Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: Impact of ocean observations

F. J. Doblas-Reyes,[1,2,3] M. A. Balmaseda,[3] A. Weisheimer,[3,4] and T. N. Palmer[3,4]

Received 29 November 2010; revised 16 July 2011; accepted 24 July 2011; published 12 October 2011.

[1] Three 10 year ensemble decadal forecast experiments have been performed with the European Centre for Medium-Range Weather Forecasts coupled forecast system using an initialization strategy common in seasonal forecasting with realistic initial conditions. One experiment initializes the ocean in a standard way using an ocean-only simulation forced with an atmospheric reanalysis and with strong relaxation to observed sea surface temperatures. The other two experiments initialize the ocean from a similar ocean-only run that, in addition, assimilates subsurface observations. This is the first time that these experiments were performed. The system drifts from the realistic initial conditions toward the model climate, the drift being of the same order as, if not larger than, the interannual signal. There are small drift differences in the three experiments that reflect mainly the influence of dynamical ocean processes in controlling the adjustment between the initialized state and the model climate in the extratropics. In spite of the drift, the predictions show that the system is able to skillfully predict some of the interannual variability of the global and regional air and ocean temperature. No significant forecast quality benefit of the assimilation of ocean observations is found over the extratropics, although a negative impact of the assimilation of incorrect expendable bathythermograph profiles has been found for the global mean upper ocean heat content and the Atlantic multidecadal oscillation. The results illustrate the importance of reducing the important model drift and the ocean analysis uncertainty.

Citation: Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer (2011), Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: Impact of ocean observations, *J. Geophys. Res.*, *116*, D19111, doi:10.1029/2010JD015394.

1. Introduction

[2] Climate change projections and near-term climate prediction (also known as decadal prediction) attempt to satisfy a growing demand for climate information for this century. It is well established that, on the basis of knowledge of the initial conditions, important aspects of regional climate are partially predictable up to a year ahead. Predictability at this time scale is primarily, though not solely, associated with the El Niño Southern Oscillation (ENSO). While climate forecasting is currently addressing the problem of climate prediction up to one year [e.g., *Doblas-Reyes et al.*, 2009; *Weisheimer et al.*, 2009], decadal prediction focuses on time scales of several years to a few decades [e.g., *Smith et al.*, 2007; *Meehl et al.*, 2009].

[3] There have been attempts to predict interannual-to-decadal climate variations using empirical models that take into account changes in boundary conditions, i.e., atmospheric composition and solar irradiance, as well as internal variability [e.g., *Lean and Rind*, 2009; *Hawkins et al.*, 2011]. Others [e.g., *Räisänen and Roukolainen*, 2006; *Roukolainen and Räisänen*, 2007; *Laepple et al.*, 2008] have employed the radiatively forced climate projections performed as part of the Third Coupled Model Intercomparison Project (CMIP3) [*Meehl et al.*, 2007], from where the part of the simulations corresponding to the first few years of the twenty-first century were used to issue climate predictions for the near term. Near-term climate will be considered in this paper as the 10–30 year period counting from a reference time, which in a forecast will correspond to the start of the prediction. As a slightly more ambitious alternative, dynamical decadal prediction explores the ability of the type of climate model employed in the Intergovernmental Panel on Climate Change (IPCC) assessments to predict regional climate changes in the near future by exploiting both initial-condition information and changes in the radiative forcing. This approach aims to take advantage of the predictability of natural climate variability to make predictions.

[1]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.
[2]Institut Català de Ciències del Clima, Barcelona, Spain.
[3]European Centre for Medium-Range Weather Forecasts, Reading, UK.
[4]National Centre for Atmospheric Science, Atmospheric, Oceanic and Planetary Physics, Department of Physics, Oxford University, Oxford, UK.

0148-0227/11/2010JD015394

[4] The relative importance of the initial conditions in climate prediction is likely to vary with the time scale, but has been assumed to be a continuous function that decreases with forecast time, becoming negligible after several decades [*Hawkins and Sutton*, 2009b]. Ocean initial conditions are more relevant than variations in atmospheric composition in seasonal forecasting [*Doblas-Reyes et al.*, 2006], except perhaps after an explosive volcanic eruption, while atmospheric composition has primary importance after several decades [*Hawkins and Sutton*, 2009b]. For the time scales ranging between a few seasons to a couple of decades, previous work [*Smith et al.*, 2007, 2010; *Keenlyside et al.*, 2008; *Pohlmann et al.*, 2009; *Mochizuki et al.*, 2010] has shown evidence that the initial state of the climate system can influence climate forecasts a decade or more ahead. Hence, for any decadal prediction system, it is critical to determine to what measure initialization improves predictions beyond the level achievable by simulating the response to naturally and anthropogenically forced climate change alone. However, the question that is addressed in this manuscript is how the quality of uncertain initial conditions, in particular those of the ocean, impacts on forecast quality.

[5] The strategy followed for the initialization of decadal predictions has so far been different from that used in seasonal forecasting. For instance, *Smith et al.* [2007, 2010] and *Mochizuki et al.* [2010] used the so-called anomaly initialization method, where ocean observations are assimilated in the form of anomalies into the coupled model taking, or not, into account the error covariance of the coupled model. In the work of *Keenlyside et al.* [2008] only observed sea surface temperature (SST) anomaly information was used to initialize the coupled system, with no further restrictions in deeper ocean temperature or salinity. In seasonal forecasting, however, it is common practice to separately initialize the ocean and the atmosphere, data assimilation being used to bring the state of each component of the coupled model close to the observed state. This approach is discussed by G. J. van Oldenborgh et al. (Decadal prediction skill in a multi-model ensemble, submitted to *Climate Dynamics*, 2011) with a multimodel experiment. *Balmaseda and Anderson* [2009] showed that for seasonal forecasting this strategy works better than an approach equivalent to that used by *Keenlyside et al.* [2008].

[6] In this study, we use the European Centre for Medium-Range Weather Forecasts (ECMWF) coupled system to investigate how the initialization strategy used in seasonal forecasting behaves in decadal forecasting. The assessment also deals with the impact of the quality of different sets of ocean initial conditions on forecast quality. Similar studies, with a different experimental design, have been undertaken by *Yasunaka et al.* [2011] and J. Krüger et al. (Impact of different ocean reanalyses on decadal climate prediction, submitted to *Climate Dynamics*, 2011). The criteria for the assessment is the extent to which atmospheric and ocean variables are skillfully predicted in the forecast range from 1 to 10 years, i.e., beyond the generally accepted limit of ENSO-related predictability. Problems related with sample and ensemble sizes in the context of the protocol set up in the ongoing Coupled Model Intercomparison Experiment, known as CMIP5 (http://www.clivar.org/organization/wgcm/references/Taylor_CMIP5.pdf), for initialized dynamical decadal predictions are also discussed.

[7] A summary of the experiment follows in section 2. The most relevant characteristics in terms of model drift and forecast quality results are given in sections 3 and 4. The main conclusions are summarized in section 5.

2. Description of the Experiment

2.1. Experimental Setup

[8] To address the key uncertainties at the source of decadal forecast error, such as uncertainties in the initial conditions and in model formulation [*Palmer*, 2000; *Anderson et al.*, 2009], ensemble methods have been proposed. They involve not only running a single model several times with slightly different initial conditions, but also employing multimodel or perturbed-parameter approaches [*Doblas-Reyes et al.*, 2009]. In this paper, three sets of single-model ensemble reforecasts have been carried out with the IFS/HOPE coupled system. The forecast system [*Anderson et al.*, 2007] was based on the atmospheric IFS cycle 35r3 [*Bechtold et al.*, 2008] with a horizontal truncation of T_L159 and 62 vertical levels extending up to 5 hPa. IFS uses a climatological annual cycle of four types of aerosol (sea salt, desert dust, organic matter, black carbon) in a scheme where only the direct aerosol effect is included. The system includes the interannual evolution of global mean annual greenhouse trace gases (CO_2, CH_4, N_2O and CFCs) and specified anthropogenic sulfate aerosols, as well as interannual variations of total solar irradiance. Carbonaceous aerosols are not included. Information on the volcanic aerosol load prior to the start date is not used during the hindcasts, in contrast with what is done in other decadal forecast systems [*Smith et al.*, 2007, 2010; *Keenlyside et al.*, 2008]. The ocean model has a horizontal resolution of 1°, with an equatorial refinement of 0.3°, and 29 levels in the vertical. There is no sea-ice module and the initial sea-ice extent is relaxed to climatological values with an e-folding time of one month. The coupler OASIS2 is used to interpolate the fields exchanged once per day between the ocean and atmospheric grids. No relaxation or flux correction was active during the forecast.

[9] The three-member ensemble reforecasts were started once every 5 years over the period 1960 to 2005, i.e., in 1960, 1965, and so on. This experimental setup is based on the decadal reforecast experiment of the ENSEMBLES (http://www.ecmwf.int/research/EU_projects/ENSEMBLES/exp_setup/stream2.html) project, which is similar to the core CMIP5 decadal experiment, although none of the experiments contributed to the ENSEMBLES multimodel (van Oldenborgh et al., submitted manuscript, 2011). The atmosphere and land surface initial conditions were taken from the ERA-40 reanalysis [*Uppala et al.*, 2005] for all start dates but for 2005, for which the operational ECMWF analyses were used. Each simulation started at 00:00 geomagnetic time on 1 November of each year and ran for 120 months. The ensemble was generated by introducing singular vector perturbations to the atmospheric initial conditions.

