Comparing the UK Met Office Climate Prediction System DePreSys with idealized predictability in the HadCM3 model

Chunlei Liu,[a*] Keith Haines,[a] Alan Iwi[b] and Doug Smith[c†]

[a]*Environmental System Science Centre, University of Reading, UK*
[b]*STFC, Rutherford Appleton Laboratory, Didcot, UK*
[c]*Met Office Hadley Centre, Exeter, UK*

*Correspondence to: C. Liu, University of Reading, Harry Pitt Building, 3 Earley Gate, Reading RG6 6AL, UK.
E-mail: c.l.liu@reading.ac.uk

The initial condition effect on climate prediction skill over a 2-year hindcast time-scale has been assessed from ensemble HadCM3 climate model runs using anomaly initialization over the period 1990–2001, and making comparisons with runs without initialization (equivalent to climatological conditions), and to anomaly persistence. It is shown that the assimilation improves the prediction skill in the first year globally, and in a number of limited areas out into the second year. Skill in hindcasting surface air temperature anomalies is most marked over ocean areas, and is coincident with areas of high sea surface temperature and ocean heat content skill. Skill improvement over land areas is much more limited but is still detectable in some cases. We found little difference in the skill of hindcasts using three different sets of ocean initial conditions, and we obtained the best results by combining these to form a grand ensemble hindcast set.

Results are also compared with the idealized predictability studies of Collins (*Clim. Dynam.* 2002; 19: 671–692), which used the same model. The maximum lead time for which initialization gives enhanced skill over runs without initialization varies in different regions but is very similar to lead times found in the idealized studies, therefore strongly supporting the process representation in the model as well as its use for operational predictions. The limited 12-year period of the study, however, means that the regional details of model skill should probably be further assessed under a wider range of observational conditions. Copyright © 2011 Royal Meteorological Society and British Crown Copyright, the Met Office

Key Words: climate prediction; DePreSys; skill; standard deviation

Received 24 November 2010; Revised 3 June 2011; Accepted 13 July 2011; Published online in Wiley Online Library 12 August 2011

Citation: Liu C, Haines K, Iwi A, Smith D. 2011. Comparing the UK Met Office Climate Prediction System DePreSys with idealized predictability in the HadCM3 model. *Q. J. R. Meteorol. Soc.* **138**: 81–90. DOI:10.1002/qj.904

1. Introduction

Initial condition problems dealing with the internal variability of the climate system are one of two distinct classes of climate prediction problems (Collins, 2002; hereafter C02). Due to the chaotic nature of climate modelling, any infinitesimal error in the initial conditions can grow and reduce the prediction skill in a coupled model. Since the atmosphere only has short time-scale memory – from days to weeks – climate prediction on longer time-scales can only be constrained by the ocean initial conditions, where the ocean has longer time-scale memory – from months to decades, even perhaps up to a century (Collins and Allen, 2002) – or also perhaps by soil moisture or snow cover on large scales over land (e.g. Douville, 2010). Initialization of these slower processes is routinely exploited in seasonal predictions. However, long-range climate prediction is also controlled by external forcing dominated by greenhouse gas and aerosol variations, which are the drivers of global warming (IPCC AR4). When assessing climate prediction on the annual to interannual time-scale (e.g. Boer, 2004; Pohlmann *et al.*, 2004), it is not easy to separate predictability of internal variability from that arising from external forcing, for the same reason that one cannot attribute any particular flood or heatwave to 'global warming', because there is only one observational reality.

C02 used the Hadley Centre coupled model HadCM3 (Gordon *et al.*, 2000) in a series of initialization experiments in order to assess the predictability of model climate on time-scales from seasonal to decadal. This idealized work does allow the predictability associated with external forcing to be separated from internal mode predictability (because the experiments are perfect twins in which the model predicts its own behaviour) and these results should therefore represent the absolute upper limit on the ability of the HadCM3 system to consistently predict the real world. C02 found some evidence for long time-scale predictability over the oceans, particularly the North Atlantic (see Table I for areas discussed in this paper), but little sign of any predictability over land areas for longer than around 1 year. Hermanson and Sutton (2010) used a different, but again idealized, approach to demonstrate that in individual cases HadCM3 predictability could last for several years and also extend over some land areas. They looked at the convergence rate of ensembles initiated with large differences in some aspect of the circulation, including regional temperature anomalies.

Some other idealized studies have also shown that aspects of internal variability may be predictable several years in advance, but few examples have actually been assessed against real observations (Palmer *et al.*, 2004; Simmons and Hollingsworth, 2002), and more work is needed in this area. It is important to know whether real observational climate anomalies show similar predictability to that seen in ideal twin experiments, and this requires assimilation of observational data into a coupled climate model.

In the EU Ensembles project (Doblas-Reyes *et al.*, 2011), a number of groups looked for seasonal to decadal skill in hindcasts based on assimilation of ocean and atmospheric data into a range of coupled seasonal forecasting models, which were enhanced by the addition of variable greenhouse gas and aerosol forcing. Three-member hindcast ensembles were initialized every 5 years from 1960 to 2000 and run out to 10 years ahead. Results from these experiments can be found in Doblas-Reyes *et al.* (2011) and Keenleyside *et al.* (2008). Some of these runs assimilated the full observed atmospheric and ocean states, thus requiring the correction of the hindcasts for climate drift (Stockdale, 1997). However, identical ensembles but without assimilation were usually not carried out, making it impossible to separate the skill present from external forcing (which should also appear in the hindcasts without assimilation) from the initial condition skill, in the results. However the UK Met Office Decadal Prediction System (hereafter DePreSys) avoids both of these drawbacks.

DePreSys is based on the HadCM3 coupled model. Smith *et al.* (2007) developed an anomaly assimilation method (Pierce *et al.*, 2004) for DePreSys, in which observed ocean and atmospheric anomalies from a mean seasonal cycle over some time period are superposed on the HadCM3 climatological seasonal cycle and assimilated over the same time period. This certainly reduces the need to correct the hindcasts for model drift. Smith *et al.* produced a set of assimilated hindcasts (referred to here as 'ASSIM') and also performed a full set of hindcasts without assimilated initial conditions (referred to here as 'NOASSIM') in order to separate skill due to initial conditions from that associated with external forcing. Ensembles of 10-year ASSIM and NOASSIM hindcasts were started every 3 months from 1982 to 2001. It was found that ASSIM had improved skill over NOASSIM in predicting changes in global mean surface air temperature, which was traced to initialization and persistence of upper ocean heat content anomalies, demonstrating that the correct initial ocean states are vital. Some areas over land also appeared to show increased skill for the assimilated hindcasts (Figure 3 in Smith *et al.*), which did seem at odds with the idealized experiments of C02. However, Robson (2010) suggested that some aspects of these differences may have been an artefact of a drifting model climatology being used.

