Two Tales of Initializing Decadal Climate Prediction Experiments with the ECHAM5/MPI-OM Model

DANIELA MATEI, HOLGER POHLMANN, JOHANN JUNGCLAUS, WOLFGANG MÜLLER,
HELMUTH HAAK, AND JOCHEM MAROTZKE

Max Planck Institute for Meteorology, Hamburg, Germany

ABSTRACT

This paper investigates the impact of different ocean initialization strategies on the forecast skill of decadal prediction experiments performed with the ECHAM5/Max Planck Institute Ocean Model (MPI-OM) coupled model. The ocean initializations assimilate three-dimensional temperature and salinity anomalies from two different ocean state estimates, the ocean reanalysis of the German contribution to Estimating the Circulation and Climate of the Ocean (GECCO) and an ensemble of MPI-OM ocean experiments forced with the NCEP–NCAR atmospheric reanalysis. The results show that North Atlantic and Mediterranean sea surface temperature (SST) variations can be skillfully predicted up to a decade ahead and with greater skill than by both uninitialized simulations and persistence forecasts. The regional distribution of SST predictive skill is similar in both initialization approaches; however, higher skill is found for the NCEP hindcasts than for the GECCO hindcasts when a combination of predictive skill measures is used. Skillful predictions of surface air temperature are obtained over northwestern Europe, northern Africa, and central-eastern Asia. The North Atlantic subpolar gyre region stands out as the region with the highest predictive skill beyond the warming trend, in both SST and upper-ocean heat-content predictions. Here the NCEP hindcasts deliver the best results due to a more accurate initialization of the observed variability. The dominant mechanism for North Atlantic climate predictability is of dynamical origin and can be attributed to the initialization of the Atlantic meridional overturning circulation, thus explaining the reoccurrence of high predictive skill within the second pentad of the hindcasts experiments. The results herein demonstrate that ocean experiments forced with the observed history of the atmospheric state constitute a simple but successful alternative strategy for the initialization of skillful climate predictions over the next decade.

1. Introduction

Creating skillful decadal climate predictions represents a major challenge (Meehl et al. 2009) because decadal prediction lies at the intersection of initial-value problems (as in seasonal forecasts) and boundary-value problems (as in long-term climate projections). To predict near-term climate changes, we have to capture both the externally forced and the internally generated climate signals. For this, the coupled climate models need to be initialized from the best estimate of the current climate state, especially of the ocean. Here we compare two different ocean initialization strategies in their effects on decadal prediction skill.

Recently, a couple of initialized decadal predictions were attempted, with partially contradictory results. In the first study, Smith et al. (2007) showed that initializing

a coupled model with observed ocean and atmospheric conditions increased the skill of global mean surface temperature and ocean heat content predictions up to a decade ahead. They predicted that the internal climate variability will partially offset the anthropogenic global warming from 2005 to 2009, after which the climate will continue to warm. By contrast, Keenlyside et al. (2008) assimilated only sea surface temperature variations and found that global mean temperature may not increase over the next decade because natural climate variations over the North Atlantic and tropical Pacific countervail the projected global warming. In addition, Keenlyside et al. (2008) showed enhanced skill in predicting decadal means of surface temperature over parts of the North Atlantic sector, including the European and North American land regions, compared to uninitialized simulations. A third study (Pohlmann et al. 2009) made use of ocean initial conditions from an ocean reanalysis [the German contribution to Estimating the Circulation and Climate of the Ocean (GECCO)] and demonstrated

Corresponding author address: Daniela Matei, Max Planck Institute for Meteorology, Bundesstrasse 53, 20146, Hamburg, Germany.
E-mail: daniela.matei@zmaw.de

enhanced predictive skill up to a decade ahead over the North Atlantic region. In a more recent study, Mochizuki et al. (2010) showed that an initialization of the upper-ocean state using historical observations was effective for successful predictions of heat content fingerprints of the Pacific decadal oscillation (PDO) over almost a decade in advance. They predicted a negative tendency for the PDO phase in the coming decade that will contribute to a slowing down of the global mean surface temperature rise.

The rather different results of the first decadal prediction systems highlight several scientific challenges of the decadal prediction efforts. Skillful predictions at decadal time scales can be achieved only if the observed decadal variability is predictable, is realistically represented by the prediction model, and is properly initialized (Keenlyside and Ba 2010). The importance of initialization is underscored by the fact that the two decadal prediction systems presented in Keenlyside et al. (2008) and Pohlmann et al. (2009) used essentially the same coupled model and yet came to some opposite outcomes, for example with respect to both the historical and future evolution of the global mean surface temperature (see Fig. 3 in Murphy et al. 2010). Therefore, one can speculate that differences between these two prediction systems arise mainly from different amounts of observed data used to initialize the decadal hindcast experiments: only sea surface temperature (SST) in the case of Keenlyside et al. (2008) compared to subsurface temperature and salinity in Pohlmann et al. (2009).

The question of how best to initialize the ocean for decadal predictions is complicated by the fact that reliable observations over the historical period are sparse, especially in the early period and for the subsurface ocean (see Fig. 5 of Hurrell et al. 2010). Some observations appear to be significantly biased (Domingues et al. 2008; Ishii and Kimoto 2009; Lyman et al. 2010), making the development and evaluation of different ocean initialization schemes difficult. For example, two recent studies (Yasunaka et al. 2011; Doblas-Reyes et al. 2011) have shown that the use of non–expendable bathythermograph (XBT)-corrected temperature data in the ocean initial conditions can negatively impact the forecast quality of SST and upper ocean heat content variations. Furthermore, there is no consensus among the different existing ocean reanalyses with respect to both the mean and long-term variability of important oceanic quantities such as the strength of the Atlantic meridional overturning circulation (AMOC) and meridional heat transport (Stammer 2006; Keenlyside and Ba 2010). Also, important dynamical relationships such as between the AMOC and North Atlantic SST are disparate between various ocean reanalyses (Munoz et al. 2011).

Here we take an alternative approach to the ocean initialization from an ocean reanalysis (Pohlmann et al. 2009) and initialize the coupled model ECHAM5/Max Planck Institute Ocean Model (MPI-OM) from an ensemble of MPI-OM ocean-only runs forced with the National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) atmospheric reanalysis. This rather simple initialization approach has the big advantage of using the same ocean model for both the generation of the ocean state estimate and for the forecast experiments; as a result of the more dynamically self-consistent representation of water masses and ocean circulation, this method might lead to improved predictive skill. Our hypothesis is supported by a recent study (Doblas-Reyes et al. 2011) that did not find any significant benefit for forecast quality by assimilating ocean subsurface observations. Moreover, ocean model experiments forced with the observed atmospheric state over the second half of the twentieth century do a very good job in reconstructing the observed variability of important oceanic dynamical quantities such as the Nordic seas overflows and Labrador Seawater formation (Olsen et al. 2008; Serra et al. 2010). These constitute the two main contributors to the North Atlantic Deep Water feeding the lower branch of the AMOC, and therefore their realistic representation in the NCEP-forced ocean state estimate might contribute to a successful initialization of the AMOC.

In this paper, we evaluate the predictive skill of decadal hindcast experiments in our two decadal prediction systems against observational data and statistical forecasts, focusing mainly on the surface climate and upper ocean heat content. While assessing the surface climate predictability, we also examine the impact of a limited set of initial starting dates used in the phase 5 of the Coupled Model Intercomparison Project (CMIP5) experimental setup on forecast skill. A discussion on the AMOC potential impact on the skill of decadal predictions is included too. A more detailed assessment of the predictive skill of a number of ocean dynamical quantities, such as the AMOC, the meridional heat transport, and the two main components of the North Atlantic Deep Water (Nordic seas overflows and Labrador Seawater transports), is presented in Matei et al. (2012) and a follow-up paper.

The paper is structured as follows: section 2 describes the coupled model and the experimental setup, while section 3 presents the statistical methods and observational data used for forecast verification. Section 4 makes a quantitative comparison between the two sets of initial conditions. Section 5 evaluates the predictive skill of global and regional surface climate, while section 6 focuses on the global and regional upper-ocean heat

content predictability. Section 7 investigates possible mechanisms for generating the predictive skill. Section 8 presents a discussion of the results, while section 9 enumerates the conclusions of our study.

2. Model and experimental setup

a. ECHAM5/MPI-OM coupled model

The coupled atmosphere–ocean–sea ice model used in this study is the Max Planck Institute for Meteorology (MPI) coupled general circulation model ECHAM5/ MPI-OM (Jungclaus et al. 2006), in the version assessed in the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4). The ECHAM5 (Roeckner et al. 2003) version employed here is the "tropospheric model," resolving the atmosphere up to the middle stratosphere (10 hPa). The model has 31 irregularly distributed vertical levels, with the highest vertical resolution in the atmospheric boundary layer. The horizontal resolution T63 corresponds to about $1.9° \times 1.9°$. The ocean model MPI-OM (Marsland et al. 2003) is a free-surface primitive equation model formulated on a C grid and orthogonal curvilinear coordinates. The model includes a Hibler-type dynamic–thermodynamic sea ice model and a river runoff scheme. To circumvent grid singularities at the geographical North Pole, the northern grid pole is shifted to Greenland, leading to high resolution in the Arctic and the high-latitude sinking regions. The horizontal resolution is about 1.5° on average and varies from a minimum of 12 km close to Greenland to a maximum of 180 km in the tropical Pacific. The model has 40 vertical nonequidistant z levels, of which 20 are distributed in the upper 700 m. The coupled model does not employ flux adjustment.

b. Experimental design of the NCEP-forced ocean simulations and the initialized hindcasts

The MPI-OM ocean model is initialized with annual-mean temperature and salinity from the Levitus et al. (1998) climatology and ocean velocities at rest. The model is then forced with daily surface fluxes of heat, freshwater, and momentum obtained from the NCEP–NCAR reanalysis (Kalnay et al. 1996). The heat flux is not taken directly from the NCEP–NCAR reanalysis but rather computed interactively using bulk formulas as described in Marsland et al. (2003) and Haak (2004). These parameterizations use both the ocean model upper-layer temperature (or sea ice–snow layer skin temperature) and a number of prescribed state variables taken from the NCEP–NCAR reanalysis (10-m wind speed, 2-m air temperature, total cloud cover, downward solar radiation flux). No additional (explicit) relaxation

toward the observed SSTs is used in our ocean-forced experiments. Freshwater forcing is calculated from precipitation minus evaporation plus river runoff; in addition, the model sea surface salinity is relaxed toward the Levitus climatology (Levitus et al. 1998) with an *e*-folding time of 30 days. The salinity relaxation is performed only in the ice-free part of the ocean; it helps to correct for unbalanced globally integrated surface freshwater flux and has the positive effect of reducing long-term model drift.

After the initialization, the ocean standalone model is consecutively integrated for 10 cycles, each covering the period 1948–2007. For each new cycle the initial ocean state is taken from the end state of the previous cycle. The first six cycles are considered to represent the spinup period. We use the ensemble mean of the last four cycles to produce the ocean state necessary for the initialization of decadal prediction experiments. More information on the iterative forcing approach can be found in Haak et al. (2003).

We then perform with the coupled model a noninitialized transient simulation of the twentieth century, forced by concentrations of radiatively active trace gases that were observed before 2000 and that follow the Special Report on Emissions Scenarios (SRES) A1B scenario thereafter. This experiment is used to estimate the predictability due to external forcing only. Then, in the so-called assimilation experiment (hereafter NCEP assimilation), we relax the three-dimensional temperature and salinity anomalies of the coupled model to the ensemble-mean temperature and salinity anomalies of the forced ocean runs to produce the initial conditions for the hindcast experiments. The upper three ocean layers are excluded from the initialization procedure in the regions that are covered by sea ice. The relaxation time scale is 1 day. This "anomaly initialization" scheme (WCRP 2011) is chosen to avoid model drift toward its own imperfect climatology as would be the case in the "full field initialized" hindcasts.

A set of 10-yr-long hindcasts is subsequently made, starting from the assimilation experiment on every 1 January between 1949 and 2008 (hereafter NCEP hindcasts). Two additional realizations starting every 5 years from 1960 to 2005 (in total 10 start dates) have been performed in order to evaluate the impact of the CMIP5 subsampling experimental setup on the skill of decadal hindcasts. The ensemble members are generated by very slightly perturbing the atmospheric horizontal diffusion coefficient (on the order of 10^{-6} m^2 s^{-1}).

c. Experimental design of the GECCO initialized hindcasts

The GECCO project (Köhl and Stammer 2008) provides the initial conditions for the GECCO initialization.