[10] The baseline experiment uses ocean initial conditions from the ORA-S3 ocean reanalysis [*Balmaseda et al.*,

2008], from which only one of the five reanalyses available has been used. All available observations of temperature and salinity, as well as altimetric sea level anomalies, have been used in this reanalysis. The atmospheric fluxes are from the ERA-40 reanalysis for the period January 1959 to June 2002 and ECMWF operational analysis thereafter. The SST is strongly relaxed to analyzed daily SST maps from the OIv2 SST [*Reynolds et al.*, 2002] product from 1982 onward. This experiment will be referred to henceforth as Assim. An alternative initialization consists in using data from an ocean simulation forced with the ECMWF atmospheric fluxes and the SST relaxation, but with no ocean data assimilated. This second experiment will be named NoOcObs. As the ocean model attractor inevitably differs from the attractor of the actual climate system, the lack of observed ocean data in NoOcObs will produce ocean initial states closer to the ocean model attractor, unlike in the case of the Assim experiment. Hence, in absence of error in the atmospheric model, the NoOcObs method would produce balanced ocean initial conditions. A third experiment has been performed using a reanalysis similar to ORA-S3, but where corrected XBT (expendable bathythermograph) profiles according to *Wijffels et al.* [2008] and *Ishii and Kimoto* [2009] have been assimilated. The third experiment will be referred to as XBT-C.

2.2. Computation of the Anomalies

[11] Various measures of forecast quality have been used to assess the differences between the experiments. The scores include the anomaly correlation coefficient and RMSE of the ensemble mean. The forecast quality measures have used different data sets. To verify near-surface temperature, a merged data set using land air temperatures from the GHCN/CAMS data set [*Fan and van den Dool*, 2008] and SST from the NCDC ERSST V3b data set [*Smith and Reynolds*, 2003], while outside the band between 60°N and 60°S the GISSTEMP data set with 1200 km decorrelation scale was used [*Hansen et al.*, 2010]. GPCP [*Adler et al.*, 2003] was taken as the reference for precipitation. For the ocean variables, the reanalysis performed with the corrected XBT profiles (ORA-XC henceforth) has been used over the period 1960–2005.

[12] Every forecast quality measure has been computed taking into account the systematic error of the forecast systems. Forecast anomalies have been estimated in cross-validation mode by removing the mean model climate for the specific forecast period using the reforecasts for which there are reference data available, as it is commonly done in seasonal forecasting. For instance, to obtain the anomalies of the average 6–10 year forecast period from the reforecast initialized in November 1970, the model climate is estimated by averaging the data for the 6–10 year forecast period from all the reforecasts for which there are reference data, except the reforecast started in November 1970. This implies that data from the 1960, 1965, 1975, 1980, 1985, 1990, 1995 and 2000 reforecasts (eight start dates) are used, because no reference data for the period 2011–2015 (i.e., the verifying dates of the reforecasts with start date in 2005) were available. The anomalies for the reference data set are estimated for the same calendar period. Figures 1a and 1b illustrate the process of a posteriori removal of the drift, an estimate of which appears in Figure 1c for global mean near-surface air temperature. The raw reforecast values appear in Figure 1a, while the anomalies resulting after the drift estimate has been removed appear in Figure 1b. The reader should be aware that this linear method assumes that there is no relationship between the model drift and the anomalies.

[13] This method of computing the anomalies is different from the approach that would be adopted in an operational context, where the anomalies would be computed using only past information. However, the shortness of the sample, with just ten reforecasts available, prevents the authors from using a more robust computation of the anomalies. In other decadal prediction experiments prediction anomalies have been estimated using a model climate estimate from a set of simulations of the twentieth century climate that do not assimilate observations [e.g., *Smith et al.*, 2007; *Doblas-Reyes et al.*, 2010a]. This is not possible here because (1) there is no twentieth-century control simulation available for our forecast system, and (2) even if a control simulation was available, as the coupled model is initialized from realistic initial conditions the model climate in the reforecasts depends on the forecast time because of the unavoidable drift, making a unique model climate estimate from a long simulation inappropriate.

3. Mean State and Model Drift

[14] Model inadequacy causes forecasts to drift away from the observed climate toward an imperfect-model climate. This drift, which can be understood as the evolving systematic error, depends on the forecast time, especially when using the full-initialization approach. Previous publications on decadal forecasting [*Smith et al.*, 2007; *Keenlyside et al.*, 2008; *Hawkins and Sutton*, 2009a] rarely illustrate the differences between the model and reference climates, probably because they used extensively referenced climate models. Sometimes model drift is not described in detail because the initialization of the reforecasts is carried out by assimilating observed anomalies into the model climate [*Smith et al.*, 2007, 2010; *Keenlyside et al.*, 2008; *Pohlmann et al.*, 2009; *Mochizuki et al.*, 2010], a method that is expected to reduce the model drift. These approaches rely on the idea that the drift is small enough to not destroy the initial-condition information. Here we consider that forecast drift is an important feature of the decadal forecasting problem when a full-state initialization is adopted and, hence, worth discussing.

[15] Figure 1 shows the global mean near-surface air temperature. The model and observed temperatures deviate from one another quickly in every reforecast (Figure 1a) and, regardless of the initialization data set, the mean error reaches around 1 K after 4 years for global mean temperature (Figure 1c), while the global mean SST error is around −0.4 K after a similar forecast time (not shown). The forecast drift in the first few years is slightly different for the Assim/XBT-C and NoOcObs experiments, showing a sharper decline in the Assim and XBT-C experiments during the first 2 years. After the third forecast year the drift differences due to the initial conditions almost completely vanish (Figure 1c). In terms of global mean land temperatures, Assim and XBT-C are slightly warmer than NoOcObs, although both experiments have a cold bias (not shown).

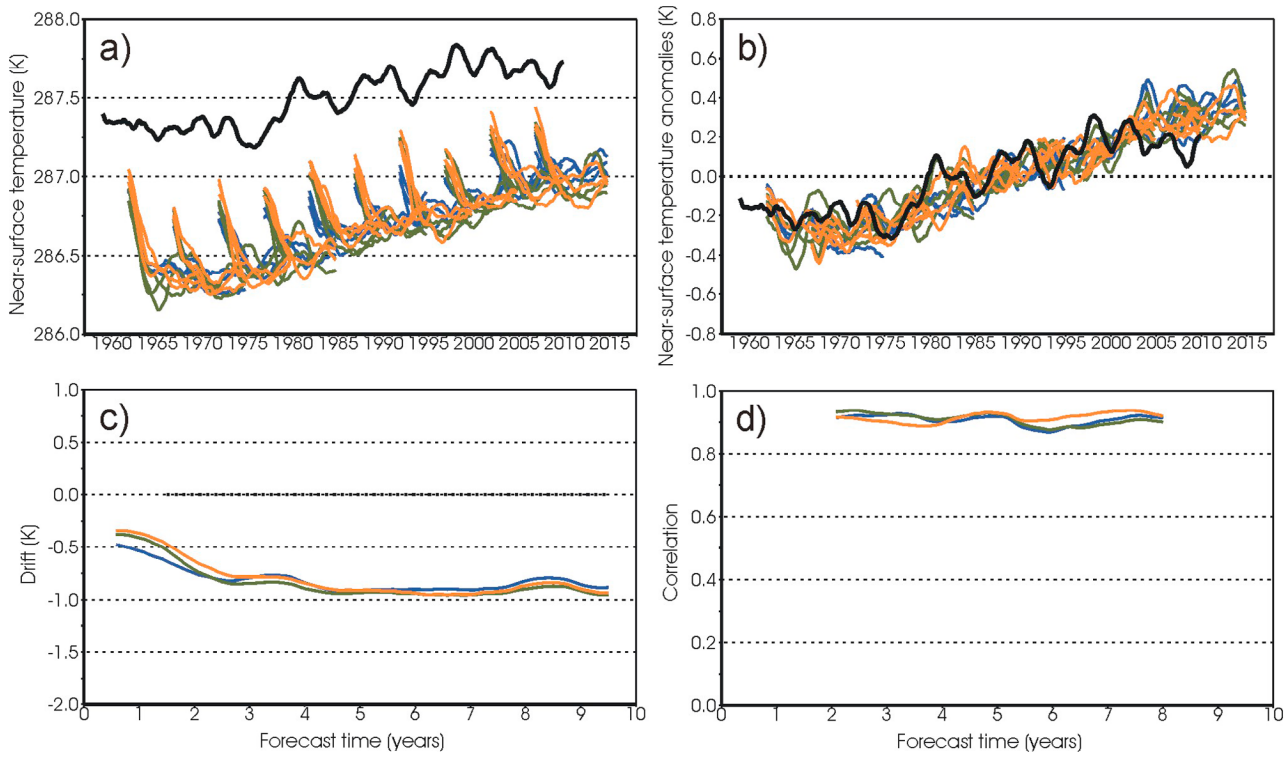

Figure 1. (a) Global mean near-surface air temperature (K) for the 10 three-member ensemble reforecasts of the XBT-C (green), NoOcObs (blue), and Assim (orange) experiments. (b) Anomalies with respect to the corresponding climate over the period 1960–2010. Data from GHCN/ERSST/GISS data set (see text for details) are shown in black. All time series have been smoothed out with a 24 month centered moving average that removes data for the first and last years of each time series. (c) Drift (K) and (d) ensemble mean correlation of the global mean near-surface air temperature for the XBT-C (green), NoOcObs (blue), and Assim (orange) experiments. The drift and the correlation have been computed using the GHCN/ERSST/GISS data set (see text for details) and three-member ensemble reforecasts for the period 1960–2000. A 12 month moving average has been applied to the drift estimates to illustrate the fast growth rate of the drift, while the correlation has been computed with a moving window of 4 year averaged anomalies to retain the interannual variability that is beyond the ENSO typical frequency.

