Decadal prediction skill in the ocean with surface nudging in the IPSL-CM5A-LR climate model

Juliette Mignot, Javier García-Serrano, Didier Swingedouw, Agathe Germe, Sébastien Nguyen, Pablo Ortega, Éric Guilyardi, Sulagna Ray

Noname manuscript No.
(will be inserted by the editor)

Decadal prediction skill in the ocean with surface nudging in the IPSL-CM5A-LR climate model

Juliette Mignot · Javier García-Serrano · Didier Swingedouw · Agathe Germe · Sébastien Nguyen · Pablo Ortega · Eric Guilyardi · Sulagna Ray

Received: date / Accepted: date

Abstract Two decadal prediction ensembles, based on the same climate model (IPSL-CM5A-LR) and the same surface nudging initialization strategy are analyzed and compared with a focus on upper-ocean variables in different regions of the globe. One ensemble consists of 3-member hindcasts launched every year since 1961 while the other ensemble benefits from 9 members but with start dates

J. Mignot
E-mail: juliette.mignot@locean-ipsl.upmc.fr
Climate and Environmental Physics and Physics Institute, Oeschger Center for Climate Change Research, University of Bern, Switzerland,

J. Mignot · J. García-Serrano · A. Germe · S. Nguyen · P. Ortega · E. Guilyardi · S. Ray
LOCEAN/IPSL (Sorbonne Universités UPMC-CNRS-IRD-MNHN), 4 place Jussieu, F-75005 Paris, France

D. Swingedouw
Environnements et Paléoenvironnements Océaniques et Continentaux (EPOC), UMR CNRS 5805 EPOC - OASU - Université de Bordeaux, Allée Geoffroy Saint-Hilaire, 33615 Pessac, France

only every 5 years. Analysis includes anomaly correlation coefficients and root mean square errors computed against several reanalysis and gridded observational fields, as well as against the nudged simulation used to produce the hindcasts initial conditions. The last skill measure gives an upper limit of the predictability horizon one can expect in the forecast system, while the comparison with different datasets highlights uncertainty when assessing the actual skill. Results provide a potential prediction skill (verification against the nudged simulation) beyond the linear trend of the order of 10 years ahead at the global scale, but essentially associated with non-linear radiative forcings, in particular from volcanoes. At regional scale, we obtain 1 year in the tropical band, 10 years at midlatitudes in the North Atlantic and North Pacific, and 5 years at tropical latitudes in the North Atlantic, for both sea surface temperature (SST) and upper-ocean heat content. Actual prediction skill (verified against observational or reanalysis data) is overall more limited and less robust. Even so, large actual skill is found in the extratropical North Atlantic for SST and in the tropical to subtropical North Pacific for upper-ocean heat content. Results are analyzed with respect to the specific dynamics of the model and the way it is influenced by the nudging. The interplay between initialization and internal modes of variability is also analyzed for sea surface salinity. The study illustrates the importance of two key ingredients both necessary for the success of future coordinated decadal prediction exercises, a high frequency of start dates is needed to achieve robust statistical significance, and a large ensemble size is required to increase the signal to noise ratio.

1 Introduction

Because of the potential socio-economic impacts, decadal climate prediction has developed as a novel topic over the last few years (Meehl et al 2014) and given rise to great expectations. The goal of this exercise is to exploit the predictability of internally-generated climate variability together with that from the externally-forced component, as well as to enhance prediction skill by correcting the forced model response. The 11^{th} chapter of the Intergovernemental Panel on Climate Change (IPCC) fifth assessment report (Kirtman et al 2013) describes the recent scientific achievements on this topic, but also emphasizes that several technical and scientific challenges remain. Although prediction skill arises mostly from external forcing (e.g. Doblas-Reyes et al 2013), initialization of the slow components of the climate system has also provided added value for the first few years of the forecast, most notably in the North Atlantic (e.g. Hazeleger et al 2013b; Corti et al 2012; Kim et al 2012; van Oldenborgh et al 2012; Swingedouw et al 2013; García-Serrano et al 2014). This is at least partly due to the initialization of the Atlantic Meridional Overturning Circulation (AMOC), which shows large inertia in climate models (e.g. Persechino et al 2013). Over the North Pacific, some signs of improved prediction skill through initialization have been found associated with the Pacific Decadal Oscillation (PDO), (Mantua et al 1997) or Interdecadal Pacific Oscillation (IPO) (Keenlyside et al 2008; Meehl et al 2010; van Oldenborgh et al 2012; Meehl and Teng 2012). Mochizuki et al (2010) and Chikamoto et al (2013) showed that models ability to follow the subsurface temperature evolution in the North Pacific increases thanks to initialization. Because of its potential effect on the atmosphere, SST has been the focus of most of these studies and is indeed commonly used as an indicator of the ocean's state in decadal prediction assessments. Nevertheless, subsurface fields are somewhat shielded from weather noise and might thus be expected to be more predictable than the surface fields (e.g. Branstator and Teng 2010), while they might still have the potential to affect the atmosphere on long time scales. Indeed, the oceanic heat content acts as a key indicator of climate perturbations on seasonal, interannual and longer time scales (e.g. Lozier et al 2008), accounting for the total amount of heat variation, through storage and transport, that could potentially be available for the atmosphere. Using a statistical analysis of control simulations, Branstator and Teng (2012) showed that initialization has the potential to improve prediction skill of the upper 300m temperature up to the first 5 years in the North Pacific and 9 years in the North Atlantic.

Initialization techniques are numerous (Kirtman et al 2013), including assimilation of surface information only (e.g. Keenlyside et al 2008; Merryfield et al 2010; Swingedouw et al 2013; Ray et al 2015), restoring to 3-dimensional data (e.g. Voldoire et al 2014; Bombardi et al 2014), forcing of the ocean model with atmospheric observations (Matei et al 2012; Yeager et al 2012) and more sophisticated alternatives based on fully coupled data assimilation schemes (Zhang 2007; Sugiura et al 2009; Karspeck et al 2014). It is yet difficult to distinguish whether one specific method clearly yields enhanced skill, as few studies have focused on comparing different techniques with a single climate model. Noteworthy is the study of Matei et al (2012), who found that hindcast experiments starting from reconstruction simulations forced with the observed evolution of the atmospheric state and associated heat flux over the ocean (including SST information although not explicitly) constitute a simple but skillful strategy for initialized climate predictions over the next decade, as compared to a 3-dimensional restoring towards ocean reanalysis. Bellucci et al (2013) highlighted the strong differences in prediction skill obtained with forecast systems using different ocean data assimilation products. Using perfect model approaches, Dunstone and Smith (2010) and Zhang et al (2010) found, as expected, an improvement in skill when subsurface information is used as part of the initialization. Nevertheless, given the uncertainty in ocean reanalysis below the surface (e.g. Ray et al 2015), several studies also focused on prediction skill using only information from the sea surface (e.g. Keenlyside et al 2008; Merryfield et al 2010). In particular, Kumar et al (2014) and Ray et al (2015) showed that SST nudging is efficient in reconstructing the observed subsurface variability in the equatorial Pacific.

Given climate models usual biases notably in terms of mean state, another question that arises regarding the generation of initial conditions for predictions is the opportunity to use full field or anomaly initialization. In the first case, the coupled model is initialized with a state close to the real-world attractor and after initialization, drifts towards its own attractor. The second case limits this shock, but leads to question the link between mean state and variability. To put it differently, is it possible to properly reconstruct, and predict ENSO variability, for example, even if the warm pool is not correctly located in the model? Magnusson et al (2012), Hazeleger et al (2013a) and Smith et al (2013) show that at decadal time scales, it is difficult to determine whether one of these two strategies is more skillful than the other.

