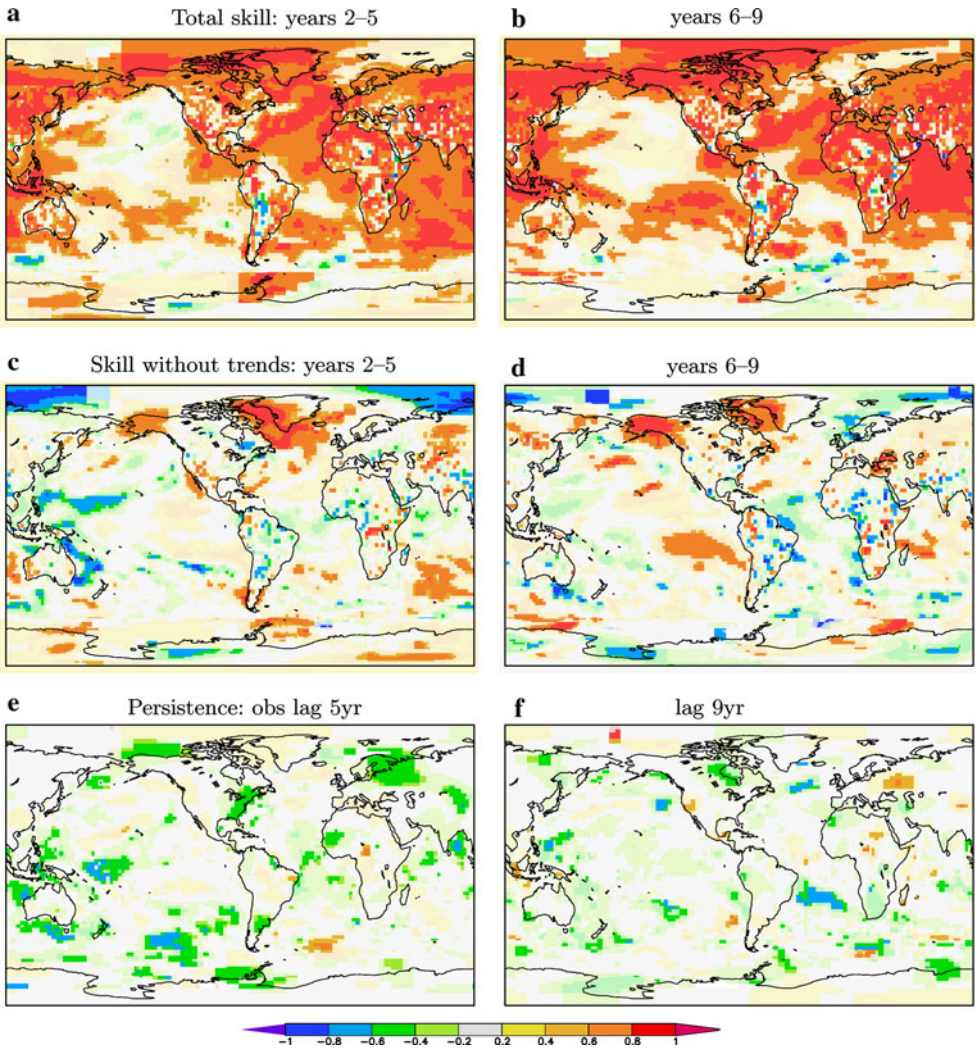


**Fig. 5** Correlation skill of T2m/SST hindcasts for years 2–5 (**a**, **c**) and years 6–9 (**b**, **d**) including the trend (**a**, **b**) and the skill that is left after subtracting the local trends (regressions on the $CO_2$ concentration) of both model and observations (**c**, **d**). For comparison the 5- and 9-year lag correlations of 4-year averaged detrended observations are given (**e**, **f**). Correlations that are not significant at $p < 0.1$ are plotted in *light colours*. SST: ERSST v3b from NCDC, T2m: GHCN/CAMS from NCEP, polar regions: GISTEMP (1200 km decorrelation)

The positive skill scores can either be due to the initialisation of the hindcasts or to forcings that are not proportional to the smooth rise of the $CO_2$ concentration. Figure 6 shows the same separation between trend and other variability for the CMIP3 multi-model ensemble, separated in the subsets with and without volcanic aerosols. Although consisting of different models, it shows areas in which climate models show skill including the trend (a, b) and in the variability around the trend (c, d). In the latter case, the highest skill scores are obtained in the western North Pacific, western North Atlantic and eastern Europe/Middle East, but only when volcanic aerosols were included in the simulation. Intriguingly, the location downstream of major aerosol emitting areas (East Asia, North America and Europe) suggests that the effect of tropospheric aerosol forcing leads to skill in these areas rather than the stratospheric volcanic aerosols. Skill in the northern North Atlantic is lower than in the initialised runs and may be related to the effect of volcanic eruptions on the overturning circulation (Stenchikov et al. 2009). The negative skill in the tropical Pacific Ocean is also unexpected given the reported influence of solar forcing in this area (Meehl and Arblaster 2009). An investigation which aspect of the forcing is responsible for these signals in the CMIP3 ensemble is beyond the scope of this article.

Comparing Figs. 5c, d with 6c, d, the initialised ensemble shows more skill than the uninitialised ones in the northern North Atlantic (years 2–5) and the eastern Pacific (mainly years 6–9). We discuss these region in Sects. 6 and 7 respectively.

## 5 Trends

From Fig. 5 we concluded that in most of the world the skill in temperature hindcasts of the ENSEMBLES multi-model ensemble is due to the trend over 1960–2009. A high correlation coefficient between observed and modelled trends does not indicate accurately how well the trends are represented in these models: as long as there is a trend in

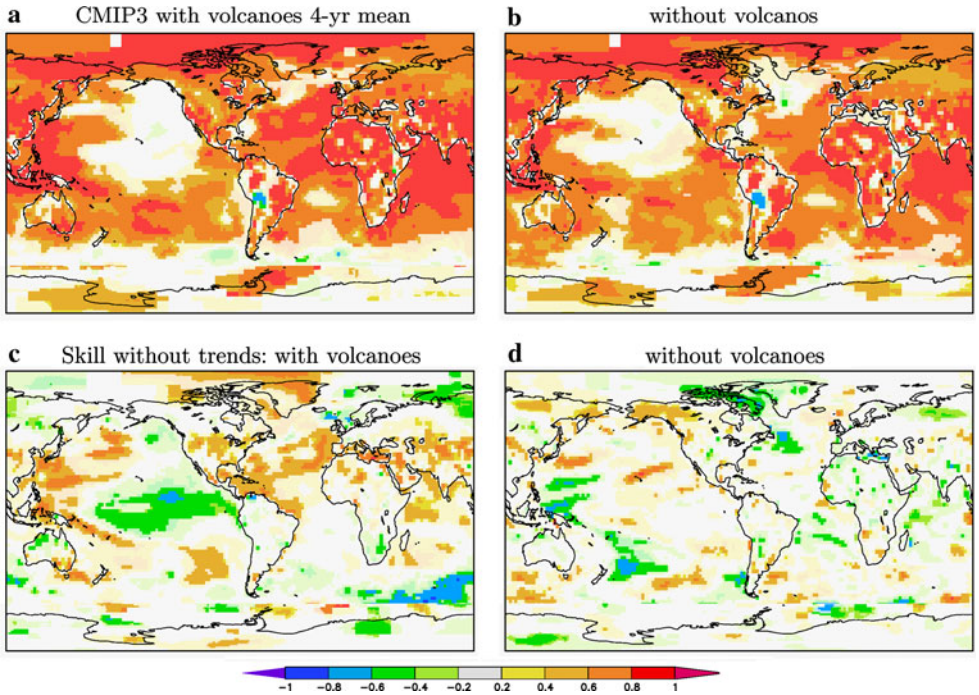


**Fig. 6** Correlation skill over 1960–2010 of 4-year running mean T2m/SST in the CMIP3 multi-model ensemble including volcanic aerosols (**a**) and without volcanic aerosols (**b**). *Panels* (**c, d**) the same after subtraction of local trends. Cf. Fig. 5. Correlations that are not significant at $p < 0.1$ are plotted in *light colours*

both that is larger than the noise the correlation coefficients will be high. A direct comparison of the modelled and observed trends over the hindcast period, defined as a regression of the nine or ten data points on the $CO_2$ concentration, is given in Fig. 7.

