Predictions of Climate Several Years Ahead Using an Improved Decadal Prediction System

JEFF R. KNIGHT, MARTIN B. ANDREWS, DOUG M. SMITH, ALBERTO ARRIBAS, ANDREW W. COLMAN, NICK J. DUNSTONE, ROSIE EADE, LEON HERMANSON, CRAIG MACLACHLAN, K. ANDREW PETERSON, ADAM A. SCAIFE, AND ANDREW WILLIAMS

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 6 January 2014, in final form 11 July 2014)

ABSTRACT

Decadal climate predictions are now established as a source of information on future climate alongside longer-term climate projections. This information has the potential to provide key evidence for decisions on climate change adaptation, especially at regional scales. Its importance implies that following the creation of an initial generation of decadal prediction systems, a process of continual development is needed to produce successive versions with better predictive skill. Here, a new version of the Met Office Hadley Centre Decadal Prediction System (DePreSys 2) is introduced, which builds upon the success of the original DePreSys. DePreSys 2 benefits from inclusion of a newer and more realistic climate model, the Hadley Centre Global Environmental Model version 3 (HadGEM3), but shares a very similar approach to initialization with its predecessor. By performing a large suite of reforecasts, it is shown that DePreSys 2 offers improved skill in predicting climate several years ahead. Differences in skill between the two systems are likely due to a multitude of differences between the underlying climate models, but it is demonstrated herein that improved simulation of tropical Pacific variability is a key source of the improved skill in DePreSys 2. While DePreSys 2 is clearly more skillful than DePreSys in a global sense, it is shown that decreases in skill in some high-latitude regions are related to errors in representing long-term trends. Detrending the results focuses on the prediction of decadal time-scale variability, and shows that the improvement in skill in DePreSys 2 is even more marked.

1. Introduction

Until relatively recently, climate prediction for time scales between those of seasonal forecasting and centennial projections had received relatively little attention. At these intermediate time scales, both the initial state of the climate system and external climate forcing factors are important in determining how climate evolves. The development of dedicated decadal prediction systems (e.g., Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2009) has filled the gap, and initialized climate prediction has become an important new component of the recent Coupled Model Intercomparison project phase 5 (CMIP5) activity (Kirtman et al. 2014). Decadal prediction is also becoming recognized as potentially a key tool for assessing strategies for adaptation to climate change in the short to medium term (Trenberth 2008), especially at the regional scale where the signal-to-noise ratio of the forced climate response is smaller (Hawkins and Sutton 2009). In spite of this, skill in current decadal prediction efforts (Doblas-Reyes et al. 2013; Kirtman et al. 2014) falls short of more idealized estimates of potential predictability (e.g., Collins 2002; Collins et al. 2006; Branstator et al. 2012). If current systems are near the limit of practical predictability then further work is needed to understand this discrepancy. More likely, given the known limitations of climate modeling, is that systems need to be improved to exploit the available predictability better. Decadal prediction systems can be improved by better initialization, more accurate external forcing, or by use of a climate model that is more accurate in predicting the climate state. Here, we examine the latter approach to producing improved decadal predictions.

The Met Office Hadley Centre Decadal Prediction System (DePreSys) was the first to attempt initialized climate predictions for a decade ahead (Smith et al. 2007).
At its core is the Hadley Centre Coupled Model version 3 (HadCM3), a coupled ocean–atmosphere model (Gordon et al. 2000), which has been highly successful in providing insights into climate and climate change (e.g., Vellinga and Wood 2002; Johns et al. 2003; Stott et al. 2004; Stainforth et al. 2005). Nevertheless, HadCM3 was developed more than a decade ago in the late 1990s and since then there have been advances in the understanding of climate processes (e.g., aerosol indirect effects) and the availability of computational resources. Both of these aspects have allowed the development of increasingly sophisticated models run at higher grid resolutions. The latest Hadley Centre climate modeling system, the Hadley Centre Global Environmental Model version 3 (HadGEM3; Hewitt et al. 2011), is a good example. In addition to having increased horizontal and vertical resolution in the ocean and atmosphere compared to HadCM3, it has the ability to represent a wide range of Earth system processes, such as chemistry–aerosol–cloud interactions. More details of the differences between the models are provided below. Analysis of HadGEM3’s performance has shown it to be an improvement on earlier Hadley Centre models (Walters et al. 2011). It is now appropriate, therefore, to exploit the benefits of climate model development in decadal prediction.

In this paper a new decadal prediction system called DePreSys 2 will be described. The principal innovation in this system is the use of the Met Office Hadley Centre’s contemporary climate model HadGEM3 to replace the HadCM3 model used in the original system. The predictive capability of the system will be demonstrated by examining the results of a new set of retrospective forecasts of climate over the last 50 yr made using DePreSys 2. The equivalent set of reforecasts from the original DePreSys system (hereafter called DePreSys 1) are those submitted to the CMIP5 project (Kirtman et al. 2014) and will be used as a baseline to assess potential improvements in skill. The paper will focus on determining the skill of DePreSys 2 compared to DePreSys 1, and understanding the sources of any differences between the systems.

### 2. Methods

DePreSys 2 uses the Hadley Centre Global Environmental Model version 3 (Hewitt et al. 2011; Walters et al. 2011). HadGEM3 includes submodels for the atmosphere, ocean, sea ice, and land surface, coupled via the Ocean Atmosphere Sea Ice Soil (OASIS3) coupler. The atmospheric model, the Hadley Centre Global Atmospheric Model version 3 (HadGAM3), is comprehensively different from the atmospheric component of the HadCM3 model (HadAM3; Pope et al. 2000) used in DePreSys 1, with a semi-Lagrangian dynamical core and entirely revised physical parameterizations. With a horizontal resolution of 1.25° in latitude by 1.875° in longitude, it has double the resolution of HadAM3 (with 2.5° by 3.75°), and 85 vertical levels to 85 km altitude (0.01 hPa) compared to 19 levels to 40 km (5 hPa) in HadAM3. HadGEM3 also uses different ocean and sea ice models than HadCM3, specifically the Nucleus for European Modelling of the Ocean (NEMO) v3.2, and the Los Alamos sea ice model (CICE), respectively. NEMO uses a tripolar ORCA1 grid with nominal resolution of 1° and has 75 vertical levels. CICE uses four thickness categories to model the evolution of sea ice. The differences between the two systems’ component models are summarized in Table 1.