This study aims at assessing prediction skill in the ocean with the IPSL-CM5A-LR climate model initialized via nudging towards observed SST anomalies. As described above, this set up lies on the side of relatively simple initialization techniques. Servonnat et al (2014) investigated the performance of this technique for the reconstruction of subsurface variability in a perfect model configuration using the same climate model. Ray et al (2015) carried similar analysis but under historical conditions and using observations, highlighting the current uncertainty in subsurface ocean variability. Swingedouw et al (2013) showed the skill of the system in reproducing the Atlantic Meridional Overturning Circulation (AMOC) variability and Séférian et al (2014) used it to demonstrate the relatively long forecasting capabilities of the primary production in the tropical Pacific as compared to SST. Here, we provide a more systematic investigation of ocean surface and subsurface predictability of the system. The model, experimental set-up and statistics are presented in section 2. Global and tropical SST prediction skills are described in section 3. Section 4 and 5 concentrate on the prediction skill in the North Atlantic and in the North Pacific respectively. Section 6 discusses issues on sea surface salinity (SSS). Conclusions are given in the final section.

2 Model and methods

2.1 The climate model

We use the Earth System Model IPSL-CM5A-LR (Dufresne et al 2013), developed at the Institut Pierre Simon Laplace (IPSL). The atmospheric model is LMDZ5 (Hourdin et al 2013), with a horizontal resolution of $1.875°$ x $3.75°$ and 39 vertical levels. The ocean model is NEMOv3.2 (Madec 2008), in ORCA2 configuration. This non-regular grid has a nominal resolution of $2°$, refined in the Tropics and the subpolar North Atlantic. The ocean grid has 31 vertical levels. NEMOv3.2 also includes the sea-ice component LIM2 (Fichefet and Maqueda 1997) and the biogeochemical module PISCES (Aumont and Bopp 2006). The performances of the oceanic component in the coupled configuration are discussed in Mignot et al (2013). The reader is referred to the special issue in Climate Dynamics (http://link.springer.com/journal/382/40/9/) for a collection of studies describing various aspects and components of the model as well as its performance for climatic studies. We emphasize here the contribution from Persechino et al (2013) who investigated the model's potential predictability.

2.2 The decadal prediction system

The set of experiments considered here is summarized in Table 1. It first includes a 3-member ensemble of non-initialized historical simulations, all available on the CMIP5 database. They use prescribed external radiative forcing from the observed increase in greenhouse gases and aerosols concentrations, as well as the ozone changes and the land-use modifications. They also include estimates of solar irradiance and volcanic eruptions, represented as a decrease in the total solar irradiance. These simulations start from year 1850. Their initial conditions come from the 1000-year long control simulation under preindustrial conditions and are each separated by 10 years. Each of these simulations was integrated until end of 2005. From January 1^{st} 2006, they were prolonged using external forcing corresponding to the RCP4.5 scenario, as described in Taylor et al (2012). This ensemble of 3 members of historical+scenario simulations will be referred to as HIST in the following.

The second set of experiments under consideration is a 3-member ensemble of nudged simulations, so called as they include a nudging towards observed anomalous SST variations. Each nudged simulation (NUDG1, NUDG2 and NUDG3 in the following) was started on January 1^{st} 1949 from one of the historical simulations, using strictly the same external forcing, and applying also a nudging, or restoring term. This term consists in an additional heat flux term Q imposed in the equation for the SST evolution and written as $Q = -\gamma(SST'_{mod} - SST'_{ERSST})$. SST'_{mod} stands for the modeled SST anomaly with respect to the climatological mean computed between 1949 and 2005 in the corresponding historical simulation. SST'_{ERSST} are the anomalous SST from the Reynolds et al (2007) dataset with respect to the same climatological period. We use a restoring coefficient γ of $40Wm^{-2}K^{-1}$, corresponding to a relaxing timescale of around 60 days over a mixed layer of 50m depth. This rather weak value as compared to previous studies using surface nudging (Keenlyside et al 2008; Dunstone and Smith 2010; Luo et al 2005) typically represents the amplitude of air-sea thermal coupling (e.g. Frankignoul and Kestenare 2002) and was justified in previous papers (Swingedouw et al 2013; Servonnat et al 2014; Ray et al 2015). Efficiency of this nudging strategy in reconstructing subsurface variability is more specifically studied in Ray et al (2015), and the reader is referred to Swingedouw et al (2013) for a focus on the AMOC. Servonnat et al (2014) investigate several aspects of surface nudging in a perfect model context. Note also that as indicated in the previous references, nudging is not applied when and where the model sea-ice cover exceeds 50%.

A set of 3-member ensembles of runs at least 10 years long where the restoring constraint is no longer applied (while the external forcing from historical and scenario simulations is used) was then launched from each nudged simulation. These simulations make up our retrospective forecasts, or hindcasts. For NUDG1 and NUDG2, hindcasts were launched on January 1^{st} 1961 and every 5 years afterwards until January 1^{st} 2006, as recommended in the CMIP5 protocol (Taylor et al 2012). These two sets of hindcasts, named DEC1 and DEC2 in the following, were both submitted to the CMIP5 near term database (e.g. García-Serrano et al 2014). Hindcasts starting from NUDG3 were launched every year from January 1^{st} 1961 until January 1^{st} 2013. These series of hindcasts, named DEC3, was not submitted to the ESG, but is now part of the multi-model decadal forecast exchange project (http://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/long-range/decadal-multimodel; Smith et al (2012)). For all ensembles, initial conditions of the individual members were obtained by applying at the first time step a perturbation to the SST field seen by the atmospheric component, chosen randomly at each grid point between $-0.05°C$ and $0.05°C$. Note that, strictly speaking, each group of 3 members in DEC9 also differ in terms of oceanic perturbation, since they originate from a different coupled simulation. Analysis of the impact of such differences in initial perturbations is beyond the scope of this paper and is not likely to have a strong effect (Du et al 2012). Note also that as in other CMIP5-type hindcasts, external forcing is exactly the same as in historical and nudged simulations. This forcing thus includes volcanic eruptions, even though this forcing would in reality not be available at the start date of the forecast in an operational context.

In the following, we evaluate the forecasting skill of the system using two ensembles of initialized hindcasts: the ensemble DEC3, on the one hand, consisting of 3 members launched every year, and the ensemble named DEC9, on the other hand, resulting from the merging of DEC1, DEC2 and a subsample of DEC3, which consists thus in a 9-member ensemble of hindcasts launched every 5 years from January 1^{st} 1961 to January 1^{st} 2006. On top of these, we consider the ensemble of HIST simulations as a benchmark for multiyear prediction skill without initialization.

2.3 Verification datasets

In order to validate the prediction skill of the system, five different datasets are used. First, we consider ERSST, the SST field from Reynolds et al (2007), which was used for the nudging. Performances are expected to be highest with this reference dataset, which, for our purposes, covers the period [1961-2013]. This dataset is represented with the dark blue color in the figures. The HadISST dataset (Rayner 2003) taken as an alternate verification dataset gave very similar results as ERSST and is thus not shown. Secondly, we consider two ocean reanalyses, namely ORAS4 (Balmaseda et al 2013, , color code orange in the figure), available until 2011, and SODA2.2.4 (SODA hereafter, color code cyan in the figures) (Carton and Giese 2008; Giese and Ray 2011; Ray and Giese 2012), available until 2005. As described in Ray et al (2015), for example, these two reanalyses are based on different ocean models, with different resolutions, different forcing datasets and different assimilation schemes, which may lead to substantial differences. They yield a consistent (significantly correlated at the 90% confidence level) reconstruction of the oceanic variability mainly down to 200m (Ray et al 2015). We use them both in order to assess the prediction skill of the system but taking into account the uncertainty in data, in particular for ocean variables hard to constrain such as the AMOC. For the AMOC, we also consider the reconstruction proposed by Latif et al (2006), using a dipole of SST between the Northern and Southern Atlantic (featured in yellow in the figures). Finally, for the subsurface temperature, integrated ocean heat content and for the salinity, we also use the EN3 set of objectively analyzed temperature and salinity profiles (color code purple) proposed by Ingleby and Huddleston (2007). This product is not optimized for SST, as it does not integrate specific surface data. All these datasets will be collectively referred to as DATA from now on in the text. Note however that these data sets are always considered individually in all computations, and not averaged out. Furthermore, for clarity of the figures, the ACC and RMSE skill scores computed for the HIST simulations with respect to each of these data sets are not identified individually with specific colors.