In this paper we have used the DePreSys system to perform a series of hindcast experiments that are arranged as closely as possible to the idealized predictability experiments of C02. The datasets, including the transient all-forcings simulations from 1860 to present, are all performed independently of the original Hadley Centre simulations (Stott *et al.*, 2000; Johns *et al.*, 2003). The aim is to try to separate the initial-condition related skill from external-forcing derived skill, and to ask whether the skill for real assimilated observational anomalies can match that found for the ideal twin experiments. Importantly, we focus on regional skill beyond the seasonal time-scale, for which there is only limited evidence in previous studies. In addition, we have examined the sensitivity of the forecasting system to assimilating different observational ocean anomalies derived from ocean reanalyses, which have much greater uncertainties than atmospheric reanalyses. The ocean initial conditions assimilated by Smith *et al.* (2007) were generated by a 4-D, multivariate optimal interpolation of salinity and temperature data (Smith and Murphy, 2007), using covariances from the HadCM3 model itself, and are therefore tuned to variability within that model. However, these complex analyses are simply nudged into the HadCM3 model to initialize the predictions. Such initial conditions are not easy to reproduce and therefore it is of considerable interest to test whether successful predictions can be made from other ocean states.

Haines *et al.* (2009) describes the porting of the DePreSys climate prediction system to run on a set of computer clusters, which make climate predictions accessible to a wider group of scientists collaborating with the Hadley Centre. We used the ported DePreSys system to perform experiments similar to Smith *et al.* (2007), but with shorter prediction lead time-scales of 2 years. We give a brief description of the DePreSys system and the experiments performed in section 2. Section 3 presents the main results comparing hindcast skill over different areas with that obtained under idealized conditions by C02. In section 4 the results are summarized and discussed.

Table I. Areas used for regional study of initial-condition skill, including those regions defined in Collins (2002, Figure 3), and additional areas identified here as having enhanced skill.

Short description	Region	Land/ocean
Global ocean and land	Global	Both
Global land	Global	Land
Global ocean	Global	Ocean
Niño3	150–90°W, 5°S–5°N	Ocean
Niño4	160°E–150°W, 5°S–5°N	Ocean
Tropical Atlantic	70–20°W, 0–20°N	Ocean
North Atlantic	50–10°W, 40–60°N	Ocean
North Pacific	160–120°W, 30–50°N	Ocean
Subtropical Pacific	120°E–110°W, 10–30°N	Ocean
Atlantic subpolar gyre	60–10°W, 50–66°N	Ocean
Labrador Sea	70–50°W, 55–70°N	Ocean
Irminger Sea	45–35°W, 60–66°N	Ocean
Nordic Sea	20°W–30°E, 65–80°N	Ocean
Indian Ocean	50–110°E, 40°S–30°N	Ocean
South Pacific	180–110°W, 75–50°S	Ocean
Northwest Europe	10°W–50°E, 30–70°N	Land
Eurasian land	0°W–180°E, 30–70°N	Land
Asia	60–120°E, 10–40°N	Land
North America and Canada	160–50°W, 20–70°N	Land
Australia	100–160°E, 40–10°S	Land
Tropical South America	90–30°W, 15°S–15°N	Land
Southern South America	90–30°W, 60–15°S	Land
Africa	20°W–60°E, 40°S–30°N	Land
Northern North America	140–50°W, 50–75°N	Land
Eastern China	90–140°E, 40–60°N	Land
South Africa	10–30°E, 30–15°S	Land
Northern Australia	120–160°E, 22.5°S–10°S	Land
Eastern Europe	40°E–70°E, 40–60°N	Land
South Asia	70–100°E, 20–40°N	Land

2. The DePreSys system and experiments

Details of DePreSys can be found in Smith *et al.* (2007), so only a brief introduction is given here. DePreSys is a newly developed decadal prediction system, based on the Hadley Centre global coupled climate model HadCM3 (Gordon *et al.*, 2000), which has reasonable representations of El Niño–Southern Oscillation (ENSO) variability and longer-term variability in important climate modes such as the North Atlantic Oscillation (NAO) and the Atlantic Meridional Overturning Circulation (MOC). The DePreSys system is based on assimilating atmospheric and oceanic anomalies only. The observational anomalies of 2-D atmospheric surface pressure, 3-D atmospheric temperature and horizontal wind components, as well as 3-D ocean temperature and salinity, are obtained by removing seasonally varying atmospheric and oceanic climatologies from observations, and then adding back the model climatologies for those quantities over for the same period. These 'analysed' total fields are then nudged strongly with a 6 h time-scale for ocean and 3 h time-scale for atmosphere into the model in order to perform DePreSys assimilation/initialization (Smith *et al.*, 2007). Different time-scales have been tested but the results are not strongly sensitive out to a few days' relaxation in the ocean.

The two sets of hindcast ensembles – NOASSIM and ASSIM – both use varying anthropogenic sources of greenhouse gases and aerosol concentrations (Johns *et al.*, 2003). These greenhouse gases, aerosols and solar variability are projected forward in time without future knowledge in performing the hindcasts, e.g. no knowledge of Pinatubo is assumed for hindcasts starting before June 1991. Solar irradiance is projected by repeating the previous 11-year solar cycle, while volcanic aerosol is projected as an exponential decay to background levels with an *e*-folding time-scale of 1 year.