[16] As an example, Figure 2 displays the mean systematic error of the boreal winter (December to February) near-surface air temperature of the XBT-C experiment for the forecast period 2 to 5 years. The experiments have in general air temperatures cooler than the reference, in agreement with Figure 1. This is particularly obvious over the tropical regions, where the cooling is slightly alleviated in Assim (not shown). This pattern is also found in summer, except for a strong warm bias over the equatorial eastern Atlantic, which is typical of both IPCC and seasonal forecast experiments (C. Caminade, personal communication, 2010). The differences in the mean systematic error between the three experiments (not shown) are much smaller than the systematic error itself, which suggests that the type of initialization has a small impact on the drift reduction. A broad similarity has also been observed for estimates of the interannual standard deviation of both experiments (not shown).

[17] Figure 2 also shows the mean systematic error of the winter (December to February) ocean temperature averaged over the top 300 m (a proxy for the upper-ocean heat content) for the forecast period 2 to 5 years. The mean error has been computed with respect to the ORA-XC ocean reanalysis. The systematic error in the ocean temperature bears some similarity with the pattern found for near-surface air temperature, although there are some differences: the tropical cooling is not as widespread and the western equatorial Pacific is actually warmer than in the reanalysis resembling a La Niña pattern. There are also differences in the western boundary currents, which are warmer in the coupled model than in the verifying reanalysis with a pattern typical of the low-resolution systems used for climate change projections [*van Oldenborgh et al.*, 2009]. Overall, the large-scale patterns in the upper-ocean heat content suggest that there are errors in the ocean circulation and not just in the surface heat fluxes. As with the near-surface air temperature, the three experiments depict a similar degree of cooling with respect to the ocean reanalysis, with some nonsignificant local differences. Small differences in the drift between the experiments (at least one order of magnitude smaller than the drift itself) can also be found for other ocean variables.

[18] To better illustrate the differences between the experiments in the upper-ocean heat content drift, Figure 3a shows the drift of the global mean upper-ocean heat content. There is a cold drift in the global mean after the first forecast year in the three experiments. The XBT-C drift is toward a

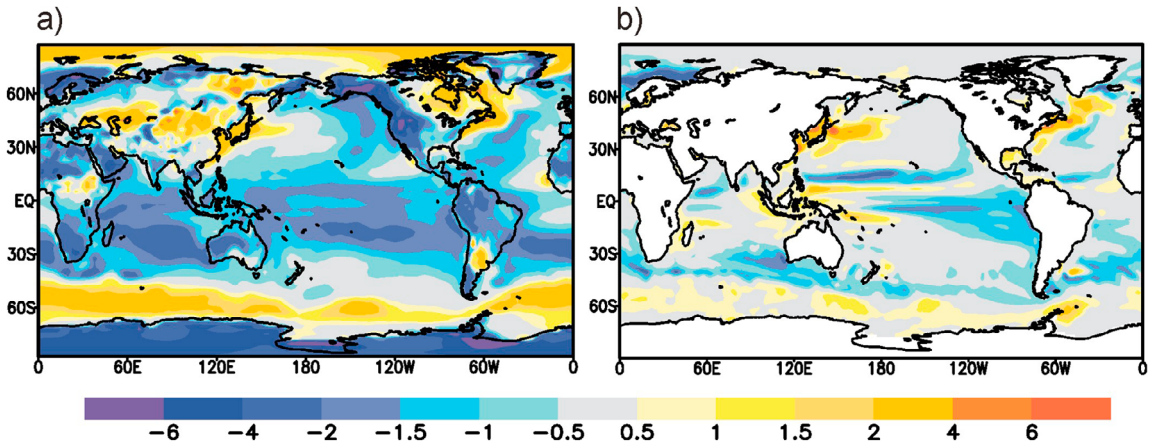

Figure 2. Winter (December to February) (a) near-surface air temperature (K) and (b) ocean temperature averaged over the top 300 m (K) mean systematic error over the forecast period 2 to 5 years of the XBT-C experiment. The systematic error has been estimated with respect to the GHCN/ERSST/GISS (Figure 2a) data set and ORA-XC (Figure 2b) (see text for details). Three-member ensemble reforecasts for the period 1960–2000 have been used.

slightly cooler state than in the other two experiments. The largest drift occurs over the Southern Hemisphere (Figure 3b), where the upper-ocean is warmer than the reanalysis in the first year, especially for NoOcObs. As the forecast time increases, the model ocean heat content is up to 0.2 K cooler than in the reanalysis, especially for Assim and XBT-C. The difference of those two experiments with NoOcObs (which is the warmest experiment in this region) is almost constant with forecast time.

[19] These results point at a strong latitudinal dependence of the upper-ocean heat content drift, which is also applicable to the associated SSTs. The drift differences between the experiments, although much smaller than the drift itself, also show a latitudinal dependence. While differences between NoOcObs and XBT-C only last for the first few forecast years in the tropical band (30°S–30°N), poleward of 30° the initial-condition differences are much more persistent and remain during the whole duration of the forecasts. The differences between NoOcObs and Assim grow with time in the tropical band and are very similar to differences between NoOcObs and XBT-C outside of the tropics. In the extratropics the differences are the result of the initial differences, although these drift differences show dynamical properties, with propagation both zonally and meridionally. The SST drift differences have an impact on the atmosphere mainly via changes in the surface latent heat flux and the surface shortwave radiation. An in-depth description of the drift differences can be found in the work of *Doblas-Reyes et al.* [2010b].

[20] Another example of how much the ocean mean state can differ between experiments is illustrated in Figure 4, which shows the zonally integrated meridional velocity across the Atlantic basin at 36°N, a proxy for the intensity of the Atlantic meridional overturning circulation (AMOC). The AMOC in the NoOcObs initial conditions is shallower and weaker (~20 Sv) than in the verifying analysis XBT-C

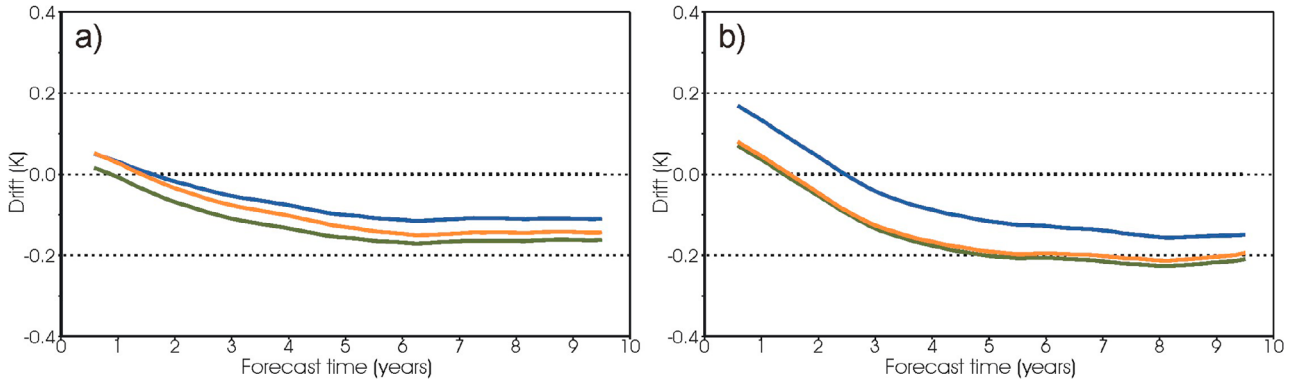

Figure 3. Drift (K) with respect to ORA-XC of (a) the global mean and (b) Southern Hemisphere (south of 20°S) ocean temperature averaged over the top 300 m for the three-member ensemble reforecasts of the (green) XBT-C, (blue) NoOcObs, and (orange) Assim experiments. Three-member ensemble reforecasts for the period 1960–1995 have been used. The time series have been smoothed out with a centered annual running mean that removes data for the first and last 6 months of each time series.