2.4 Data processing

As discussed for example in van Oldenborgh et al (2012), a large part of the skill in decadal temperature forecasts is due to the trend. In order to study the predictability of the variability around the trend, it is important to remove the effect of the trend as cleanly as possible. A good definition of the trend is nevertheless difficult to obtain, given the non-linearity of the forcing (see discussion in García-Serrano et al (2014)). Furthermore, estimates of local trends are subject to large sampling variability because of the lower signal to noise ratio for smaller spatial scales. Therefore, we focus here on spatial averages over relatively large domains (typically, the North Atlantic Ocean between 30°N and 60°N) in order to maximize the signal to noise ratio (Goddard et al 2012).

The treatment of data is then done as follows. Firstly, all ensemble sets (HIST, NUDG3, DEC and DATA) are organized mimicking the hindcasts outputs, that is as a function of start dates (from 1961 to 2013 or 2006 depending on the DEC system under consideration) and lead times (from 1 to 10 years). Secondly, anomalies are computed. The reference period is estimated as the overlapping period between the observational records and the hindcasts, i.e. [1961 - 2005] if the SODA reanalysis is included. Results were also tested against the use of a longer reference period, namely 1961-2011. This implies excluding the SODA reanalysis, but main results were unchanged. We then consider, for each dataset, anomalies with respect to the linear trend. This trend is estimated separately for each forecast time over the reference period. The simulated trend is computed separately for each individual member and the same methodology is applied both for DEC3 and DEC9. Observational trend is also considered as forecast-time dependent. Note that this procedure includes a correction of a bias in the mean state as well as of the linear response to external forcings. We assume that the residual signal represents the unforced variability, but we know that this is just an assumption as the external forcing is not linear. Note that the IPSL-CM5A-LR coupled model has a climate sensitivity of 3.9K for a doubling of CO_2 (Dufresne et al 2013), which places it at the 4^{th} out of 11 models of the CMIP5 ranked per decreasing climate sensitivity (Vial et al 2013) and is stronger than the newest estimates of climate sensitivity around 3K (Collins et al 2014).

To ensure having the same number of verification years at each forecast time in DEC3, we consider the verification period [1966 - 2005] when the SODA dataset is included. Following the four-year average approach this implies that the common verification period spans from 1966/69 to 2002/05,with a total of 37 values per forecast lead time. Results are also tested against the common verification period 1966/69 to 2008/11, when SODA is excluded. Except if discussed in the text, results are generally similar. Note that the use of such common verification framework yields the same number of degrees of freedom for all lead times for a single time series (e.g. García-Serrano et al 2012); this enables a consistent comparison of forecast skills at different lead times. Furthermore, given that the non-initialized simulations are in fact a re-organization of the outputs from three long-term simulations (HIST1, HIST2, HIST3), the time series constructed for the different lead times are identical and thus the statistical metrics are constant. The same applies to the DATA time series following this approach. Note furthermore that this common verification framework was not used for DEC9 due to the few start dates available.

2.5 Forecast quality assessment

Multi-annual prediction skill is measured in terms of anomaly correlation coefficients (ACC) and root mean square errors (RMSE). ACC and RMSE are calculated based on the ensemble mean of the hindcasts. Both measures are computed for DEC and HIST respectively, against DATA, and for each lead time. Significance of the correlation is tested with a one-sided Student t-test at the 90% confidence level. The number of degrees of freedom takes into account the autocorrelation of each time series, as suggested in Bretherton et al (1999). We also test the significance of the ACC difference between HIST and DEC. The purpose of this additional test is to evaluate the added-value of initialization for the prediction skill. Significance of the difference between the RMSE of initialized (DEC) versus non-initialized (HIST) ensembles is evaluated using a Fisher test. Note that a fair estimation of the continuous ranked probability score (Ferro 2014) was found to yield very similar conclusions as the RMSE. Given that the evaluation of probability distribution might be problematic in DEC9 which only counts 8 realizations, we decided to show only RMSE here.

All ACC and RMSE are also computed against the NUDG3 (simply named NUDG in the following) outputs, and significance is tested similarly. The point of evaluating prediction skill against both DATA and NUDG is to compare actual and potential predictability, respectively. Such assessment is particularly relevant when initial conditions have been constructed through nudging rather than directly taken from an independent dataset. In this case, indeed, the correlation and RMSE of hindcasts with respect to NUDG is expected to be higher than computed against DATA, as NUDG contains effectively the initial conditions from which the hindcasts were launched, and these can then be substantially different from the data (e.g. Ray et al 2015). The forecasting skill against NUDG gives an idea of the upper limit of possible skill in the system, while the one computed against DATA measures the actual skill against a particular reconstruction of reality. The potential prediction skill defined here is inspired from Boer et al (2013) but not fully equivalent: for Boer et al (2013) potential forecast skill is analogous to actual forecast skill, but with the divergence of the forecast from the observed evolution being replaced by a measure of the divergence of model results from each other. Here, we rather use a different reference, namely the NUDG simulation. Note also that only one nudged simulation is used, and not the average of the three. Indeed, the nudging only has a limited impact on the ocean subsurface, so that the three nudged simulations do slightly differ after a certain depth (Ray et al 2015). As a result, averaging the three nudged simulations in these regions would risk to blur the reconstructed variability at depth. Note however that it would not change the results regarding the SST prediction skill.

We also compare the skill of the forecasts with the performance of a first order auto-regressive model (e.g. Ho et al 2012). Initial conditions are taken from the last year before the beginning of the hindcasts, that is the last year with supposedly known conditions. The time constant involved in the auto-regressive model is estimated from the fit of the autocorrelation function of the considered time series taken in the long-term control run by a decreasing exponential (e.g. Mignot and Frankignoul 2003).

Finally, while the metrics presented above focus on the ensemble mean, it is also important to consider the dispersion of the hindcasts around this mean, in order to estimate their reliability. A forecast system is considered as reliable when the forecast probabilities of a certain variable match the observed ones. These questions have been extensively tackled for seasonal forecasts (e.g. Weisheimer et al 2011; Batté and Déqué 2012), and much less for the decadal predictions (Corti et al 2012; Ho et al 2013). Here, since our analysis only uses one prediction system, the error primarily comes from uncertainty in initial conditions. In this respect, the spread of the set of predictions can be used as a measure of the prediction error. This ensemble spread is compared to the RMSE of the forecast ensembles with respect to DATA or NUDG. For a prediction to be reliable, or trustworthy, the time-mean ensemble spread about the ensemble mean should equal the time-mean RMSE of the ensemble mean forecast. The system is said overdispersed if the spread significantly exceeds the RMSE. In this case, the probabilistic forecasts are unreliable as the individual forecasts may produce too different results. On the contrary, if the spread is significantly smaller than the RMSE (system underdispersive), especially at short forecast ranges, it may indicate that the initial perturbation of the probabilistic forecast is too weak to realistically sample the uncertainty of the system. The system can then be characterized as overconfident, and it is in any case also poorly reliable. Note nevertheless that caution is required when assessing the reliability in DEC3, given the very low number of members.

3 Global and tropical SST prediction skill

3.1 Global SST prediction skill

Fig. 1a shows the time series of detrended global-mean SST anomalies averaged over the forecast years 2-5 in the DEC3 ensemble mean and the corresponding non-initialized hindcasts HIST. Outputs from the NUDG simulation and ERSST are also shown. These time series highlight the decadal climate variability at global scale and the cooling signatures of the major volcanoes which have erupted over the last 50 years: Mt Agung in 1963, El Chichon in 1982 and Mt Pinatubo in 1991. Because of the strong negative radiative forcing of these volcanic eruptions, ACC of the hindcasts with both NUDG and the DATA is not significantly different from that obtained with the non-initialized hindcasts (Fig 1b). The global mean SST indeed primarily responds to external forcing, and this figure illustrates the weak added value of initialization for predicting this climate quantity over the period considered here (which includes rather strong volcanic eruptions). Consistently with Mehta et al (2013), volcanic eruptions are one of the important sources of decadal prediction skill for global SST. When computed against NUDG and ERSST (the dataset used for the nudging) ACC remains significant at all lead times. SODA and more clearly ORAS4 yield lower scores. This illustrates the uncertainty in available datasets, and how it hampers hindcast verification. Note that the AR1 predictive method started from DEC3 and computed with respect to

NUDG is not skillful. This is consistent with an important role of external forcing, which may appear after the date when the hindcast was launched.