In GECCO hydrographic and satellite data are used in an adjoint method to derive a dynamically consistent oceanic reanalysis. The GECCO reanalysis was not corrected for the biases in XBT data, but it does not show the spurious increase in global heat content in the 1970s followed by a ''cold period'' in the late 1980s (Stammer et al. 2010), in contrast to other ocean state estimates based on uncorrected XBT datasets. This is because the dynamically self-consistent GECCO solution rejected the heat content increase in the 1970s as being dynamically inconsistent and thus identified that period as one with enhanced data errors (Stammer et al. 2010). Temperature and salinity anomalies from GECCO are assimilated into the coupled model with a relaxation time of 10 days to provide the initial conditions for the hindcast simulations (hereafter GECCO assimilation). The assimilation procedure is not performed in the top ocean model layer or in the regions covered by sea ice. The GECCO assimilation also uses the anomaly initialization technique to minimize drift problems. One decadal hindcast experiment is started on 1 January each year between 1952 and 2001 (hereafter GECCO hindcasts). The experimental setup is explained in more detail in Pohlmann et al. (2009). In contrast to the CMIP5 near-term climate predictions experimental protocol, we did not include any information about historical volcanic eruptions in the experiments analyzed in this paper.

3. Data and evaluation of hindcast skill

SST observations are taken from the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003), while land surface temperatures are taken from the NCEP Global Historical Climatology Network (GHCN)–Climate Anomaly Monitoring System (CAMS) dataset (Fan and van den Dool 2008). Upper-ocean (700 m) heat content observations are taken from the National Oceanographic Data Center (NODC) dataset (Levitus et al. 2009). The surface climate variations in the two sets of initialized decadal hindcasts are evaluated over the period 1953–2010. For the analyses of upper-ocean heat content variations, we focus on the period 1965–2009, since the data coverage has greatly improved since the mid-1960s when XBT measurements of the upper ocean began (Levitus et al. 2009). However, time series plots show results from the whole length of the hindcasts experiments. Note that a somewhat higher noise level of the initialized hindcasts is to be expected when results from unfiltered single member hindcasts are displayed.

As a measure of hindcast skill we use the anomaly correlation coefficient (COR) and the root-mean-square error (RMSE). RMSE, a widely used forecast skill measure, is sensitive to outliers and therefore can produce large error (Collins and Allen 2002). The COR is not sensitive to outliers, but since it measures only the phase difference between the forecast and observations, a high COR does not guarantee an accurate forecast. Moreover, COR is sensitive to the presence of a trend and thus may overestimate the performance of the forecast system. For example, a high COR skill score may result from a strong global warming trend hiding the decadal climate variability signal supposed to be improved by the initialization of the climate predictions (Murphy et al. 2010). We compare the predictive skill of the initialized hindcasts experiments against the benchmark skill of the noninitialized model simulations and of the persistence forecast, a commonly used statistical forecast, which assumes that future conditions will be the same as past conditions. For comparing the RMSE in different forecasts experiments, we apply the RMSE skill score (Jolliffe and Stephenson 2003), using as reference forecast the noninitialized model simulations:

$$\text{RMSE}_{\text{skill}} = 1 - \frac{\text{RMSE}_{\text{hindcast}}}{\text{RMSE}_{\text{NonINIT}}}.$$

The 5% statistical significance levels used to display the global COR skill are computed for every grid point using a two-sided t test. The computation takes into consideration both the length and the decorrelation time scales of the two time series (as in Von Storch and Zwiers 1999). For regional SST, upper ocean heat content, and AMOC transport indices a one-sided t test is employed.

In the following we refer to the experiments in which we nudge the temperature and salinity anomalies from the GECCO reanalysis and the ensemble of NCEP-forced MPI-OM integrations as the GECCO assimilation and NCEP assimilation, respectively.

4. NCEP and GECCO surface climate and upper-ocean heat content initial conditions

Before assessing the predictive skill of surface climate and upper ocean heat content, we evaluate the fidelity with which the two sets of initial conditions are following the observed climate variations. Figure 1 shows that there are similarities but also substantial differences between the accuracy of the two ocean reconstructions that might impact the predictive skill of the initialized hindcasts. We see a very good correspondence between the SST variations in both reanalyses and observations over most of the global ocean (Figs. 1a,b). However, only the NCEP initialization accurately reproduces the observed SST variations over the North Atlantic subpolar region. On the other hand, the best resemblance to the observations
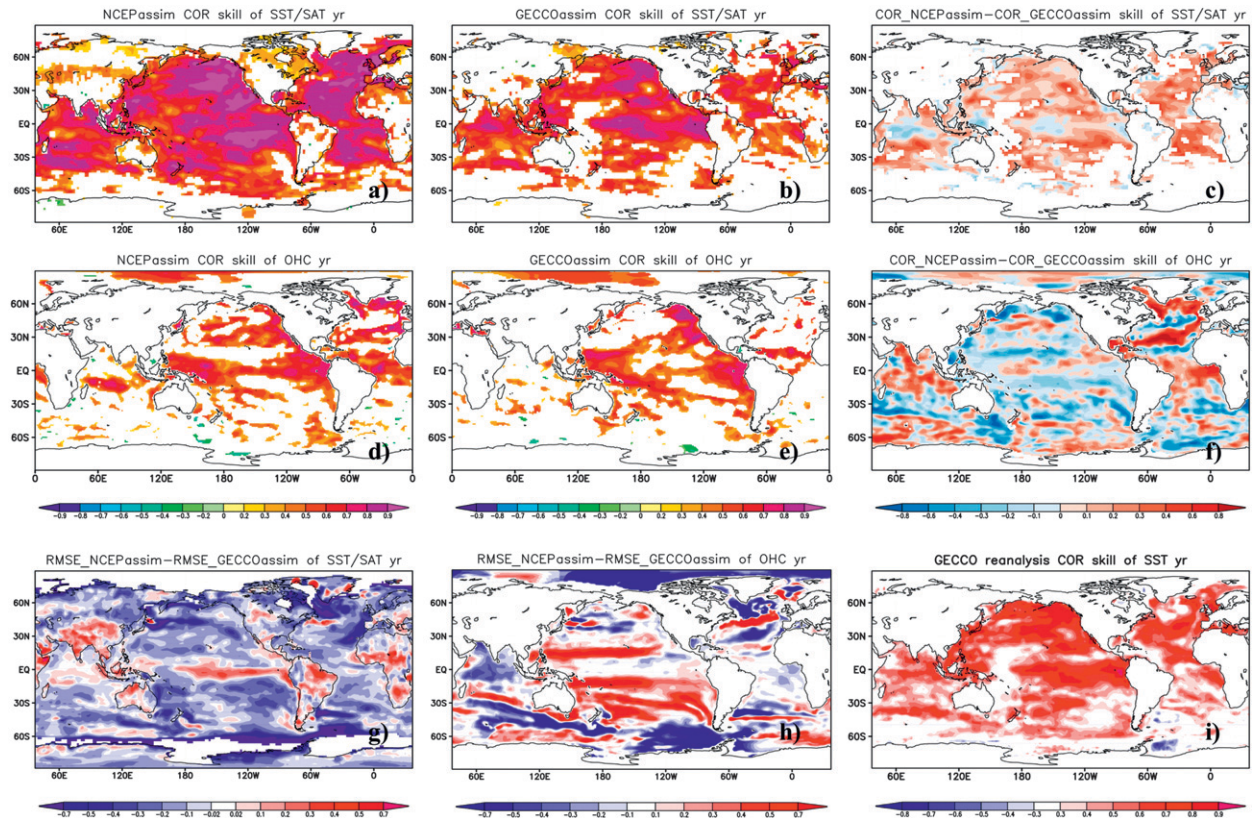
FIG. 1. Fidelity of assimilation results. COR between the surface temperature (SAT) variations from observations and (a) NCEP assimilation and (b) GECCO assimilation. (c) The SAT COR difference of the two assimilation runs, and COR between the upper-ocean heat content (OHC) variations from observations and (d) NCEP assimilation and (e) GECCO assimilation. (f) The OHC COR difference of the two assimilation runs. (g) Difference in SAT RMSE of the NCEP and GECCO assimilation experiments. (h) OHC RMSE difference of the two assimilation experiments. (i) COR between the SST variations in observations and GECCO reanalysis. The observations are taken from HadISST for the SST and GHCN–CAMS for land temperature, while OHC observed variations are taken from the NODC dataset.

over the Indian Ocean and the central-eastern equatorial Pacific is found in the GECCO assimilation (Figs. 1c,g). The lack of significant COR skill found in the GECCO assimilation over a number of oceanic regions such as the North Atlantic subpolar gyre (SPG), the western tropical Pacific east of Australian continent, the southeastern South Pacific, and the South Atlantic (Fig. 1b) is not due to the exclusion of the top ocean model layer from the GECCO initialization procedure (not shown); rather, it is the result of a comparatively large SST error allowed in the dynamically self-consistent GECCO ocean estimation (cf. Figs. 1b and 1i). Both assimilation experiments simulate very low COR over the land regions, with the exception of the British Isles, the Iberian Peninsula, and the northern parts of Africa and South America. This generally low COR over land can be partially explained through the fact that neither the atmospheric nor the land components have been initialized. On the other hand, to reproduce the observed variability of land temperature, ensemble experiments might be required.

Upper ocean heat content variability shows good agreement between both assimilation experiments and the NODC observational estimate, over large areas of the tropical and northern Pacific and tropical Atlantic (Figs. 1d,e). Again, only the NCEP assimilation accurately follows the observed OHC fluctuations over the North Atlantic subpolar gyre and in addition over the Indian Ocean (Fig. 1d). However, the NCEP assimilation also shows a less realistic OHC variability over the subtropical Pacific and along the path of the North Atlantic Current than does the GECCO assimilation (Figs. 1f,h).

5. Decadal prediction skill of surface temperature

a. Spatial distribution of SST predictive skill

We first examine the global distribution of SST predictive skill in the two sets of initialized decadal hindcasts. We consider three time scales: the first year (yr1),
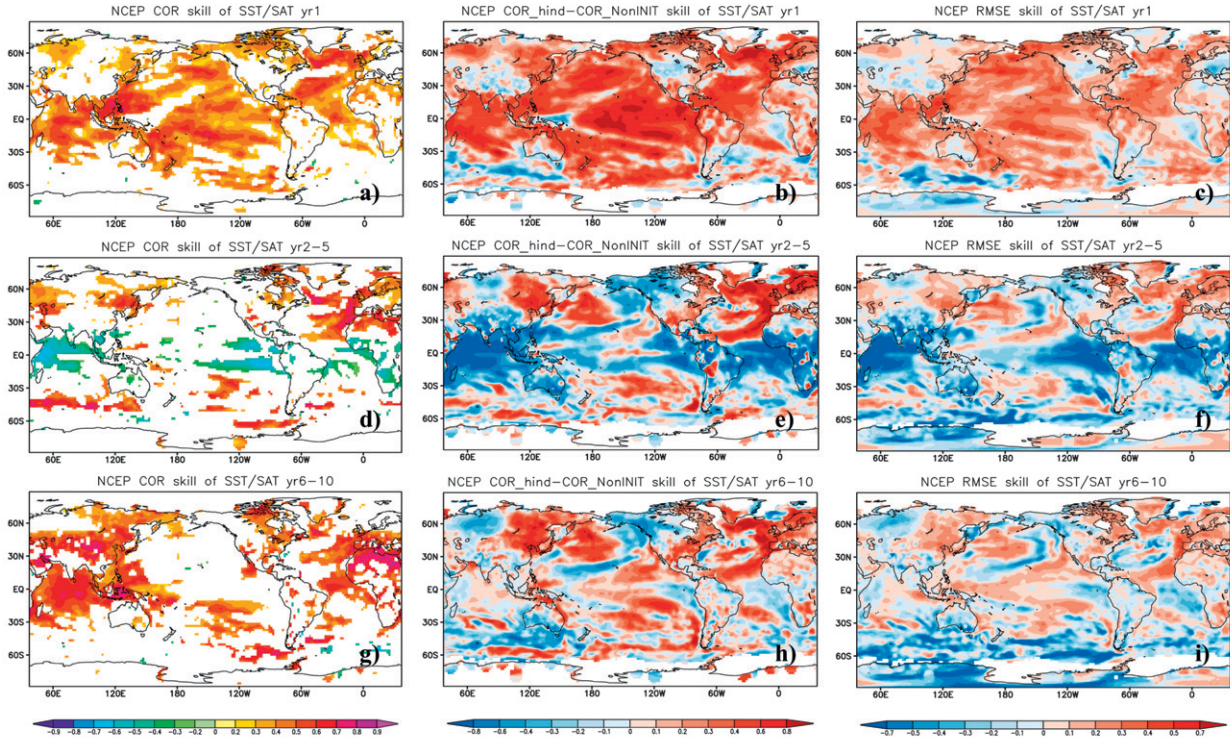
FIG. 2. Surface temperature predictive skill of the NCEP initialized hindcasts at lead times (top) yr1, (middle) yr2–5, and (bottom) yr6–10. The observations are taken from HadISST for the SST and GHCN–CAMS for SAT. (left) Anomaly correlation coefficient (COR) between the observations and hindcasts. Only the significant COR (at 5% level) are plotted. (middle) Difference in COR skill score of hindcasts and NonINIT experiments. (right) RMSE skill score of hindcasts referenced to the RMSE of the NonINIT experiments.

the mean over years 2 to 5 (yr2–5), and the mean over years 6 to 10 (yr6–10). The separation of time scales is done to differentiate between the predictive skill associated with persistence and ENSO predictability (for the first year) and the longer-term predictability.