We also compare the trends in the ENSEMBLES decadal forecast models with those in the CMIP3 multi-model ensemble mean. Again, these figures do not depend strongly on the definition of the trend. A linear trend gives virtually the same results, as the variability per grid point is much larger than the difference between a linear increase and the accelerating increase implied by using the $CO_2$ concentration.

The trends are fairly similar in years 2–5 and 6–9 (cf. Fig. 7a, b). The differences with observed trends are shown in Fig. 7e, f). For comparison we also show the trend in the full CMIP3 ensemble multi-model mean, and its deviation from the observed trend over the same period. The subsets with and without volcanoes have similar trends.

The agreement with the observed trends is similar in the initialised ENSEMBLES ensemble and the uninitialised CMIP3 ensemble mean outside the polar regions: the spatial standard deviations of the trend differences averaged over the ocean 60°S–60°N are indistinguishable between the three maps Fig. 7e,f, g. The same holds for the spatial standard deviations over the land trend biases. The patterns are also similar, with a common failure to reproduce the absence of a heating trend in the North Pacific Ocean and around Florida. Over land, the lack of temperature rise over central North America is not simulated, whereas temperature trends in Europe (van Oldenborgh et al. 2009a) and China are underestimated by all ensembles. Note that the IFS33r1 model does not include a sea ice model, which can explain part of the poor performance of the initialised ensemble in the Arctic.

We conclude that the initialised ENSEMBLES hindcasts do not simulate the observed temperature trends better or worse than the uninitialised CMIP3 hindcasts, except in the Arctic. The poor representation of this main predictable signal is cause for caution in using climate models for local climate forecasts.

## 6 Atlantic multi-decadal oscillation

Sea surface temperature in the North Atlantic shows variability on time scales of 20 years and more, known as the Atlantic Multi-decadal Oscillation (AMO, Schlesinger and Ramankutty 1994). On these time scales, global warming also affects North Atlantic SST. In model studies the effect of AMO fluctuations on global mean temperature is fairly small; van Oldenborgh et al (2009b) find a maximum correlation of 0.25 in the MPI ECHAM5/OM-1 model and in the CCSM 3 control run the correlations are even lower (Hofer et al. 2011).[1] We therefore use the AMO index proposed by Trenberth and Shea (2006): SST anomalies averaged over EQ–60°N, 80–0°W minus global SST anomalies averaged over 60°S–60°N. By coincidence, this index is also almost orthogonal to the global mean temperature on the short period 1960–2009 and the response of

[1] Priv. Comm., C. C. Raible

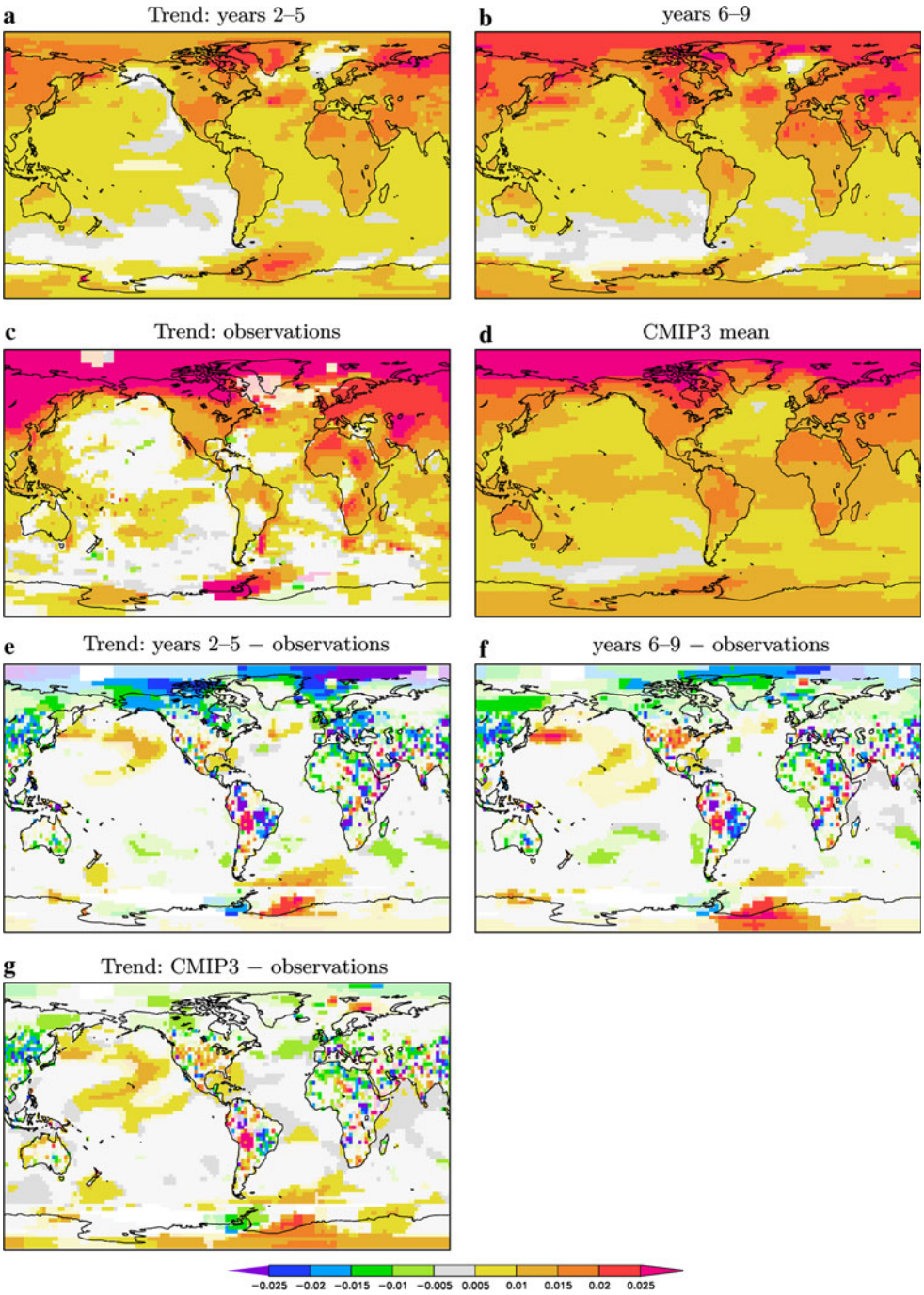


**Fig. 7** The SST/T2m trend [K/ppm] in the ENSEMBLES multi-model ensemble years 2–5 (**a**) and 6–9 (**b**), for the observations (**c**) and for the full CMIP3 multi-model mean over 1960–2010 (**d**, T2m only). The difference between the ENSEMBLES multi-model trend and the observed one is shown in (**e, f**), the same for the CMIP3 ensemble 1960–2010 in (**g**). *Grid boxes* in which the trend (difference) is not significant at $p < 0.1$ are plotted in *light colours*

the full CMIP3 ensemble. This justifies *a posteriori* the method to separate the trend from the variability described in Sect. 2.4 for this region, even in the presence of multi-decadal variability. Note that no bias or trend is subtracted beyond the definition of the AMO index itself.