Despite the underlying climate model being entirely new, the methodology of producing decadal forecasts with DePreSys 2 is closely based on that used in DePreSys 1 (Smith et al. 2007). This consists of an anomaly initialization method where an estimate of the observed climatology is subtracted from observational data to produce observational anomalies. These anomalies are added to an estimate of the model climatology to provide an initial state in model hindcasts and forecasts. This is considered beneficial since model climatology is usually biased compared to observations, with biases often exceeding the size of the variability that we

<table>
<thead>
<tr>
<th>Table 1. Comparison of the underlying models for DePreSys 1 and DePreSys 2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributive grid</td>
</tr>
<tr>
<td>Top level</td>
</tr>
<tr>
<td>Dynamical scheme</td>
</tr>
<tr>
<td>Time step</td>
</tr>
<tr>
<td>Coupling frequency</td>
</tr>
</tbody>
</table>
are attempting to predict. As a result, a realistic initial state will drift toward the model climatology over time. In the anomaly approach, therefore, the model is initialized with a state that is much more likely to be within the limits of its own variability, so avoiding large drift, but preserving observational information, at least in an anomaly sense. The disadvantage of the anomaly method is that the absolute values of the model’s initial climate variables are further from the observed initial state, which could generate errors if there is substantial nonlinearity in the climate system. The relative benefits of initialization with and without using the anomaly method (full-field initialization) are still to be fully explored in a range of decadal prediction systems, although some initial results suggest that the difference in predictive skill is minimal (Pohlmann et al. 2013; Hazeleger et al. 2013). Given this, and the fact that DePreSys 1 (Smith et al. 2007) demonstrates that anomaly initialization produces skillful results, this approach is adopted for DePreSys 2. As a consequence, any differences in levels of skill in DePreSys 2 and DePreSys 1 will be primarily attributable to the use of the new underlying model.

As already indicated, anomaly initialization requires the production of comparable estimates of the observed and simulated climatologies of the atmosphere and ocean. For observations, it is desirable to have as long a data series as possible to minimize the effect of sampling of climate variability on the estimate obtained. Given that very few subsurface ocean data are available before the 1960s, our best estimate is for the period from 1960 to near the present day [the climatology used here is 1960–2006 following Smith et al. (2007)]. To obtain the equivalent estimate for the model requires climate integrations of the same period using transient climate forcing factors, since global climate was not stable in this period. Since the only control simulation with HadGEM3 uses near-present-day levels of climate forcing, a spinup simulation of 86 years using values of atmospheric trace gases, aerosols, ozone, solar irradiation and volcanic aerosols appropriate to 1960 is produced to provide a stable initial state for these transient simulations (Fig. 1). During this spinup integration it was found that sea ice thickness in certain parts of the Canadian Archipelago in HadGEM3 became unrealistically large, leading to numerical failures in the simulation through unrealistic local ice–ocean interactions. To prevent this, the thickness was artificially limited to 4 m globally on 1 September every year in all HadGEM3 model integrations in this study. Separate investigations revealed the cause of the ice build-up to be an error in coding in the shortwave radiation scheme at coastal grid points.

**FIG. 1.** Schematic diagram illustrating the experimental design.
This has been corrected in subsequent versions of HadGEM3.

Three transient simulations are performed from 1960 to 2009 using initial conditions at 10 yr intervals from near the end of the spinup. The simulations use CMIP5 historical (to 2005) and representative concentration pathway (RCP) 8.5 time-evolving forcings appropriate for the period 1960 to 2009 (Jones et al. 2011). This includes well-mixed greenhouse gases \([\text{CO}_2, \text{CH}_4, \text{N}_2\text{O}, \text{chlorofluorocarbons (CFCs)}, \text{and hydrofluorocarbons (HFCs)})\], tropospheric aerosols— including sulfates, soot, biomass aerosols and organic carbon from fossil fuels—varying tropospheric and stratospheric ozone concentrations, solar variability, and volcanic eruptions. Solar variability is held constant after December 2008.

The results of the three transient simulations are combined to produce climatologies of key quantities for the HadGEM3 model for the period 1960–2006. These are used in combination with observational datasets to produce initial conditions for the DePreSys 2 hindcast ensembles. To obtain initial conditions for the full range of variables required to initialize the model from a limited number of observed fields, an “assimilation” integration of the model is performed (Fig. 1). This is identical to the transient simulations but with relaxation of atmospheric (three-dimensional zonal and meridional winds and temperatures), oceanic (three-dimensional temperature and salinity), and sea ice (spatial concentration) data derived from observations to constrain the climate state. The atmospheric data used in the assimilation are derived from a combination of 6-hourly 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) and Interim ECMWF Re-Analysis (ERA-Interim; Dee et al. 2011) data. The ERA-40 data are adjusted to be consistent with ERA-Interim by using the difference in the seasonally varying climatology in a period of overlap (January 1989–August 2002). A combined observationally based climatology for the period 1960–2006 is then calculated from the adjusted data. The anomalies of the 6-h adjusted ERA data for 1960–2009 are calculated by subtracting this climatology, and added onto the model climatology for 1960–2006 to provide input into the assimilation run. Newtonian relaxation of the relevant model fields toward these data is performed with a time scale of 6 h (Telford et al. 2008).