We start by analyzing the spatial distribution of the anomaly correlation coefficient between observed and hindcast SST variations. For the first year, skillful predictions of SST are found in the NCEP hindcasts over almost all ocean basins, with the exception of the extratropical South Atlantic (Fig. 2a). On this time scale, particularly high SST COR skill is found over the North Atlantic, the western Pacific–eastern Indian Oceans, the North Pacific, and the off-equatorial tropical Pacific. Skill for year one comes mostly from persistence and improved ENSO predictability during the first 6 months of the hindcasts (not shown). The lack of SST predictive skill over the North Atlantic subpolar frontal zone–Gulf Stream path in the NCEP hindcasts can be attributed to a not very accurate initialization of SST variations over this area in the NCEP assimilation (Fig. 1a). The lower SST COR skill scores in yr1 of the GECCO initialized hindcasts (Fig. 3a) are a consequence of a relatively large

SST error allowed in the GECCO adjoint solution that leads to a less accurate SST initialization (section 4, Pohlmann et al. 2009).

We now explore the relative skill of the initialized hindcasts and the radiatively forced, noninitialized simulations (hereafter NonINIT). We find large differences in the SST COR skill in the initialized hindcasts compared to the NonINIT simulations (Fig. 2b) over almost the whole global ocean. This increase in SST COR skill is much stronger in the NCEP than in the GECCO hindcast experiments, with the latter even showing a decrease in the predictive skill over the Indian and South Atlantic Oceans (cf. Figs. 2b and 3b). Considering now the RMSE as a skill measure instead of COR, we find that both sets of initialized hindcasts exhibit a significant increase in SST RMSE skill at a lead time of one year when compared to NonINIT simulations. The strongest enhancement in RMSE skill is achieved in the extratropical Northern Hemisphere regions and the tropical Pacific (Figs. 2c and 3c).

The COR skill picture changes as the lead time of the forecast increases to yr2–5 and yr6–10. For yr2–5, skillful SST predictions are found in both approaches over the North Atlantic and the subtropical South Pacific (Figs. 2d
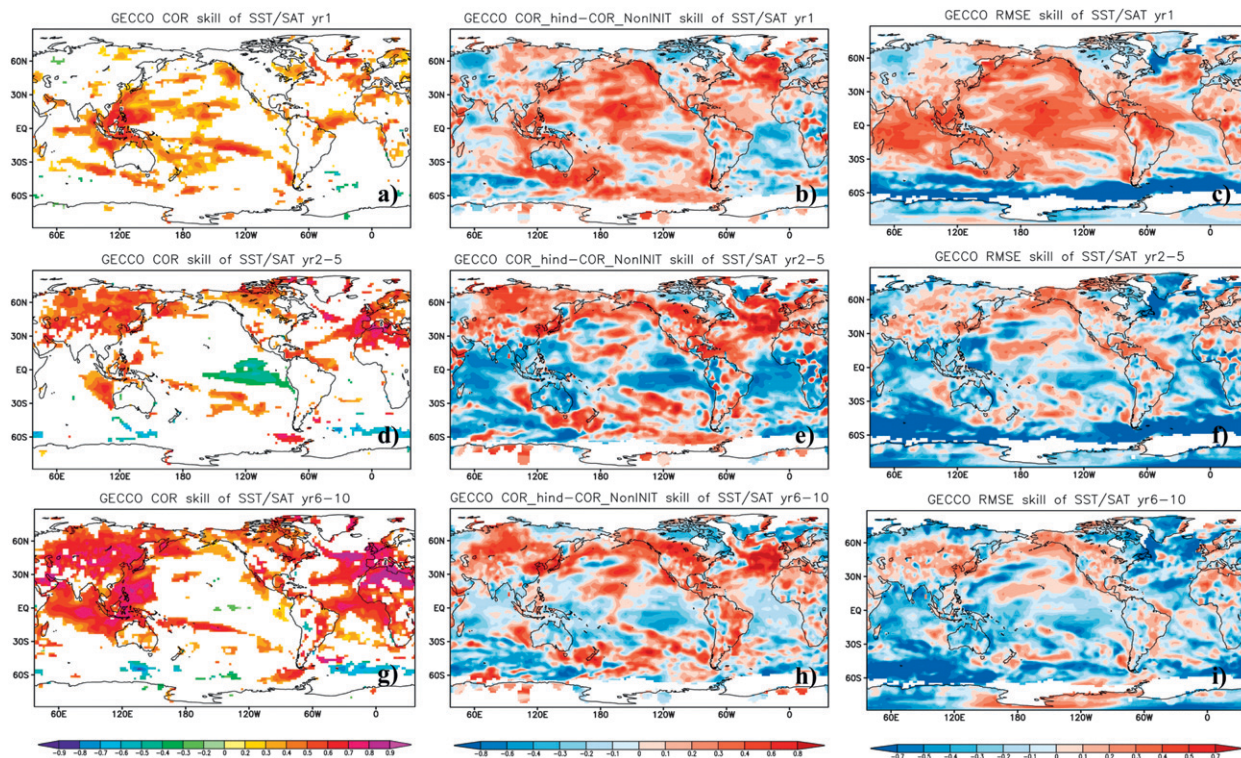
FIG. 3. As in Fig.2, but for the GECCO initialized hindcasts.

and 3d). In addition, the GECCO hindcasts show predictive skill over the southeastern Indian Ocean/Maritime Continent (Fig. 3d), while the NCEP hindcasts exhibit substantial, albeit marginally significant, COR skill at around 30°N in the central North Pacific (not shown). However, there are also regions where the initialization causes a significant decrease in predictive skill compared to the NonINIT. This is the case for most of the tropical oceans (Figs. 2e and 3e). The remarkable correspondence between the regions exhibiting a large negative COR skill and the regions that in the model world are strongly impacted by teleconnections from the eastern tropical Pacific (not shown) suggests that the loss of skill is a direct consequence of failing to predict the tropical Pacific SST variations beyond the first year. When one looks at the relative COR and RMSE skill maps (Figs. 2e,f), both the large skill improvement over the North Atlantic and North Pacific and the degradation of skill over the tropical ocean basins are clearly illustrated for the NCEP-initialized hindcasts. Therefore, in the NCEP approach, the RMSE skill results are consistent with the relative COR skill scores. In contrast, for the GECCO approach, the increase in COR skill over the North Atlantic is associated with a decrease in RMSE skill over most of the same region (Figs. 3e,f). We will return to this aspect in section 5b.

Analyzing the COR skill score maps over the second pentad (yr6–10) of the initialized hindcasts, we find that the region of skillful predictive skill is now extended to the whole Atlantic Ocean (Figs. 2g and 3g). In addition, high correlation skill scores are obtained over the Indian Ocean, the western tropical North Pacific, and over the central subtropical South Pacific. Contrasting now the results of the hindcasts and NonINIT, we see that both ocean initializations lead to a significant improvement in COR skill score over the North Atlantic (Figs. 2h and 3h). In the NCEP hindcasts the area of improved COR skill when compared to the NonINIT simulation is extended to the North Pacific and central subtropical South Pacific, too. In both prediction systems, the COR skill scores over the western tropical North Pacific and Indian Ocean, albeit very high, are at the same level as those of the NonINIT experiments and therefore must arise from the external radiative forcing. One noticeable exception is the Indian Ocean region west of the Australian coast, where the GECCO initialization leads to a substantial skill enhancement (Fig. 3h). We generally obtain also for yr6–10 a consistent picture given by the two skill measures used in our study (Figs. 2h,i and 3h,i). An exception is again found in the GECCO hindcasts that show increased COR and decreased RMSE skill scores over the North Atlantic (Figs. 3h,i). In addition we find that in
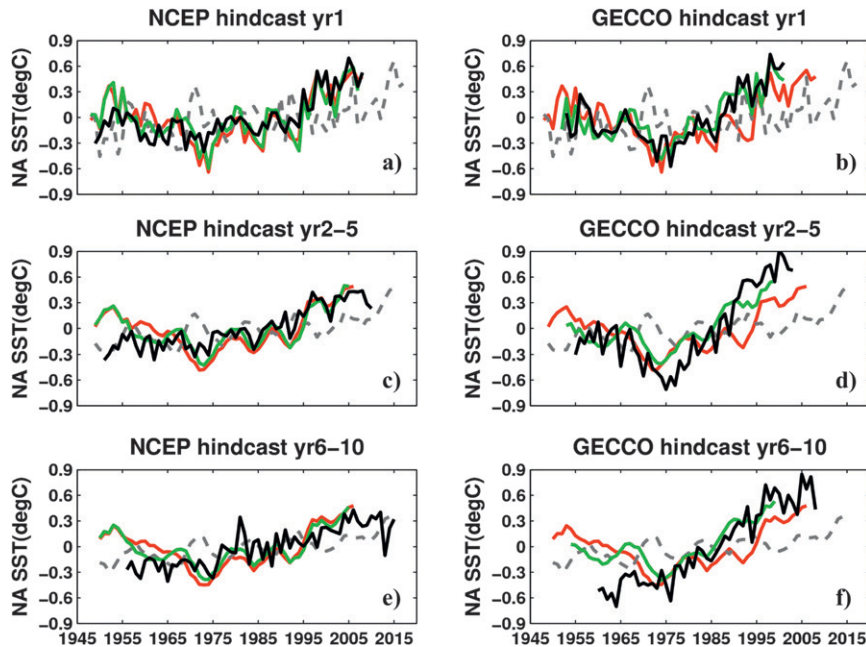
FIG. 4. Time series of the North Atlantic SSTA in observations (red, from HadISST), assimilation (green), NonINIT (gray dashed), and hindcast (black) experiments at lead times (top) yr1, (middle) yr2–5, and (bottom) yr6–10. North Atlantic SSTA is defined as area averaged over the region 20°–60°N, 50°–10°W. From observations, assimilations and NonINIT experiments 4-yr (5-yr) means, centered at year 2 (year 3), are shown in the middle (bottom) panels; for the yr2–5 lead time, the hindcast time series are plotted centered in yr3; for the yr6–10 lead time, the hindcast time series are plotted centered in yr8. Results are shown from the (left) NCEP and (right) GECCO initialized hindcasts.

the second pentad the errors in forecasting tropical SSTs have been greatly reduced in the NCEP prediction system, while not changing too much between the first (yr2–5) and second pentads (yr6–10) in the GECCO prediction system.

The fact that anthropogenically forced climate change plays an important role in near-term tropical climate variations, as shown by the results of our decadal predictions over the Indian–western tropical North Pacific, comes as no surprise. Analyzing very long control integrations with 21 state-of-the-art coupled climate models, Boer and Lambert (2008) concluded that the forced response is largest over the tropical oceans, while internal variability dominates the decadal potential predictability over the extratropical and high-latitude oceans. Furthermore, the "crossover" time at which forced response predictability becomes more important than the initial value predictability was estimated by a recent potential predictability study (Branstator and Teng 2010) to be much shorter for the tropical Pacific (~2 yr) than for the midlatitude ocean regions (7–11 yr). This can be due to a different signal-to-noise ratio of interannual-to-decadal climate variability in the tropical and extratropical regions, but also to the different mechanisms involved. The longest time scale natural potential predictability (not

radiatively forced) is found over the mid- to high-latitude ocean regions where the surface climate is connected to the deeper ocean (Boer 2008). This is indeed the case for the ECHAM5/MPI-OM coupled model, too (not shown).

b. Regional improvement in SST predictive skill

So far we have dealt with the predictive skill of pointwise SST variations. We now turn to the discussion of regionally aggregate SST predictions, evaluating the impact of the two different ocean initializations on the North Atlantic and Mediterranean SST predictive skill. By focusing on the predictability of spatially aggregate quantities, we hope to improve the signal-to-noise ratio of the hindcasts.