Fig. 1e further illustrates the influence of non-linear external forcing in the DEC9 predictive system. Because hindcasts are launched every 5 years only in this set, their specific timing with respect to the volcanic eruptions listed above is very important. More precisely, one should note that the start dates used in DEC9 (following the CMIP5 protocol) are in phase or slightly leading the eruptions. As a result, for the forecast range 2-5 years for example, two start dates (1982-1985 and 1992-1995) are very strongly influenced by the eruptions (since the radiative impact typically lasts 3 years, Robock (e.g. 2000)). This highly contrasts with the forecast range 4-7 years, which is, for each start date, only impacted by the last year of the volcanic radiative effect (see also Figure 10 in Germe et al (2014)). As a result, the main source of predictability for global SST is partly lost for the forecast range 4-7 years and the correlation skill drops. Impact of the main volcanic eruptions in the last 60 years falls again in the time window of the predictions at lead times 6-9 years, thereby contributing to enhance the correlation skill again. Such specific sampling issue does not occur in DEC3 (Fig. 1b). A subsampling analysis of the start date frequency in DEC3 confirms that the drop of skill from forecast ranges 3-6 years until 5-8 years, followed by a recovery at the forecast range 6-9 years essentially comes from the specific choice of start dates every 5 years starting from 1961 (Fig. 2).

Benefits of the system's initialization in bringing together the different members are yet visible from the fact that the spread of the initialized hindcasts is initially smaller than for non-initialized hindcasts (Fig 1c.). Afterwards, it increases with forecast time, towards the level of the non-initialized hindcasts spread, illustrating the decreased influence of initialization with forecast time. Eventually, the spread of DEC3 is even slightly larger than that of HIST. Note however that differences are not significant. The spread of HIST hindcasts is slightly lower than the RMSE with respect to the NUDG simulation, suggesting that the potential non-initialized forecast system is overconfident (underdispersive). This feature is worse for the initialized system (Fig 1c.). This lack of reliability is reduced in the DEC9 system (Fig 1f) for which the RMSE is reduced. We recall that DEC9 differs from DEC3 in terms of start dates frequency and ensemble size. Fig. 2 shows that the reduction of the RMSE in DEC9 does not arise from a decrease in the start date frequency. It is thus due to the increase in the number of members which indeed is expected to yield a better estimate of RMSE through a more accurate estimation of the ensemble mean. Nevertheless, Fig. 2 also shows that a reduction of the start date frequency yields more noisy and therefore less robust statistics, which can lead to spurious results. The RMSE of DEC3 is larger than that of HIST, whatever the reference set (Fig 1c.) This feature is reduced in DEC9, probably as a result of the better estimation of the RMSE. Still, this result is relatively surprising, given the expected added value from initialization to correct part of the errors in the unforced model response and put the model in phase with the unforced variability, thereby decreasing the RMSE similarly for DEC3 and DEC9. These differences are nevertheless not significant, and this feature disappears for other regions investigated below.

Fig. 3 shows the potential ACC skill score of the HIST and DEC3 ensembles computed grid-pointwise for detrended SST for the lead times 1 year, 2-5 years and 6-9 years. The added-value of initialization for the first lead time is clearly illustrated on the top panel: for a lead time of 1 year, SST is skillfully predicted over all oceanic regions in the initialized hindcasts. For longer lead times, fewer regions remain skillfully predicted in the initialized runs. The subpolar North Atlantic, the extratropical North Pacific, the northern Indian Ocean and the western tropical Pacific, as well as localized areas of the Southern Ocean stand out. In the CCSM4 experimental decadal prediction system, Karspeck et al (2014) found that the subpolar North Atlantic was the only region where the initialized predictions outperform the non-initialized ones. The maps shown here are a bit more encouraging, but they only show potential skill. Note that the maps computed against ERSST rather than NUDG are very similar (not shown). In the following, we focus on specific regions and discuss both the potential and actual prediction skill, including uncertainty arising from observational datasets.

3.2 Tropical SST prediction skill

In the tropical band, forecasting skill is investigated using individual forecast years, instead of multi-year averages. Both potential and actual SST predictions are skillful for the first lead time only (Fig. 4b). The non-initialized ensemble, on the other hand, is never significantly skillful (ACC is always negative), indicating that the prediction skill at 1 year lead time has been enabled by the initialization of the coupled model. For this first lead time, RMSE of DEC3 is smaller (but not significantly) than that of HIST, further highlighting the impact of initialization. This effect is lost for longer forecast ranges, with the spread of DEC3 reaching the level of HIST. All statistics (both actual and potential) thus nicely highlight a prediction skill of 1 year over the tropical band, thanks to the better initial conditions, an effect that is lost afterwards. Actual and potential ACC skills also loose significance after the first lead time, but the decrease is more gradual in DEC9, this may be due to sampling effects. Furthermore, DEC9 is roughly reliable for the first two lead times. As above, a subsampling analysis of the start dates frequency in DEC3 shows that these improvements of DEC9 performances over DEC3 comes from the increase in the number of members (not shown). However, for lead times longer than 3 years, the evolution of skills with the lead time in DEC9 is, again, very noisy. This ACC recovery at lead time 7 years in DEC9 (Fig. 4e) gives another illustration of possible spurious predictions and conclusions when too few start dates are used. Another sampling impact is noticeable in the RMSE of DEC9 with two peaks at lead time 4 and 9 years, separated by the start date frequency of 5 years (Fig. 4f).

Further analysis shows that skill at lead time 1 is also found when considering the tropical Atlantic or the tropical Pacific separately (Fig. 3 right). In the tropical Pacific, the skill of year 1 in this region is consistent with the literature: in theory, ENSO is believed to be predictable on the order of 1 or 2 years in advance because of the self-sustained nature of the tropical Pacific coupled ocean-atmosphere system (e.g. Neelin et al 1998). In practice, however, this predictability is reduced because of the influence of stochastic atmospheric forcings, such as surface wind bursts in the western equatorial Pacific (e.g. Kleeman and Moore 1997; Perigaud and Cassou 2000; Fedorov et al 2003). Thus, ENSO predictability is usually limited to a few months, reaching two years only in some specific studies (Luo et al 2008; Volpi et al 2013). This general result seems to hold for our specific forecast system.

4 Prediction skill in the North Atlantic Ocean

As indicated above, the North Atlantic Ocean is often found to be the most predictable region of the world's ocean when compared to non-initialized predictions (e.g. Hazeleger et al 2013b; Corti et al 2012; Kim et al 2012; van Oldenborgh et al 2012; Doblas-Reyes et al 2013). We focus first on the North Atlantic variability, by looking at the linearly detrended SST average over the Atlantic region [0-60°N] (Fig. 5). Note that this index slightly differs from the canonical definition of Atlantic Multidecadal Oscillation (AMO, e.g Sutton and Hodson 2005) as it is not low pass filtered. It is only computed using a four-year running mean, as forecast ranges of 4 years are considered. It is used here to characterize the Atlantic Multidecadal Variability (AMV). The variability in HIST is strongly dominated by the model's bidecadal variability described in Escudier et al (2013) and Ortega et al (2015b). This internal variability is partly phased by external forcings, as shown in Swingedouw et al (2013, 2015). However, according to these studies, the Mt Agung eruption (1963) induces a phasing of the AMOC (see below) only 15 years later and thus of the North Atlantic SSTs after about 20 years, i.e. from the mid-1980s. This phasing can indeed be seen around the end of the period in Fig 5a and is confirmed by a positive correlation between the North Atlantic SST from HIST and from ERSST for the period [1987-2005] (not shown). Before this, the variability in HIST is strong and completely un-phased with data.