Figure 8 shows the AMO observations and hindcasts. The interannual variability of the AMO is not captured well in the first year of the decadal forecasts. In contrast, the slower variations are simulated well in years 2–5 ($r = 0.74$, $p \approx 0.03$ taking serial correlations into account) and years 6–9 ($r = 0.57$, $p \approx 0.05$). These numbers are similar to the ones obtained by Pohlmann et al. (2009) ($r \approx 0.7$ for years 1–5, 0.6 for years 5–10). The amplitude of the variations is underestimated by the multi-model mean, however. The uninitialised CMIP3 ensemble captures some of the cooling trend around 1960, but does not capture the warming trend of the last two decades. Consequently, the correlation is much lower than for the initialised ensemble, −0.1 for all years 1960–2010 in the full ensemble, 0.4 for the subset that includes volcanic aerosols

(CMIP3v). The difference is explained by the subset that does not include volcanic aerosols (CMIP3n) simulating a decline of the AMO index throughout the interval.

From theoretical arguments and model analyses it is expected that the skill in forecasting the AMO is to some extend based on predictable fluctuations of the Atlantic Meridional Overturning Circulation (AMOC) (Delworth and Mann 2000; Knight et al. 2005; Dijkstra et al. 2006). However, direct observations of the AMOC are only available since 2004. An intercomparison of the AMOC hindcasts does not show much coherency beyond a few years (not shown). This may explain the relatively short lead time of skilful forecasts compared with the time scales of the AMO.

The predictability of the AMO with a lead time of around five years opens the possibility to regional decadal forecasts beyond the trend using AMO teleconnections, although the combination of imperfect skill in the AMO forecast and the weakness of the teleconnections ($r < 0.5$ in all but a few land areas) may not lead to useful skill. An estimate of AMO teleconnections over 1901–2006 with a 4-year running mean is shown in Fig. 9a. The comparison with Fig. 5c shows that the positive skill in northern Africa and the Middle East may be related to the AMO teleconnection to these regions, although the Middle East also shows a clear aerosol signature (Fig. 6c, d). The AMO teleconnection to central and eastern US temperature does not lead to skill in temperature hindcasts in these regions in the initialised ENSEMBLES multi-model ensemble. Attributing skill to teleconnections requires a much more detailed analysis of the physical mechanisms, and is beyond the scope of this study.

## 7 Decadal ENSO

To investigate the skill in forecasting low-frequency variability in the Pacific Ocean we define a decadal ENSO index as the normalised principal component of the first EOF of detrended SST in the region 50°S − 50°N, 100°E − 70°W for each model separately. For year 1 we take 12-month averaged (Nov–Oct) SST, for years 2–5 and 6–9 we taken 4-year running means of Nov–Oct SST before computing the EOFs. The EOFs are taken to be the eigenvalues of the correlation matrix rather than the covariance matrix, i.e., the SST variability is normalised at each grid point prior to the computation. The resulting patterns are similar to the Interdecadal Pacific Oscillation (Power et al. 1999), but not constrained to be orthogonal to the trend by construction. The regressions of the associated time series on SST are shown in Fig. 10 for the observations and the four decadal forecast models.

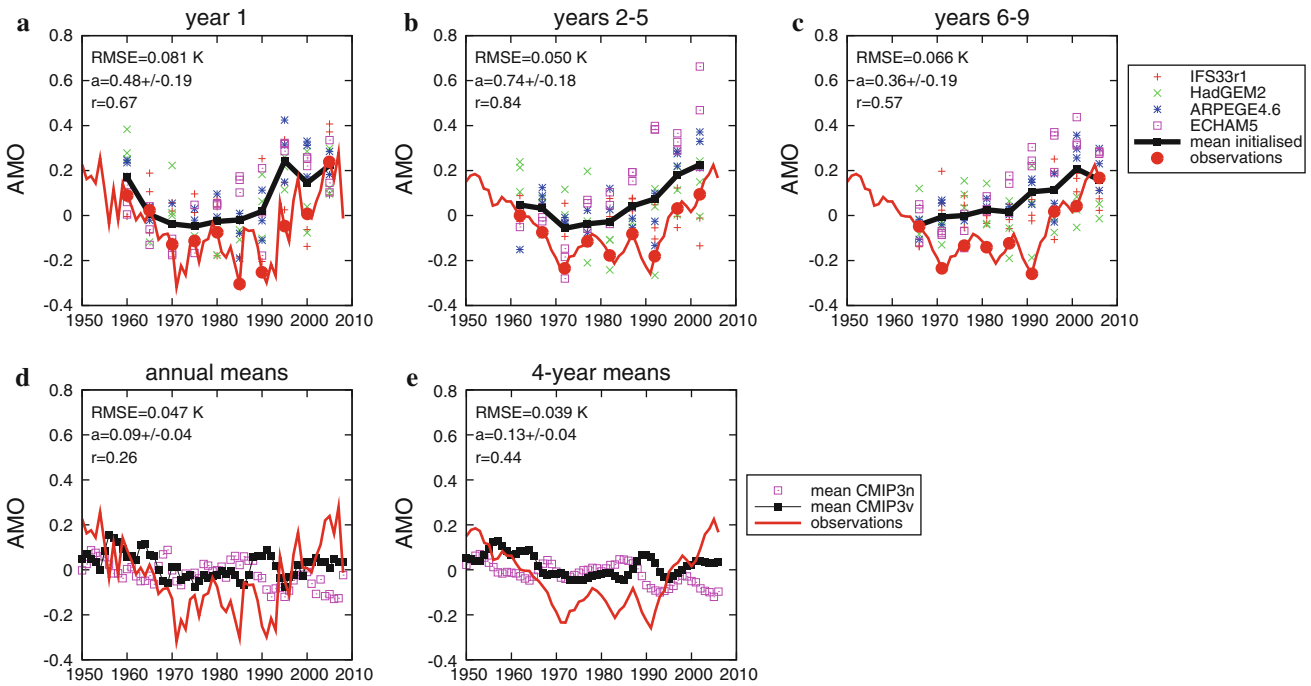


**Fig. 8** Comparison of predicted AMO index with observed ones based on ERSST v3b for year 1 (**a**), years 2–5 (**b**) and years 6–9 (**c**). *Panels* **d** and **e** show the same for the CMIP3 ensemble with and without volcanic aerosols. No bias or trend correction has been applied. The correlation coefficient, RMSE and regression $a$ (with $1\sigma$ error) are given for the ENSEMBLES multi-model ensemble mean in (**a**–**c**) and for the CMIP3 models with volcanic aerosols in (**d**–**e**)

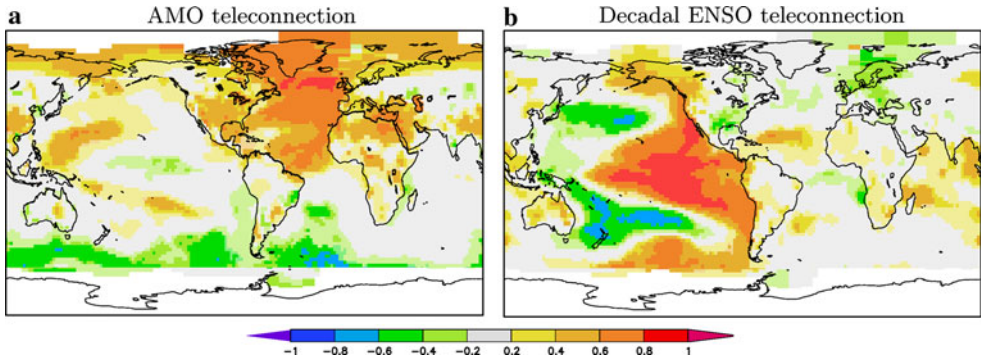


**Fig. 9** Observed **a** AMO and **b** decadal ENSO teleconnections to temperature based on correlations of the AMO and decadal ENSO indices defined in the text with detrended CRU TS 3 temperatures, ERSST v3b SST and GISTEMP 1,200 km (polar regions) using a 4-year running mean over 1901–2006. The detrending was against the observed $CO_2$ distribution. Areas with correlations with $p < 0.1$ are denoted by *light colours*