We used three different ocean datasets for assimilation and validation. First was the ocean data assimilated by Smith *et al.* (2007) using HadCM3 covariances (Smith and Murphy, 2007), hereafter denoted 'DePreSys ocean'. For a second set of ocean conditions we used gridded *in situ* ocean temperature and salinity analyses from the UK Met Office developed within the EU ENSEMBLES (EN3) project. These data are quality controlled using a comprehensive set of objective checks developed at the Met Office Hadley Centre (http://hadobs.metoffice.gov.uk/en3/) (Ingleby and Huddleston, 2007) and are developed into a gridded dataset by successive objective analysis from 1960 to present, starting from WOA01 climatology background fields. This analysis is therefore model independent and is denoted 'EN3 ocean'. We also used an ocean model reanalysis based on the global ORCA1 version of the NEMO model (Madec *et al.*, 1998), forced by ERA40 (Smith and Haines, 2009), which assimilates the same hydrographic data as goes into EN3 ocean. This dataset is termed 'ORCA1 ocean'. Atmospheric observational data are always taken from the ERA40 reanalysis (Uppala *et al.*, 2005), similarly adjusted to produce anomalies for the required time periods.

We performed hindcasts over the period 1990–2001 (12 years) and this same period was chosen to define the observational and model climatologies used for the anomaly assimilation. This ensures that the mean anomalies from all the hindcasts remains small, which is not the case if a much longer period climatology is chosen because of the global impact of externally driven changes in greenhouse gases. Four member ensemble hindcasts were carried out with lead times out to 2 years, starting twice per year (1 May and 1 November) from 1990 to 1999, giving a total of 20 start dates. Four members are usually sufficient to give confidence in initial condition hindcasts (Collins and Allen, 2002). We had available in total 11 members without data assimilation (called NOASSIM), consisting of six separate all-forcings transient runs of HadCM3 from 1860 to 2008 (as in Stott *et al.*, 2000) and five more based on small sea surface temperature (SST) perturbations made from one of these runs in 1945. The initial states of these transient runs in 1860 are from a 1500-year pre-industrial control HadCM3 run (900-year run in UK Met Office and 600-year run on the Reading PC cluster). The plan of the runs is given schematically in Figure 1. After careful checking, both salinity and temperature drift in the deep ocean are very small in this control run and can be neglected. Note there are significant differences in this respect from the transient runs used by Smith *et al.* (2007), as noted by Robson (2010).

The seasonally varying climatology for the anomaly assimilation is defined from the NOASSIM transient all-forcings runs over the same period as the hindcasts, i.e. 1990–2001, in order to avoid the possibility of large anomalies associated with the external forcing changes on decadal timescales. The assimilation of atmosphere and ocean data is carried out using one of these transient all-forcings HadCM3 runs. Four ensemble member initial states for the ASSIM experiment are obtained by adding SST noise (pointwise amplitude 0.05°C) to the initial states (see Figure 1). To simplify notation, ensemble hindcast runs using initial states generated with assimilation of DePreSys, EN3 and ORCA1 ocean data will be called ASSIM-DePreSys, ASSIM-EN3 and ASSIM-ORCA1, respectively.

3. Hindcast skill evaluation

If we wish to identify the effectiveness of initializing the hindcasts, it is critical to compare the skill of the assimilated with the non-assimilated hindcasts. Also, since the skill evaluations must be performed against observed anomalies, it is impossible to separate the impact of external forcing on these observations, and thus the clean separation achieved by C02 is not possible. We have chosen to use the standard deviation (STD) of the anomaly errors (anomaly differences with observations, Eq. (1)) as a measure of skill, and to look for smaller errors in the assimilated runs than the non-assimilated runs:

$$\mathrm{STD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(E_i - \overline{E})} \qquad (1)$$

where N is the total number of samples in the time series, E_i is the anomaly error and $\overline{E}$ is the mean of the anomaly error. The standard deviation has the advantage of being insensitive to the choice of climatology in the anomaly assimilation scheme (Haines *et al.*, 2009) – an issue which does not arise in the ideal studies of C02. The exactly equivalent measure in C02 is the root mean square error (RMSE), but where Collins looks for the RMSE to be smaller than the standard deviation of anomalies from a long control run we compare with the anomaly errors from the NOASSIM run in order to identify significantly enhanced skill.

Figure 2 shows the time evolution of the ASSIM-DePreSys hindcast mean, for the surface air temperature (SAT)

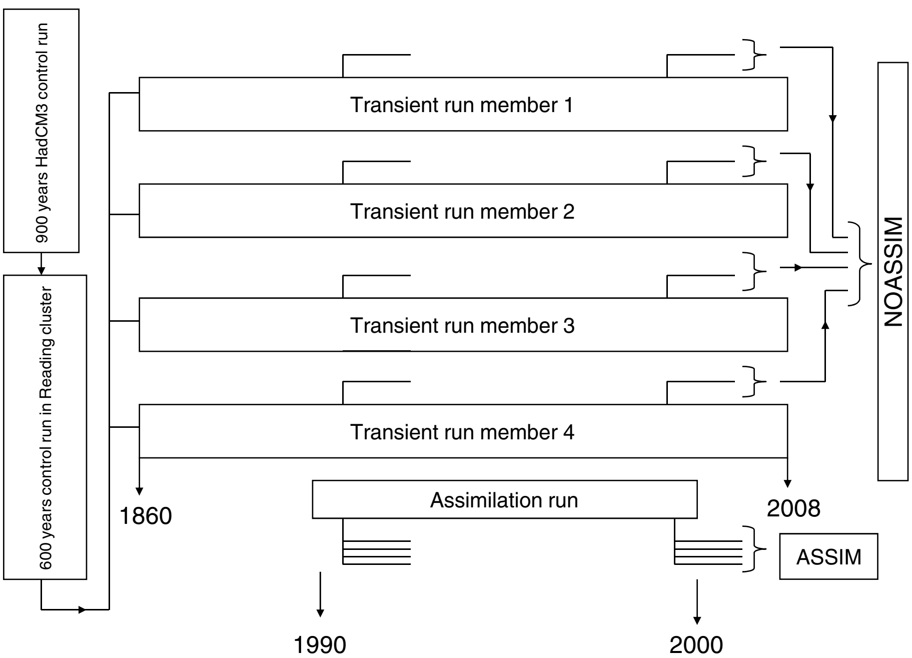

Figure 1. Schematic showing control run, transient run, assimilation run, NOASSIM and ASSIM ensemble runs.

superimposed on the observed SAT anomalies from ERA40 (black line). The largest discrepancies between observations and hindcasts starting from May and November 1990 are due to not including aerosol changes from the Pinatubo volcanic eruption. The hindcast starting from November 1991 successfully follows the initial SAT decrease and later increase during the volcanic-aerosol affected period. Other particular cases of successful anomaly hindcasts include: the continuous increase of global SAT from 1993 to 1994, the decrease from 1995 to 1996 (in the hindcast initialized May 1995), and then the large increase due to the 1997–1998 ENSO, which is well captured by the model. This ENSO is forecast successfully from as early as November 1996 (thick red line). The recovery of SAT from the decrease after the ENSO event is also well captured by the forecast starting from May 1999.