Both potential and actual prediction skill are significant for all forecast ranges for DEC3, contrary to HIST (Fig 5b). The statistical prediction based on an AR1 process is also significantly correlated with the NUDG, but only for the forecast range 1-4 years, which is consistent with previous findings showing that dynamical predictions out-perform statistical predictions based on persistence over large parts of North Atlantic for longer lead times (e.g. Ho et al 2012). This suggests that the additional skill potentially coming from ocean dynamics, beyond the thermal inertia, is noticeable after about 1-4 years ahead (e.g. Matei et al 2012). We also note that ACC computed against NUDG is generally slightly higher than the ones computed against DATA, in particular for shortest forecast ranges, and it shows a skill decrease with forecast time. The degradation in the North Atlantic SST multi-year skill is even more clearly seen in DEC9, and it has also been found in recent studies using start dates every 5 years, in particular with the ENSEMBLES decadal re-forecasts ensemble (van Oldenborgh et al 2012; García-Serrano and Doblas-Reyes 2012) and the CMIP5 ensemble (Kim et al 2012). This is less obvious from yearly start dates, but it was reported in the DePreSys system by García-Serrano et al (2012). In DEC9, significance of actual skill is lost at forecast ranges longer than 4-7 yrs.

As for ACC (Fig 5b), RMSE of the initialized hindcasts (with respect to the NUDG simulation) is significantly smaller than for the non-initialized ones for all forecast ranges (Fig 5c). The difference is no longer significant when RMSE is computed against all other datasets, except for ORAS4. This can indicate a weak impact of initialization or a weak signal to noise ratio. In DEC9, RMSE is reduced as compared to DEC3, but given the reduced degrees of freedom, it is not significantly different from that of HIST, even when assessed against NUDG (Fig 5f). Furthermore, as above, while DEC3 is strongly overconfident (underdispersive), DEC9 is a more reliable prediction system thanks to the increased number of members.

Fig. 6 compares the prediction skill of SST anomalies in the North Atlantic midlatitude ([30°N - 60°N]) and low-latitude ([0 - 30°N]) regions respectively. As for the total North Atlantic SST variability, correlation with the NUDG simulation is significant at all lead times for the extratropical North Atlantic, both in DEC3 (Fig. 6b) and in DEC9 (not shown). Furthermore, the correlation skill score with NUDG is almost constant for all forecast ranges, as in Fig. 5. On the contrary, for the low-latitude part, the potential skill score is significant and significantly different from non-initialized hindcasts only until the forecast range 2-5 to 3-6 years in DEC3 (Fig. 6d and in DEC9, not shown). As discussed in García-Serrano et al (2012), this finding illustrates that the added-value from initialization in the AMV skill during the second half of the hindcast is likely dominated by midlatitudes in the SST area average. The skill of the AR1 model is also very different in the two regions: while it is pretty skillful at midlatitudes, it does not provide any skillful information at lower latitudes. This suggests that the long prediction skill at midlatitudes is linked to the long persistence of SST anomalies. It is consistent with the observed autocorrelation functions shown for the two regions in García-Serrano et al (2012). This difference between low and mid-latitudes skill as a function for short and long forecast ranges can be carried over to actual prediction skill in DEC3, although details in the significance of ACC depend on the dataset and forecast range that is considered for verification. On the contrary, ACC significance decays with forecast time at lower latitudes. The picture is consistent but more noisy in DEC9, in particular in the northern region (not shown).

Fig. 7(a and b) shows the correlation maps of the observed SST averaged over the northern Atlantic [0-60°N] with SST anomalies in observations and NUDG. All time series have been averaged over four consecutive years prior to computing the correlation. These maps compare the representation of the observed variability averaged over the North Atlantic in the nudged simulation and in the observations. The patterns are both well significant over the whole North Atlantic, except primarily along the Gulf Stream path, similarly to what is found in other studies (e.g. Marini and Frankignoul 2013). The pattern in the bottom panel (Fig. 7c) is different with observations and NUDG: in the non-initialized simulations (HIST), correlation against the AMV variability is only significant equatorward of 15°N and in the western subtropical part of the North Atlantic. This suggests that SST variability in the extratropical North Atlantic mainly relies on the internal variability rather than the response to radiative forcing Comparing Fig. 7(b) and Fig. 7(c) shows the nudging efficiency to bring North Atlantic variability close to observations. Nevertheless, at subpolar latitudes, the NUDG pattern shows non significant areas, unlike what is found in ERSST (Fig. 7a and b). These areas are quite small, but they indicate that locally, the nudging is not always sufficiently strong with respect to the model's deficiencies and internal variability to constrain the SST anomalies. As previous studies have suggested that this area is crucial for predictability in the north and tropical Atlantic (e.g Dunstone et al 2011), this may explain the lack of actual predictability in our model. Specific reasons for this poor constraining of SST in this region is probably linked to the strong internal variability of this area in the model Escudier et al (2013); Ortega et al (2015a) and/or a particular sensitivity to external radiative forcing as in other CMIP5 models (e.g. García-Serrano et al 2014). The correlation of the predicted SST at forecast range 2-5 years with the observed North Atlantic variability (Fig. 7 d) largely resembles the one found for HIST (panel c): it is hardly significant in the extratropical North Atlantic and the significant domain extends only slightly poleward as compared to HIST. In other words, the nudging works correctly in the

North Atlantic but it yields a gain of predictability only between 15°N and 30°N in the North Atlantic. It does not constrain sufficiently the subpolar SSTs. At the forecast range 6-9 years (Fig. 7 e), though, areas of significant correlation in the northern and eastern subpolar Atlantic emerge. This is consistent with enhanced actual predictability seen in Fig. 6b. This cannot be due to external forcing in the model, as the structure in HIST is very different. Oceanic dynamics is a plausible explanation, as it may bring the DEC structure closer to the one of NUDG in spite of a lack of predictability in the subpolar North Atlantic. Predictability gained thanks to oceanic dynamics in the North Atlantic has already been invoked by previous studies (e.g. Matei et al 2012). Another candidate is the effect of the initialization in correcting the model's response to external forcing, identified as one of the premises of decadal climate prediction (Meehl et al 2014), and its persistence along the hindcast period (Fig. 6b). In IPSL-CM5-LR probably both effects are at play.

Given the impact of the AMOC on the North Atlantic temperatures (e.g. Knight et al 2005), we also attempt to evaluate its prediction skill. The major limitation for this assessment is the poor consistency of reanalyses in terms of AMOC variability (Reichler et al 2012; Pohlmann et al 2013). As an illustration, the time series of the maximum of the AMOC at 48°N from the ORAS4 and SODA reanalyses have a correlation coefficient of 0.24 over the common period [1961-2012], and 0.25 at 26°N. Both values are significant at the 90% level (1-sided) but explain only 6% of the covariance. Correlation for the absolute maximum in latitude is close to 0. Swingedouw et al (2015) have evidenced the influence of the volcanic forcing on the timing of bi-decadal variability in the North Atlantic in data and simulations. In particular, volcanic eruptions were found to induce an acceleration of the AMOC with a delay of roughly 15 years after the eruption. Swingedouw et al (2013) showed that the SST nudging still plays an important role, as they translate the role of atmospheric forcing such as the persistent NAO events in the 1980s and 1990s. This might explain the slightly delayed AMOC maximum around the end of the 1990s in NUDG as compared to HIST (Fig. 8 a and b), but this effect is weaker in the present analysis than in Swingedouw et al (2013) as only one realization of NUDG is used here.

Fig. 8 shows that our system has no skill in predicting the AMOC reconstructed by either of these reanalyses. By contrast, potential predictability as measured using ACC is significant at all lead times (Fig. 8b), in agreement with the long AMOC internal predictability (Persechino et al 2013). Although these values start higher than for the non-initialized hindcasts at the first two forecast ranges, the difference is not significant. The same conclusion holds for the RMSE although initialization has also helped to reduce the spread of the initialized hindcasts.

In order to better understand the impact of the initialization on the North Atlantic ocean and its predictability, we investigate the predictability of vertically averaged ocean heat content in DEC3 (Fig. 9). In the North Atlantic midlatitudes, there is practically no actual skill for the heat content integrated down to 300m or below which is consistent with the lack of actual SST skill in the same region (Figs. 6b, 7d,e). The potential skill is significant for all forecast ranges. It is higher than the skill obtained for non-initialized hindcasts until the forecast range 2-5 years, but the difference is not significant. As for the AMOC, the ocean heat content is found to be strongly impacted by the model's internal variability, characterized by a 20 year time scale.