The observational constraint on the ocean state in the assimilation comes from the dataset created by Smith et al. (2007), who developed a monthly mean ocean analysis using a four-dimensional, multivariate optimal interpolation of subsurface temperature and salinity observations, along with sea surface temperature from the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST; Rayner et al. 2003). Their technique uses the global-scale covariance of these fields estimated initially from the HadCM3 model to provide globally far more complete estimates of the historical ocean state than is available from methods that use only the local-scale covariance. In producing the assimilation for DePreSys 2, we utilize the Smith et al. analysis interpolated onto the HadGEM3 NEMO ocean model grid. As with the atmospheric quantities, the anomalies from this analysis are added to the model climatology. Sea ice concentration data for assimilation were obtained from HadISST and interpolated onto the HadGEM3 grid. As with the atmosphere, the ocean and sea ice inputs are assimilated using a 6-h Newtonian relaxation time scale. Sea ice thickness was not constrained but allowed to evolve in the model in response to the imposed concentration changes. Data for each of the model subcomponents (atmosphere, ocean and sea ice) are assimilated into a single coupled climate integration, allowing for the possibility that observational data made available to one of these subcomponents might improve the representation of the state of the other subcomponents via the model’s coupling. As a result of using essentially the same datasets for assimilation, the analysis produced for DePreSys 2 is very similar to that for DePreSys 1 in an anomaly sense, with absolute values differing primarily as a result of the different climatologies of the constituent models.

Output from the assimilation integration is used to provide atmosphere–ocean–sea ice initial conditions for a set of reforecast (hindcast) simulations from 1960 through to 2008 (Fig. 1). In this paper we will focus on these hindcasts, rather than the initialization or production of forecasts, since comparing the hindcasts with past observations allows the skill of the prediction system to be assessed. Nevertheless, the initialization of forecast simulations is performed in the same way, using up-to-date observational data to extend the assimilation integration in near–real time. The hindcasts consist of four-member ensembles initialized on 1 November in each year from 1960 to 2008. The initial conditions in each ensemble member are identical, so ensemble spread is obtained by using the stochastic kinetic energy backscatter (SKEB2) scheme (Tennant et al. 2011) within the atmospheric component of HadGEM3. SKEB2 is designed to improve the upscale cascade of energy from unresolved scales by adding stochastically generated wind increments at every atmospheric model time step. It is used here to introduce random changes to the model’s evolution as a result of selecting different values of an initial “seed” value. As a consequence of its use, different ensemble members quickly diverge as the
small perturbations are amplified by chaotic atmospheric dynamics. This method of producing ensemble spread differs from that used in DePreSys 1, where initial conditions from consecutive days preceding and including the hindcast start date were used, but is the same as is successfully used in the Met Office operational seasonal prediction system (Arribas et al. 2011). Note that SKEB2 is utilized here as a component of the HadGEM3 model, so it is active in the transient and assimilation simulations as well as continuously throughout the hindcasts.

Each hindcast member is run for 5 yr, since Smith et al. (2007, 2010) demonstrated that in DePreSys 1 most of the effect of initialization was found within the first half of the decade. The hindcasts use the same climate forcings as are used in the transient simulations of the HadGEM3 model to mirror the experimental protocol for decadal prediction used in CMIP5. This means the hindcasts are not true retrospective forecasts, as they include foreknowledge of forcings such as the occasional abrupt cooling from volcanic eruptions. We compare the DePreSys 2 hindcasts with those from the DePreSys 1 hindcast submitted to CMIP5. This was also initialized on 1 November of each year of our hindcast period, with 10-member ensembles available for each initial date. For comparability, we will use only the first four members of the DePreSys 1 ensemble in calculating measures of the relative skill in the prediction systems.

3. Results

a. Comparison of skill between DePreSys versions 1 and 2

Model biases prevent direct comparison of the output from the decadal prediction system with observations. The anomaly initialization method, however, ensures that a priori the results are expected to be unbiased with respect to the climatology of the underlying model for all prediction time scales. As a result, we are able to compare simulated anomalies (with respect to model climatology) with observed anomalies (with respect to the observational climatology). Two commonly used measures to compare skill in decadal prediction systems are used here: the anomaly correlation coefficient (ACC) and the root-mean-square error (RMSE). The former is a measure of the similarity of the anomalies in terms of the sequence of anomalous events, whereas the latter measures the overall deviation of the prediction, including magnitude, from observations. Both of these statistics are calculated using ensemble means of the time series obtained in the hindcast simulations along with the corresponding observational data. Probabilistic measures, which evaluate the hindcast distribution in addition to the ensemble mean, are another way to assess skill. To obtain estimates of these types of statistics with usable uncertainties, however, likely requires larger ensemble sizes than the four members available here (Corti et al. 2012). In addition to RMSE, the standardized RMSE (SRMSE; Eade et al. 2012) is used in all spatial maps so that geographical comparisons of skill can be more easily made. SRMSE estimates are RMSE normalized by the standard deviation of the variability in the observations to provide a sense of scale. SRMSE is less appropriate for globally integrated skill measures, however, so in these instances RMSE is retained.

To get a first impression of the relative skill in DePreSys 1 and DePreSys 2, we calculate ACC and SRMSE for the mean near-surface temperature in years 2–5 of the hindcasts (Fig. 2), using observed anomalies from a combination (simple average) of the Hadley Centre Climate Research Unit (HadCRUT3; Brohan et al. 2006), National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS; Hansen et al. 2010), and National Climatic Data Center (NCDC; Smith et al. 2008; Vose et al. 2012) datasets. We use this approach to verification since we view each individual dataset as an equiprobable estimate of past surface temperatures. The year 2–5 metrics focus on the skill beyond seasonal time scales. Significance is estimated from 1000 random resamples of the hindcast data from each system’s ensemble (Smith et al. 2013). The aim is to test the uncertainty arising from the use of a finite ensemble size. Each resampling is performed by selecting a random ensemble member (1 to 4) for each consecutive (randomly selected) 5-yr block of start times to construct a “pseudo-ensemble” member. Repeating this process four times for each system’s ensemble gives a pair of pseudo-ensembles and corresponding pseudo-ensemble means. By chance, these pseudo-ensembles contain subsets of the original ensembles’ data and so allow pairs of alternative estimates of our chosen skill metrics to be made. Over many trials, this process provides distributions of skill estimates that tend to cluster around the original (full ensemble data) values. The 5%–95% ranges of resampled ACC and RMSE values give estimates of uncertainty. For ACC, where these uncertainty ranges do not include zero, we determine that significant skill has been detected. For SRMSE on the other hand, we test for overlap with unity, since we are interested in where the hindcast is significantly more accurate than the observed variability. The significance of differences between the ensembles is assessed by examining the range of differences in ACC and SRMSE derived from the resampled pseudo-ensembles. Here we
test for overlap with zero for both quantities to ascribe significance.