The North Atlantic SST (NA SST) interannual fluctuations are well captured by the initialized hindcasts (Figs. 4a,b), with the highest COR skill being obtained in the NCEP hindcasts (COR skill 0.75 for NCEP vs 0.59 for GECCO). In contrast, the NonINIT runs do simulate NA SST variations of comparable amplitude to the observed ones on top of a continuous warming trend, but of completely different phase (COR 0.05). Therefore, the initialized hindcasts are considerably more skillful in predicting the North Atlantic SST variations not only
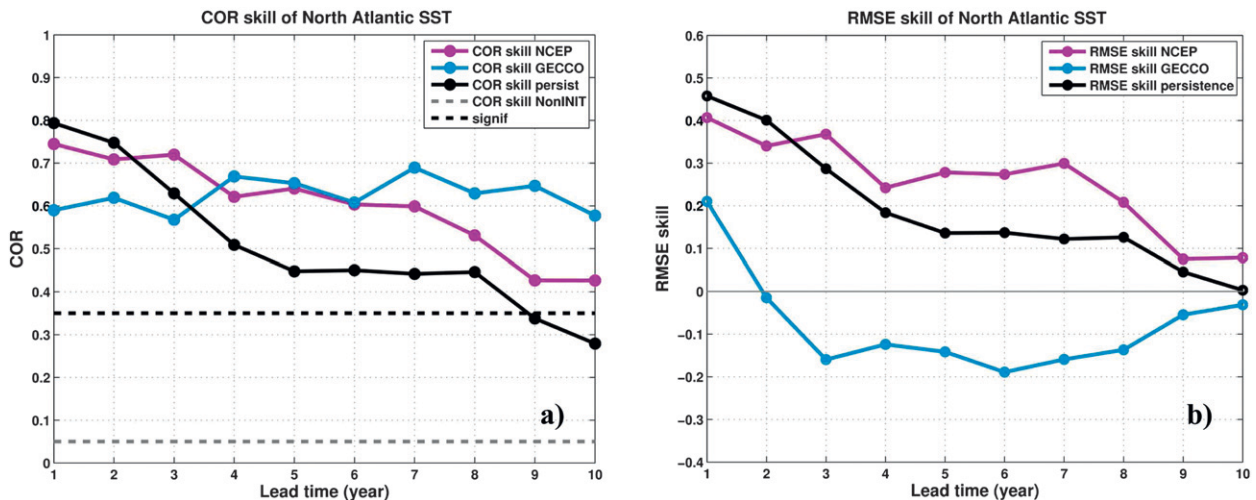
FIG. 5. (left) Anomaly correlation coefficient between the North Atlantic SST in observations (from HadISST) and hindcasts at lead times 1 to 10 years from the NCEP system (violet) and GECCO system (blue). North Atlantic SSTA is defined as area averaged over the region 20°–60°N, 50°–10°W. The predictive skill of persistence forecast is shown in solid black and of NonINIT experiments in dashed gray. The dashed black line represents the 5% significance level computed according to a one-sided t test. (right) RMSE skill score of hindcasts at lead times 1 to 10 years from the NCEP system (violet) and GECCO system (blue). The predictive skill of persistence forecast is shown in solid black. The RMSE skill score is referenced to the RMSE of the NonINIT experiments.

because of the initialization of natural fluctuations, but also because of a much more accurate representation (or a correction) of the climate system response to the external radiative forcing that includes both greenhouse gases and sulfate aerosol. The latter, through both the direct and indirect aerosol effects, was suggested to play an important role in driving the twentieth-century observed variability within the North Atlantic (Booth et al. 2012). Since we are using a coupled model that does not include any of the aerosol indirect effects, our non-initialized historical simulations do not accurately simulate the externally forced climate variations over the North Atlantic region.

At longer lead times, NA SST variations are predicted with significantly enhanced COR skill by both NCEP (yr2–5: 0.85; yr6–10: 0.73) and GECCO (yr2–5: 0.80; yr6–10: 0.88) initialized hindcasts, compared to the NonINIT experiments (~0) (Figs. 4c–f). However, there are important differences between the results of the two initialization approaches. The highest COR skill over the first pentad (yr2–5) is found in the NCEP-forced hindcasts and over the second pentad (yr6–10) in the GECCO hindcasts (Figs. 4c,f). While the GECCO hindcasts follow the observed variability much better than do the NonINIT simulations, the GECCO hindcasts are too cold during the 1970s (in the yr2–5 hindcasts) and too warm after the mid-1980s (Fig. 4d). This will result in a much bigger RMSE of the NA SST index in the GECCO hindcasts compared to the NCEP-forced hindcasts. The less good agreement between the GECCO NA SST

hindcasts and the observations is partially due to an overestimation of the warming trend from the mid-1970s onward in the GECCO assimilation experiment. This overestimated trend, which can be traced back to a strong positive trend in the AMOC in the GECCO ocean reanalysis, leads to a high COR skill score accompanied by a high RMSE.

We now investigate the NA SST predictability at lead times from 1 to 10 years ahead, using as a comparison the skill not only of the NonINIT experiments but also of the persistence forecast. The NA SST COR skill scores of the initialized hindcasts are far superior to those of the NonINIT simulation for all lead times (Fig. 5a). Again, the highest COR skill is found in the NCEP hindcasts over the first pentad and in the GECCO hindcasts during the second pentad. The NCEP hindcast has a similar COR skill as the persistence forecast during the first two years, confirming that the effects of the persistence are strong at shorter lead times. The fact that the COR skill of the GECCO initialized hindcasts is slightly below the level of persistence during the first three years is once more attributable to the less successful SST initialization in the GECCO approach. At longer lead times (yr4 to yr10), both ocean initializations lead to a significant enhancement in the predictive skill compared to the persistence forecast. Using RMSE as a measure of predictive skill, we find substantial differences between the two sets of NA SST initialized hindcasts (Fig. 5b). The COR and RMSE results are consistent only in the case of the
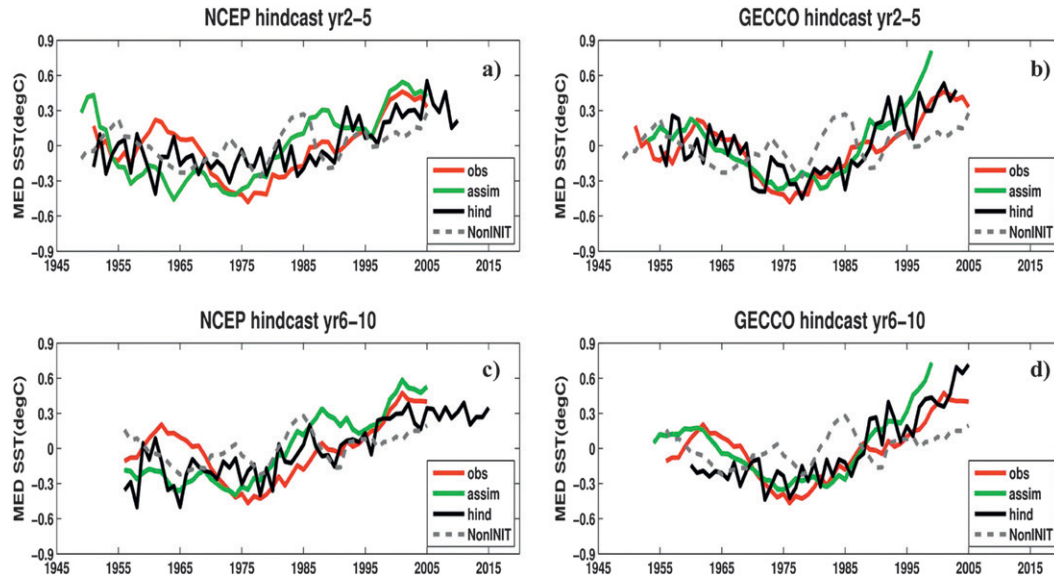
FIG. 6. As in Fig. 4, but for the Mediterranean SSTA. Mediterranean SSTA is defined as area averaged over the region 29°–42°N, 0°–35°E. From observations, assimilation and NonINIT experiments 4-yr (5 yr) means are shown in the top (bottom) panels.

NCEP initialized hindcasts. The RMSE skill score of the GECCO hindcasts is lower than the RMSE skill of the persistence forecast for all lead times; it is even below the RMSE skill of the NonINIT experiment (the zero line in Fig. 5b) at lead times longer than two years. Again, the reason for this poor RMSE skill is the overestimated warming trend in the second half of the GECCO assimilation.

In a second example of regional predictive skill, we focus on one of the projected "hot spot" areas under future climate change conditions: the Mediterranean basin (Meehl et al. 2007). Mariotti et al. (2008) found both observational and modeling evidence that Mediterranean SST variations can greatly alter the Mediterranean water cycle characteristics and thus affect the land dryness of the neighboring regions and potentially precipitation in remote regions (e.g., the Sahel). The results of our initialized hindcasts demonstrate that the Mediterranean SST variations can be skillfully predicted up to a decade ahead (Fig. 6), with far greater skill than by the NonINIT simulations. The highest COR skill score over the first pentad is found in the GECCO hindcasts (0.84 compared to 0.74 in NCEP hindcasts). For the second pentad, the two initialization methods lead to comparable COR skill scores: 0.87 in GECCO and 0.84 in NCEP. Investigating the sensitivity of the Mediterranean predictive skill on the index choice, we find that the bulk of the skill comes in both prediction systems from the western part of the Mediterranean Sea (not shown), a region strongly influenced in the model

by AMOC and SPG variations. However, the Mediterranean SST hindcasts in the GECCO approach have a smaller RMSE due to a better representation of the observed variability in the GECCO assimilation compared to the NCEP assimilation experiment over the eastern Mediterranean (Figs. 1c,g). The substantial differences between the two assimilation experiments demonstrate that a direct initialization of the ocean subsurface (as in the case in the GECCO approach) is required to maximize the skill of Mediterranean SST predictions.

Another interesting aspect is represented by the time evolution of the Mediterranean SST predictive skill, with the modest skill of yr1 being followed by a strong increase in skill for longer lead times (Figs. 2a and 3a vs Figs. 2d,g and 3d,g). This leads to the question of whether the Mediterranean SST predictive skill is due to local processes or is remotely driven. There is a striking similarity between the observed evolution of both the North Atlantic and Mediterranean SST, indicative of a strong linkage between the SST fluctuations in both regions. Indeed by analyzing the observed SST records, Marullo et al. (2011) found that the North Atlantic Ocean, and particularly the subpolar gyre region, is the only area of the World Ocean that exhibits the same multidecadal oscillation observed in the Mediterranean SST. This suggests that the good decadal prediction skill of the Mediterranean SST might be a consequence of skillfully predicting the North Atlantic SST fluctuations, both internally generated [the Atlantic multidecadal

oscillation (AMO)] or externally driven (the recent warming trend). However, the mechanism responsible for this teleconnection from the North Atlantic sector to the Mediterranean region is at present not fully understood. A couple of recent studies (Chronis et al. 2011; Bladé et al. 2011; Marullo et al. 2011) have presented observational evidence for strong atmospheric teleconnection from the North Atlantic to the Mediterranean region during both the winter and summer seasons. This atmospheric teleconnection has the biggest impact on the eastern Mediterranean region and involves the dominant mode of atmospheric variability in the Atlantic sector, the North Atlantic Oscillation (NAO).

c. SAT predictive skill

Predicting SST variations is important since it is through the ocean–atmosphere interface that ocean fluctuations are transmitted to the atmosphere and, through atmospheric teleconnections, influence the climate of remote regions. Therefore SST predictions can be successfully used in constructing statistical prediction models or forcing AGCM-only experiments. However, the level of climate predictability over land has a more direct relevance to society and decision makers.

Figures 2a and 3a show that in the first year the land surface air temperature (SAT) fluctuations can be predicted with modest skill over North Africa, over the high latitudes of the North American continent, and in the NCEP hindcasts over midlatitude East Asia. For years 2–5, there is considerable intensification of skill over northern Africa, the eastern Mediterranean–Middle East region, eastern Asia, and a limited region in the southwestern part of the United States (Figs. 2d and 3d). In addition to these regions, we find significant skill scores in the GECCO hindcasts over western Europe, Scandinavia, and the northern–northeastern parts of North America. As in the case of SST predictability, the negative SAT skill (or lack of skill) over the Southern Hemisphere land regions can be attributed to the negative predictive skill of the tropical SST at the same lead time. At longer lead times (yr6–10), the picture becomes more coherent, with both prediction systems showing high predictive skill for many regions: northern and western Africa, western and central Europe, central and eastern Asia, and northeastern and parts of southwestern North America (Figs. 2g and 3g). Again, as in the case of Mediterranean SST, the counterintuitive increase in the predictive skill with lead time suggests that the skill originates from both atmospheric teleconnections from the North Atlantic and the radiatively forced warming trend.

One limitation of our study is represented by the fact that our sets of initialized hindcasts, albeit sampling the whole range of initial conditions between 1953 and 2008,

consist of only a single realization for every initial state and therefore might provide a lower limit of predictability. This may be especially true for a quantity with such a high noise level as SAT. We check the impact of an increased ensemble size on SAT predictability by further analyzing the skill scores of a three-member ensemble of NCEP hindcasts that follow the CMIP5 protocol (i.e., initialized every five years between 1960 and 2005). Comparing the results of yearly initialized hindcasts versus the CMIP5 subsampled hindcasts, one needs to be aware of the substantially difference in the significance level between the two sets of hindcasts, resulting from almost 5 times more initial dates in the former ensemble than in the latter, and different levels of autocorrelation.