5 Prediction skill in the North Pacific Ocean

Prediction skill of the tropical Pacific was discussed in section 3.2. The northern Pacific Ocean is usually one of the regions with the lowest actual skill in near-term temperature forecasting (Guemas et al 2012; Kim et al 2012; Branstator and Teng 2012; Bellucci et al 2013), although hints of improved predictability in the North Pacific temperatures by initialization have been found by Mochizuki et al (2010), Chikamoto et al (2013) and Magnusson et al (2012). After a trend analysis, Bellucci et al (2014) suggest that the poor skill in the extra-tropical North Pacific reflects the inability of the models to correctly reproduce the observed ratio between forced and unforced variability in this region, where the warming trend only explains a small fraction of the total variability. Fig. 3 nevertheless reveals potential prediction skill in our system in the North Pacific midlatitudes. One can identify three skilful regions in the North Pacific in our model, at lead-time 2-5 years (middle right panel): Firstly, a skilful region is found between 5°N and 15°N in the western Pacific, which also appears in HIST, thereby suggesting that it is associated to external forcing. A second skilful region is found between 15°N and 30°N in the western to central Pacific. This region is not skilful in HIST. Thus it has been positively affected by the initialization. It looses skill at lead time 6-9 years (Fig. 3 bottom right). Consistently, ACC for SST averaged over the low latitudes ($[0-30°N]$) in the Pacific is only significant when computed against NUDG (potential predictability), and only over the forecast range 1-4 years (not shown). This is less than what was described for the tropical to subtropical North Atlantic above. As discussed previously, this is due to the dominant influence of ENSO in the Pacific, poorly predictable beyond one year, while the tropical Atlantic benefits from the influence of subpolar latitudes and cross-equatorial heat transport by the AMOC. Finally, the maps also show a skilful region between 30°N and 45°N extending almost through the whole Pacific basin, which is still significantly correlated with NUDG at forecast range 6-9 years, while no skill is found in HIST. This region bears similarity with the skilful region highlighted in Kim et al (2012, 2014); Doblas-Reyes et al (2013). Fig. 10 confirms that in our system, the potential skill averaged over the northern extratropical Pacific from 30°N to 45°N is significant for all forecast ranges and significantly different from the skill obtained for non-initialized hindcasts. Interestingly, actual prediction skill is also significant for all lead times so that although scores are slightly lower, actual prediction skill practically equals potential skill in this region. Furthermore, the actual skill is at least as good as for the Atlantic (Fig. 5b 6b). Note that the shape of the ACC evolution with increasing forecast ranges 1-4 years, as computed against NUDG and DATA contrasts with the skill of the statistical AR1 process. The latter yields a significant correlation only for the shortest forecast range, and it decreases quickly afterwards. This suggests a role of the oceanic circulation on this predictability beyond thermal inertia. RMSE of DEC3 is not significantly different from HIST, and neither is the spread (Fig. 10c). In general, DEC3 appears to be reliable, with the ensemble mean RSME matching the ensemble spread, while DEC9 can be rather considered as overdispersive.

The correlation between SST averaged over this region ([30°N-45°N]) and the first empirical orthogonal function of annual mean SSTs between 20°N and 75°N amounts to -0.94 (significant at the 95% level, not shown) in the control simulation. This indicates that the SST average shown in Fig.10 can be taken as a measure of the negative phase of the Pacific Decadal Oscillation (PDO) in IPSL-CM5A-LR, in a manner similar to the definition in Mantua et al (1997). Fig. 11 shows that in observations, SSTs averaged in the area also project on the typical PDO pattern (a), and that this is well represented in NUDG (b). However, the spatial pattern associated in the model with the observed variations of SST in the North Pacific ([30°N-45°N], Fig. 11c) is not a PDO-like pattern. It rather bears similarity with the second least damped mode of North Pacific SST variability found by Newman et al 2007. The predicted pattern related to the observed time series (d and e) captures some of the positive anomalies in the central North Pacific, but not in the latitude band between 30°N and 45°N. Furthermore, the predicted pattern is positive in the whole subtropics, near the eastern coast and in the north. This also resembles the second least-damped mode of North Pacific SST variability found by Newman (2007), except for the tropical and eastern subtropical part. Newman (2007) and Newman (2013) suggested that the observed PDO represents the sum of several stochastic phenomena rather than a single physical process, and they showed that long term predictability in the North Pacific is primarily due to the second least-damped mode. The fact that the observed PDO time series projects onto this mode in the historical simulation may explain the relatively long predictability in the North Pacific found in the model. The North Pacific climate has experienced several climate shifts over the past decades, in particular in 1976/1977 (e.g. Trenberth and Hurrell 1994; Mantua et al 1997; Deser et al 2004; Yeh et al 2011), in 1988/89 (Hare and Mantua 2000; Trenberth and Hurrell 1994) and in 1998/99 (Minobe 2000; Di Lorenzo et al 2008; Ding et al 2013). In the context of the PDO being represented by the sum of several stochastic processes, Newman (2007) explain that these shifts may only be predictable within the timescale of the most rapidly decorrelating noise, i.e. around 2 years. The ERSST curve in Fig.

10a shows how these shifts translate in terms of SST averaged of the North Pacific midlatitudes. The three transitions are reasonably reproduced in the NUDG simulation, and the 1976 and the 1998 ones are reasonably predicted 2-5 years in advance. This may again be explained by the dominance in the model of one specific mechanism for the PDO, as opposed to what is found in Newman (2007). The late 1980's event is rather well predicted with a 1 year lead time (not shown), while it is missed with at a 2-5 years forecast range. Note also that in the model, SST average between 30°N and 45° in the Pacific is strongly correlated with the SSTs in the North Atlantic low-latitudes (r=0.45, significant at the 95% level, not shown). Although this statistical link is not realistic (see for example Marini and Frankignoul (2013)), it may also explain the relatively long predictive skill detected in the North Pacific in our model.

We turn now to the investigation of the OHC, a key variable for ocean memory and thus predictability. Ocean heat content integrated down to 300m over the extratropical Pacific shows surprisingly good potential prediction skill, as compared to literature (Fig. 12). Initialized predictions are potentially skillful for all forecast ranges, and ACC measured against SODA (i.e. actual skill) is significant and significantly different from non-initialized hindcasts up to the forecast range of 5-8 years. For ORAS4 and EN3, ACC is in general significant as well, although not significantly different from the skill obtained in HIST. Time series for the forecast range 2-5 years (Fig. 12a) confirm the relatively good reconstruction of the ocean heat content variability in NUDG with respect to EN3. These performances are overall striking and good and contrast with the general idea that decadal predictability over the North Pacific is quite low. Nevertheless, Chikamoto et al (2013) reported prediction skill over almost a decade for subsurface temperatures in the

North Pacific, which is in agreement with our actual skill assessment. The potential predictability of our system suggests that even longer skillful forecasts might be achieved in the future. Interestingly, once again, the AR1 statistical model yields significant prediction skill for lead times 1-4 years, but the ACC drops rapidly as forecast times increases. This clearly suggests a role of ocean processes for the long predictability detected in ocean heat content in IPSL-CM5A-LR.

6 Results on salinity

In a perfect model framework, Servonnat et al (2014) showed a good ability of SST nudging in reconstructing SSS variability in the tropics. It is therefore interesting to evaluate the prediction skill of this variable in the same region for our set of experiments (Fig. 13). Note however that given the lack of long-term satellite measurements, SSS reconstructions and reanalysis are subject to much higher uncertainty than temperature, so that actual prediction skill (or the lack of) has to be interpreted with care. Potential prediction skill of SSS over the tropical band ($20^{\circ}S$-$20^{\circ}N$) is significant for the first three forecast years, but both ACC and RMSE are significantly different in DEC3 and HIST only the first year. SSS has thus been impacted by the nudging in the Tropics, as described in (Servonnat et al 2014) and given its relatively longer persistence than SST (e.g. Mignot and Frankignoul 2003), it is potentially predictable over relatively longer forecast ranges too. The AR1 model yields potential skill for 1-year lead time. In terms of actual prediction skill, ACC is low but significant only when computed against ORAS4. NUDG is indeed significantly correlated with ORAS4 at the 90% confidence level ($r = 0.70$), suggesting that SSS has been reconstructed with some agreement as compared to ORAS4. Note that these results primarily come from the tropical Pacific, while potential skill is only significant for the first two lead times in the tropical Atlantic. Séférian et al (2014) found similar results in the tropical Pacific for the nutrient primary productivity.