DePreSys 1 is significantly skilful in most regions of the globe, in particular in the tropics and subtropics, except for the northern central Pacific. In the high latitudes there is less ACC skill, especially in the Southern Hemisphere, as might be expected given the paucity of historical ocean observations available for initialization in this region. DePreSys 2 shows a similar pattern of ACC to that in DePreSys 1, but with generally higher values and slightly broader significance. Over some parts of the North Atlantic, the Mediterranean, North Africa, and South Asia, the DePreSys 2 year 2–5 ACC exceeds 0.9. In addition, DePreSys 2 substantially improves the poor ACC skill over many parts of the tropical Pacific and North Pacific, albeit without eliminating this feature entirely. The difference in ACC between the two systems also shows this general improvement, and highlights specific regions where skill has significantly improved, such as the tropical west Pacific, midlatitude North Atlantic, and northern Asia. SRMSE patterns are also similar between the two systems, with tropical errors tending to be smaller than those at high latitudes. DePreSys 2 shows significantly larger errors than DePreSys 1 over the Southern Ocean, however, and also has large errors over the northern North Atlantic and Pacific Oceans that do not appear in DePreSys 1. The difference in SRMSE between the two systems shows that tropical skill is generally significantly improved in DePreSys 2, in agreement with the results obtained using ACC. In the mid and high latitudes, however, the SRMSE reveals a reduction in predictive skill that is only partly apparent in the ACC.

An equivalent view of the relative skill in mean precipitation for years 2–5 also shows some improvements in DePreSys 2 (Fig. 3). For this analysis we use the Global Precipitation Climatology Centre (GPCC) historical rainfall dataset (Schneider et al. 2011) to compare with the hindcast simulations. Use of this dataset means we can only analyze performance over land, since marine precipitation is unavailable before the satellite era and thus not throughout our hindcast period from 1960. While the difference between the systems is noisier than for temperature, more areas show improvement than deterioration. In particular, the semi-arid regions bordering the Sahara Desert show significantly more skill, as do parts of southern Africa and North America. Central Asia, on the other hand, appears to have significantly less precipitation skill in DePreSys 2 than DePreSys 1, as indicated by the higher SRMSE values in this region.

To summarize the differences in skill between the systems, we compute the area-weighted global average ACC for a range of key variables and lead times in the
hindcasts from each system (Fig. 4). Significance of the computed differences is estimated in the same way as for each grid point in the skill maps (Figs. 2 and 3), that is, from the range of differences in ACC obtained from large sets of random resamples of each ensemble. Where these uncertainty ranges do not include zero, we determine that a significant difference has been detected. For near-surface temperature, DePreSys 2 is consistently more skilful than DePreSys 1 for all time scales from year 1 to years 2–5. These differences are significant at the 90% confidence limit for most lead times. Skill for precipitation is much lower than for temperature, as is usually found on these time scales (Eade et al. 2012). Nevertheless, DePreSys 2 shows a marked and significant increase in skill for precipitation in year 1. For longer time scales, the skill in DePreSys 2 is consistently higher than in DePreSys 1, although significant improvements are not generally seen. For mean sea level pressure (MSLP), however, skill is generally higher and the change in ACC between DePreSys 1 and DePreSys 2

Fig. 3. As in Fig. 2, but for precipitation, with a linear color scale for SRMSE differences.

Fig. 4. Global mean area-weighted ACC for near-surface temperature (dots), precipitation (triangles), and mean sea level pressure (diamonds) as a function of hindcast lead time. Values for DePreSys 1 are in blue and values for DePreSys 2 are in red. The 5%–95% significance ranges of differences (DePreSys 2 – DePreSys 1) are shown by the vertical black bars.
is significantly positive for year 1 and the averages over years 2–4 and 2–5.

Global average RMSE can also be used to characterize the differences in skill between the systems (Fig. 5). For near-surface temperature, DePreSys 2 appears to have slightly larger errors, and hence less skill, than DePreSys 1. At shorter time scales the differences are not significant, but are marginally significant at longer time scales (years 2–4 and 2–5). This is in contrast to the results for ACC, which indicate that DePreSys 2 is more skilful. The results are consistent, however, with the maps in Fig. 2, which show extratropical regions in which the RMSE is much larger in DePreSys 2 than in DePreSys 1. Outside of these regions RMSE is generally better in the new system, and is better over many (although not all) global land regions. As a result, area-average RMSE values in the tropical and midlatitude zone between 40°N and 40°S have the opposite tendency, with significantly improved skill in DePreSys 2 compared to DePreSys 1 (not shown). The global average RMSE for precipitation shows that DePreSys 1 again significantly exceeds the skill of DePreSys 2. This deterioration also appears to arise from the representation of parts of the extratropics, namely over much of Asia. For MSLP, differences between the two systems are not generally statistically significant. The fact that global improvements in RMSE skill appear to be hampered by large errors in specific geographical regions will be examined in section 3b.

Global mean near-surface temperature (abbreviated as GMNST) is a key output from decadal prediction systems (e.g., Smith et al. 2007; Kirtman et al. 2014). The ACC of the mean GMNST over years 2–5 of the forecasts is 0.96 for DePreSys 1 and 0.95 for DePreSys 2. These values are very high because strong positive GMNST trends observed over the hindcast period are also found in the hindcasts themselves. An analysis of detrended GMNST skill will be provided in section 3c. The difference in ACC between the systems is not significant at the 90% level, as inferred by the same resampling methods used above. The RMSE of years 2–5 GMNST is 0.069 K for DePreSys 1 and 0.075 K for DePreSys 2. Again, this difference is not significant at the 90% level.

b. Sources of differences in skill

The results in section 3a show that some of the largest improvements in predictive skill occur in the tropics and subtropics. Given its importance to the climate of these regions, we ask whether the tropical Pacific El Niño–Southern Oscillation (ENSO) phenomenon is linked with the improved skill in DePreSys 2. Both the representation of the ENSO process itself and its teleconnections with remote regions have the scope to improve in the HadGEM3 model, and we will examine both of these aspects.