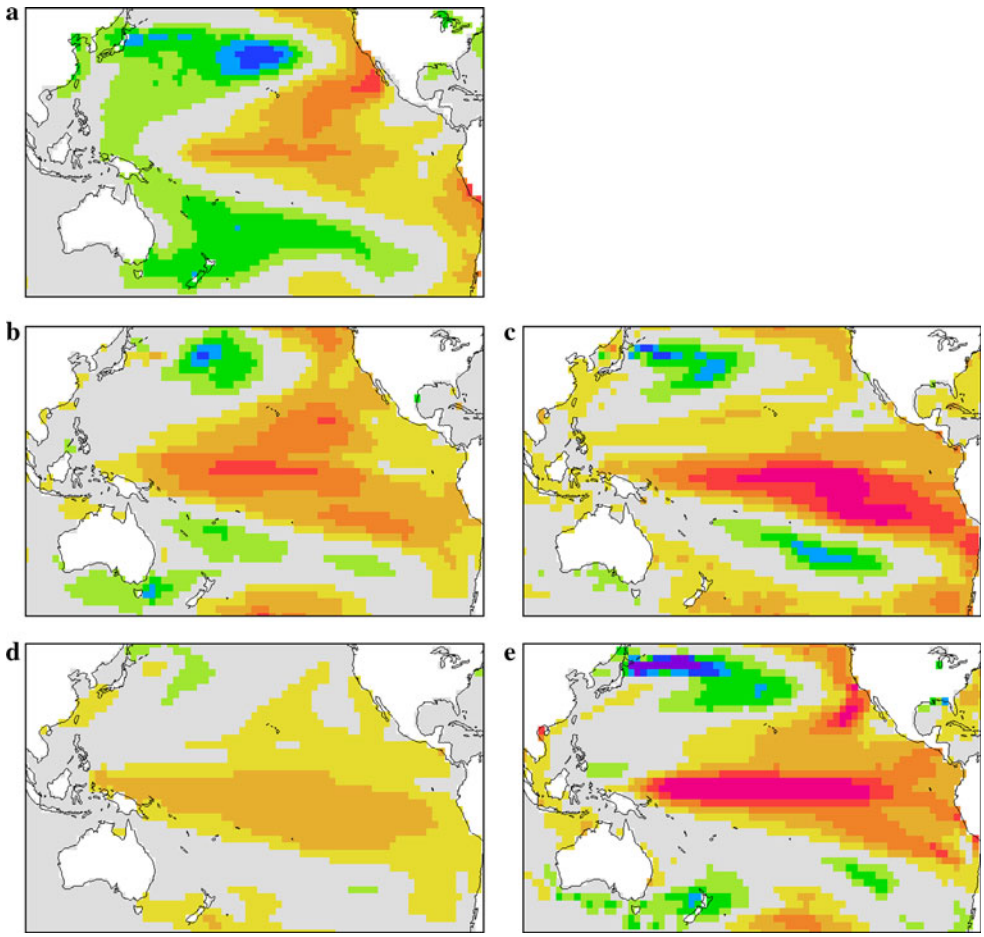


**Fig. 10** First EOF of 4-year mean detrended Pacific SST (50°S–50°N, 100°E–70°W) in **a** the observations (ERSST), the **b** IFS33r1, **b** HadGEM2, **c** ARPEGE4.6 and **d** ECHAM5 decadal forecast models at years 2–5. The accompanying Principal Components (time series) have been normalised to one

For one-year means this decadal ENSO index is highly correlated to the Niño3.4 index ($r \approx 0.9$). For 4-year means the pattern becomes much wider meridionally and the correlation drops to $r \approx 0.6$. The decadal ENSO index of years 2–5 and 6–9 is more similar to the Pacific Decadal Oscillation ($r \approx 0.8$). Like the PDO, our decadal ENSO index is almost orthogonal to global warming as both are characterised by a dipole SST pattern. The orthogonality also holds for the short verification period 1960–2009.

For the ENSEMBLES initialised multi-model ensemble, we computed the EOFs for each model separately in order to capture the differences in the patterns of the different models. The normalised time series were then averaged into a multi-model mean. For the CMIP3 ensembles we

used the observed pattern to define time series of decadal ENSO variability and checked the results with the patterns of the initialised ensemble, Fig. 10.

Figure 11 compares the decadal ENSO indices in the hindcasts and observations. As ENSO can be predicted well for the first half year from November, the good skill ($r = 0.67$) of the initialised models in year 1 is not unexpected. In years 2–5 and 6–9 there is an indication of possible skill, $r \approx 0.4$, in agreement with Fig. 5c, d. Statistically this is not significant at $p < 0.1$, but it is in agreement with other reports of skill in decadal hindcasts (Mochizuki et al. 2010; Yasunaka et al. 2011). As expected from Fig. 6c, d, the uninitialised models do not show skill in simulating this Pacific-wide pattern with correlation coefficients ranging from $-0.2$ to $+0.1$ depending on the pattern used. The subset with volcanic aerosols does not show more skill ($-0.25$ to $+0.25$).

The strongly positive skill scores over Alaska in Fig. 5c, d is probably for only a small part due to the teleconnection from decadal ENSO ($r \approx 0.5$ in the observations, Fig. 9b) combined with the low skill in predicting decadal ENSO itself. The high skill score results from the correct hindcasts of only three cold events that coincided with extended La Niña events, indicating that chance fluctuations played a major part.

Conversely, some of the the negative skill scores in the western Pacific can be understood from the difference of modelled decadal ENSO patterns with he observed ones. In the observations SST in this area is strongly anti-correlated to the eastern Pacific (Fig. 10a). Most climate models extend the equatorial cold tongue too far into the central Pacific (e.g., Guilyardi 2006) and hence represent ENSO activity too far to the west (e.g., van Oldenborgh et al. 2005b). In all ENSEMBLES models this results in a decadal ENSO pattern in which the region positively correlated to the eastern Pacific extends all the way to the tropical West Pacific, three even into the maritime continent (Fig. 10b–e). A point-wise SST verification hence produces negative correlations in these areas. Taking model pattern biases into account (e.g., with a procedure as in Coelho et al. 2006 or Shongwe et al. 2006) could transform them into positive scores.

## 8 Local precipitation forecast skill

The skill of the ENSEMBLES multi-model ensemble precipitation hindcasts is shown in Fig. 12a, b. In precipitation the trends are less important than in temperature as they are smaller than the natural interannual and decadal variability over 1960–2009, with the exception of small areas such as Scandinavia, northern Canada and the south coast of West Africa. We therefore show the total skill without subtracting the trends first (Fig. 13).