Figure 3 shows the standard deviation errors for SAT anomalies as a function of hindcast lead time for the whole globe, for the global land, the global ocean and other selected ocean regions where long lead time skill is clearly present. The NOASSIM errors in each case are also shown and these effectively define the climatology forecast for the system; the anomaly persistence errors are also shown (dashed). In this case we show the NOASSIM errors based on all 11 independent runs, allowing better comparisons with the ASSIM Grand ensemble results which use all 12 hindcasts based on different ocean conditions. If a smaller number of NOASSIM runs is used the ASSIM errors can be lower because of greater smoothing from averaging a larger ensemble, rather than from the initialization. It is clear that the errors in the ASSIM hindcasts are converging to the NOASSIM climatological values at longer lead times. The persistence errors rapidly become higher than the climatology because these do not have smoothing from the ensemble mean. The anomaly standard deviations are given as seasonal, i.e. 3 months; means and the lead times for which the ASSIM errors are significantly lower than the NOASSIM errors according to a one-sided *F*-test are marked with dots, as in C02. A 5% confidence limit is used; the degrees of freedom are $20 \times 4 - 1 = 79$ for each individual ensemble and for the ocean variables, and $3 \times 20 \times 4 - 1 = 239$ for the Grand ensemble applied to the SAT values, where 20 is the number of start dates and 4 is the number of members in each ensemble. We use the ensemble mean of each hindcast to assess skill because this is the best estimate of the true anomalies. We show three individual ASSIM hindcasts on each plot using the three different ocean anomalies, as well as a Grand mean where all 12 hindcast ensemble members are combined, irrespective of the assimilated ocean anomalies. In addition, each plot shows the persistence errors for the DePreSys ocean anomalies (dashed). Our conclusion from the individual ocean hindcast results for both SAT and ocean quantities – SST and top 113 m ocean heat content (OHC) – is that using ocean anomalies from different datasets makes little difference to the period of enhanced skill (possibly due to the reasonably good ocean sampling in the 1990s study period). We also investigated the difference between the May and November start dates (not shown) and did not find any robust difference in forecast skill. The Grand mean clearly gives the smoothest results for the error growth and therefore in all further results the Grand ensemble skill results only are shown.

The global mean SAT hindcast skill is dominated by ocean areas with skill over and above the NOASSIM runs typically out to about four seasons, i.e. 1 year ahead. The skill period in the Niño3 area, ∼15 months, where we expect most enhanced skill, also reflects the period of enhanced skill over the ocean in the global mean. Comparing with C02 these results are similar in many respects. Differences in the amplitude of the standard deviation errors (they are smaller here than in Collins) is entirely attributable to our using ensemble means as the predictor. This is not a foregone conclusion because the amplitude of observational anomalies can be different from the ideal anomalies studied by C02. The key time periods for enhanced skill are very similar to C02, which is encouraging considering that we are now forecasting real observational anomalies rather than model anomalies as in C02. We do not, however, see the

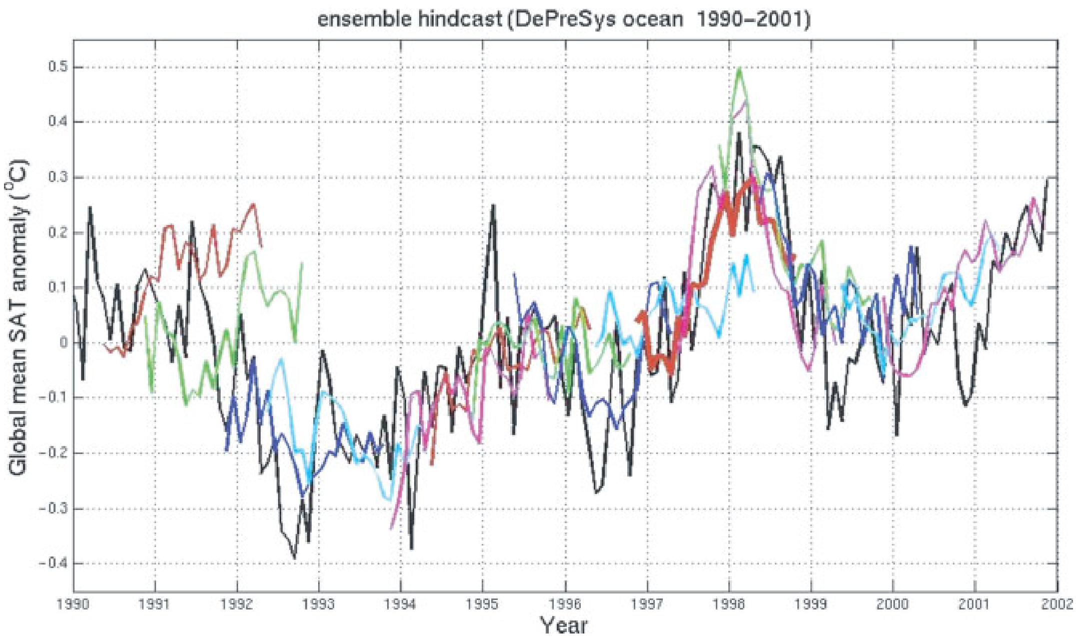

Figure 2. The black line is the observed SAT anomaly relative to climatology over the period 1990–2001. Coloured lines are ASSIM–DePreSys ensemble hindcast anomalies from the four-member ensemble mean.