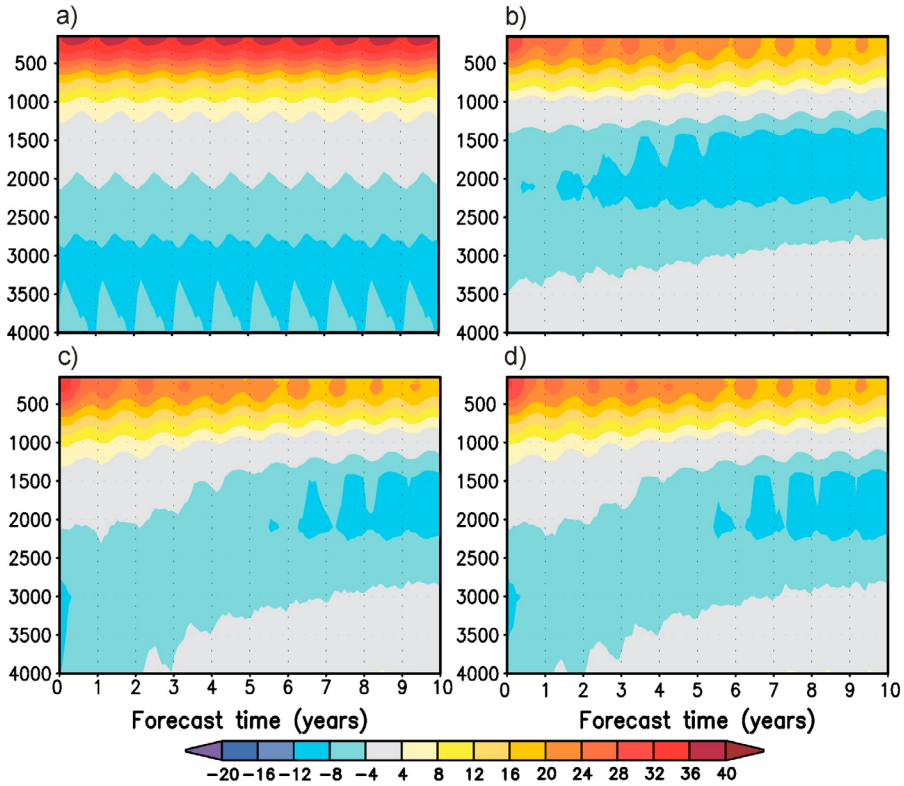

Figure 4. Zonally integrated meridional water velocity (10^3 m^2 s^{-1}) across the Atlantic basin at 36°N for the (a) ORA-XC ocean reanalysis, (b) NoOcObs, (c) XBT-C, and (d) Assim experiments as a function of depth and forecast time. The horizontal axis covers 120 months and represents the mean seasonal cycle of the reanalysis repeated ten times and the drift from the reforecasts. The vertical axis starts at 150 m and goes to 4000 m. All estimates have been computed using three-member ensemble reforecasts for the period 1960–1995.

(~24 Sv), in agreement with *Balmaseda et al.* [2007]. The AMOC in the ocean reanalyses with ocean data assimilation also show a stronger interannual variability (not shown). The reader should be aware that the differences in the AMOC characteristics between the three reanalyses mentioned here reflect the large uncertainty of these measures, for which observational evidence is scarce (Krüger et al., submitted manuscript, 2011). The Assim and XBT-C simulations experience a transition after the first 2 forecast years toward a shallower (by ~500 m) AMOC cell, weakening the northward branch and strengthening the southward one. Such a drift is slower in NoOcObs (not necessarily a positive feature in a full-state initialization context), for which the AMOC cell is already shallower than in the ORA-XC analysis at the beginning of the simulations. As for the atmospheric variables, the three experiments evolve toward a similar state at the end of the integrations. As a consequence of the shallowing of the meridional overturning cell and the reduction of its vertical gradient, the AMOC intensity, estimated as the maximum of the vertically integrated meridional transport (the integration is done from the surface to the bottom layer by layer) across the Atlantic at 36°N (which is the latitude where the maximum intensity occurs in the reanalyses), decreases to around 10 Sv. Other model results [e.g., *Drijfhout et al.*, 2008; *Lozier*, 2010] suggest that whereas the interannual variability in the ocean overturning is largely driven by surface winds, variability on decadal and longer time scales (and probably the drift too) is primarily driven by buoyancy fluxes. Contributions to the buoyancy fluxes comprise fresh water forcing by precipitation, evaporation, runoff and sea-ice melting or formation, and thermal forcing by turbulent fluxes (sensible and latent heat), radiative fluxes and the latent heat of fusion associated with formation or melting of sea ice. It is possible that errors in most of these processes, in particular the missing ones such as those related to sea-ice melting and formation of unresolved ocean eddies, are responsible for the drift of the AMOC in this model. However, it is difficult to assign the error to specific processes.

4. Forecast Quality

4.1. Atmospheric Variables

[21] A subset of the reforecasts has been used to estimate the forecast quality of the experiments. This is because there is not a complete reference data set available to compare to the 2005 start date simulation beyond 2010. In other words, at the time of writing there is no full verification available for the forecast period 6–10 years of the 2005 start date.

[22] The near-surface air temperature anomalies obtained using the verification data available to compute the model climate are shown in Figure 1b. No substantial differences between the reforecasts of the three experiments are found at first sight. Every experiment reproduces the upward trend

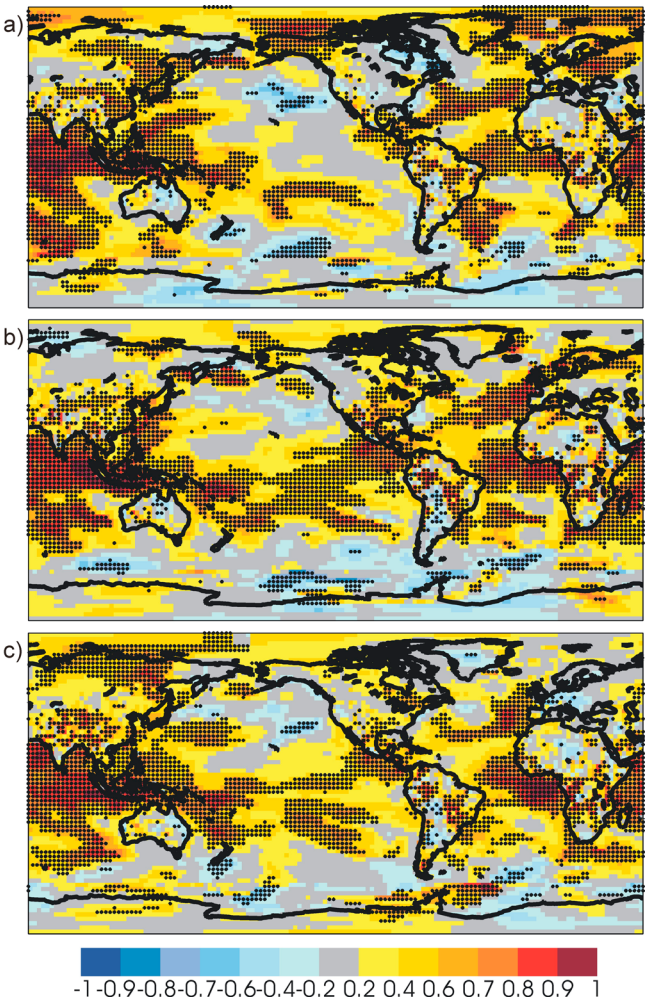

Figure 5. Ensemble mean correlation for near-surface air temperature with respect to the GHCN/ERSST/GISS data set (see text for details) for winter (December to February) over the forecast period 2 to 5 years of the (a) NoOcObs, (b) XBT-C, and (c) Assim experiments. Three-member ensemble reforecasts for the period 1960–2005 have been used. The black dots depict the grid points where the correlation is significantly different from zero with 95% confidence.

that is especially noticeable from 1975 onward. The ensemble mean correlation for 4 year mean predictions computed using a centered moving window is shown in Figure 1d, where no meaningful difference can be found between the three experiments. Note that only a reduced range of forecast times is available because a 4 year mean allows predictions for forecast periods ranging from months 25 to 96 (i.e., forecast time 3 to 8 years). In spite of the nonnegligible drift, the correlation of the anomalies computed linearly with respect to the drift estimates is high and statistically significant with 95% confidence. There are already examples in the literature [*Smith et al.*, 2007, 2010] that suggest that, to first order, this high skill is due to the projection of forced climate change rather than the impact of forecast initialization. An explicit separation of the initialized and forced component of the skill would require parallel uninitialized reforecast ensembles which, as explained in section 2, are not available for these experiments. Hence, the quantification of the relative contributions of the two effects is beyond the scope of this paper. Contrary to the conventional dependence of forecast skill with lead time, where skill decreases with lead time in short-range, medium-range, subseasonal, and seasonal prediction, the correlation does not decrease with forecast time. This feature, a likely consequence of the relevance of radiatively forced long-term trends, has also been observed [*Doblas-Reyes et al.*, 2010a] when the same calculation is carried out on the predictions described by *Smith et al.* [2010] and *Keenlyside et al.* [2008].

[23] Figure 5 shows the ensemble mean correlation for near-surface temperature computed for the 4 year winter averages covering the forecast period 2 to 5 years. Large areas with positive skill appear in all experiments, a consequence of both correctly projected climate change and forecast initialization. Assim and XBT-C have both higher skill than NoOcObs over the tropical Pacific and Atlantic, while NoOcObs gives better skill over the tropical Indian Ocean. Results for other seasons are similar. The different tropical skill is consistent with the results found in a similar seasonal reforecast experiment [*Balmaseda and Anderson*, 2009]. Instead, several extratropical regions such as Europe and the Arctic show NoOcObs as having slightly higher skill than the two other experiments. It is important to bear in mind that the differences in skill between the experiments should be considered in the context of the correlations being computed with very small samples and, hence, not statistically significant.