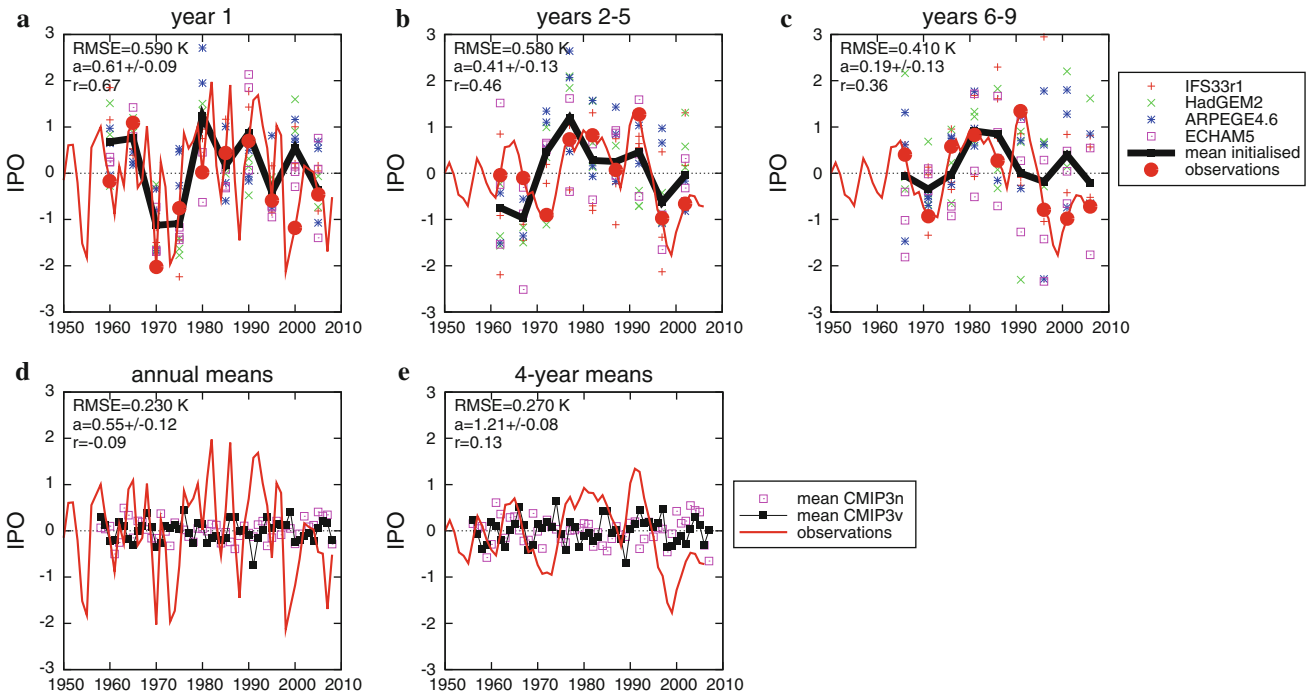


**Fig. 11** Comparison of predicted decadal ENSO index with observed ones based on ERSST v3b for year 1 (**a**), years 2–5 (**b**) and years 6–9 (**c**). *Panels* (**d**) and (**e**) show the same for the CMIP3 ensemble with and without volcanic aerosols. The correlation coefficient, RMSE and regression $a$ (with $1\sigma$ error) are given for the multi-model ensemble mean in (**a**–**c**) and for the CMIP3 models with volcanic aerosols in (**d**–**e**)

The skill scores are compared with teleconnections of 4-year averaged precipitation with the AMO and decadal ENSO over the period 1901–2007 from observations. Given the skill in forecasting the AMO and to some extent decadal ENSO, one expects that the areas with positive (orange/red) or negative (green/blue) teleconnections in Fig. 12c,d will translate to some extent into areas of positive (orange) skill in Fig. 12a, b. As was argued in the case of temperature teleconnections, the added effects of the unexplained variance in the AMO forecast and the weakness of the AMO teleconnection implies that one should not expect correlations higher then roughly 0.5, which are both at the edge of statistical significance in the limited data sample and of limited practical value. For comparison, the same skill scores are also plotted for the uninitialised CMIP3v and CMIP3n ensembles (Fig. 12e, f).

In the Sahel there is positive but pointwise not significant skill in hindcasting rainfall both in years 2–5 and 6–9, both in the initialised ENSEMBLES hindcasts and in the uninitialised CMIP3v ensemble with volcanic aerosols. The ENSEMBLES multi-model mean area-averaged rainfall over 10°−20°N, 18°W−20°E has a correlation skill of $r = 0.38$ ($p \approx 0.2$ taking serial correlations into account) in years 2–5, $r = 0.46 \pm 0.20$ ($p \approx 0.1$) in years 6–9. For the CMIP3v ensemble with volcanic aerosols we obtain $0.54^{+0.13}_{-0.28}$ for 4-year means 1960–2010, which is significantly different from zero at $p < 0.06$. The CMIP3n ensemble without volcanic aerosols however does not show any skill ($r = -0.11^{+0.36}_{-0.11}$). The errors are $1\sigma$ errors determined with a non-parametric bootstrap method taking the serial correlations into account.

Although not or barely statistically significant, combined with the expected physical teleconnection to the AMO (e.g., Zhang and Delworth 2006; Ting et al. 2009) and PDO (Fig. 12c,d) and the effect of aerosol cooling of SST (Rotstayn and Lohmann 2002) these numbers indicate that probably 4-year mean Sahel rainfall is to some extent predictable with a lead time of one year with a relatively low skill. The slightly higher skill in the CMIP3v ensemble including volcanic eruptions cannot be regarded as evidence for predictability as these eruptions are not predictable and the CMIP3n ensemble without them does not

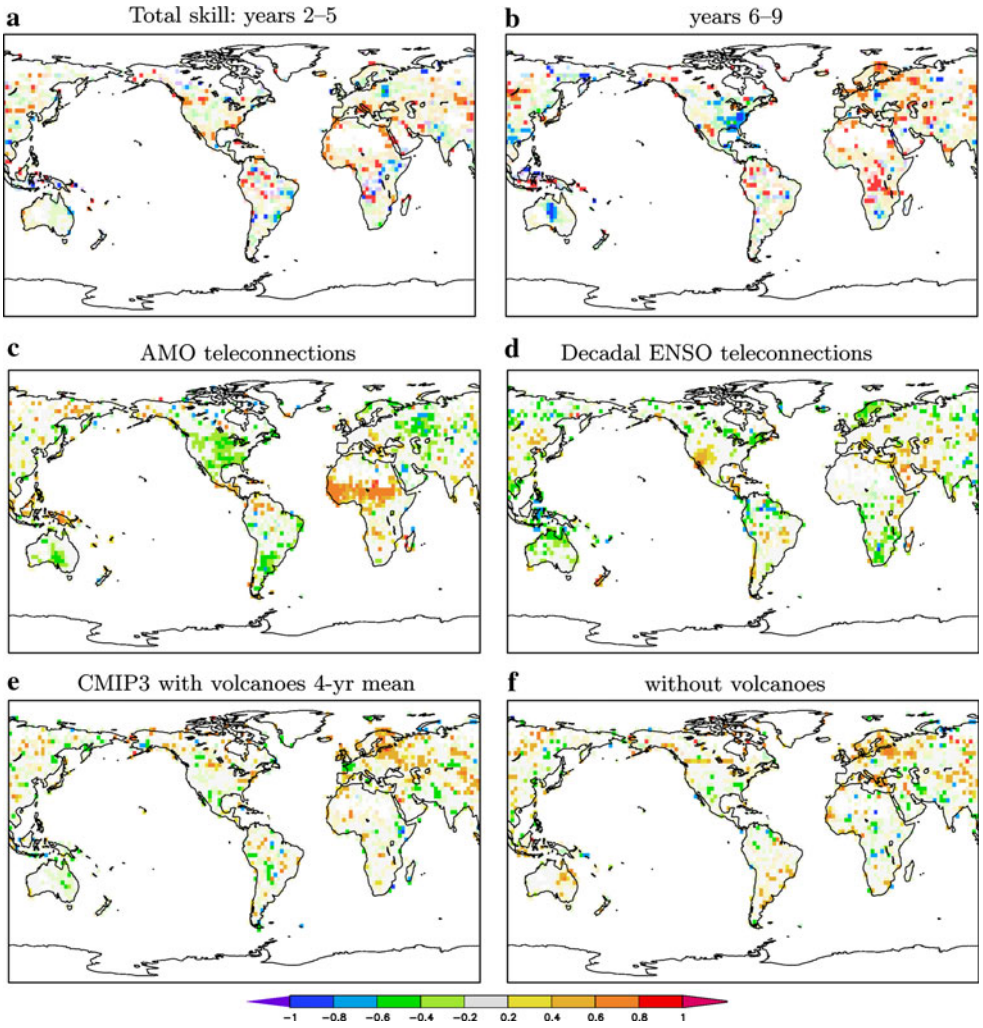


**Fig. 12** Correlation skill of precipitation hindcasts for years 2–5 (**a**) and years 6–9 (**b**) over 1960–2007. This is compared with observed teleconnections: the correlation of 4-year averaged precipitation with the AMO index (**c**) and decadal ENSO index (**d**) over 1901–2007. Also shown are the corresponding skill maps of the full CMIP3 ensemble (**e**) and the subset implementing volcanic aerosols (**f**) over 1960–2010. Areas in which the correlation were not significant at $p < 0.1$ are plotted in *light colours*. Precipitation is taken from the GPCC v5 analysis 1901–2007, demanding at least one observation per 2.5° *grid box*