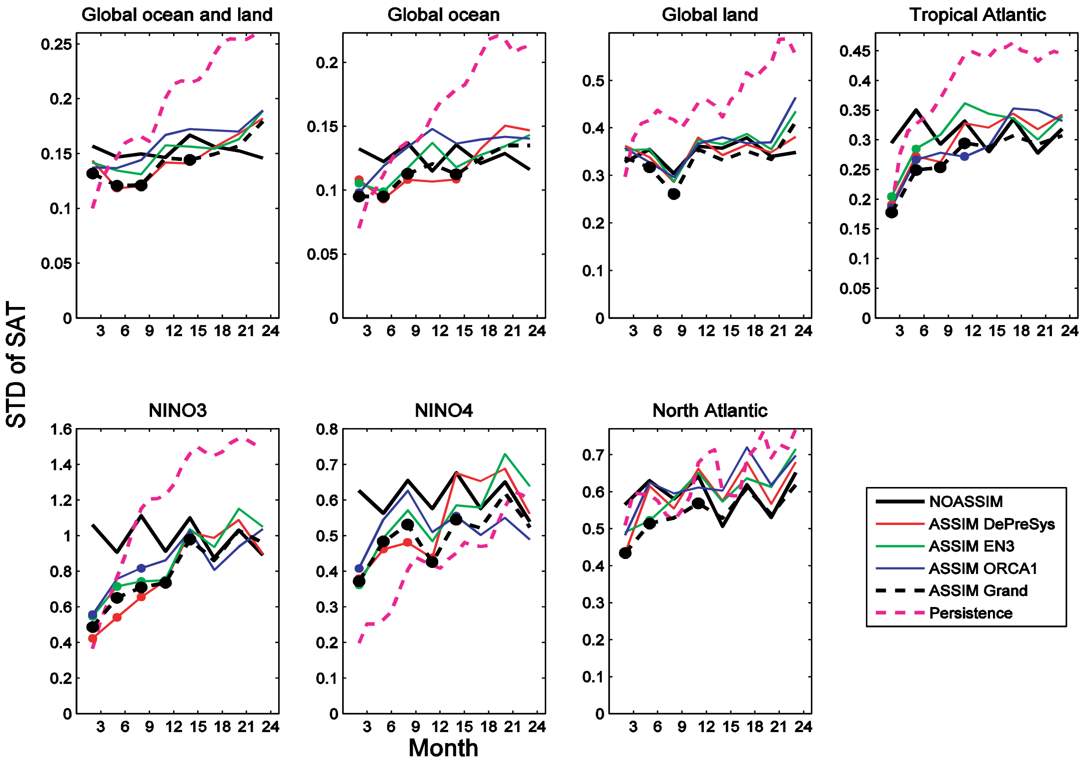

Figure 3. Surface air temperature (SAT) anomaly error standard deviations over the whole globe, the global ocean, the global land, the Niño3,4 areas, the North and tropical Atlantic, as labelled, for the 20 2-year hindcasts during the period 1990–2001. The solid black line is the 11-member ensemble mean error of the NOASSIM runs. The coloured solid lines are the four-member ensemble mean error using the different ocean initial conditions, as labelled, and the dashed black line is the Grand ensemble mean using all ocean initial conditions. Persistence errors are also shown in the mauve dashed lines. Seasonal mean (3-month) errors are assessed, with significant differences between the NOASSIM and ASSIM results marked with a solid dot using a one sided *F*-test.

slightly enhanced skill for the global ocean lasting beyond the 2-year hindcast period, which in C02 is attributable to decadal variability in the North Atlantic. We will return to this later when we look at individual regions.

Figures 4(a,b) shows maps of the difference in SAT standard deviations, NOASSIM–ASSIM for the first and second years of the hindcasts, (similar to Figure 3 of Smith *et al.*, 2007), being a mapped expression of the error gap between the NOASSIM and ASSIM errors in the graphs in Figure 3. A second atmospheric variable, the sea-level pressure (SLP), standard deviation differences are also shown in Figure 4(g,h). The Grand ensemble is used for the ASSIM errors and 11 NOASSIM members are used with the one sided *F*-test to establish the significant differences for SAT and SLP. Note that the *F*-test already accounts for differences in regional variability of the temperature and pressure anomalies in assessing where ASSIM has significantly smaller STD errors. For the ocean variables, SST and OHC, in Figure 4(c)–(f), only four NOASSIM members are used because the ASSIM errors cannot be lowered by averaging when three different versions of the ocean truth are being assessed. The first-year additional skill in ASSIM is dominated by the ENSO region, along with associated areas with SAT and SLP known to be influenced by ENSO such as northern Australia, the southwest USA and eastern China (e.g. Trenberth *et al.*, 1998), with areas defined as in Table I, based on C02 wherever possible. We did test whether the single 1997–1998 ENSO provided all the year 1 skill by removing the relevant start dates, but we still found mean skill in the same locations from the rest of the decade.

There is also enhanced first-year SAT skill in the Atlantic subpolar gyre and around the periphery of the Atlantic subtropical gyre and in the Nordic seas. These areas are potentially important to the European climate as they may affect the behaviour of the North Atlantic Oscillation (Rodwell *et al.*, 1999) and storm tracks, although a higher-resolution model may be required to show this. In the second year, areas of enhanced skill in ASSIM are greatly reduced, with the ENSO signal in the Pacific almost disappearing. There are small areas of greater SAT skill in the Indian ocean and South Pacific but nothing in the North Atlantic. Over land there is slightly greater skill over India, South Africa, northern Australia and tropical South America. These areas of enhanced skill are also significant in the individual hindcast ensembles (using separate ocean datasets), although there are equally many areas where initialized skill is reduced. There are no obvious regions of enhanced SLP skill anywhere in the second year (Figure 4(h)).