Both decadal systems have at least moderate predictive skill for the Niño-4 index well beyond the seasonal time scale, with a correlation of greater than 0.3 or more for approximately the first 18 months after initialization (Fig. 6). Where there exists this level of skill, DePreSys 2 appears to be consistently more skilful than DePreSys 1. The significance of this result is tested using the same resampling method introduced above. The ACC is not generally significantly larger in DePreSys 2 than in DePreSys 1 at the 90% level, except where it is marginally so toward the ends of the first boreal spring.
and autumn. On the other hand, the difference in RMSE is consistently negative and significant, indicating lower RMSE values in DePreSys 2 between the late boreal summer of the first year and the summer of the second year (approximately between months 10 and 20 of the hindcasts). As a result, there is evidence for better prediction of ENSO in DePreSys 2. The reason for the lack of significance in the differences in ACC is investigated using the full 10-member DePreSys 1 ensembles (as opposed to the first four members used for comparison with the four-member DePreSys 2 ensembles). It is found that the first four members of the DePreSys 1 ensemble are particularly unrepresentative of the typical four-member seasonal ACC (lying at greater than the 90th percentile of the distribution) for Niño-4 between 1- and 2-yr lead times. Average ACC in four-member subsets of the 10-member DePreSys 1 ensemble is up to 0.06 lower than the ACC derived from the first four members. This suggests it is likely that there is a larger and more confident difference in skill, although we do not know how representative the four DePreSys 2 members are compared to a larger ensemble. The improvement in ENSO skill is linked to better representation of the intrinsic pattern of ENSO variability in HadGEM3. The newer model corrects much of the error in the ENSO SST pattern in HadCM3, which stretches too far into the western Pacific (Collins et al. 2001). In contrast, HadGEM3 has a much more realistic westerly extent in free-running simulations (not shown).

Having established the intrinsically better skill of the ENSO process in the new system, it is also possible to examine teleconnections between ENSO and other parts of the world (Fig. 7). These play an important role in year-to-year climate variability in many regions, especially in the tropics. The teleconnections depend on season, so we calculate correlations of near-surface temperature with the Niño-4 index for both December–February (DJF) and June–August (JJA). Observational teleconnections show a familiar pattern in the Pacific Ocean, with anomalies of one sign in the tropical east and central Pacific flanked by anomalies of the opposite sign to the north, south, and west, particularly in boreal

![Fig. 6. Skill of 3-month average Niño-4 index (mean SST for 5°N–5°S, 160°E–150°W) as a function of lead time. The first points on the x axes correspond to the first DJF season in the hindcast ensembles. Successive points correspond to successive 3-month seasons (i.e., January–March, February–April, etc.) until the third DJF in the hindcasts. (top) The ACC for DePreSys 1 (black curve) is compared to the ACC for DePreSys 2 (red curve), and the difference DePreSys 2 – DePreSys 1 (dashed red curve). The 5%-95% range of ACC differences from a large set of random resamples of the data from each system is shown as the gray region. (bottom) As at top, but for RMSE (K). The verifying SST data are taken from the HadISST dataset (Rayner et al. 2003).](image-url)
winter. In addition, there are remote associations in the Indian and Atlantic Oceans and over Africa, South America, South Asia, Australia, and parts of North America. To assess the systems’ “errors” in reproducing observed teleconnections, we show the difference between the correlation computed for each system and the correlation in observations. The system teleconnections are derived using data for all lead times and members of all ensembles in the respective hindcasts. These have been found to provide a good estimate of the teleconnections in the underlying model (not shown).

For DePreSys 1, it is clear that there are errors in the representation of ENSO within the tropical Pacific itself. In particular, the westernmost equatorial Pacific is correlated with Niño-4, not anticorrelated as in observations. This indicates that the HadCM3 model used in this system tends to extend the ENSO signal too far into the west Pacific. In addition, there are dipolar errors suggesting displacement of the South Pacific convergence zone (SPCZ). In DePreSys 2, using the HadGEM3 model, these errors are much reduced. There remains a relatively small overextension of ENSO into the west Pacific in JJA, but in DJF this feature is almost eliminated. Beyond the Pacific Ocean, HadGEM3 has more accurate Niño-4 teleconnections to the South Atlantic Ocean and South America, especially in JJA. An overall estimate of the faithfulness of the models in reproducing global teleconnections is obtained from area-weighted pattern correlations between the model teleconnection patterns and the observed teleconnection patterns. For the HadCM3 model used in DePreSys 1, these are 0.71 ± 0.04 (DJF; 5%–95% range) and 0.61 ± 0.05 (JJA), whereas in the HadGEM3 model used in DePreSys 2 these are 0.81 ± 0.03 (DJF) and 0.74 ± 0.05 (JJA). Consequently, it can be concluded that DePreSys 2 has better ENSO teleconnections. This, along with the intrinsically better representation of ENSO itself, contributes to the improvement in skill over DePreSys 1 seen in many regions in years 1 and 2 (the period for which there is significant ENSO skill). Beyond the 2-yr time scale, other factors must be responsible for the improved skill. One possibility is improvements in representing trends, which contribute to skill, especially for temperature, at longer time scales. The role of trends will be examined further in section 3c.