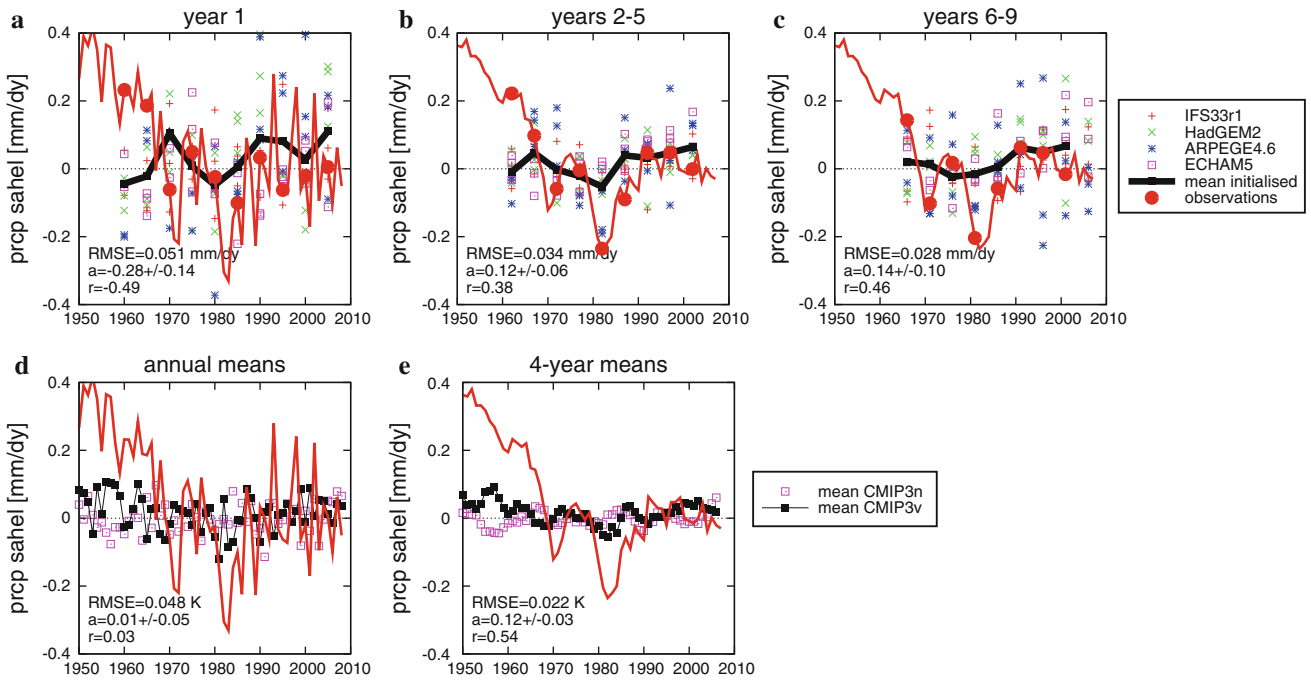


**Fig. 13** Comparison of predicted Sahel rainfall anomalies (mm/dy) against observed anomalies (GPCC V5 and monitoring analysis) for year 1 (**a**), years 2–5 (**b**) and years 6–9 (**c**). *Panels* **d** and **e** show the same for the CMIP3 ensemble with and without volcanic aerosols. The correlation coefficient, RMSE and regression $a$ (with $1\sigma$ error) are given for the ENSEMBLES multi-model ensemble mean in (**a**–**c**) and for the CMIP3 models with volcanic aerosols in (**d**–**e**)

show any skill. However, there are other differences between the three ensembles that make it impossible to draw firm conclusions. Only experiments with the same multi-model ensemble with and without different forcings and initialisation can show where the skill comes from. A good independent check on the skill would be to perform hindcasts in the 1950s, although the limited ocean data for this period is problematic. A simulation of the high anomalies in that decade would strengthen confidence in the skill.

The skill is even lower in the central and western USA, where we expect predictability due to weak teleconnections to the AMO and stronger ones to decadal ENSO (McCabe et al. 2004). In spite of the fact that these teleconnections were also active over the period 1960–2009, the multi-model ensemble does not show skill in these regions, nor in other regions with AMO or decadal ENSO teleconnections.

In Scandinavia there is a strong trend in the observed precipitation ($0.15 \pm 0.03$ %/ppm averaged over the land points north of 60°N and west of 30°E over 1901–2007) that is reproduced to some extent by the ENSEMBLES ($0.05 \pm 0.02$ %/ppm in years 2–5, $0.09 \pm 0.02$ %/ppm in years 6–9) and CMIP3 ($0.07 \pm 0.01$ %/ppm) multi-model means, giving rise to positive skill. The underestimation of the trend in this area is common to most climate models (Zhang et al. 2007; Bhend and von Storch 2008).

To summarise, there seems to be some skill in forecasting 4-year averaged Sahel rainfall with a lead time of one year, but it is unclear whether this is due to the ocean initialisation or the aerosol forcing. In Scandinavia the trend in (winter) precipitation gives skill, although the magnitude of the trend is underestimated. Other regions do not show skill in the precipitation hindcasts of this ensemble.

## 9 Conclusions

A 4-model 12-member ensemble of 10-year hindcasts has been analysed for skill in SST, 2 m temperature and precipitation. The main source of skill in temperature is the trend, which is primarily forced by greenhouse gases and aerosols. This trend contributes almost everywhere to the skill. Variation in the global mean temperature around the trend do not have any skill beyond the first year. However, regionally there appears to be skill beyond the trend in the two areas of well-known low-frequency variability: SST in parts of the North Atlantic and Pacific Oceans is predicted better than persistence. A comparison with the CMIP3 ensemble shows that the skill in the northern North Atlantic and eastern Pacific is most likely due to the initialisation, whereas the skill in the subtropical North Atlantic and western North Pacific are probably due to the forcing.

In the Atlantic, the ensemble shows clear skill in predicting an AMO index that is orthogonal to the trend in years 2–5, and reasonable skill in years 6–9. The skill in decadal ENSO is lower, not statistically significant, but in agreement with other studies. The CMIP3 ensemble shows less skill in both these indices. There is also an indication of skill in hindcasting decadal Sahel rainfall variations, which are known to be teleconnected to North Atlantic and Pacific SST. The uninitialised CMIP3 ensemble that includes volcanic aerosols reproduces these variations as well, but the models without volcanic aerosols do not. It therefore remains an open question whether initialisation improves predictions of Sahel rainfall.

The modelled trends agree well with observations in the global mean, but the agreement is not so good at the local scale.