[24] A clearer picture of the differences in forecast quality between the three experiments for different forecast periods is shown in Figure 6, which depicts the anomaly correlation of near-surface air temperature over the Northern Hemisphere and the tropical band for different forecast periods. To obtain each anomaly correlation coefficient, the spatial variance/covariance between the ensemble mean and the corresponding reference is computed for each one of the reforecasts available. The set of variances and covariances is then averaged over the set of start dates before the final correlation is computed. Confidence intervals for the scores have been computed using a bootstrap method, where the reforecast/reference pairs were resampled with replacement 1000 times [*Lanzante*, 2005; *Jolliffe*, 2007]. The scores were then computed for each of the 1000 samples, ranked and the intervals for specific confidence levels estimated [*Doblas-Reyes et al.*, 2009]. The reader should bear in mind that these correlations are lower than those obtained for global mean variables due to the additional requirement of an adequate spatial distribution of the signal. Most cases display positive skill, with typically higher values for the tropical band than for the northern extratropics. Estimates for predictions of the first forecast year have been included to illustrate the impact of the time averaging. The scores for the tropical region are all statistically significantly different from zero, which is not always the case over the northern extratropics. The extratropical scores show a hint of seasonal variation of skill, with higher scores for boreal summer than for winter. Similar results have been found for the Southern Hemisphere (not shown). The skill is larger for longer forecast times, a likely consequence of the stronger impact of climate change for longer forecast times in this

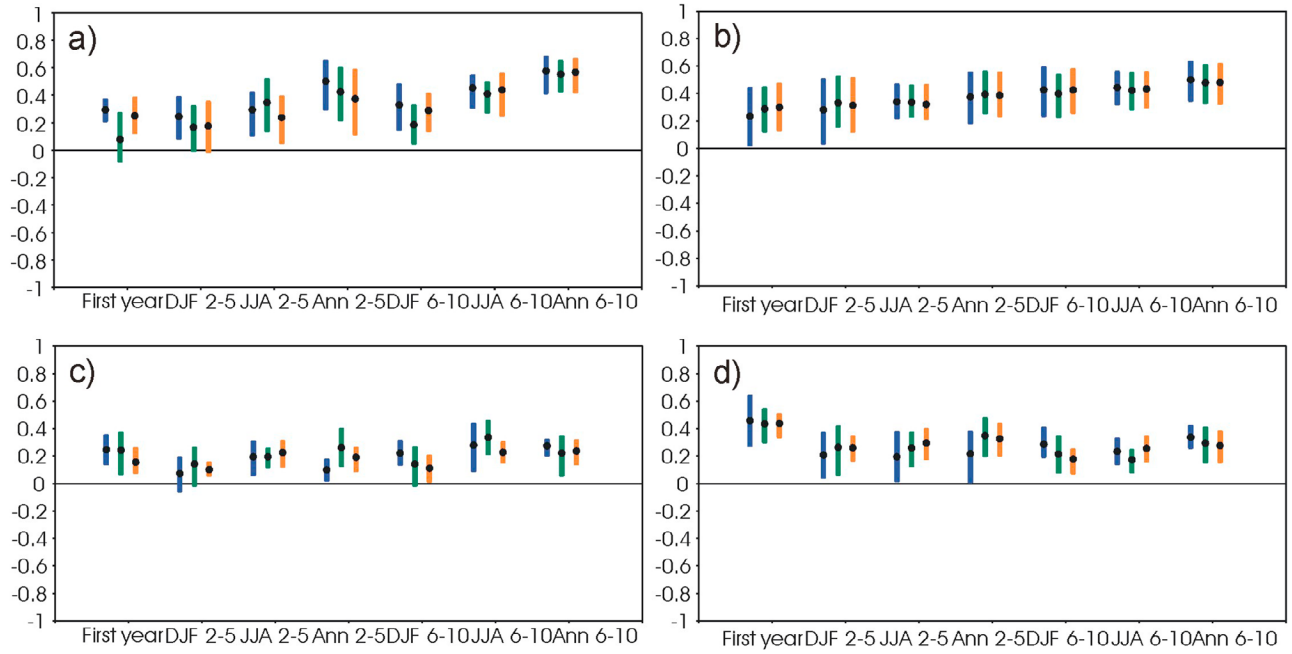

Figure 6. (a and b) Ensemble mean correlation and (c and d) perfect-model correlation for near-surface air temperature over the Northern Hemisphere (north of 30°N, Figures 6a and 6c) and the tropical band (20°N–20°S, Figures 6b and 6d) for different forecast periods of the NoOcObs (blue bars), XBT-C (green bars), and Assim (orange bars) experiments. The sets of three bars correspond, from left to right, to the first calendar year of forecast (months 3–14, computed with reforecasts for the period 1960–2005), the winter, summer, and annual mean of the 2–5 year forecast period (computed with reforecasts for the period 1960–2005), and the winter, summer, and annual mean of the 6–10 year forecast period (computed with reforecasts for the period 1960–2000). The black dots depict the sample values and the bars show the 95% confidence intervals. The estimates have been computed using the GHCN/ERSST/GISS data set (see text for details) and three-member ensemble reforecasts.

type of experiment. As in previous examples, the experiments display a similar forecast quality, with almost identical scores over the tropical band and small differences for the northern extratropics. Correlations for precipitation are much lower and in most cases nonsignificant.

[25] Figure 6 also contains information about the behavior of the ensemble. The ratio between the spread, computed as the standard deviation of the ensemble members around the ensemble mean, and the RMSE has traditionally been used as a measure of the degree of calibration of the ensemble [*Palmer et al.*, 2007]. However, the small ensemble size of these experiments prevents the spread estimates from being robust enough to obtain meaningful results. Instead, the so-called perfect-model anomaly correlation has been used. This estimator measures the spread relative to the variability of the predictions and should not be interpreted as an upper level of skill because it is model dependent. The perfect-model anomaly correlation is computed as the ensemble mean anomaly correlation mentioned above, but in this instance taking one ensemble member as the reference. In other words, this estimator assumes that the reference is drawn from the same population as the reforecasts (a hypothesis that is rarely true in an actual context), hence the use of the words "perfect model." The set of variances and covariances is computed taking each one of the ensemble members as reference in turns and then averaging the other two, prior to the computation of the correlation.

[26] The higher values found in Figure 6d for the first year suggest that the ensemble spread over the tropics is smaller at the beginning of the reforecasts, the spread increasing with time, something that in most single-model forecast systems is a desirable feature because of their tendency to underestimate the spread [*Weigel et al.*, 2008]. The decrease in perfect-model correlation from the first year is not found for the northern extratropics, where a slight seasonality is found with similar characteristics to that found for the correlation against the observations. Also for this measure, no substantial differences are found between the experiments. However, as mentioned above, the current experimental setup, which is shared with the one proposed for CMIP5, makes it difficult to make conclusive statements about the ensemble spread because of the small sample. Larger ensemble sizes will be needed to address the question of an appropriate ensemble generation that would take into account the specific characteristics of decadal forecasting.

[27] The perfect-model anomaly correlation has been used here as a measure of ensemble spread. It is sometimes also considered as a measure of the upper limit of the skill of a forecast system. This use is not appropriate because, apart from this interpretation being valid only in the case of unbiased models, the perfect-model anomaly correlation could only be considered as an upper estimate of the skill of an imperfect system, that unavoidably misses important processes to formulate skillful predictions, in a stationary