We now examine the prediction skill, both potential and actual, of the SSS in the North Atlantic ([30°N-60°N] Fig. 14). As indicated by the weak correlation between NUDG and the DATA (Table 2, top), SSS has not been properly reconstructed in these regions as compared to reanalysis. SSS typical variability in all simulations is much stronger than in the DATA (Table 2, top, first column), probably as a result of the strong bi-decadal variability in this region in the model. Nevertheless, SSS has been influenced by the nudging, as correlations between HIST and NUDG are also very weak. Note that the same applies to SST (Fig. 6a). In the North Atlantic, the resulting SSS variability both in the NUDG and DEC3 time series is strongly correlated with the corresponding SST. It was also the case in the non-initialized runs HIST. This strong link between SST and SSS in the North Atlantic in this model has been extensively described in Escudier et al (2013). The correlation of SST and SSS in the NUDG shows that SST nudging has strongly impacted the SSS through the 20-yr cycle. Significant skill score and correlation of the DEC3 time series of SST and SSS for the forecast range 2-5 years shows that this phasing in the NUDG carries on in the hindcasts and yields potential predictability for the SSS in the northern North Atlantic. Given the role of SSS anomalies for deep convection and the AMOC, this type of mechanism for SSS predictability is encouraging for AMOC predictability. Unfortunately, actual prediction skill is not significant. Nevertheless, since SSS is not properly constrained in this region in data and reanalysis, large uncertainties remain concerning large-scale SSS observation products. Reasons for these discrepancies are beyond the scope of the present study.

In the model, SSS and SST are not as tightly linked in the North Pacific as in the North Atlantic. Nevertheless, the salinity is also affected by the nudging, as seen from the weak correlations between HIST and NUDG time series (Table 2). The high (although not significant at the 90% confidence level) correlation between NUDG and DEC3 can thus be attributed to the SSS internal persistence, with makes it potentially predictable in the model.

7 Conclusions

Two decadal prediction ensembles, based on hindcasts performed with the same model and the same simple initialization strategy have been analyzed. The initialization consists of surface nudging to ERSST anomalies, with a relatively weak nudging strength, namely 40 $W.m^{-2}.K^{-1}$. The first ensemble consists of 3 members of hindcasts launched every year between 1961 and 2013. The second ensemble consists of 9 members launched every 5 years between 1961 and 2006. The focus of this study has been on assessing multi-year prediction skill of the ocean in these two decadal prediction ensembles.

The first important outcome of this study is precisely the difficulty to assess the actual skill, because of data uncertainty. For SST, ACC and RMSE measured from one observational dataset (ERSST) and two reanalysis (ORAS4 and SODA) led in general to similar conclusions in terms of predictability horizon, but with different values for the ACC and the RMSE. For the salinity and the ocean heat content, EN3, ORAS4 and SODA could also lead to different predictability horizons. For the AMOC, the three reconstructions considered here were found to be very weakly correlated. Understanding the reasons for these particularities are beyond the scope of this study. We suggest nevertheless that forthcoming assessments of decadal predictions should be performed against several -at least more than one -datasets, as a measure of the uncertainty of the data.

A second major conclusion is the importance of increasing the number of members and start dates in decadal prediction systems. This idea is not new (e.g. Kirtman et al 2013) and in the literature, the issue of the small size of ensembles has been overcome by using multi-model ensembles (e.g. van Oldenborgh et al 2012; Bellucci et al 2014). We showed here that 3 members are usually not enough to estimate consistently the ensemble mean, and thus yield biased estimates of the RMSE. Increasing the ensemble size to 9 members helps in reducing this problem. It leads to overall more reliable predictions, as the ensemble mean is more accurately estimated, so that the RMSE is reduced and it becomes comparable to the spread. Probabilistic skill scores yield similar conclusions (not shown), although the estimation of a probability density function with 9 members could only be tested with a start date interval of 5 years (DEC9) and should be considered with care. Increasing the number of start dates also appeared crucial in order to obtain robust prediction skill scores. With only 8 or 9 start dates to verify against, prediction scores are very noisy and thus poorly trustworthy. The major influence of non-linear effects of external forcing as well as background decadal variability has been illustrated.

A third particularity of the present study as compared to previously published evaluations of decadal prediction systems is the parallel assessment of both potential and actual prediction skill. Computing skill scores against observations and reanalysis datasets is of course crucial for practical applications. From a technical point of view, this is also important in order to evaluate the efficiency of the initialization strategy. However, from a pure scientific point of view, potential prediction skill gives a robust insight in the maximum predictive horizon which can be expected for a particular forecast system, thereby suggesting possible mechanisms responsible for the predictability, and areas where specific efforts on measurement systems and/or model improvements should be made. In the case of DEC3, particularly long potential prediction skill has been found for the AMOC, the upper 300m ocean heat content and the SSS in the North Atlantic, and could be interpreted in terms of the internal mode variability of the IPSL-CM5A-LR model. Even if this does not translate in terms of actual skill it gives hope for future systems using more efficient initialization techniques, and provides physical explanation for predictive skill.

For linearly detrended SST, both potential and actual prediction skill is of the order of 10 years at the global scale, and this is essentially due to the non-linear response to external forcing. Regionally, the horizon of the potential skill is 1 year in the tropical band, 10 years at mid latitudes in the North Atlantic and in the North Pacific and 5 years at low latitudes in the North Atlantic. These results are generally consistent with previously published single and multi-models analysis, even yielding longer predictability in the North Pacific midlatitudes. This is a particularly important result given the relatively simple initialization strategy used here, namely a weak nudging to observed SST anomalies. This score may come from the model's specific spatial pattern associated to the observed SST variability in the North Pacific, and/or spurious correlation between SST variability in the North Atlantic and North Pacific. Regarding the North Atlantic, we have shown that the nudging helps phasing the SST but in hindcast mode, it is not strong enough to constrain it with respect to the strong internal variability of the model. Few studies analyzed in detail the prediction skill of integrated ocean heat content in such systems. Here, we find surprisingly high actual skill for this variable in the extratropical North Pacific. Over the North Atlantic, it has no actual skill, and neither does the AMOC, but we also underlined very strong discrepancies among the different datasets for this variable, illustrating the difficulties to observe or reconstruct this large-scale feature. The particularly long prediction skill obtained in surface and subsurface over the extratropical North Pacific will deserve a dedicated future study.

Surface SST nudging also proved relatively efficient to induce significant potential predictability of sea surface salinity in the tropics for about three years, which is longer than the prediction skill on SST. In the extratropical North Atlantic, our analysis also showed distinctive behavior resulting from a dominant internal mode of variability at the 20-year timescale in our model. SST nudging indeed exerts a strong influence on SSS, which induces a strong phasing of this variable in the nudged simulation. This leads to a surprisingly long potential predictability of SSS in the extratropical North Atlantic. Comparison with other systems should be performed in order to better understand the robustness and the reasons for this result. Although the mechanism is encouraging, this effect did not induce significant actual skill for SSS. Given promising results regarding the realism of this 20-year timescale in the North Atlantic (e.g. Swingedouw et al 2015), next steps on the path of investigating the performance of surface initialization will consist of testing SSS and surface wind stress initialization. Data uncertainty is presently a strong limitation regarding the use of SSS for decadal prediction initial conditions but hope may come from recent satellite missions.

Acknowledgements This work was supported by the EU project SPECS funded by the European Commissions Seventh Framework Research Program (FP7) under the grant agreement 308378. J.G.-S. was supported by the FP7-funded NACLIM (ENV-308299) project. Computations were carried out at the CCRT-TGCC supercomputing centre. We are grateful to both reviewers for their constructive comments which helped improved the manuscript.