Figure 4(c,d) and (e,f) shows the equivalent maps to Figure 4(a,b) but for the SST and the top 113 m OHC, respectively. Both SST and OHC show similar regions of enhanced skill over the oceans as for SAT and show that skill is linked to the presence of predictable OHC anomalies. Some areas have more ocean temperature skill that SAT skill in the second year, including the tropical Pacific and the Nordic seas and the eastern boundary of the Atlantic. The only area showing significantly poorer ASSIM results is in the Atlantic subpolar gyre in the second year. This area requires further attention, as noted by Robson (2010), where large changes occur during the mid 1990s. The OHC does not show the large areas of enhanced skill in the southern

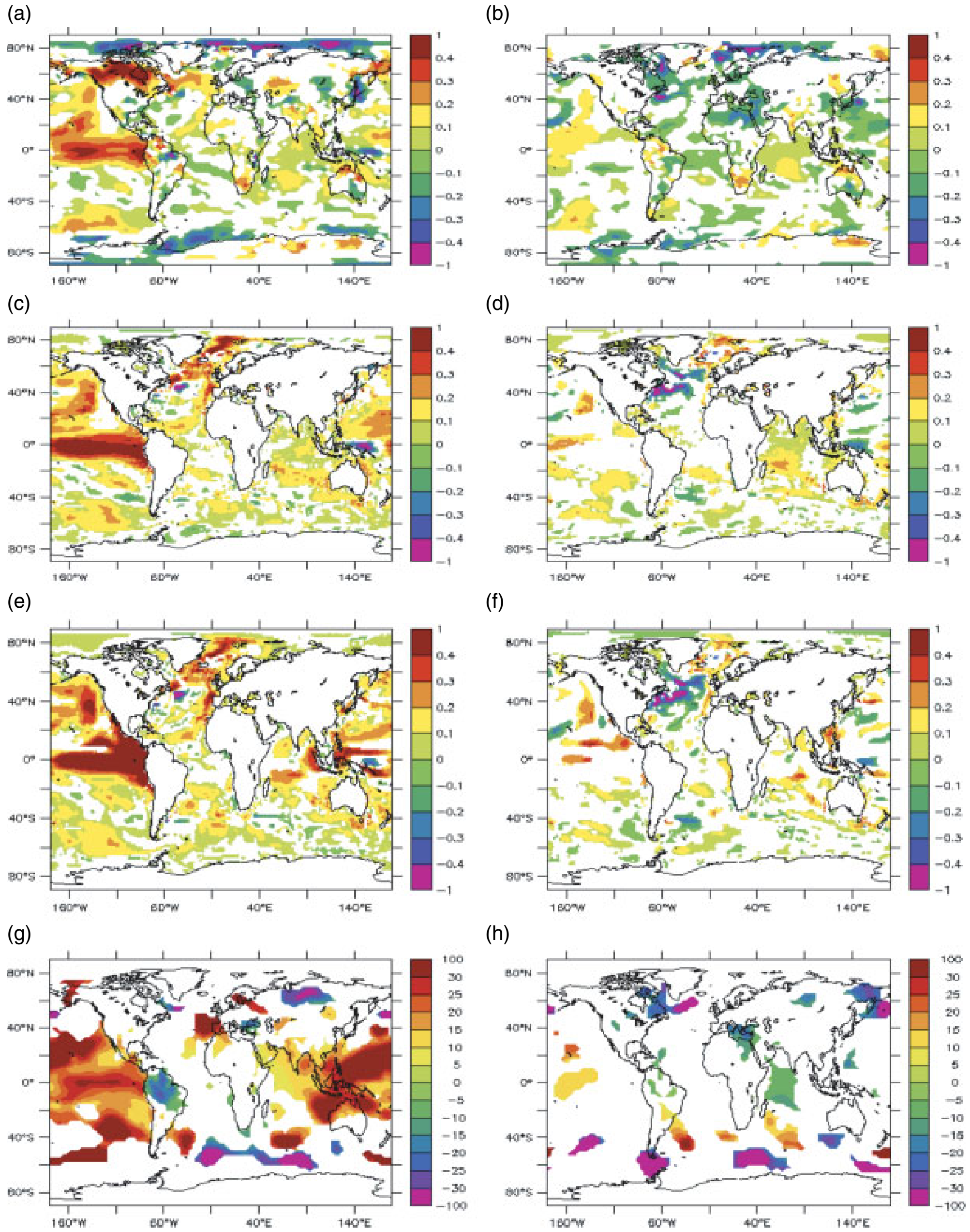

Figure 4. Maps of NOASSIM−ASSIM standard deviation errors for the SAT (a,b), SST (c,d), top 113 m OHC (e,f) and SLP (g,h). Only significant differences are shown using the *F*-test. The first-year anomaly standard deviation errors are on the left and the second-year standard deviation errors are on the right. Units are°C or Pa.

oceans as shown in Smith *et al.* (2007), which are likely an artefact of model drift (Robson, 2010).

Returning now to the comparison with C02, Table I indicates the regions identified in Figure 4 of C02 for assessing error growth. In addition, the mapped NOASSIM−ASSIM differences in our Figure 4 suggest identifying several additional areas, including the Indian Ocean, the Nordic seas, South Africa and eastern China – all areas clearly defined in Table I.

Figure 5 shows the SAT standard deviation errors in each of the ocean regions and Figure 6 for each of the land regions identified in Table I. Over the ocean (see Figures 3 and 5) the tropical Atlantic and the subtropical Pacific show the most consistently enhanced skill going out to 1 and 2 years ahead, respectively. This is entirely consistent with the C02 results and is also consistent with Hermanson and Sutton (2010), who looked at an idealized case study of predictability in HadCM3 over the tropical Atlantic. In the North Atlantic and the Atlantic subpolar gyre there is also some enhanced skill for the first year. All these regions are identifiable with enhanced skill in the individual ASSIM hindcast ensembles using the single-ocean datasets. However, over land areas (Figure 6) there is very little evidence of enhanced skill from the initialized runs, and this is also consistent with the C02 results. Although the maps in Figure 4(a, b) do show some regions of extra skill, these regions are small and may represent the expected level of noise in the results.

Enhanced skill in the North Atlantic (see Figure 3) is particularly relevant to the European climate, although the enhanced skill we find (Figure 4) is mostly focused on the Atlantic subpolar gyre and Nordic seas and tends only to persist for the first year. We tested the robustness of this result by looking at the NOASSIM−ASSIM error maps for the different ocean initialization states separately, and