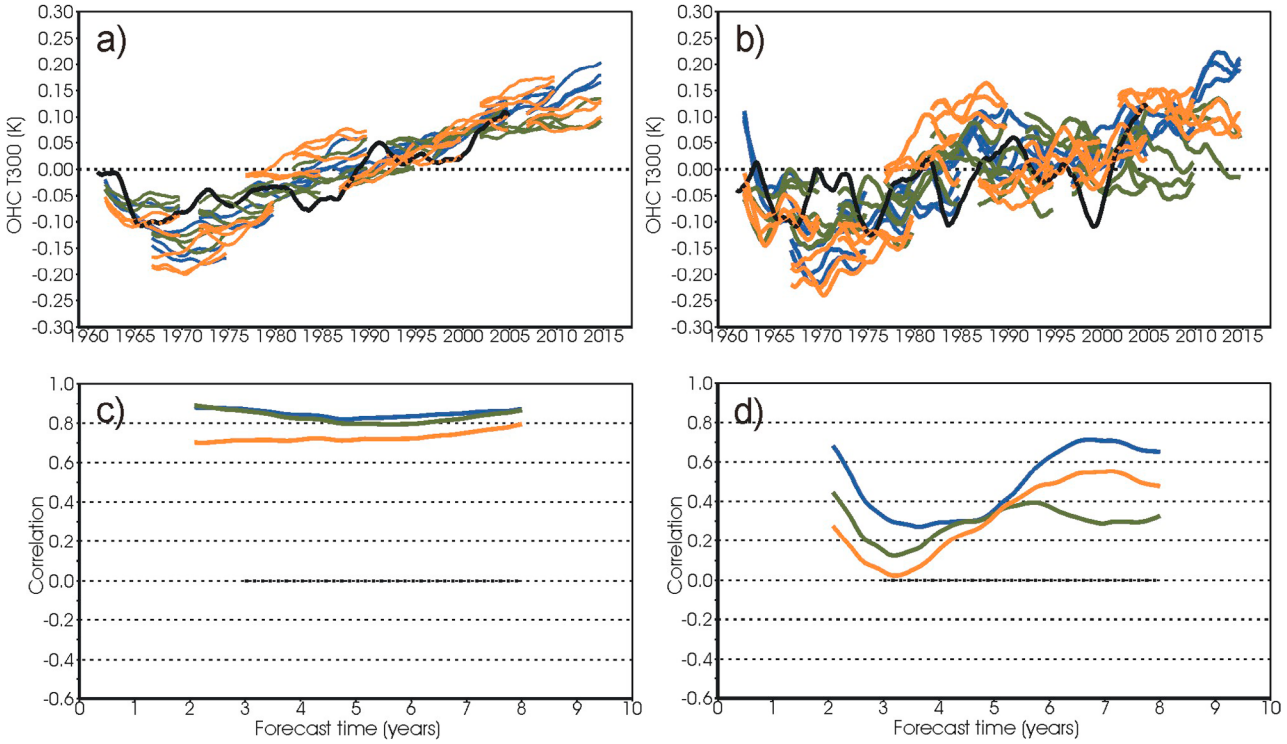

Figure 7. Anomalies of (a) global mean and (b) tropical (20°N–20°S) ocean temperature (K) averaged over the top 300 m for the 10 three-member ensemble reforecasts of the XBT-C (green), NoOcObs (blue), and Assim (orange) experiments. Anomalies are computed with respect to the corresponding climate over the period 1960–2005 (eight reforecasts). Each reforecast is illustrated with lines of a different color. Anomalies from the ORA-XC ocean reanalysis are shown in black solid lines. All time series have been smoothed out with a 24 month centered moving average that removes data for the first and last years of each time series. Also shown are the ensemble mean correlation of the (c) global mean and (d) tropical averaged ocean temperature averaged over the top 300 m of the XBT-C (green), NoOcObs (blue), and Assim (orange) experiments and have been computed using ORA-XC data and three-member ensemble reforecasts for the period 1960–1995. The correlation has been computed with a moving window of 4 year averaged anomalies.

climate. By construction, the perfect-model correlation cannot take into account the effect of the long-term trends on the skill, which is a substantial contribution to the forecast quality. As an indication, the reader will note that in most cases in Figure 6 the anomaly correlation estimates are larger than the corresponding perfect-model anomaly correlations suggesting that the initial-condition predictability in this system is modest and that the correlation against observations has an important contribution from the correctly reproduced long-term trends and low-frequency variability. This also suggests that some estimates of decadal predictability based solely on ensemble agreement measures [e.g., *Boer and Lambert*, 2008] underestimate the actual skill.

4.2. Ocean Variables

[28] Early results suggest that skillful projections of ocean heat content are one obvious mean by which initializing climate models may increase some aspects of decadal predictability. Figure 7 displays anomalies of the global mean upper-ocean heat content from the three experiments. Anomalies for the XBT-C ocean reanalyses are also displayed in Figure 7 as a reference. There are differences between the three ocean reanalyses used to initialize the experiments of up to 20% in the interannual variability, which are due to the large uncertainty in the estimates of the ocean state (not shown). The reanalysis shows an upward trend after 1970 that is matched by the reforecast anomalies. The main difference between the experiments is that the Assim predictions have a larger variability from one reforecast to the next in the early half of the period, to the point that they do not always overlap as the ensembles of the other two experiments do. This agrees with the caveat of the ORA-S3 reanalysis, which has been affected by the error in the dropping rate of the XBTs that caused spurious interannual variations in the ocean heat content estimates.

[29] The ensemble mean correlation of the global mean upper-ocean heat content with respect to ORA-XC is higher for NoOcObs and XBT-C than for the Assim experiment (Figure 7c), which illustrates the impact of the observational error in decadal forecast initialization in agreement with *Yasunaka et al.* [2011]. The low correlation for Assim is in contrast with what has been found for the global mean near-surface air temperature (Figures 1 and 6) and SST (not shown). The differences between the experiments are also clear in the tropical upper-ocean heat content reforecasts

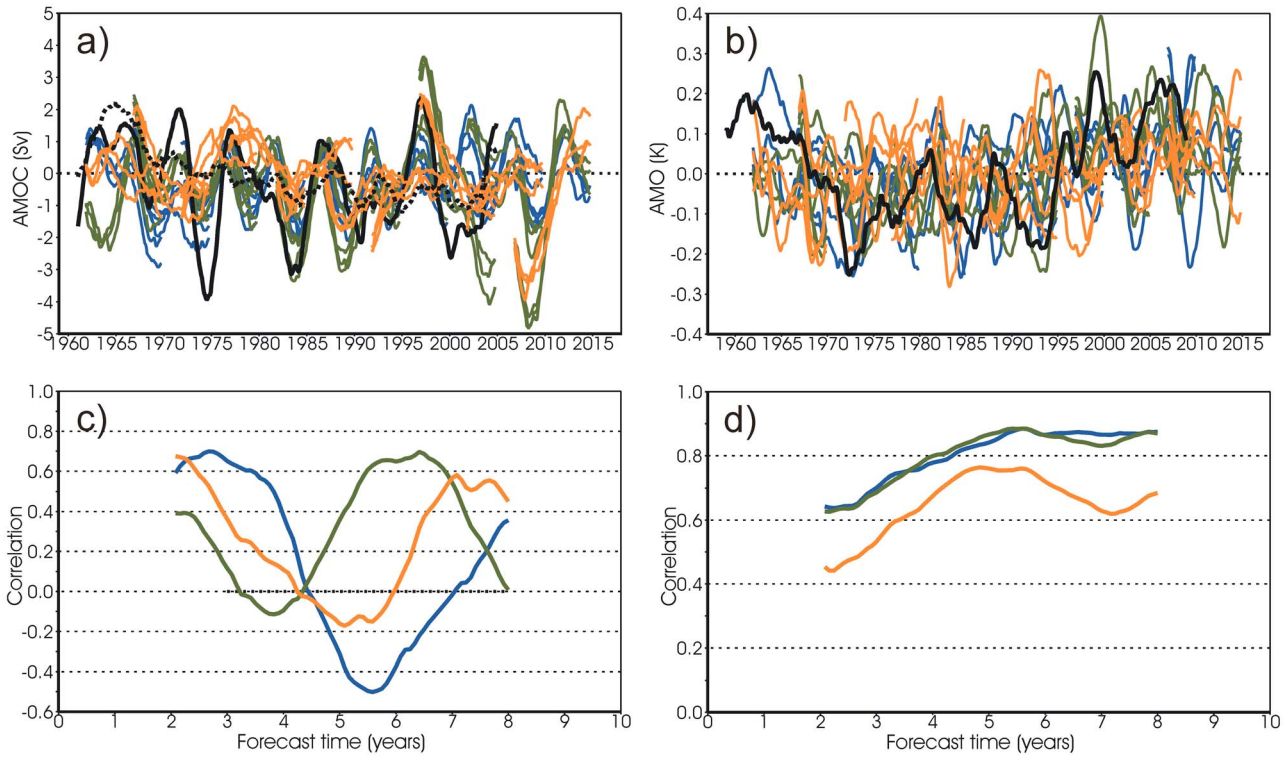

Figure 8. (a) Atlantic meridional overturning circulation (AMOC) intensity (Sv) and (b) Atlantic multidecadal oscillation (AMO) index in K, computed as the North Atlantic, 10°N–65°N, average SST minus the global average 60°S–65°N SST) anomalies for the ten decadal three-member ensemble reforecasts of the XBT-C (green), NoOcObs (blue), and Assim (orange) experiments. Anomalies are computed with respect to the corresponding climate over the period 1960–2005. Anomalies from the ORA-XC (NoOcObs) ocean reanalysis are shown in black solid (dashed) lines in Figure 8a. Data from ERA-40/ERA-Interim are shown in black in Figure 8b. All time series have been smoothed out with a 24 month centered running mean that removes data for the first and last years of each time series. Also shown are the ensemble mean correlation of the (c) AMOC and (d) AMO with respect to ORA-XC and ERA-40/ERA-Interim for XBT-C (green), NoOcObs (blue), and Assim (orange), which has been computed with three-member ensemble reforecasts for the period 1960–1995 on 4 year running mean anomalies.

(Figure 7b), which show a less-pronounced upward trend than the global mean. The ensemble mean correlation for the tropical band (Figure 7d) is noticeably lower than for the global average, and not statistically significant with 95% confidence for most of the forecast range. The highest skill is found for the NoOcObs experiment, a contradictory result when considering that no substantial differences in skill were found for near-surface air temperature. This type of result suggests that this issue requires more investigation with larger samples.

[30] Previous studies [e.g., *Collins et al.*, 2006] suggest that an accurate initialization of the AMOC could allow skillful predictions of the Atlantic multidecadal variability (AMDV) a few years in advance. However, past AMOC fluctuations have been poorly observed and a large uncertainty in ocean reanalyses exists [*Doblas-Reyes et al.*, 2010a; Krüger et al., submitted manuscript, 2011]. This uncertainty is found even when ocean reanalyses carried out using the same ocean model are considered [*Doblas-Reyes et al.*, 2010b], as seen in Figure 8 where the AMOC from the reanalyses used to initialize the XBT-C and NoOcObs experiments is shown. The uncertainty in the ocean reference implies that an assessment of the forecast quality of the AMOC predictions would necessarily give highly uncertain estimates.