In addition to the widespread improvement in skill between DePreSys 1 and DePreSys 2, there are some high-latitude areas which show lower skill through markedly higher SRMSE values in DePreSys 2 (e.g., northern North Atlantic and North Pacific and the Arctic and Atlantic sector of the Southern Ocean; see Fig. 2). Here we will investigate the source of these larger errors. We take the subpolar gyre (SPG) region of the northern North Atlantic as an example, since this region has been the focus of many efforts to understand decadal time-scale variability via interactions with ocean
circulation (Pohlmann et al. 2013). Area-averaged year 2–5 surface temperature in this region from the DePreSys 2 hindcast ensemble is compared with values in the transient ensemble and observations in Fig. 8. It is immediately clear that there is a large cold bias in the transient (uninitialized) model simulations of about 5°C. While undesirable, this is not a fundamental obstacle to prediction using the anomaly method since initial and predicted anomalies are referenced to the model climatology. So long, therefore, as biases do not change the variability too much, then the approach is viable. The initialized ensembles do not, however, appear to have realistic variability. In the 1970s and 1980s, temperatures tend to be even lower (~3°C) than the model climatology, whereas after 2000 temperatures are about 1°C to 2°C higher than the model climatology. The magnitude of the resulting trend (~5°C in 40 yr) is greatly in excess of that seen in observations, despite the well-documented rise in SPG SSTs in the mid-1990s (Robson et al. 2012). This difference in trend clearly accounts for the large RMSE found in this region in Fig. 2.

The source of the very large SPG warming trend is explained by the interaction of initialization with sea ice feedbacks. Figure 9 shows sea ice concentration in the Atlantic sector in March, when Northern Hemisphere sea ice extent reaches its annual peak, comparing data from two periods: 1968–87 and 1988–2009. In observations, the mean ice edge lies almost entirely outside the SPG region used to calculate the temperature curves in Fig. 8. Some retreat is visible between these periods, especially in the Norwegian Sea. For the HadGEM3 transient ensemble, from which the model climatology is derived, there is, in contrast, substantially more peak sea ice coverage, consistent with the large cold bias. In the DePreSys 2 ensembles the sea ice extent is even greater in the earlier period than it is in the transient simulations, covering much of the SPG region. There is also a very large retreat (to levels less than in the transient simulations) by the later period.

The presence of substantial sea ice in the SPG of the model explains the source of the large trend in near-surface air temperature. Air temperature over the ocean generally has only a small deviation from local SST, which is limited by the freezing point of seawater (~1.8°C). Air temperature over near complete coverage of sea ice, however, can be substantially lower, due to the insulating effect of ice cover. As a result, the fractional coverage of sea ice within the domain is an important factor in determining its mean air temperature. As the ice retreats, more of the area is covered by the relatively warm ocean surface and the average temperature of the SPG greatly increases. These ice changes are
largest in the initialized hindcasts because of the temperature anomalies in the initial conditions and the feedback between temperature and sea ice growth; that is, when SSTs are colder, sea ice extent increases, reducing SSTs in adjacent regions through export of cold air and causing further growth of sea ice, and vice versa. The periods when the hindcast temperatures for years 2–5 are lowest correspond to times when the observationally based SPG ocean temperature anomalies used in the initialization of the hindcast ensembles are negative. This can be seen in Fig. 8 where the observed temperatures (albeit for March) show the multidecadal variability known to be a feature of SPG temperatures in all seasons (Robson et al. 2012). Although relatively small (<1°C), in the anomaly initialization method used by DePreSys 2 these anomalies are added to the cold-biased model climatology, promoting even greater sea ice growth than is seen in the uninitialized simulations. Reflection of solar radiation further enhances the cycle of sea ice growth and cooling, eventually leading to much larger cold anomalies in the initialized ensembles. On the other hand, warm anomalies in the initial conditions of the ensembles diminish sea ice and drive the sea ice–temperature feedback to produce further warming. The size of the difference in temperature compared to the initialized simulations is smaller than in the cold case (see the difference between transient and DePreSys 2 temperatures after 2000). This is because a relatively small fraction of the SPG area has near-total sea ice coverage in the model climatology; reducing this further gives less scope, therefore, for substantial temperature changes. Overall, it can be seen that the model cold bias in the northern North Atlantic influences the low skill in this region, through unrealistic sea ice feedbacks that degrade the skill of surface temperature signals.

c. Skill of detrended predictions

In the previous section, we showed that spurious trends related to model errors can cause degradation in our metrics of skill. In decadal prediction, however, our focus is more on the prediction of variability on shorter time scales than the integrated change over the last 50 yr, which has been shown to be largely related to long-term changes in external forcings (Stott 2003). In contrast, the potential strength of decadal prediction systems is capturing internal fluctuations mainly through initializing the ocean state. As a result, it is useful to examine the skill of systems once linear trends over the hindcast period have been removed (detrending) from the hindcasts and observations.

The ACC of detrended year 2–5 average near-surface temperature is shown in Fig. 10. Skill in both systems is considerably smaller (cf. Fig. 2) as trends act to inflate
correlations. Nevertheless, correlations over many regions, such as the North Atlantic, western Pacific, and parts of the Arctic, North Africa, and South Asia are significant and relatively large. The effect on the difference between systems is, to first order, to “stretch” the range of correlations rather than change the pattern. That is to say, where DePreSys 2 improved (deteriorated) from DePreSys 1, this improvement (deterioration) appears larger. Detrending emphasizes the improvement from representation of tropical Pacific processes (at least in the central and west Pacific).

SRMSE improvements in DePreSys 2 (over South America, the South Atlantic, South Africa, India, and the western Pacific) also become stronger and more significant. Importantly, the high-latitude regions of poor SRMSE skill seen in Fig. 2 are greatly reduced in extent and intensity, confirming the finding in section 3c that erroneous trends make a large contribution to these errors. These regions are not eliminated, however, suggesting that part of the erroneous variability is not captured by a linear trend. For the SPG region, this can be seen in Fig. 8 as the less cold temperatures in the mid-1960s, for example.

Changes in precipitation skill (Fig. 11) as a result of detrending are not as marked as for temperature as precipitation does not generally show such strong trends. SRMSE is improved in both systems, particularly over tropical West Africa. The low skill (high SRMSE) seen in central Asian rainfall in DePreSys 2 is also improved by detrending, as can be seen in the difference between systems, which is much less over this region than before. This implies that incorrect rainfall trends are partly responsible for the original error. A potential source of this spurious trend is the trend in North Atlantic and Arctic temperatures discussed in section 3c, which may alter moisture availability in this region, although this has not been demonstrated. Global area-averaged RMSE for the detrended ensemble data (Fig. 12) shows that DePreSys 2 near-surface temperature is marginally more skillful than that for DePreSys 1. This is opposite to the result found with trends retained (Fig. 5), although the differences in the detrended case are not statistically significant. In the averages for years 2–4 and 2–5 a significant deficit in skill in DePreSys 2 is eliminated, resulting in no significant difference. This shows that when we control for long-term trend errors the impact of high-latitude errors in DePreSys 2 on overall skill is effectively mitigated. This result is in addition to the improvements in skill seen in many individual regions in DePreSys 2.