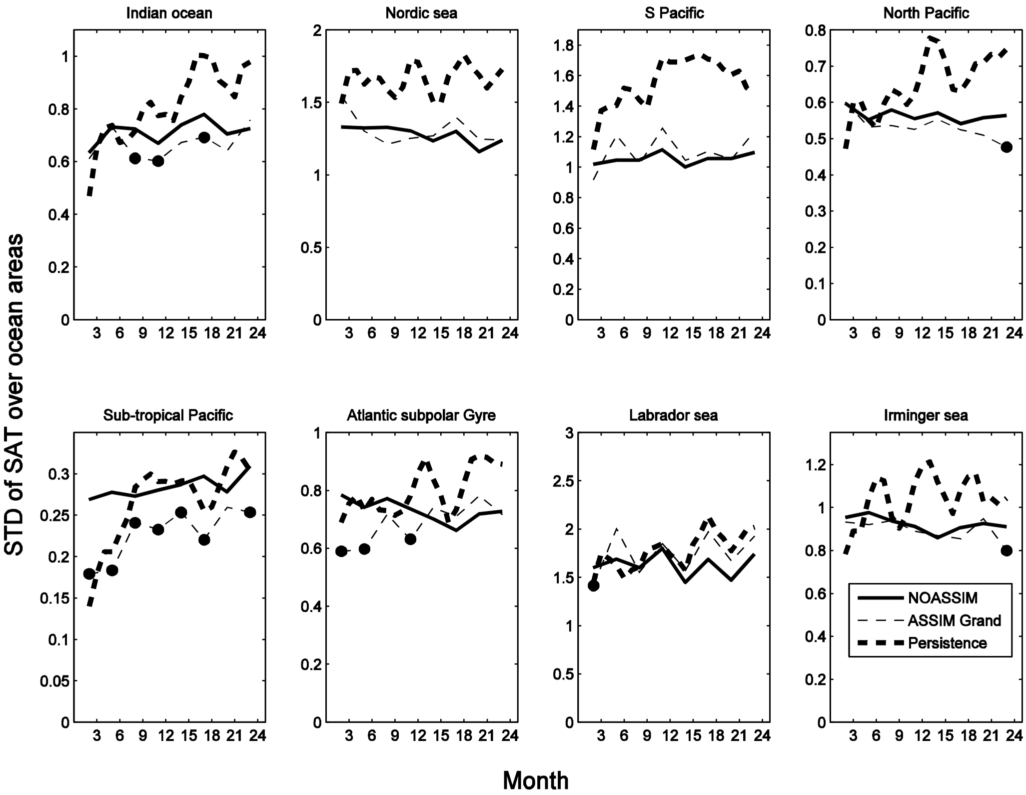

Figure 5. Ocean SAT anomaly error standard deviations for different regions as labelled. Regions are defined in Table I and taken from Collins (2002), along with additional regions chosen from Figure 4. The solid line is the NOASSIM ensemble mean errors, thicker dashed lines are persistence errors, and thin dashed lines are the Grand ensemble ASSIM errors, marked with a dot where the seasonal mean is significantly lower. Units are °C.

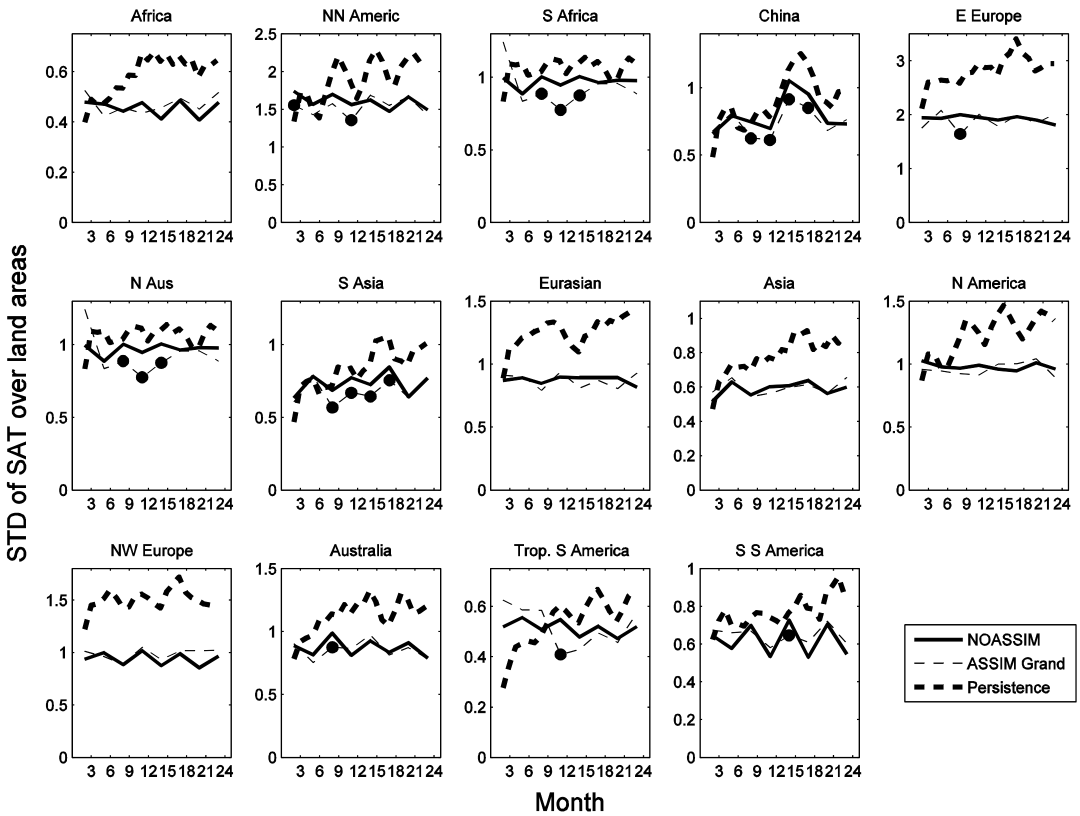

Figure 6. As Figure 5 but for land SAT.

found that in all three hindcasts there was some additional significant error reduction in the Atlantic subpolar gyre and the Nordic seas, although these are not as clear as in the Grand ensemble. In the second year only the Nordic seas appear still to show enhanced skill and this may be associated with predictability of the ice edge, which plays a crucial role in air–sea interaction in this region, although further studies of this are needed. The enhanced skill we find in the North Atlantic is therefore shorter lived and focused in different regions from the decadal skill identified in C02 and Pohlmann *et al.* (2009). We think this is due to the much longer time period sampled by these other authors with the decadal hindcasts from a wider range of decadal periods (whereas here we are reproducing only a single 12-year period of observational anomalies).