[31] The XBT-C experiment has been initialized with ocean states that are expected to give the most reliable estimate of our initialized estimates of true AMOC variability because it uses the most complete and correct set of ocean data, while the ocean initial conditions used in the NoOcObs experiment underestimate both the mean AMOC intensity and its variability (Figure 8a). This is in agreement with the ocean circulation results shown in Figure 4. Figure 8 shows the anomalies of the AMOC intensity and of an index of the AMDV known as the Atlantic multidecadal oscillation (AMO) index. The AMO index is calculated in this manuscript as the average North Atlantic SSTs north of 10°N from which the contemporaneous global mean SST between 60°S and 65°N has been removed [*Trenberth and Shea*, 2006]. The AMOC reforecast anomalies show interannual oscillations of amplitude similar to those from the ORA-XC analysis. The Assim reforecasts show interannual variations of similar amplitude as those of ORA-XC, and their corresponding ensembles agree well in several instances, especially for the 2005 start date. This behavior has to be put in the context of the strong AMOC

drift described in section 4.1. The ensemble mean AMOC correlation (Figure 8c) has positive values in the first few forecast years although lower than 0.6 for most forecast times, with large differences between the three experiments but with the highest skill for NoOcObs in the first half of the forecast and XBT-C during most of the second one. As suggested by *Hawkins and Sutton* [2008], some predictability of this order of the AMOC is expected from persistent changes in the thermohaline circulation at decadal time scales in the North Atlantic. It is very difficult though, on the basis of these results, to claim that one experiment is better than any other as the largest difference in correlation has a confidence of 85%, a low value even before taking into account the obvious serial correlation of the time series.

[32] As suggested by *Dijkstra et al.* [2006], some agreement is expected between the AMOC and AMO indices from the reference data sets. The AMO has been considered as a proxy indicator for the intensity of the AMOC, and some indications along these lines can be observed when comparing Figures 8a and 8b. However, the correspondence between the AMO and AMOC reference estimates is low in the reference data used in this paper, the simultaneous correlation between the two time series being 0.35. The SSTs averaged over the North Atlantic (not shown) are highly skillful, with correlations above 0.6, because they correctly reproduce the upward trend observed since the mid 1970s. Although the AMO index used here discounts for most of this warming by removing the global mean SSTs, the ensemble mean correlation of the 4 year average predictions is higher than 0.4 in all instances, increasing with forecast time. A comparison of the AMOC and AMO correlations as a function of forecast time (Figure 8) also suggests that in the systems and time scales dealt with here, a strong relationship between the two indices should not be expected. The AMO skill is similar for the NoOcObs and XBT-C experiments, and systematically lower for the Assim experiment, a result that again points at the detrimental impact on the decadal forecast quality of the assimilation of incorrect XBT data. More pessimistic skill scores have been found with an alternative AMO index estimated as SST averaged over a northern (40°N–60°N, 60°W–10°W) and a southern (40°S–60°S, 50°W–0°W) Atlantic box [*Latif et al.*, 2006]. In this case, the strong SST drift over the southern Atlantic Ocean (Figure 2) might adversely affect the simulations and limit the reproducibility of that specific AMO index.

5. Summary and Discussion

[33] The drift and forecast quality of three sets of decadal reforecasts carried out with the IFS/HOPE coupled system have been analyzed. The three experiments are different in their initial conditions. The Assim experiment has been initialized with an ocean reanalysis that includes ocean data assimilation, the XBT-C experiment is initialized with data from an ocean reanalysis similar to the one used for Assim but with an important correction in the XBT data assimilated while the NoOcObs experiment has been initialized with data from an ocean-only simulation where no subsurface ocean data have been assimilated. All ocean reanalyses were performed with a strong relaxation to observed SSTs. This is the first time that a parallel set of decadal predictions has been carried out in an attempt to assess the relevance of ocean data assimilation in decadal prediction with realistic initialization.

[34] As the reforecasts are initialized with a realistic state of the climate system, a sizable drift develops during the forecast time in both ocean and atmospheric variables. There are many possible reasons for the existence of the drift, starting from the lack of balance between the radiation budget used to produce the ocean and atmospheric reanalyses, and including the intrinsic systematic errors of the atmospheric and ocean models due to missing processes. For instance, the atmospheric model does not properly simulate the tropical Sc clouds and has excessively weak trade winds, while the ocean model does not correctly represent the most relevant eddies. We made efforts to reduce the model drift, and that was the reason for introducing changes in the cloud microphysics that reduced the atmospheric model cold bias by increasing the surface solar radiation, but much more remains to be done. Furthermore, in a seamless climate prediction spirit [*Hurrell et al.*, 2009], reducing the drift for decadal forecasting benefits monthly and seasonal forecast systems, which are affected by a similar problem.

[35] Although all experiments have a similar drift, cold over the tropical oceans and warm over certain areas of the northern continents and the southern oceans, there are differences of the order of tenths of a degree, and hence much smaller than the drift itself, between the three experiments. The tropical SST drift becomes virtually equal in the three experiments following an increased surface latent heat flux into the atmosphere over the west Pacific in Assim and XBT-C with respect to NoOcObs, which is linked to an increase in outgoing top of the atmosphere net radiation. In contrast with the tropics, the small extratropical drift differences persist for the whole duration of the simulations, the differences concentrating on specific basins as forecast time increases. This suggests that small-amplitude signals in the extratropical ocean initial conditions can have an impact a long time after the forecast is started.

[36] In spite of the model drift and the fact that several climate processes, such as those related to sea-ice formation, export and melting, are not represented in the model, the decadal prediction experiments described here show a positive forecast quality that is statistically significant over several areas and that is comparable to experiments published previously. Positive correlation with observations is found for tropospheric air temperature and upper-ocean heat content, the correlation increasing with forecast time in most cases. The regions with significant skill and the skill level obtained will depend strongly on the reforecast period considered because of the different phases of the low-frequency variability to be predicted and the strength of the anthropogenic climate change, which is the likely source of the increase of skill with forecast time. Precipitation does not show significantly positive skill beyond the first year.

[37] The experiments show very similar forecast quality. In those cases where some differences appear, the differences are not statistically significant with a high confidence level. This leads to apparent contradictory results that might not hold with longer samples and larger ensembles. For instance, while no substantial skill differences between the experiments are found for near-surface air temperature, the predictions of the upper-ocean averaged temperature are for

some regions more skillful for NoOcObs than for XBT-C or Assim. The ability to predict interannual variations of the AMOC is difficult to assess because of the uncertainty in AMOC estimates from ocean reanalyses [*Doblas-Reyes et al.*, 2010a]. The reforecasts show anomalies with oscillations that resemble those observed in an ocean reanalysis and encourage a more in-depth analysis of the skill in predicting interannual variations of the AMOC. An estimate of the AMO has been used as a proxy of the AMOC intensity. The AMO index also shows positive skill, as for the AMOC intensity, that increased with forecast time.

[38] A negative impact of the assimilation of corrupted XBT data in the ocean reanalysis used to initialize the decadal forecasts has been found in the AMO index and the global mean upper-ocean heat content. This is in agreement with results described by *Yasunaka et al.* [2011]. However, many other instances have been found where no clear signs of the negative effect of the corrupted XBT data are evidenced. In fact, in some cases such as the tropical averaged upper-ocean heat content or the Arctic near-surface air temperature, the reforecasts initialized using a reanalysis without ocean data assimilation show a consistently better skill.

[39] The insignificant differences found between the three experiments might be disappointing, but have an important aspect. It is difficult to obtain statistical significance with limited samples, but this does not necessarily mean that the differences do not have a physical basis. The finding that the less-complete initialization method can apparently give better results at times is interesting, and points to the need for more detailed investigation. The small forecast quality differences might also be due to the important model drift that, as shown in the seasonal forecasting context [*Balmaseda and Anderson*, 2009], could prevent the model from making the most of the additional information available in the experiment initialized from reanalysis that use ocean data assimilation. The use of an a posteriori linear bias-correction scheme is just a very simple approach to make the reforecasts tractable because the interaction between the forecast anomalies and the drift can be highly nonlinear. The option of initializing the reforecasts in anomaly mode [*Smith et al.*, 2007] with the same forecast system for the realistic initialization should be explored.

[40] The experiments described in this paper use the experimental setup defined in the ENSEMBLES project. This setup shares many characteristics with the CMIP5 decadal prediction exercise. One of the main caveats of the results in this paper is that the differences between experiments are so subtle and the uncertainty in the forecast quality estimates so large that it is difficult to extract significant conclusions with the short samples, high interval between start dates and small ensemble size considered. Unfortunately, both the sample and ensemble sizes are, instead, limited by the important computing resources required to run even the experiments described here. We believe that similar difficulties are likely to be found when this type of experiment will be used to determine how the decadal prediction forecast quality is improved when using initialization with respect to uninitialized predictions.

[41] The results of the few initialized decadal forecast experiments carried out to date as well as the results shown here suggest that although there is some skill in predicting air temperature, the skill for other variables is rather limited. However, some decadal predictability and gain in skill of multiyear averages of atmospheric variables and ocean circulation from initializing with respect to uninitialized predictions has been found [*Smith et al.*, 2010]. As happened already in the field of seasonal forecasting, a reduction of the model drift, improved observational reference data sets and a better understanding of the processes at the origin of the interannual and decadal predictability should produce more skillful multiyear useful predictions in the future, as well as an increased benefit from a better informed initialization of the predictions.