Figure 7 reveals that the spatial distribution of the SST–SAT COR skill for the CMIP5 ensemble-mean hindcasts is in general similar to the one obtained using only one hindcast per initial state. However, the ensemble mean delivers a much higher skill score over the North Atlantic region, while the skill over the eastern tropical Pacific is somewhat reduced (cf. Figs. 7 and 2). Moreover, starting every fifth year with a multimember ensemble provides the lowest RMSE over the whole globe (not shown). The ensemble mean also shows significant predictability over areas where there was none in the one-member ensemble. This is the case for surface temperature variations over the central western United States at lead time one year and northwestern–central Europe at lead time yr2–5. Very interestingly, the spatial distribution of ensemble-mean SAT predictive skill at intermediate lead times and its enhancement over the NonINIT simulations (Figs. 7c,d) closely resemble the AO–NAO fingerprints (Hurrell et al. 2003). To better understand the mechanism through which the North Atlantic climate variability controls the predictability of land surface temperature further investigations are required, involving, for example, seasonal stratified analyses of the predictive skill. However, this is beyond the scope of the present paper.

In conclusion, our results indicate that a multimember ensemble of initialized hindcasts leads to a higher signal-to-noise ratio of the internal variability and consequently to a higher level of surface climate predictability, especially over the extratropical ocean and land regions. However, the very limited sample size of the CMIP5 near-term experimental setup makes it very difficult to obtain statistical significant results. Moreover, the CMIP5 subsampling may also overestimate the contribution of the forced component to the decadal predictive skill of climate quantities strongly impacted by the global warming trend, such as surface temperature or OHC, particularly when using COR as an evaluation metric (not shown).
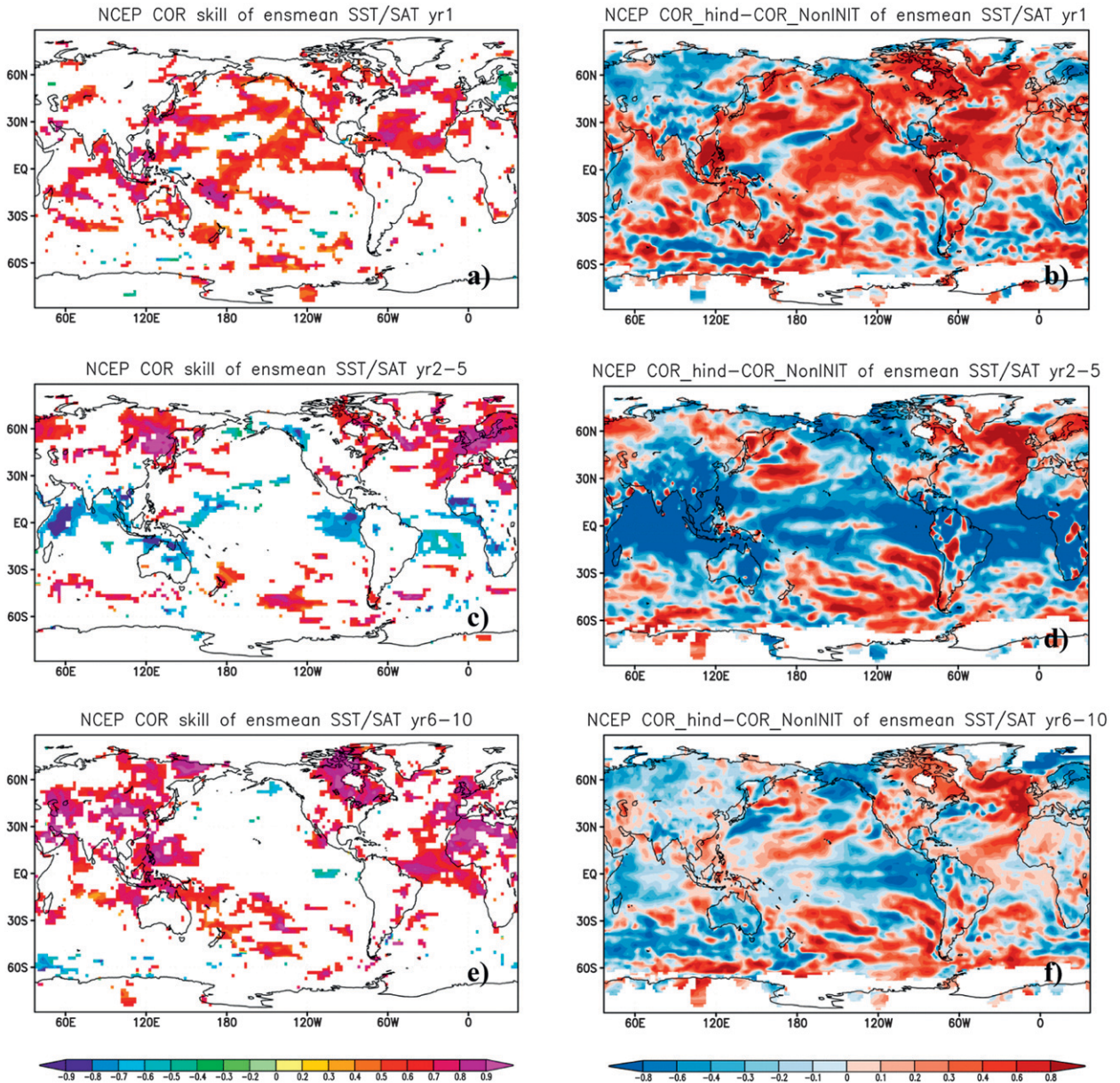
FIG. 7. Ensemble mean surface temperature predictive skill of the NCEP initialized hindcasts at lead times (top) yr1, (middle) yr2–5, and (bottom) yr6–10. Results are shown from a three-member ensemble of hindcast experiments that are initialized every fifth year following the CMIP5 experimental setup. (left) Anomaly correlation coefficient (COR) between the observations and hindcasts. Only the significant COR (at 5% level) are plotted. (right) The difference in COR skill score of hindcasts and NonINIT experiments.

Therefore, an optimal experimental setup for the decadal predictions will consist of multimember ensemble hindcast experiments that sample as many initial conditions as possible (i.e., initialized at least once every year), thus leading to a proper initialization of interannual climate variations (like those associated with ENSO and NAO) and a reduced uncertainty in forecast quality estimates due to the increased sample size (Doblas-Reyes et al. 2011).

6. Decadal prediction skill of upper ocean heat content

a. Global distribution of skill

The evolution of global upper-ocean heat content (OHC) during the past 50 years exhibits clear decadal–multidecadal variations in addition to the strong warming trend attributed to the increase in anthropogenic greenhouse gases (Levitus et al. 2009). Therefore,

there is the expectation that through ocean initialization a more accurate prediction of the next decade might be achieved. In the following, we explore the predictability of the OHC anomalies, focusing at first on the global distribution of skill.

Skillful OHC predictions can be made at 1-yr lead time over several regions of the Atlantic and Pacific Oceans (Figs. 8a,b). We find similarities but also differences between the two initialization approaches. On the one hand, both initialized hindcasts show increased COR skill over the western tropical Pacific and central North Pacific. On the other hand, only the NCEP hindcasts show large COR skill over the tropical and extratropical North Atlantic (Fig. 8a). While the tropical and northern Atlantic OHC COR skill improvement can be attributed to the persistence of the heat content anomalies, the skill enhancement in the western tropical Pacific is due to the improved ENSO predictions during the first half year (not shown). The RMSE skill in predicted upper-ocean heat content variations is significantly higher than the RMSE skill in the NonINIT simulations over almost the whole global ocean, with the exception of the Southern Ocean (Figs. 8e,f).

At longer lead times (yr2–5), high OHC COR skill is found over the tropical and extratropical North Atlantic and over the central North Pacific, with the largest predictive skill shown by the NCEP hindcasts (Figs. 8c, d). The spatial distribution of the COR skill resembles a superposition of the heat-content fingerprints of the dominant modes of decadal variability in the two ocean basins, the AMO and PDO. This does not necessarily imply predictive skill for the two well-known decadal phenomena, but rather shows that the persistence of the heat content variations associated with them may give rise to predictability up to several years in the future. Investigating now the OHC RMSE skill at the same lead time yr2–5, we see that areas showing an increase in RMSE skill in the initialized experiments relative to NonINIT experiment are limited to the extratropical North Atlantic and North Pacific (Figs. 8g,h). In the tropical regions, the ocean initialization leads to either no improvement or even a degradation of RMSE skill. At even longer lead times (yr6–10), regions with high COR values are still found over the extratropical North Atlantic in the NCEP experiments and over the central North Pacific in the GECCO experiments (not shown).

b. Hindcasting North Atlantic upper-ocean heat content

We now examine in more detail the regional predictive skill for the upper-ocean heat content, focusing on the North Atlantic Ocean. Our results indicate that through ocean initialization a significant improvement in the predictive skill of North Atlantic OHC (NA OHC) can be obtained in comparison to the NonINIT experiments (Figs. 9a,b). The high NA OHC COR skill scores in both hindcast experiments clearly exceed the hindcast skill of NonINIT for all lead times (Fig. 9c). However, even though the skill scores of NA OHC in the initialized hindcasts are high, they lie just below the COR level of the persistence forecast. The generally high OHC persistence is amplified by the strong warming trend that has started around 1970. Nevertheless, the NCEP-initialized hindcasts are successful in reproducing the evolution of NA OHC during the last decade: the plateau around 2004 and the strong decrease thereafter (Fig. 9a). However, all our model simulations, both the hindcasts and the assimilation experiment, seem to underestimate the NA OHC trend seen in the NODC estimate. On the other side, the very strong increase in the NA OHC around the year 2001 that coincides with the transition period from an ocean temperature record mostly consisting of XBT data to the one dominated by the ARGO floats around 2001 might also be overestimated in the NODC OHC estimate.

When comparing the results between the two sets of hindcasts, the largest predictive skill for the North Atlantic OHC is obtained by the NCEP experiments for lead times up to four years and by the GECCO experiments for the second pentad (Fig. 9c). Therefore, in our decadal prediction systems the North Atlantic SST and OHC have similar levels of predictability. This raises the question of whether North Atlantic SST predictability arises from the initialization of the OHC or whether both the SST and the OHC skills originate from the initialization of the AMOC. We address this issue further in the next section.