4. Summary and discussion

In this study we have quantified the hindcast skill of the HadCM3 climate model to predict the evolution of observed surface air temperature, sea surface temperature and heat content anomalies over 2-year periods. The enhanced skill of the system due to the initialization with observed ocean temperature and salinity anomalies is assessed by looking for significantly smaller standard deviation errors than equivalent forecasts without initialization. We show that, at least over the period of study (1990–2001), different realizations of the ocean initial conditions (which have considerable uncertainty in places) make little difference to the mean skill of the anomaly forecasting system (although individual forecasts will differ), and so we show most of the results as a Grand ensemble using all the ocean conditions in the hindcasts.

We show that the period of enhanced skill from the initialization is in many places remarkably similar to the idealized perfect twin results found in Collins (2002) with the same HadCM3 model. The areas noted as having enhanced skill in Collins (2002), such as the tropical Atlantic, show up clearly in this study. It is a remarkable result and very encouraging that the hindcasting of real observational anomalies can be improved, by using assimilation and initialization methods, to the same extent as for hindcasting idealized anomalies, and it strongly supports the use of these models in the development of real operational long-term forecasting systems (e.g. Smith *et al.*, 2007). Furthermore, our results provide important further evidence of regional prediction skill beyond the seasonal time-scale arising from initial conditions. We anticipate that the level of skill will be further improved in future with continued model improvements and sustained ocean observations.

However, in contrast to the predictions in C02, we are only evaluating against observational anomalies over a single 12-year period (1990–2001). In consequence it could be argued that these results should be regarded more as a case study of predictability over this particular period rather than a definitive measure of the mean predictability available from the DePreSys system. Hermanson and Sutton (2010) noted that individual cases of enhanced predictive skill may arise over many regions, e.g. long-range skill over land areas, in particular periods, when on average the enhanced skill attainable by initialization may be much less.

Although we find enhanced skill in the North Atlantic, by the second year this is focused at high latitudes over the Nordic seas. We do not find greatly enhanced skill further to the south as in C02 or Figure 2(c,d) of Pohlmann (2009), but we are not assessing the long decadal hindcast periods they studied, and we have only assessed skill against a very limited 12-year period of observational anomalies. The results are encouraging and suggest that many more detailed studies are possible, particularly looking at hindcasts over longer time periods.

Acknowledgements

Chunlei Liu and Alan Iwi were supported by the NERC e-Science program under the GCEP project, which also provided for the computer cluster used to perform the integrations. Doug Smith was supported by the Joint DECC and Defra Integrated Climate Programme – DECC/Defra (GA01101)UK. We particularly thank Rowan Sutton, Leon Hermanson and Jon Robson for many useful discussions.

References

Boer GJ. 2004. Long timescale potential predictability in an ensemble of coupled climate models. *Clim. Dynam.* **23**: 29–44.

Collins M. 2002. Climate predictability on interannual to decadal time scales: the initial value problem. *Clim. Dynam.* **19**(8): 671–692.

Collins M, Allen MR. 2002. Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate* **15**(21): 3104–3109.

Doblas-Reyes FJ, Balmaseda MA, Weisheimer A, Palmer TN. 2011. Decadal climate prediction with the ECMWF coupled forecast system: impact of ocean observations. *J. Geophy. Res. A*, DOI: 10.1029/2010JD015394.

Douville H. 2010. Relative contribution of soil moisture and snow mass to seasonal climate predictability: a pilot study. *Clim. Dynam.* **34**: 797–819.

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA. 2000. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dynam.* **16**: 147–168.

Haines K, Hermanson L, Liu C, Putt D, Sutton R, Iwi A, Smith D. 2009. Decadal climate prediction (project GCEP). *Phil. Trans. R. Soc. A* **367**: 925–937.

Hermanson L, Sutton RT. 2010. Case studies in interannual to decadal climate predictability. *Clim. Dynam.* **35**(7–8): 1169–1189.

Ingleby B, Huddleston M. 2007. Quality control of ocean temperature and salinity profiles: historical and real-time data. *J. Marine Syst.* **65**: 158–175.

Johns TC, Gregory JM, Ingram WJ, Johnson CE, Jones A, Lowe JA, Mitchell JFB, Roberts DL, Sexton DMH, Stevenson DS, Tett SFB, Woodage MJ. 2003. Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. *Clim. Dynam.* **20**: 583–612.

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.

Madec G, Delecluse P, Imbard M, Levy C. 1998. 'OPA 8.1 Ocean General Circulation Model reference manual'. Notes du pôle de modélisation, Institut Pierre Simon Laplace (IPSL). http://www.lodyc.jussieu.fr/opa/

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

Pierce, DW, Barnett TP, Tokmakian R, Semtner A, Maltrud M, Lysne J, Craig A. 2004. The ACPI project, element 1: initializing a coupled climate model from observed conditions. *Climatic Change* **62**: 13–28.

Pohlmann H, Botzet M, Latif M, Roesch A, Wild M, Tschuck P. 2004. Estimating the decadal predictability of a coupled AOGCM. *J. Climate* **17**: 4463–4472.

Pohlmann H, Jungclaus J, Koehl A, Stammer D, Marotzke J. 2009. Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic. *J. Climate* **22**(14): 3926–3938.

Robson J. 2010. 'Understanding the performance of a decadal prediction system'. PhD thesis, University of Reading, UK.

Rodwell MJ, Rowell DP, Folland CK. 1999. Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Nature* **398**: 320–323.

Simmons AJ, Hollingsworth A. 2002. Some aspects of the improvement in skill in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**: 647–677.

Smith G, Haines K. 2009. Evaluation of the S(T) assimilation method with the Argo dataset. *Q. J. R. Meteorol. Soc.* **135**: 739–756.

Smith DM, Murphy JM. 2007. An objective ocean temperature and salinity analysis using covariances from a global climate model. *J. Geophys. Res.* **112**: C02022, DOI: 10.1029/2005JC003172.

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.

Stockdale TN. 1997. Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Weather Rev.* **25**: 809–818.

Stott PA, Tett SFB, Jones GS, Allen MR, Mitchell JFB, Jenkins GJ. 2000. External control of 20th century temperature by natural and anthropogenic factors. *Science* **290**: 2133–2137.

Trenberth KE, Branstator GW, Karoly D, Kumar A, Lau N-C, Ropelewski C. 1998. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.* **103**: 14291–14324.

Uppala SM, Kållberg PW, Simmons AJ, Andrae U, Da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, Van De Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, Mcnally AP, Mahfouf J-F, Morcrette J-J, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**: 2961–3012.