Detrended values of year 2–5 mean GMNST possess ACCs of 0.63 (DePreSys 1) and 0.69 (DePreSys 2). These correlations are considerably smaller than ACC values obtained before detrending (>0.95). Such high values demonstrate that ACC is a poor metric for discriminating between the systems in the presence of strong trends. The differences between the detrended ACCs might be expected to be a better measure of predictive ability. Nevertheless, we find that the apparent 0.06 improvement in ACC skill is not significant at the 90% level. RMSE values for detrended temperatures are very similar to those before detrending, namely 0.069 K for DePreSys 1 and 0.075 K for DePreSys 2. This difference is also not significant.
While some useful improvements in hindcast skill can be obtained by detrending, forecasts require the combination of both trend and variability to be meaningful. An option for removing forecast bias introduced by inaccurate trends in the prediction system is the method of Kharin et al. (2012), in which best-fit linear trends to hindcast variables are substituted by the corresponding trends from observations. While this is beneficial to the forecast, skill measures for the hindcast (e.g., Figs. 10–12) calculated after applying this procedure are not shown, since this method artificially improves hindcast skill by including (trend) information from the verifying data. These measures would therefore be an inflated estimate of forecast skill, corresponding to the case where historical trends continue unchanged through the forecast period. As a result, while DePreSys 2 will routinely use the Kharin et al. (2012) procedure in the production of forecasts, their skill is best gauged using the full (Figs. 2–5) and detrended (Figs. 10–12) metrics included above.

**d. Impact of initialization**

Initialization of climate models with observationally based data has been demonstrated to improve the skill of decadal time-scale climate predictions beyond the level obtained using boundary forcing information alone (Kirtman et al. 2014). Here, we examine the improvement in skill provided by initialization in DePreSys 2, using the three HadGEM3 transient simulations (see

![Fig. 11. As in Fig. 3, but for detrended precipitation.](image)

![Fig. 12. As in Fig. 5 but RMSE for detrended fields.](image)
methods in section 2) as a baseline to compare with the initialized hindcast ensembles. To focus on the difference in skill in predicting decadal variability (as opposed to the long-term trends), we compute skill indices based on detrended data (see section 3c). The ACC of year 2–5 mean temperatures in the boundary-forced transient ensemble (Fig. 13a) shows significant skill over much of the North Atlantic Ocean, North Africa, South Asia, and parts of the western Pacific. In addition, there are regions of no or negative skill, particularly over the tropical and northern subtropical Pacific Ocean. These results are consistent with the finding that the Hadley Centre Global Environment Model, version 2–Earth System (HadGEM2-ES), which shares a very similar aerosol representation with the HadGEM3 model used here, can reproduce a large fraction of the observed variability in twentieth-century North Atlantic SSTs through indirect aerosol effects (Booth et al. 2012). It also highlights that the ENSO region is unpredictable on these time scales based on external forcings alone. The ACC for the initialized hindcast ensembles (Fig. 13b) shows similar patterns to those in the uninitialized transient ensemble over the Atlantic–African–Asian sectors, but a rather different pattern over the Pacific. The difference in ACC (Fig. 13c) highlights this, with a large increase in skill over the tropical and northern subtropical Pacific. Partly, this offsets negative ACC values, but it also introduces moderate skill over a much wider area of the central western Pacific. The ACC difference also reveals more widespread improvement as a result of initialization, including in the South Atlantic and Indian Oceans. In the North Atlantic Ocean there are improvements in the tropical and northeastern parts of the basin, but deterioration in the northwestern part. The problems arising from model biases in this region (see section 3b) are likely to have hampered the skill in the initialized hindcasts. It is notable that the North Atlantic region as a whole does not appear to be a major center for improved decadal skill in DePreSys 2 as it does in other systems, including DePreSys 1 (Smith et al. 2010). We note the fact that high levels of skill already exist in the uninitialized ensemble due to aerosol effects, and as such initialization has less scope to make an impact. A similar comparison of skill between the initialized and uninitialized cases is made for SRMSE (Figs. 13d–f). Overall, there is a fairly close correspondence between the effect of initialization on ACC and SRMSE. The results also highlight the findings of section 3c, namely that initialization acts to exacerbate high-latitude errors. In the case of the subpolar gyre region of the North Atlantic, this was found to be due to nonlinearities in the response of sea ice. It is likely that this issue also affects other sea ice regions, explaining in part why initialization is of widespread detriment to SRMSE skill in high latitudes.

4. Conclusions

In this paper, we have demonstrated that DePreSys 2, a new version of the Met Office Hadley Centre decadal
prediction system (DePreSys), provides skilful forecasts of climate for several years ahead. DePreSys 2 uses HadGEM3, the current Hadley Centre climate model, in preference to the HadCM3 model that is used in DePreSys 1. HadGEM3 benefits from many improvements in climate modeling, including improved grid resolution in the ocean and atmosphere. Like DePreSys 1, DePreSys 2 uses an anomaly initialization method, so simulations to establish the climatology of the HadGEM3 model and to assimilate observed oceanic and atmospheric anomalies relative this model climatology were performed. These produced initial conditions for a large set of reforecasts of climate over the last 50 yr. Comparing this set of reforecasts with the DePreSys 1 set submitted to the CMIP5 archive shows that overall, ACC skill of DePreSys 2 significantly exceeds that in DePreSys 1. The RMSE skill is similar or marginally worse, however, in DePreSys 2 compared to DePreSys 1. Analysis of regional SRMSE shows that the improvements in DePreSys 2 tend to occur at low to midlatitudes, whereas mid to high latitudes deteriorate.