7. Mechanisms giving rise to predictive skill

a. Predictability beyond the global warming trend

We now consider the predictive skill of the linearly detrended surface temperature variations, focusing on the longer lead times (yr2–5 and yr6–10). The extratropical North Atlantic Ocean and particularly the subpolar gyre stand out as regions where ocean initialization leads to significant additional predictability beyond that arising from the global warming trend (Figs. 10a–d). Mediterranean SST fluctuations beyond the trend can be skillfully predicted, too, especially by the GECCO hindcasts at intermediate forecast lead times. Besides the skill in predicting SST variations, both initialized hindcasts indicate some skill in predicting SAT fluctuations over northern Africa and the Arabian Peninsula, western Europe and the British Isles, and over parts of eastern Asia.
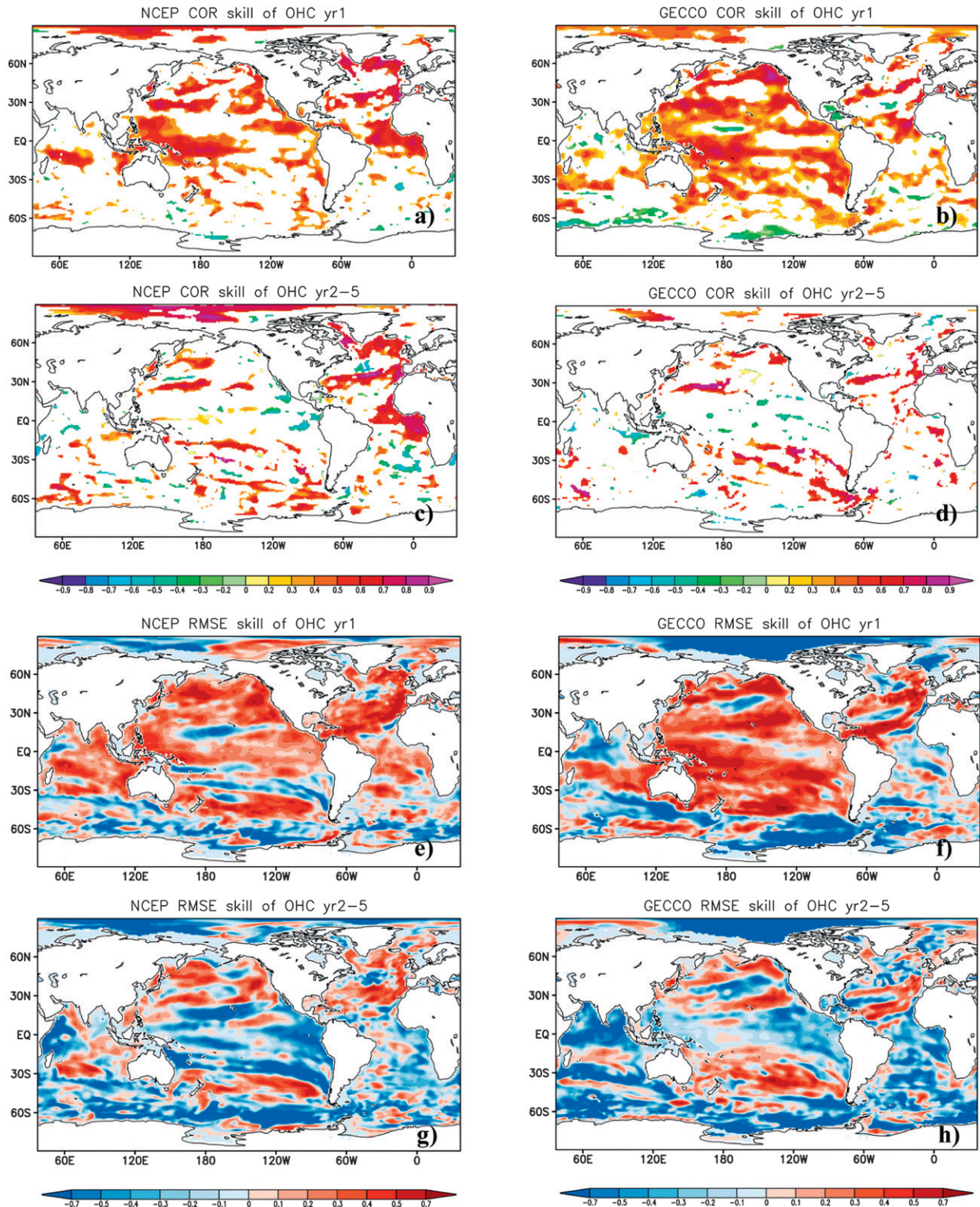
FIG. 8. Upper-ocean heat content predictability in the initialized hindcasts. Results are from the (left) NCEP and (right) GECCO initialized hindcasts. (a)–(d) Anomaly correlation coefficient (COR) between the observations (from NODC) and hindcasts at lead times (top) yr1 and (second row) yr2–5. Only the significant COR (at 5% level) are plotted. (e)–(h) RMSE skill score of hindcasts referenced to the NonINIT experiments at lead time (third row) yr1 and (bottom) yr2–5.
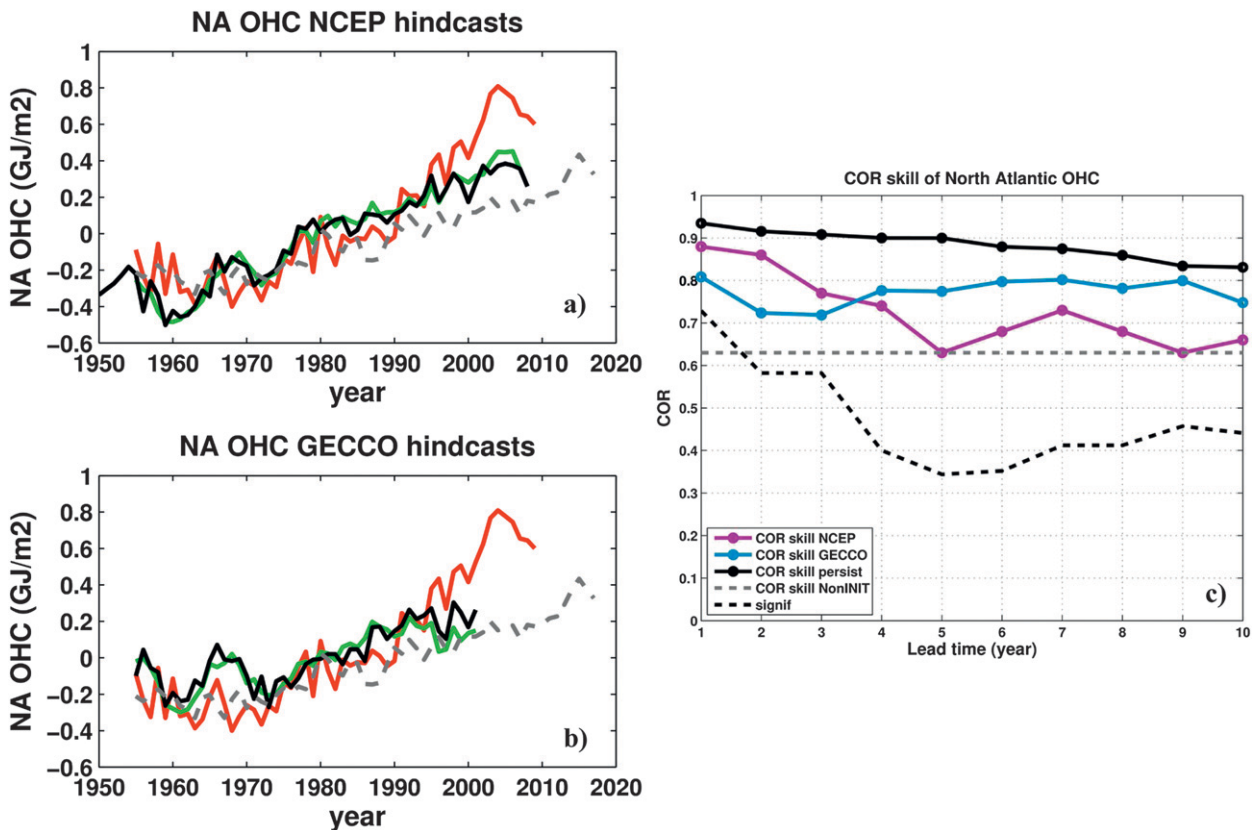
FIG. 9. North Atlantic upper-ocean heat content (NA OHC) variability in observations and hindcasts. North Atlantic OHC is defined as area averaged over the region 0°–60°N, 80°W–0°. (top left) NA OHC in NODC observations (red), NonINIT simulation (dashed gray), NCEP assimilation (green), and NCEP hindcasts at lead time one year (black). (bottom left) NA OHC in NODC observations (red), NonINIT simulation (dashed gray), GECCO assimilation (green), and GECCO hindcasts at lead time one year (black). (right) Anomaly correlation coefficient between the NA OHC variations in observations and hindcasts at lead times 1 to 10 years from the NCEP system (violet) and GECCO system (blue). The predictive skill of persistence forecast is shown in solid black and of NonINIT experiments in dashed gray. The dashed black line represents the 5% significance level computed according to a one-sided t test.

The regional distribution of COR skill resembles the well-known fingerprints of the observed (Knight et al. 2005, 2006) and modeled AMO (not shown). Therefore, we investigate in more detail the predictive skill of the SST variations over the subpolar gyre region, a very good proxy for AMO. We find that the detrended SST fluctuations over the North Atlantic subpolar gyre (SPG SSTD) can be skillfully predicted for a decade ahead by the NCEP prediction system (Fig. 10e). While the predictive skill for SPG SSTD lies at the same level as the persistence forecast for the first 4 years, it is well above the persistence level during the second pentad. An intriguing feature is the reoccurrence in the predictive skill during the second pentad, with a second maximum at 7-yr lead time. An even more peculiar behavior is exhibited by the evolution of SPG SSTD predictive skill in the GECCO hindcasts, where we find relatively modest skill during the first pentad that lies only slightly above the skill level of the NonINIT experiments,

followed by a strong increase in skill during the second pentad.

To gain more understanding of the mechanism behind the decadal hindcast skill, we examine the predictability of OHC variations beyond the linear trend. Figures 11a and 11b identify the subpolar North Atlantic as a distinct region of high predictive skill out to 10 years ahead. Outside the North Atlantic, high predictability is found only in a limited area around 30°N in the western North Pacific. Therefore we further analyze the predictive skill of natural (linearly detrended) SPG upper ocean heat content fluctuations (SPG OHCD), suggested to constitute a distinct fingerprint of the AMOC strength (Zhang 2008). The uninitialized hindcasts, albeit having a level of variability similar to the observed one, do not follow the observed history of the SPG OHCD (COR: −0.42). The results show that in the NCEP-initialized hindcasts the SPG OHCD variations can be skillfully predicted up to a decade ahead at levels well above that
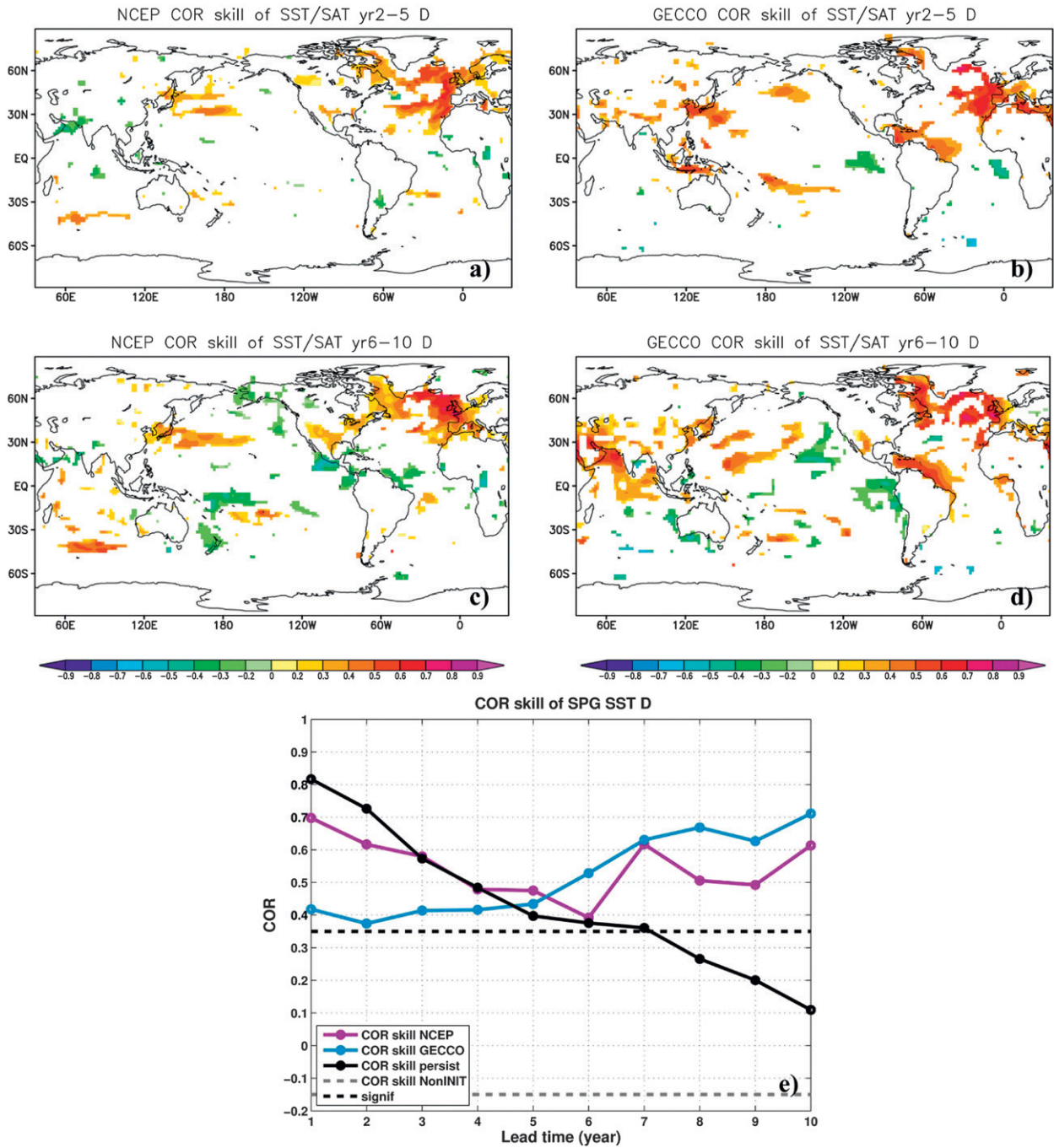
FIG. 10. Surface temperature predictability beyond the global warming trend. Anomaly correlation coefficient (COR) between the linearly detrended observations and hindcasts at lead times (a),(b) yr2–5 and (c),(d) yr6–10. The observations were taken from HadISST for the SST and GHCN–CAMS for SAT. Only the significant COR (at 5% level) are plotted. Results are shown from the (left) NCEP and (right) GECCO initialized hindcasts. (e) Anomaly correlation coefficient between the linearly detrended North Atlantic Subpolar Gyre SST (50°–70°N, 60°W–0°) in observations (from HadISST) and hindcasts at lead times 1 to 10 years from the NCEP system (violet) and GECCO system (blue). The predictive skill of persistence forecast is shown in solid black and of NonINIT experiments in dashed gray. The dashed black line represents the 5% significance level computed according to a one-sided *t* test.
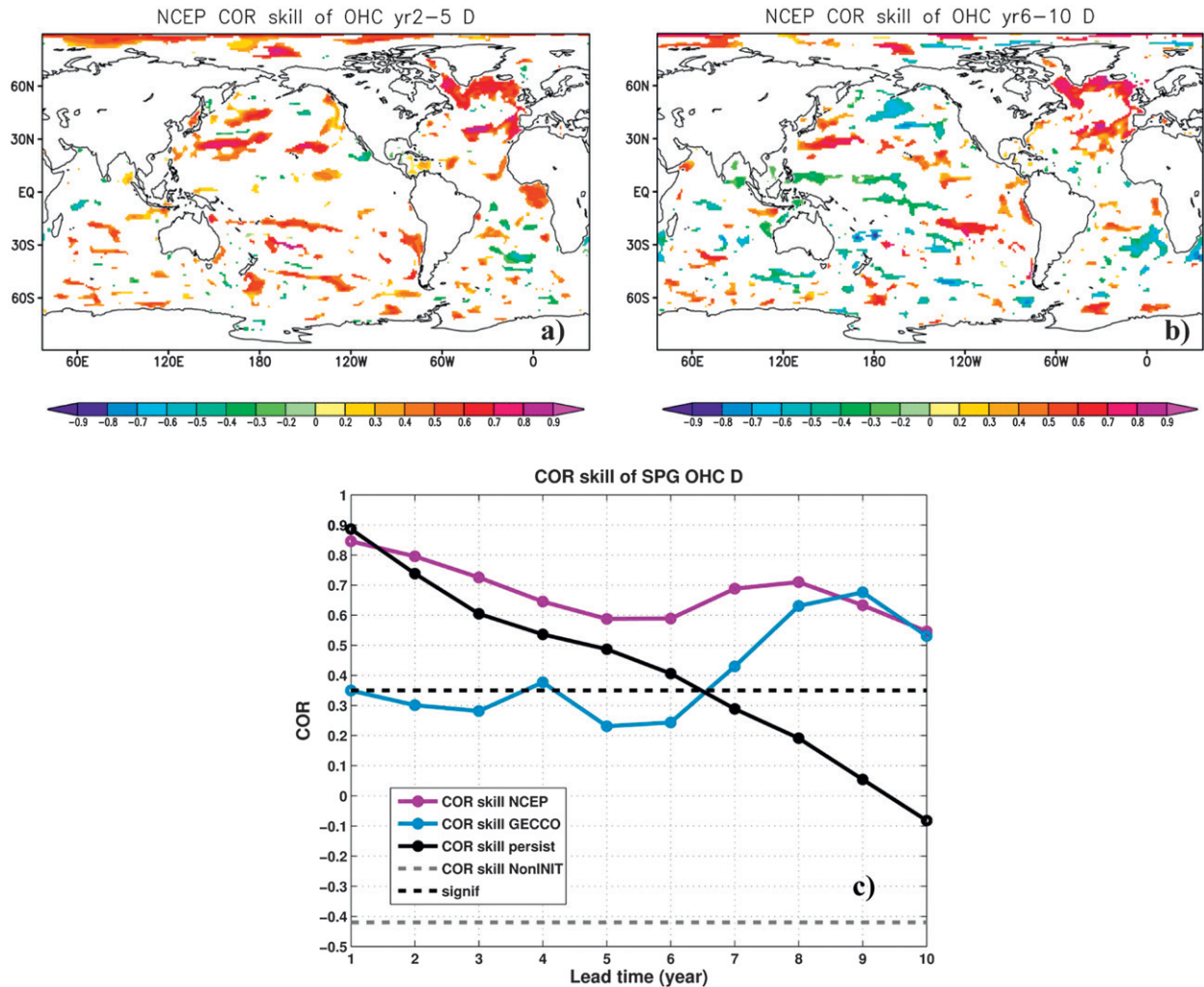
FIG. 11. Upper-ocean heat content predictive skill beyond the global warming trend. Anomaly correlation coefficient (COR) between the linearly detrended observations from NODC and the NCEP hindcasts at lead times (a) yr2–5 and (b) yr6–10. Only the significant COR (at 5% level) are plotted. (c) Anomaly correlation coefficient between the linearly detrended North Atlantic Subpolar Gyre OHC (50°–70°N, 60°–15°W) in observations (from NODC) and hindcasts at lead times 1 to 10 years from the NCEP system (violet) and GECCO system (blue). The predictive skill of persistence forecast is shown in solid black and of NonINIT experiments in dashed gray. The dashed black line represents the 5% significance level computed according to a one-sided t test.