[42] **Acknowledgments.** This work was supported by the ENSEMBLES (FP6-GOCE-CT-2003-505539) and the QWeCI (FP7-ENV-2009-1-243964) projects. The authors acknowledge the significant contributions by Richard Forbes, Kristian Mogensen, Jean-Jacques Morcrette, Franco Molteni, Geert Jan van Oldenborgh, and Tim Stockdale. This paper has also benefited from fruitful discussions with David Anderson, Thomas Jung, James Murphy, Doug Smith, and Noel Keenlyside.

References

Adler, R. F., et al. (2003), The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), *J. Hydrometeorol.*, *4*, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Anderson, D. L. T., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, F. Molteni, F. J. Doblas-Reyes, K. Mogensen, and A. Vidard (2007), Development of the ECMWF seasonal forecast system 3, *ECMWF Tech. Memo. 503*, 56 pp., Euro. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Anderson, D. L. T., F. J. Doblas-Reyes, M. Balmaseda, and A. Weisheimer (2009), Decadal variability: Processes, predictability and prediction, *ECMWF Tech. Memo. 591*, 47 pp., Euro. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Balmaseda, M., and D. L. T. Anderson (2009), Impact of initialization strategies and observations on seasonal forecast skill, *Geophys. Res. Lett.*, *36*, L01701, doi:10.1029/2008GL035561.

Balmaseda, M. A., G. C. Smith, K. Haines, D. Anderson, T. N. Palmer, and A. Vidard (2007), Historical reconstruction of the Atlantic meridional overturning circulation from the ECMWF operational ocean reanalysis, *Geophys. Res. Lett.*, *34*, L23615, doi:10.1029/2007GL031645.

Balmaseda, M. A., A. Vidard, and D. L. T. Anderson (2008), The ECMWF ocean analysis system: ORA-S3, *Mon. Weather Rev.*, *136*, 3018–3034, doi:10.1175/2008MWR2433.1.

Bechtold, P., M. Köhler, T. Jung, F. J. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo (2008), Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales, *Q. J. R. Meteorol. Soc.*, *134*, 1337–1351, doi:10.1002/qj.289.

Boer, G. J., and S. J. Lambert (2008), Multi-model decadal potential predictability of precipitation and temperature, *Geophys. Res. Lett.*, *35*, L05706, doi:10.1029/2008GL033234.

Collins, M., et al. (2006), Interannual to decadal climate predictability in the North Atlantic: A multimodel-ensemble study, *J. Clim.*, *19*, 1195–1203, doi:10.1175/JCLI3654.1.

Dijkstra, H. A., L. te Raa, M. Schmeits, and J. Gerrits (2006), On the physics of the Atlantic multidecadal oscillation, *Ocean Dyn.*, *56*, 36–50, doi:10.1007/s10236-005-0043-0.

Doblas-Reyes, F. J., R. Hagedorn, T. N. Palmer, and J.-J. Morcrette (2006), Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts, *Geophys. Res. Lett.*, *33*, L07708, doi:10.1029/2005GL025061.

Doblas-Reyes, F. J., A. Weisheimer, M. Déqué, N. Keenlyside, M. McVean, J. M. Murphy, P. Rogel, D. Smith, and T. N. Palmer (2009), Addressing model uncertainty in seasonal and annual dynamical seasonal forecasts, *Q. J. R. Meteorol. Soc.*, *135*, 1538–1559, doi:10.1002/qj.464.

Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith (2010a), Forecast quality assessment of the ENSEMBLES seasonal-to-decadal Stream 2 hindcasts, *ECMWF Tech. Memo. 621*, 45 pp., Euro. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer (2010b), Decadal climate prediction with the ECMWF coupled forecast system: Impact of ocean observations, *ECMWF Tech. Memo. 633*,

24 pp., Euro. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Drijfhout, S., W. Hazeleger, F. Selten, and R. Haarsma (2008), Future changes in internal variability of the Atlantic meridional overturning circulation, *Clim. Dyn.*, *30*, 407–419, doi:10.1007/s00382-007-0297-y.

Fan, Y., and H. van den Dool (2008), A global monthly land surface air temperature analysis for 1948–present, *J. Geophys. Res.*, *113*, D01103, doi:10.1029/2007JD008470.

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

Hawkins, E., and R. Sutton (2008), Potential predictability of rapid changes in the Atlantic meridional overturning circulation, *Geophys. Res. Lett.*, *35*, L11603, doi:10.1029/2008GL034059.

Hawkins, E., and R. Sutton (2009a), Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling, *J. Clim.*, *22*, 3960–3978, doi:10.1175/2009JCLI2720.1.

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

Hawkins, E., J. Robson, R. Sutton, D. Smith, and N. Keenlyside (2011), Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach, *Clim. Dyn.*, doi:10.1007/s00382-011-1023-3, in press.

Hurrell, J., G. A. Meehl, D. Bader, T. L. Delworth, B. Kirtman, and B. Wielicki (2009), A unified modeling approach to climate system prediction, *Bull. Am. Meteorol. Soc.*, *90*, 1819–1832, doi:10.1175/2009BAMS2752.1.

Ishii, M., and M. Kimoto (2009), Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections, *J. Oceanogr.*, *65*, 287–299, doi:10.1007/s10872-009-0027-7.

Jolliffe, I. T. (2007), Uncertainty and inference for verification measures, *Weather Forecasting*, *22*, 637–650, doi:10.1175/WAF989.1.

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

Laepple, T., S. Jewson, and K. Coughlin (2008), Interannual temperature predictions using the CMIP3 multi-model ensemble mean, *Geophys. Res. Lett.*, *35*, L10701, doi:10.1029/2008GL033576.

Lanzante, J. R. (2005), A cautionary note on the use of error bars, *J. Clim.*, *18*, 3699–3703, doi:10.1175/JCLI3499.1.

Latif, M., C. Böning, J. Willebrand, A. Biastoch, J. Dengg, N. Keenlyside, U. Schweckendiek, and G. Madec (2006), Is the thermohaline circulation changing?, *J. Clim.*, *19*, 4631–4637, doi:10.1175/JCLI3876.1.

Lean, J. L., and D. H. Rind (2009), How will Earth's surface temperature change in future decades?, *Geophys. Res. Lett.*, *36*, L15708, doi:10.1029/2009GL038932.

Lozier, M. S. (2010), Deconstructing the conveyor belt, *Science*, *328*, 1507–1511, doi:10.1126/science.1189250.

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell (2007), The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Meehl, G. A., et al. (2009), Decadal prediction: Can it be skillful?, *Bull. Am. Meteorol. Soc.*, *90*, 1467–1485, doi:10.1175/2009BAMS2778.1.

Mochizuki, T., et al. (2010), Pacific decadal oscillation hindcasts relevant to near-term climate prediction, *Proc. Natl. Acad. Sci. U. S. A.*, *107*, 1833–1837, doi:10.1073/pnas.0906531107.

Palmer, T. N. (2000), Pedicting uncertainty in forecasts of weather and climate, *Rep. Prog. Phys.*, *63*, 71–116, doi:10.1088/0034-4885/63/2/201.

Palmer, T. N., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith (2007), Ensemble prediction: A pedagogical perspective, *ECMWF Newsl.* *106*, pp. 10–17, Euro. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

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

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

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang (2002), An improved in situ and satellite SST analysis for climate, *J. Clim.*, *15*, 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

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

Smith, T. M., and R. W. Reynolds (2003), Extended reconstruction of global sea surface temperature based on COADS data (1854–1997), *J. Clim.*, *16*, 1495–1510, doi:10.1175/1520-0442-16.10.1495.

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

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife (2010), Skillful multi-year predictions of Atlantic hurricane frequency, *Nat. Geosci.*, *3*, 846–849, doi:10.1038/ngeo1004.

Trenberth, K. E., and D. J. Shea (2006), Atlantic hurricanes and natural variability in 2005, *Geophys. Res. Lett.*, *33*, L12704, doi:10.1029/2006GL026894.

Uppala, S. M., et al. (2005), The ERA-40 reanalysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012, doi:10.1256/qj.04.176.

van Oldenborgh, G. J., S. S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger, and H. Dijkstra (2009), Western Europe is warming much faster than expected, *Clim. Past*, *5*(1), 1–12, doi:10.5194/cp-5-1-2009.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Q. J. R. Meteorol. Soc.*, *134*, 241–260, doi:10.1002/qj.210.

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel (2009), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, *36*, L21711, doi:10.1029/2009GL040896.

Wijffels, S. E., J. Willis, C. M. Domingues, P. Barker, N. J. White, A. Gronell, K. Ridgway, and J. A. Church (2008), Changing expendable bathythermograph fall rates and their impact on estimates of thermosteric sea level rise, *J. Clim.*, *21*, 5657–5672, doi:10.1175/2008JCLI2290.1.

Yasunaka, S., M. Ishii, M. Kimoto, T. Mochizuki, and H. Shiogama (2011), Influence of XBT temperature bias on decadal climate prediction with a coupled climate model, *J. Clim.*, doi:10.1175/2011JCL14230.1.

M. A. Balmaseda, T. N. Palmer, and A. Weisheimer, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK.

F. J. Doblas-Reyes, Institut Català de Ciències del Clima, Doctor Trueta 203, E-08005 Barcelona, Spain. (f.doblas-reyes@ic3.cat)