The origin of the changes in skill between the systems is likely due to many differences between the underlying climate models, but given the particular relevance of tropical Pacific variability to the climate of the tropics, we examined the relative performance of the systems in predicting the Niño-4 index and its teleconnections. We find that the DePreSys 2 system has significant predictive skill for this index for about 18 months, much longer than seasonal time scales, which is consistent with findings using other coupled prediction systems (Luo et al. 2008). DePreSys 2 is found to be significantly more skilful in predicting Niño-4 than DePreSys 1. In addition, we show that the teleconnections of the Niño-4 index with other parts of the world are more accurate in the HadGEM3 model used by DePreSys 2. This combination strongly suggests that the tropical Pacific is a major factor in explaining the improvements in skill between the two systems in years 1 and 2 of the forecast. Further improvements may have arisen from better representation of local processes in DePreSys 2 compared to DePreSys 1. For example, it might be imagined that the much better vertical resolution in the ocean in HadGEM3 compared to HadCM3 could improve the treatment of ocean mixed layers and hence ocean heat content anomalies initialized in the DePreSys 2. As highlighted above, however, the underlying climate models differ in many ways, so it would require detailed further investigation to discover whether this or any other model difference actually had a positive impact on skill.

Increased high-latitude errors in DePreSys 2 were found to be a result of spurious long-term warming trends in the DePreSys 2 hindcasts. Specific comparison of the performance of each system in the Northern Hemisphere revealed that trends in this region are associated with a cold bias in the HadGEM3 model that results in excessive sea ice in the North Atlantic Ocean. The anomaly initialization approach preserves these biases in the initial conditions of the forecast and increases the sensitivity to initial anomalies via enhanced ice–albedo feedback. As climate warms over the 1960–2008 period, large shrinkage of the ice area occurs, introducing unrealistically large warming trends. The effect of model biases near the sea ice edge likely has an impact on wider high-latitude errors in DePreSys 2, but other potential model errors, such as in the amount of polar warming over the hindcast period, could also be an important factor. Furthermore, the simple linear relaxation of sea ice concentration to observed values used in initializing the hindcasts does not guarantee accurate initial sea ice thicknesses (Lindsay and Zhang 2006), which could potentially contribute to the high-latitude surface temperature errors.

The effect of long-term trend errors in the prediction system is to shift the focus away from prediction of decadal time-scale climate features. To compensate for this, the hindcast results were detrended and skill metrics recomputed. In this framework, the improvement in skill in DePreSys 2 compared to DePreSys 1 is more marked, with less impact from those regions that were originally less skilful in DePreSys 2. The increase in global average RMSE between DePreSys 1 and DePreSys 2 is essentially eliminated in the detrended analysis. This shows that the negative effect of regions with trend errors can effectively be mitigated. The improvement in high-latitude temperature skill as a result of detrending might be expected given model temperature biases, but the improvement of the relatively poor precipitation skill over Asia in DePreSys 2 is more surprising. We speculate that this change may be linked to the improvements in high-latitude temperature skill through evaporation and transport of moisture. The detrended analysis also shows that improved ACC for temperature beyond year 2 is a result of more accurate long-term trends overall, despite the specific issues in the high latitudes discussed above. No statistically significant differences in skill between the systems were found for prediction of global mean temperatures, either before or after detrending.

Comparison of skill measures between the uninitialized transient ensemble and the initialized hindcast ensembles confirms previous findings that initialization has a positive impact on overall decadal prediction skill. Widespread benefits are found in DePreSys 2, most prominently in the tropical and subtropical west Pacific Ocean. Some regions, however, show less skill in the
initialized ensembles, particularly those in high latitudes. We note once again the issues found with high-latitude model errors and the nonlinearity of the sea ice response to initialization. Interestingly, the North Atlantic Ocean is not found to be a center for additional predictability from initialization as it has been in previous studies. This is likely to be a result of the relatively high level of predictability in North Atlantic temperatures in the uninitialized transient simulations, which arises through the response of Atlantic SSTs to temporal variations in anthropogenic aerosol forcing. If this forced response is correct, then it suggests that previous indications of enhanced skill by initialization of the North Atlantic are misleading, since in this case initialization merely acts to compensate for errors in forcing. Alternatively, if the forcing is too large, it may conceal the benefit of initialization through its ability to (spuriously) reproduce historical North Atlantic temperatures. Other processes unrelated to forcings (e.g., ocean circulation) may in reality have contributed to generating North Atlantic decadal temperature anomalies, implying that decadal hindcasts should benefit from initialization. Even if forcings can account for much of the past surface variability, they may not be able to account for subsurface ocean variability and its potential effect on climate in coming decades.

The Atlantic meridional overturning circulation (AMOC) is often considered important in decadal prediction studies (e.g., Pohlmann et al. 2013). We have not performed an analysis of its predictability in DePreSys2 because of the issues we have found with the HadGEM3 Northern North Atlantic cold bias. Nevertheless, we note that the AMOC and the interior of the North Atlantic Ocean do not particularly appear to be compromised. It is still possible, of course, that the cold bias may have prevented additional skill arising from initializing AMOC variations. A cold North Atlantic bias is common in climate models (Flato et al. 2014) and is related to errors in the simulation of the path of the Gulf Stream (Keeley et al. 2012). Model improvements are needed to deal with this: development of a version of HadGEM3 with even higher ocean and atmospheric grid resolution than used here has removed this problem to first order. It will take time, however, for such improvements to feed into the development of the decadal prediction system. Regardless of the North Atlantic bias, DePreSys 2 offers improved decadal prediction skill compared to DePreSys 1 because it is better able to exploit other sources of skill. This shows the importance of a global outlook, especially since most predictability in decadal systems is for years rather than decades ahead. Furthermore, the ability of detrending to improve some aspects of the DePreSys 2 hindcast illustrates that better interpretation of the information produced by decadal prediction systems could help to produce better decadal predictions.

**Acknowledgments.** This work was supported by a project funded by UK aid from the UK Department for International Development (DFID) for the benefit of developing countries. The views expressed are not necessarily those of DFID. Additional support was provided by the joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). We also acknowledge the input of three anonymous reviewers whose comments and suggestions led to numerous improvements in this article.

**REFERENCES**