of persistence (Fig. 11c). A reoccurrence in skill at longer lead times, albeit slightly weaker, can be also seen here. The hindcasts initialized from the GECCO reanalyses exhibit again a different evolution of SPG OHCD hindcast skill when compared to the NCEP hindcasts, with modest skill [COR $\sim$ (0.3–0.4)] in the first pentad followed by a strong increase in skill in the second pentad. The significant difference in the predictive skill of the two hindcast systems can be traced back to the relatively large disagreement between the SPG OHC in observations and in the GECCO assimilation experiment during the 1985–95 period (not shown), which results in inaccurate initial conditions for the hindcast experiments.

b. Predictive skill and the AMOC

The fact that both the North Atlantic SST and OHC have a similar evolution of the forecast skill points to a common physical mechanism that extends the predictive skill in the second pentad beyond the skill of persistence. One potential candidate is the AMOC, which is known to have a strong impact on climate variability of the extratropical North Atlantic in both observations and models (Latif et al. 2006; Knight et al. 2005). Moreover, the AMOC strength shows robust potential predictability at 45°N (Pohlmann et al. 2012, manuscript submitted to *Climate Dyn.*) and can be
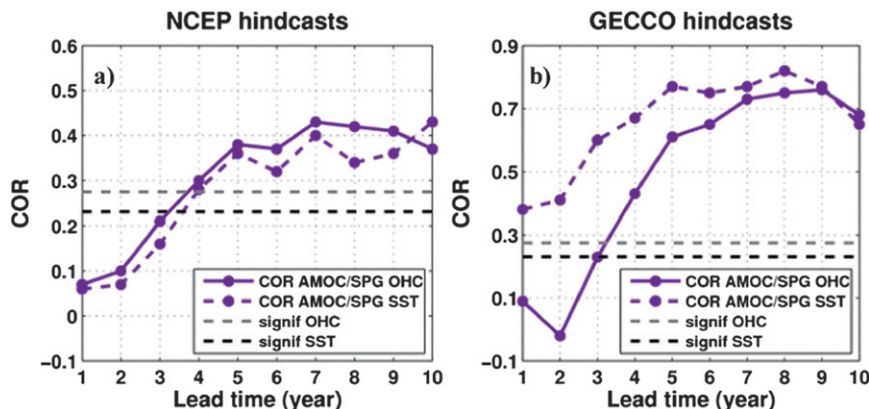
FIG. 12. (a),(b) COR between the AMOC variations and the SPG SST are shown in solid violet, while the COR between AMOC variations and the SPG OHC are depicted by the dashed violet line. The dotted gray lines represent the 5% significance level computed according to a one-sided t test for SST (lighter gray) and OHC (darker gray). Results are from the (left) NCEP and (right) GECCO initialized hindcasts. AMOC strength is the zonally integrated northward flow at 26.5°N latitude above 1000-m depth.

predicted up to 4 years ahead at 26.5°N (Matei et al. 2012). In both prediction systems we find a significant correlation between the yr1 AMOC at 26.5°N and SPG SST and OHC in year 4 and later (Fig. 12). This correlation implies that a correct initialization of AMOC variations significantly impacts the SPG SST and OHC fluctuations 5 to 10 years later. Furthermore, there is a remarkable correspondence between the evolutions of COR curves displayed in Fig. 12 and of the SPG predictive skill during the second pentad of the hindcasts (Figs. 10e and 11c). Accordingly, in the NCEP hindcasts the AMOC impact on the SPG peaks after about 7 years. On the other hand, the strong impact of AMOC variations on the surface and subsurface variability of the SPG in the GECCO hindcasts explains the sharp increase in skill seen during the second pentad of the GECCO hindcasts. These results strongly indicate that AMOC-induced meridional heat transport variations are responsible for the decadal predictive skill in the North Atlantic.

Therefore, we suggest that in our decadal prediction systems, the predictive skill of SST and OHC fluctuations at longer lead times is a consequence of initializing the AMOC variability, while the skill in the first five forecast years can be attributed to SST and OHC persistence. The relatively low predictive skill of SPG SSTD in the first pentad of the GECCO hindcasts is a result of the GECCO initialization leading to only a modest correction of the initial state over the SPG region, when compared to the NonINIT experiments (Figs. 1b,e). On the other hand, the NCEP initialization captures well the observed variability of both SST and OHC over the SPG region (Figs. 1a,d), thus leading to very good prediction skill over the first five forecast years.

8. Summary and discussion

We have presented two different approaches for the initialization of decadal climate predictions with the ECHAM5/MPI-OM atmosphere–ocean–sea ice coupled model: in the first case, by an ocean reanalysis (namely GECCO), and in the second, by an ensemble of MPI-OM ocean experiments forced with the observed atmospheric state taken from the NCEP–NCAR reanalysis. The latter initialization method is currently employed at the MPI-M for the CMIP5 near-term experiments. We have evaluated the results of the initialized decadal hindcasts of the past 60 years against the observed variations of surface climate (namely SST and SAT) and of the upper-ocean heat content.

We show that the assimilation of ocean surface and subsurface data greatly increases the predictive skill of SST up to a decade ahead over the North Atlantic, central North Pacific, and the Mediterranean region, when compared to both the uninitialized simulations and persistence forecasts. At first order, the predictive skill of GECCO and NCEP-forced hindcasts is regionally similar. However, as a consequence of a more accurate initialization of the North Atlantic SST and OHC variations in the NCEP assimilation, a higher SST skill is obtained for NCEP-forced hindcasts in the first pentad.

There are also regions, namely the tropical oceans, where at lead times longer than one year the initialized hindcasts do not outperform the uninitialized simulations (as is the case of the Indian Ocean–western tropical North Pacific region during the second pentad) or even show a substantial degradation in skill (as for the tropical eastern Pacific at intermediate lead times). This is partially due to

the fact that the anthropogenically forced climate change seems to play a dominant role in the evolution of tropical climate at the decadal time scale (Boer 2009) and therefore most of the tropical predictive skill at lead times beyond one year might arise from predicting the warming trend. On the other hand, the lack of tropical SST predictive skill found in the initialized hindcasts at intermediate lead times (yr2–5) can be related to deficiencies, common to all non–flux-adjusted coupled models, in simulating tropical climate variability, such as a too strong and too far westward extended equatorial cold tongue or a too strong (as in our coupled model) or too weak ENSO amplitude (Guilyardi et al. 2009). All these systematic model errors contribute to reducing the potential predictability of tropical SST variations and can also lead to initialization imbalances in both surface and subsurface layers of the tropical oceans when the observed anomalies are superimposed on the imperfect model climatology. Initializing the atmospheric component in addition to the ocean might lead to a more balanced atmosphere–ocean initial state and subsequently to enhanced tropical predictability.

Another remarkable aspect is the different picture given by the two evaluation metrics (COR skill and RMSE skill) in the case of the GECCO hindcasts. For example, at longer lead times a substantially higher COR skill for the North Atlantic SST is accompanied by a decrease in RMSE skill over the same area and especially over the North Atlantic subpolar gyre. These differences between skill measures are due to an overestimation of the global warming trend from the mid-1970s onward in the GECCO assimilation experiment. Therefore, our study underlines the necessity of using more than one verification metric to evaluate the predictive skill of decadal predictions experiments, since different metrics capture different aspects of forecast quality.

We demonstrate predictive skill of land surface air temperature up to a decade ahead over northwestern Europe, the eastern Mediterranean–Middle East region, northern Africa, and the central and eastern parts of Asia. In general, the SAT skill increases with forecast lead time, suggesting that the skill originates from both atmospheric teleconnections from the North Atlantic region and the radiatively forced warming trend. The regional distribution of SAT predictive skill, especially at intermediate lead times, clearly resembles the AO–NAO fingerprints. One can speculate that this result might imply some predictability for the multiyear modulation of the AO–NAO. Further investigations are, however, necessary to fully understand the mechanism through which the North Atlantic climate variability controls the predictability of land surface temperature.

We find a much higher SAT predictive skill of the ensemble-mean initialized hindcasts than for individual hindcasts. This is perhaps counter to the expectation that ensemble averaging would act as a filter for the internally generated climate fluctuations, as is the case for long-term climate simulations. However, this is not the case for our decadal prediction system, in which the ensemble averaging increases the signal-to-noise ratio of the interannual climate variability. This underlines the urgent need to develop methods for large-ensemble generation that can span both the uncertainty in the initial conditions and model errors.