These experiments are only a first step towards decadal forecasting using non-optimised methods from seasonal forecasting. The skill assessment does not take into account the considerable biases and drift of the models. It is based on only nine or ten data points and hence suffers from large statistical uncertainties. Larger ensembles sizes per model and more frequent and earlier starting dates will be required to characterise the skill of decadal forecasts better. The verification of decadal hindcasts can then be used to improve the climate models, their forcings and initialisation procedures to give more reliable and skilful climate forecasts.

**Acknowledgments** This work was supported by the EU FP7 large-scale collaborative project THOR (GA212643, 2008-2012) and the QWeCI project (ENV-FP7-2009-1-243964). We acknowledge the FP6 ENSEMBLES project (contract GOCE-CT-2003-505539) for the decadal forecasts and PCMDI for archiving and distributing the CMIP3 data. All data used are available from the ECMWF ENSEMBLES data server and/or the KNMI Climate Explorer.

## References

Bechtold P, Kohler M, Jung T, Doblas-Reyes F, Leutbecher M, Rodwell M, Vitart F, Balsamo G (2008) Advances in simulating atmospheric variability with the ECMWF model: from synoptic to decadal time-scales. Quart J R Meteor Soc 134:1337–1352. doi:10.1002/qj.289

Bhend J, von Storch H (2008) Consistency of observed winter precipitation trends in northern Europe with regional climate change projections. Clim Dyn 31:17–28. doi:10.1007/s00382-007-0335-9

Boer GJ (2004) Long time-scale potential predictability in an ensemble of coupled climate models. Clim Dyn 23(1):29–44. doi:10.1007/s00382-004-0419-8

Brohan P, Kennedy J, Haris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12,106. doi:10.1029/2005JD006548

Coelho CAS, Stephenson DB, Balmaseda M, Doblas-Reyes FJ, van Oldenborgh GJ (2006) Toward an integrated seasonal forecasting system for South America. J Clim 19:3704–3721. doi:10.1175/JCLI3801.1

Collins M (2002) Climate predictability on interannual to decadal time scales: the initial value problem. Clim Dyn 19:671–692. doi:10.1007/s00382-002-0254-8

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

Delworth TL, Mann ME (2000) Observed and simulated multidecadal variability in the northern hemisphere. Clim Dyn 16:661–676. doi:10.1007/s003820000075

Dijkstra HA, Te Raa LA, Schmeits M, Gerrits J (2006) On the physics of the Atlantic multidecadal oscillation. Ocean Dyn 56:36–50. doi:10.1007/s10236-005-0043-0

Doblas-Reyes FJ, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—II calibration and combination. Tellus A 57(3):234–252. doi:10.1111/j.1600-0870.2005.00104.x

Doblas-Reyes FJ, Weisheimer A, Palmer TN, Murphy JM, Smith D (2010) Forecast quality assessment of the ENSEMBLES seasonal-to-decadal stream 2 hindcasts. Tech Memo 619, ECMWF. http://www.ecmwf.int/publications/

Dunstone NJ, Smith DM (2010) Impact of atmosphere and sub-surface ocean data on decadal climate prediction. Geophys Res Lett 37:L02,709. doi:10.1029/2009GL041609

Fan Y, van den Dool H (2008) A global monthly land surface air temperature analysis for 1948-present. J Geophys Res 113:D01,103. doi:10.1029/2007JD008470

Guilyardi E (2006) El Niño—mean state—seasonal cycle interactions in a multi-model ensemble. Clim Dyn 26:329–348. doi:10.1007/s00382-005-0084-6

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57(3):219–233. doi:10.1111/j.1600-0870.2005.00103.x

Hansen J, Ruedy R, Sato M, Lo K (2010) Global surface temperature change. Rev Geophys 48:RG4004. doi:10.1029/2010RG000345

Hawkins E, Sutton RT (2009) The potential to narrow uncertainty in regional climate predictions. Bull Am Meteorol Soc 90:1095–1107. doi:10.1175/2009BAMS2607.1

Hofer D, Raible CC, Stocker TF (2011) Variations of the Atlantic meridional overturning circulation in control and transient simulations of the last millennium. Clim Past 7(1):133–150. doi:10.5194/cp-7-133-2011

IPCC (2007) Climate Change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change (IPCC) [Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M and Miller HL (eds)]. Cambridge University Press, Cambridge

Johns TC, Durman CF, Banks HT, Roberts MJ, McLaren AJ, Ridley JK, Senior CA, Williams KD, Jones A, Rickard GJ, Cusack S, Ingram WJ, Crucifix M, Sexton DMH, Joshi MM, Dong BW, Spencer H, Hill RSR, Gregory JM, Keen AB, Pardaens AK, Lowe JA, Bodas-Salcedo A, Stark S, Searl Y (2006) The new Hadley Centre climate model (HadGEM1): evaluation of coupled simulations. J Clim 19:1327–1353. doi:10.1175/JCLI3712.1

Jungclaus JH, Keenlyside N, Botzet M, Haak H, Luo JJ, Latif M, Marotzke J, Mikolajewicz U, Roeckner E (2006) Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM. J Clim 19:3952–3972. doi:10.1175/JCLI3827.1

Keeling CD, Bacastow RB, Bainbridge AE, Ekdahl CA, Guenther P, Waterman LS (1976) Atmospheric carbon dioxide variations at Mauna Loa observatory, Hawaii. Tellus 28:538–551. doi:10.1111/j.2153-3490.1976.tb00701.x

Keenlyside NS, Ba J (2010) Prospects for decadal climate prediction. Wiley Interdiscip Rev Clim Change 1(5):627–635. doi:10.1002/wcc.69

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

Knight JR, Allan RJ, Folland cK, Vellinga M, Mann ME (2005) A signature of persistent natural thermohaline circulation cycles in observed climate. Geophys Res Lett 32:L20,708. doi:10.1029/2005GL024233

Knight JR, Folland CK, Scaife AA (2006) Climate impacts of the Atlantic multidecadal oscillation. Geophys Res Lett 33:L17,706. doi:10.1029/2006GL026242

Knutson TR, Delworth TL, Dixon KW, Held IM, Lu J, Ramaswamy V, Schwarzkopf MD, Stenchikov G, Stouffer RJ (2006) Assessment of twentieth-century regional surface temperature trends using the gfdl cm2 coupled models coupled models. J Clim 19(9):1624–1651. doi:10.1175/JCLI3709.1

van der Linden P, Mitchell JFB (eds) (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, Fitzroy Road, Exeter EX1 3PB, UK

Livezey RE, Chen WY (1983) Statistical field significance and its determination by Monte Carlo techniques. Mon Wea Rev 111:46–59. doi:10.1175/1520-0493(1983)1110046:SFSAID2.0.CO;2

McCabe GJ, Palecki MA, Betancourt JL (2004) Pacific and Atlantic Ocean influences on multidecadal drought frequency in the united states. Proc Nat Acad Sci 101:4136–4141. doi:10.1073/pnas.0306738101

Meehl G, Teng H, Branstator G (2006) Future changes of El Niño in two coupled climate models. Clim Dyn 26:549–566. doi:10.1007/s00382-005-0098-0

Meehl GA, Arblaster JM (2009) A lagged warm event–like response to peaks in solar forcing in the Pacific region. J Clim 22(13):3647–3660. doi:10.1175/2009JCLI2619.1

Meehl GA, Covey C, Delworth TL, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Met Soc 88:1383–1394. doi:10.1175/BAMS-88-9-1383

Meehl GA, Goddard L, Murphy JM, Stouffer RJ, Boer G, Danabasoglu G, Dixon KW, Giorgetta MA, Greene AM, Hawkins E (2009) Decadal prediction. Can it be skillfull?. Bull Am Met Soc 90(2009):1467–1485. doi:10.1175/2009BAMS2778.1

Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high resolution grids. Int J Climatol 25:693–712. doi:10.1002/joc.1181

Mochizuki T, Ishii M, Kimoto M, Chikamoto Y, Watanabe M, Nozawa T, Sakamoto TT, Shiogama H, Awaji T, Sugiura N, Toyoda T, Yasunaka S, Tatebe H, Mori M (2010) Pacific decadal oscillation hindcasts relevant to near-term climate prediction. PNAS 107(5):1833–1837. doi:10.1073/pnas.0906531107

van Oldenborgh GJ, Balmaseda MA, Ferranti L, Stockdale TN, Anderson DLT (2005a) Did the ECMWF seasonal forecast model outperform statistical ENSO forecasts models over the last 15 years. J Clim 18:3240–3249. doi:10.1175/JCLI3420.1

van Oldenborgh GJ, Philip SY, Collins M (2005b) El Niño in a changing climate: a multi-model study. Ocean Sci 1:81–95. doi:10.5194/os-1-81-2005

van Oldenborgh GJ, Coelho CAS, Doblas-Reyes FJ (2008) Exploratory analysis and verification of seasonal forecasts with the KNMI climate explorer. ECMWF Newslett 116:4–5. http://www.ecmwf.int/publications/newsletters/pdf/116.pdf

van Oldenborgh GJ, Drijfhout SS, van Ulden AP, Haarsma R, SterlC A Severijns, Hazeleger W, Dijkstra HA (2009) Western Europe is warming much faster than expected. Clim Past 5:1–12. doi:10.5194/cp-5-1-2009

van Oldenborgh GJ, te Raa LA, Dijkstra HA, Philip SY (2009) Frequency- or amplitude-dependent effects of the atlantic meridional overturning on the tropical pacific ocean. Ocean Sci 5(3):293–301. doi:10.5194/os-5-293-2009

Pohlmann H, Botzet M, Latif M, Roesch A, Wild M, Tschuck P (2004) Estimating the decadal predictability of a coupled AOGCM. J Clim 17(22):4463–4472. doi:10.1175/3209.1

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. doi:10.1175/2009JCLI2535.1

Power S, Casey T, Folland C, Colman A, Mehta V (1999) Inter-decadal modulation of the impact of ENSO on Australia. Clim Dyn 15:319–324. doi:10.1007/s003820050284

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in FORTRAN: the art of scientific computing. Cambridge University Press, New York

Räsänen J, Ruokolainen L (2006) Probabilistic forecasts of near-term climate change based on a resampling ensemble technique. Tellus A 58:461–472. doi:10.1111/j.1600-0870.2006.00189.x

Robock A (2000) Volcanic eruptions and climate. Rev Geophys 38:191–219. doi:10.1029/1998RG000054

Robson JI (2010) Understanding the performance of a decadal prediction system. PhD thesis, University of Reading

Rotstayn LD, Lohmann U (2002) Tropical rainfall trends and the indirect aerosol effect. J Clim 15(15):2103–2116. doi:10.1175/1520-0442(2002)015<2103:TRTATI>2.0.CO;2

Ruckstuhl C, Norris JR (2009) How do aerosol histories affect solar "dimming" and "brightening" over Europe?: IPCC-AR4 models versus observations. J Geophys Res 114:D00D04. doi:10.1029/2008JD011066

Rudolf B, Becker A, Schneider U, Meyer-Christoffer A, Ziese M (2010) The new "GPCC Full Data Reanalysis Version 5" providing high-quality gridded monthly precipitation data for the global land-surface is public available since December 2010. Tech. rep., GPCC. http://gpcc.dwd.de

Ruokolainen L, Räsänen J (2007) Probabilistic forecasts of near-term climate change: sensitivity to adjustment of simulated variability and choice of baseline period. Tellus A 59(3):309–320. doi:10.1111/j.1600-0870.2007.00233.x

Salas Mélia D (2002) A global coupled sea ice-ocean model. Ocean Model 4(2):137–172. doi:10.1016/S1463-5003(01)00015-4

Schlesinger ME, Ramankutty N (1994) An oscillation in the global climate system of period 65–70 years. Nature 367:723–726. doi:10.1038/367723a0

Shongwe ME, Landman WA, Mason SJ (2006) Performance of recalibration systems for GCM forecasts for southern Africa. Int J Climatol 26:1567–1585. doi:10.1002/joc.1319

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. doi:10.1126/science.1139540

Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H, Scaife AA (2010) Skilful multi-year predictions of Atlantic hurricane frequency. Nat Geosci 3(12):846–849. doi:10.1038/ngeo1004

Smith T, Reynolds R, Peterson T, Lawrimore J (2008) Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). J Clim 21:2283–2296. doi:10.1175/2007JCLI2100.1

Solomon A, Goddard L, Kumar A, Carton JA, Deser C, Fukumori I, Greene AM, Hegerl GC, Kirtman B, Kushnir Y, Newman M, Smith D, Vimont D, Delworth TL, Meehl GA, Stockale TN (2011) Distinguishing the roles of natural and anthropogenically

forced decadal climate variability: implications for prediction. Bull Am Met Soc 92:141–156. doi:10.1175/2010BAMS2962.1

Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, Piani C, Sexton D, Smith LA, Spicer RA, Thorpe AJ, Allen MR (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433(7024):403–406. doi:10.1038/nature03301

Stenchikov G, Delworth TL, Ramaswamy V, Stouffer RJ, Wittenberg A, Zeng F (2009) Volcanic signals in oceans. J Geophys Res 114:D16,104. doi:10.1029/2008JD011673

Sterl A, van Oldenborgh GJ, Hazeleger W, Burgers G (2007) On the robustness of ENSO teleconnections. Clim Dyn 29:469–485. doi:10.1007/s00382-007-0251-z

Stott PA (2003) Attribution of regional-scale temperature changes to anthropogenic and natural causes. Geophys Res Lett 30:1728. doi:1029/2003GL017324

Stott PA, Kettleborough JA (2002) Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. Nature 416:723–726. doi:10.1038/416723a

Thompson DWJ, Kennedy JJ, Wallace JM, Jones PD (2008) A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. Nature 453(7195):646–649. doi:10.1038/nature06982

Ting M, Kushnir Y, Seager R, Li C (2009) Forced and internal twentieth-century SST trends in the North Atlantic. J Clim 22(6):1469–1481. doi:10.1175/2008JCLI2561.1

Trenberth KE, Shea DJ (2006) Atlantic hurricanes and natural variability in 2005. Geophys Res Lett 33:L12,704. doi:10.1029/2006GL026894

Weigel AP, Liniger MA, Appenzeller C (2008) Can muti-model combination really enhance the prediction skill of probabilistic ensemble forecasts. Quart J R Meteor Soc 134:241–260. doi:10.1002/qj.210

Weisheimer A, Doblas-Reyes F, Palmer T, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P (2009) ENSEMBLES: a new multi-model ensemble for seasonal-to-annual predictions—skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophys Res Lett 36:L21,711. doi:10.1029/2009GL040896

Wild M (2009) Global dimming and brightening: a review. J Geophys Res 114:D00D16. doi:10.1029/2008JD011470

Yasunaka S, Ishii M, Kimoto M, Mochizuki T, Shiogama H (2011) Influence of XBT temperature bias on decadal climate prediction with a coupled climate model. J Clim 24:5303–5308. http://dx.doi.org/10.1175/2011JCLI4230.1

Zhang R, Delworth TL (2006) Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. Geophys Res Lett 33(17):L17,712. doi:10.1029/2006GL026267

Zhang X, Zwiers FW, Hegerl GC, Lambert FH, Gillett NP, Solomon S, Stott PA, Nozawa T (2007) Detection of human influence on twentieth-century precipitation trends. Nature 448:461–465. doi:10.1038/nature06025