In our decadal prediction experiments, there is good agreement between the regions of greatest predictability and those identified as "potentially predictable" through diagnostic predictability studies (Boer 2009; Ting et al. 2009). The latter try to estimate in a perfect-model framework the ratio between the predictable component of climate variations and the unpredictable "noise" background. Accordingly, the extratropical Northern Hemisphere oceans (especially the North Atlantic and Mediterranean regions) exhibit higher predictability than the tropical oceans. In addition, SAT predictability over land is weaker and restricted to regions strongly impacted by atmospheric teleconnections from the North Atlantic region. However, in the case of the North Pacific SST, the high potential predictability has not been translated in realized predictive skill in our initialized hindcasts. One reason for this might be that the relative strong North Pacific decadal variability in ECHAM5/MPI-OM (not shown) is generated by random processes and therefore is not predictable, or the lack of skill over the tropical Pacific has negatively impacted the predictive skill over the North Pacific region.

The dominant mechanism for North Atlantic surface climate predictability in our decadal prediction system is of dynamical origin and can be attributed to the initialization of the AMOC variations. This explains why the SST variations over the North Atlantic subpolar gyre show a clear reoccurrence of predictive skill during the second pentad—the lead times that corresponds in our model to the maximum impact of the AMOC on SST variations in that region. Therefore, the high predictive skill of the extratropical North Atlantic SST variations can be attributed to the persistence of SST and OHC fluctuations during the first pentad, and to a delayed response of the subpolar North Atlantic to changes in the North Atlantic northward heat transport associated with AMOC variations in the second pentad.

In our decadal prediction system, the North Atlantic subpolar gyre region stands out as the region with the highest predictive skill in the upper-ocean heat content. Especially the NCEP ocean state estimate, which captures the ocean-integrated response to the dominant modes of large-scale atmospheric variability such as the

North Atlantic Oscillation and eastern Atlantic pattern, reproduces very well the observed history of SPG OHC variations and therefore renders the best hindcast skill. The evolution of the detrended SPG OHC predictive skill also reveals an intensification of skill in the second pentad (yr6–10 of the hindcast) when compared to the first pentad, again as a consequence of initializing the AMOC variations.

Predictions of the whole North Atlantic upper-ocean heat content are significantly improved in the initialized hindcasts over the uninitialized simulations. However, in both decadal prediction systems, the NA OHC skill scores of the initialized hindcasts lie slightly below the high persistence level (with COR around 0.9). The very high persistence of the observed NA OHC reflects the strong warming trend of the upper North Atlantic Ocean since the beginning of the 1970s that is super-imposed on rather high-frequency variability that might have very limited predictability.

In this study we have used the GECCO reanalysis as an example for the initialization of decadal predictions from an ocean reanalyses. Therefore, the results pre-sented here should be considered specific to this par-ticular ocean synthesis and not generalized to other ocean reanalyses, since the results might depend on both the fidelity with which the individual ocean reanalysis reproduces the observed variability and the degree of compatibility between the ocean reanalysis and the model used to perform the prediction. Kröger et al. (2012) have investigated the impact of using different ocean reanalyses [GECCO, Simple Ocean Data As-similation (SODA), Ocean Reanalysis System 3 (ORA-S3) XBT corrected] on the skill of decadal predictions performed with a lower-resolution version of our model and a slightly different radiative forcing. They found that the fidelity with which the variations in the three ocean state estimates are adopted by our coupled model is high for both the North Atlantic SST and OHC, es-pecially when using the ORA-S3 ocean reanalysis. However, the assimilation procedure distorts to some degree the AMOC interannual variability and its trend. In their experiments, the GECCO initialization leads to the highest North Atlantic SST predictive skill at time leads longer than one year and the highest AMOC po-tential predictability, while the North Atlantic OHC is best predicted in the ORA-S3 hindcasts.

9. Conclusions

The purpose of this study has been to investigate how the two different ocean initializations—from the GECCO ocean reanalysis and an ensemble of ocean-forced experiments—impact the quality of decadal hindcasts performed with the ECHAM5/MPI-OM coupled model. We conclude the following.

- Both initialization approaches significantly enhance the skill for North Atlantic and Mediterranean SST up to a decade ahead.
- Over land, SAT skill improvement is found over northwestern Europe, the eastern Mediterranean–Middle East region, northern Africa, and central-eastern Asia. The largest skill improvement is obtained with the initialized ensemble mean.
- The North Atlantic subpolar gyre region stands out as the region with the highest predictive skill beyond the warming trend, in both SST and OHC predictions. Here the NCEP hindcasts deliver the best results because of a more accurate initialization of the observed variability.
- The dominant mechanism that extends the predictive skill of North Atlantic SST and OHC in the second pentad, beyond the skill of persistence, is of dynamical origin and can be attributed to the initialization of the AMOC.
- Our results demonstrate that ocean experiments forced with the observed history of the atmospheric state constitute a simple but successful alternative strategy for the initialization of skillful climate pre-dictions over the next decade.

REFERENCES

Bladé, I., B. Liebmann, D. Fortuny, and G. J. van Oldenbourgh, 2011: Observed and simulated impacts of the summer NAO in Europe: Implications for projected drying in the Mediterra-nean region. *Climate Dyn.,* **39,** 709–727, doi:10.1007/s00382-011-1195-x.

Boer, G. J., 2009: Changes in interannual variability and decadal potential predictability under global warming. *J. Climate,* **22,** 3098–3109.

——, and S. J. Lambert, 2008: Multi-model decadal potential pre-dictability of precipitation and temperature. *Geophys. Res. Lett.,* **35,** L05706, doi:10.1029/2008GL033234.

Booth, B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature,* **484,** 228–232.

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

Chronis, T., D. E. Raitsos, D. Kassis, and A. Sarantopoulos, 2011: The summer North Atlantic oscillation influence on the eastern Mediterranean. *J. Climate,* **24,** 5584–5596.

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

Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer, 2011: Decadal climate prediction with the ECMWF coupled forecast system: Impact of ocean observations. *J. Geophys. Res.,* **116,** D19111, doi:10.1029/2010JD015394.

Domingues, C. M., J. A. Church, N. J. White, P. J. Gleckler, S. E. Wijffels, P. M. Barker, and J. R. Dunn, 2008: Improved estimate of upper-ocean warming and multidecadal sea level rise. *Nature,* **453,** 1090–1093.

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.

Guilyardi, E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G. J. van Oldenborgh, and T. Stockdale, 2009: Understanding El Niño in ocean–atmosphere general circulation models: Progress and challenges. *Bull. Amer. Meteor. Soc.,* **90,** 325–340.

Haak, H., 2004: Simulation of low-frequency climate variability in the North Atlantic Ocean and the Arctic. MPI Rep. on Earth System Science No. 1, 115 pp. [Available online at http://www.mpimet.mpg.de/en/science/publications/reports-on-earth-system-science.html.]

——, J. Jungclaus, U. Mikolajewicz, and M. Latif, 2003: Formation and propagation of great salinity anomalies. *Geophys. Res. Lett.,* **30,** 1473, doi:10.1029/2003GL017065.

Hurrell, J. W., Y. Kushnir, M. Visbeck, and G. Ottersen, 2003: An overview of the North Atlantic oscillation. *The North Atlantic Oscillation: Climate Significance and Environmental Impact, Geophys. Monogr.,* Vol. 134, Amer. Geophys. Union, 1–35.

——, and Coauthors, 2010: Decadal climate prediction: Opportunities and challenges. *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society,* J. Hall, D. E. Harrison, and D. Stammer, Eds., ESA Publication WPP-306. [Available online at http://www.oceanobs09.net/blog/?p=97.]

Ishii, M., and M. Kimoto, 2009: Revaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanogr.,* **65,** 287–299.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, 240 pp.

Jungclaus, J. H., and Coauthors, 2006: Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM. *J. Climate,* **19,** 3952–3972.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Keenlyside, N. S., and J. Ba, 2010: Prospects for decadal climate predictions. *WIREs Climate Change,* **1,** 627–635.

——, M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature,* **453,** 84–88.

Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann, 2005: A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.,* **32,** L20708, doi:10.1029/2005GL024233.

——, C. K. Folland, and A. A. Scaife, 2006: Climate impacts of the Atlantic multidecadal oscillation. *Geophys. Res. Lett.,* **33,** L17706, doi:10.1029/2006GL026242.

Köhl, A., and D. Stammer, 2008: Variability of the meridional overturning circulation in the North Atlantic from the 50-year GECCO state estimation. *J. Phys. Oceanogr.,* **38,** 1913–1930.

Kröger, J., W. Müller, and J. S. von Storch, 2012: Impact of different ocean reanalyses on decadal climate prediction. *Climate Dyn.,* **39,** 795–810, doi:10.1007/s00382-012-1310-7.

Latif, M., C. Böning, J. Willebrand, A. Biastoch, J. Dengg, N. Keenlyside, U. Schweckendiek, and G. Madec, 2006: Is the thermohaline circulation changing? *J. Climate,* **19,** 4631–4637.

Levitus, S., and Coauthors, 1998: Introduction. *World Ocean Database 1998,* Vol. I, NOAA Atlas NESDIS 18, 18–42.

——, J. I. Antonov, T. P. Boyer, R. A. Locarnini, H. E. Garcia, and A. V. Mishonov, 2009: Global ocean heat content 1955–2008 in light of recently revealed instrumentation problems. *Geophys. Res. Lett.,* **36,** L07608, doi:10.1029/2008GL037155.

Lyman, J. M., S. A. Good, V. V. Gouretski, M. Ishii, G. C. Johnson, M. D. Palmer, D. M. Smith, and J. K. Willis, 2010: Robust warming of the global upper ocean. *Nature,* **465,** 334–337.

Mariotti, A., N. Zeng, J. Yoon, V. Artale, A. Navarra, P. Alpert, and L. Z. X. Li, 2008: Mediterranean water cycle changes: Transition to drier 21st century conditions in observations and CMIP3 simulations. *Environ. Res. Lett.,* **3,** 044001, doi:10.1088/1748-9326/3/4/044001.

Marsland, S. J., H. Haak, J. Jungclaus, M. Latif, and F. Röske, 2003: The Max Planck Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modell.,* **5,** 91–127.

Marullo, S., V. Artale, and R. Santoleri, 2011: The SST multidecadal variability in the Atlantic–Mediterranean region and its relationship to AMO. *J. Climate,* **24,** 4385–4401.

Matei, D., J. Baehr, J. Jungclaus, H. Haak, W. Müller, and J. Marotzke, 2012: Multiyear prediction of monthly mean Atlantic meridional overturning circulation at 26.5°N. *Science,* **335,** 76–79.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–845.

——, and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Mochizuki, T., and Coauthors, 2010: Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proc. Natl. Acad. Sci. USA,* **107,** 1833–1837.

Munoz, E., B. Kirtman, and W. Weijer, 2011: Varied representation of the Atlantic meridional overturning across multidecadal ocean reanalyses. *Deep-Sea Res. II,* **58,** 1848–1857.

Murphy, J., and Coauthors, 2010: Towards prediction of decadal climate variability and change. *Procedia Environ. Sci.,* **1,** 287–304.

Olsen, S. M., B. Hansen, D. Quadfasel, and S. Osterhus, 2008: Observed and modelled stability of overflow across the Greenland–Scotland ridge. *Nature,* **455,** 519–522.

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. Climate,* **22,** 3926–3938.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of seas surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM5. Part I: Model description.

Max-Planck-Institut für Meteorologie Rep. 349, 140 pp. [Available online at http://www.mpimet.mpg.de/fileadmin/publikationen/Reports/max_scirep_349.pdf.]

Serra, N., R. H. Käse, A. Köhl, D. Stammer, and D. Quadfasel, 2010: On the low-frequency phase relation between the Denmark Strait and the Faroe–Bank Channel overflows. *Tellus,* **62,** 530–550.

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.

Stammer, D., 2006: Report of the first CLIVAR Workshop on Oceanic Reanalysis. WCRP Informal Publication No. 9, ICPO Publication Series Vol. 93, 40 pp.

——, and Coauthors, 2010: Ocean information provided through ensemble ocean syntheses. *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society,* Vol. 2, J. Hall, D. Harrison, and D. Stammer, Eds., ESA Publication WPP-306, 979–989. [Available online at http://www.oceanobs09.net/proceedings/cwp/cwp85/.]

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

Von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

WCRP, 2011: Data and bias correction for decadal climate predictions. International CLIVAR Project Office Publication Series, No. 150, 5 pp.

Yasunaka, S., M. Ishii, M. Kimoto, T. Mochizuchi, and H. Shiogama, 2011: Influence of XBT temperature bias on decadal climate prediction with a coupled climate model. *J. Climate,* **24,** 5303–5308.

Zhang, R., 2008: Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation. *Geophys. Res. Lett.,* **35,** L20705, doi:10.1029/2008GL035463.