References

Aumont O, Bopp L (2006) Globalizing results from ocean in situ iron fertilization studies. Global Biogeochemical Cycles 20, DOI 10.1029/2005GB002591

Balmaseda MA, Mogensen K, Weaver AT (2013) Evaluation of the ECMWF ocean reanalysis system ORAS4. Quarterly Journal of the Royal Meteorological Society 139(674):1132–1161, DOI 10.1002/qj.2063

Batté L, Déqué M (2012) A stochastic method for improving seasonal predictions. Geophysical Research Letters 39(9), DOI 10.1029/2012GL051406

Bellucci A, Gualdi S, Masina S, Storto A, Scoccimarro E, Cagnazzo C, Fogli P, Manzini E, Navarra A (2013) Decadal climate predictions with a coupled OAGCM initialized with oceanic reanalyses. Climate Dynamics 40(5-6):1483–1497, DOI 10.1007/s00382-012-1468-z

Bellucci A, Haarsma R, Gualdi S, Athanasiadis PJ, Caian M, Cassou C, Fernandez E, Germe A, Jungclaus J, Kröger J, Matei D, Müller W, Pohlmann H, Salas y Melia D, Sanchez E, Smith D, Terray L, Wyser K, Yang S (2014) An assessment of a multi-model ensemble of decadal climate predictions. Climate Dynamics DOI 10.1007/s00382-014-2164-y

Boer GJ, Kharin VV, Merryfield WJ (2013) Decadal predictability and forecast skill. Climate Dynamics 41(7-8):1817–1833, DOI 10.1007/s00382-013-1705-0

Bombardi RJ, Zhu J, Marx L, Huang B, Chen H, Lu J, Krishnamurthy L, Krishnamurthy V, Colfescu I, Kinter JL, Kumar A, Hu ZZ, Moorthi S, Tripp P, Wu X, Schneider EK (2014) Evaluation of the CFSv2 CMIP5 decadal predictions. Climate Dynamics DOI 10.1007/s00382-014-2360-9

Branstator G, Teng H (2010) Two Limits of Initial-Value Decadal Predictability in a CGCM. Journal of Climate 23:6292–6311, DOI 10.1175/2010JCLI3678.1

Branstator G, Teng H (2012) Potential impact of initialization on decadal predictions as assessed for CMIP5 models. Geophysical Research Letters 39(12), DOI 10.1029/2012GL051974

Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Bladé I (1999) The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field. Journal of Climate 12(7):1990–2009, DOI 10.1175/1520-0442(1999)012¡1990:TENOSD¿2.0.CO;2

Carton Ja, Giese BS (2008) A Reanalysis of Ocean Climate Using Simple Ocean Data Assimilation (SODA). Monthly Weather Review 136(8):2999–3017, DOI 10.1175/2007MWR1978.1

Chikamoto Y, Kimoto M, Ishii M, Mochizuki T, Sakamoto TT, Tatebe H, Komuro Y, Watanabe M, Nozawa T, Shiogama H, Mori M, Yasunaka S, Imada Y (2013) An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. Climate Dynamics 40(5-6):1201–1222, DOI 10.1007/s00382-012-1351-y

Collins M, Knutti R, Dufresne JL, Fichefet T, Friedlingstein P, Gao X, Gutowski WJ, Johns T, Krinner G, Shongwe M, Tebaldi C, Weaver AJ, Wehner M (2014) Long-term Climate Change: Projections, Commitments and Irreversibility. In: Stocker T, D Qin GK, Plattner M, Tignor S, Allen J, Boschung A, Nauels Y, Xia Y, Bex P, Midgley V (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

Corti S, Weisheimer A, Palmer TN, Doblas-Reyes FJ, Magnusson L (2012) Reliability of decadal predictions. Geophysical Research Letters 39(21), DOI 10.1029/2012GL053354

Deser C, Phillips AAS, Hurrell JWJ (2004) Pacific interdecadal climate variability: Linkages between the tropics and the North Pacific during boreal winter since 1900. Journal of Climate 17(16):3109–3124, DOI 10.1175/1520-0442(2004)017¡3109:PICVLB¿2.0.CO;2

Di Lorenzo E, Schneider N, Cobb KM, Franks PJS, Chhak K, Miller AJ, McWilliams JC, Bograd SJ, Arango H, Curchitser E, Powell TM, Rivière P (2008) North Pacific Gyre Oscillation links ocean climate and ecosystem change. Geophysical Research Letters 35(8):L08,607, DOI 10.1029/2007GL032838

Ding H, Greatbatch RJ, Latif M, Park W, Gerdes R (2013) Hindcast of the 1976/77 and 1998/99 Climate Shifts in the Pacific. Journal of Climate 26(19):7650–7661, DOI 10.1175/JCLI-D-12-00626.1

Doblas-Reyes FJ, Andreu-Burillo I, Chikamoto Y, García-Serrano J, Guemas V, Kimoto M, Mochizuki T, Rodrigues LRL, van Oldenborgh GJ (2013) Initialized near-term regional climate change prediction. Nature communications 4:1715, DOI 10.1038/ncomms2704

Du H, Doblas-Reyes FJ, García-Serrano J, Guemas V, Soufflet Y, Wouters B (2012) Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. Climate

Dynamics 39(7-8):2013–2023, DOI 10.1007/s00382-011-1285-9

Dufresne JL, Foujols Ma, Denvil S, Caubel A, Marti O, Aumont O, Balkanski Y, Bekki S, Bellenger H, Benshila R, Bony S, Bopp L, Braconnot P, Brockmann P, Cadule P, Cheruy F, Codron F, Cozic A, Cugnet D, Noblet N, Duvel JP, Ethé C, Fairhead L, Fichefet T, Flavoni S, Friedlingstein P, Grandpeix JY, Guez L, Guilyardi E, Hauglustaine D, Hourdin F, Idelkadi A, Ghattas J, Joussaume S, Kageyama M, Krinner G, Labetoulle S, Lahellec A, Lefebvre MP, Lefevre F, Levy C, Li ZX, Lloyd J, Lott F, Madec G, Mancip M, Marchand M, Masson S, Meurdesoif Y, Mignot J, Musat I, Parouty S, Polcher J, Rio C, Schulz M, Swingedouw D, Szopa S, Talandier C, Terray P, Viovy N, Vuichard N (2013) Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Climate Dynamics 40(9-10):2123–2165, DOI 10.1007/s00382-012-1636-1

Dunstone NJ, Smith DM (2010) Impact of atmosphere and sub-surface ocean data on decadal climate prediction. Geophysical Research Letters 37(2), DOI 10.1029/2009GL041609

Dunstone NJ, Smith DM, Eade R (2011) Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. Geophysical Research Letters 38(14), DOI 10.1029/2011GL047949

Escudier R, Mignot J, Swingedouw D (2013) A 20-year coupled ocean-sea ice-atmosphere variability mode in the North Atlantic in an AOGCM. Climate dynamics DOI 10.1007/s00382-012-1402-4

Fedorov AV, Harper SL, Philander SG, Winter B, Wittenberg A (2003) How Predictable is El Niño? Bulletin of the American Meteorological Society 84(7):911–919, DOI 10.1175/BAMS-84-7-911

Ferro CAT (2014) Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society 140(683):1917–1923, DOI 10.1002/qj.2270

Fichefet T, Maqueda MAM (1997) Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. J Geophys Res 102:12,609–612,646, DOI 10.1029/97JC00480

Frankignoul C, Kestenare E (2002) The surface heat flux feedback. Part I: estimates from observations in the Atlantic and the North Pacific. Climate Dynamics 19(8):633–647, DOI 10.1007/s00382-002-0252-x

García-Serrano J, Doblas-Reyes FJ (2012) On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. Climate Dynamics 39(7-8):2025–2040, DOI 10.1007/s00382-012-1413-1

García-Serrano J, Doblas-Reyes FJ, Coelho CaS (2012) Understanding Atlantic multi-decadal variability prediction skill. Geophysical Research Letters 39(18), DOI 10.1029/2012GL053283

García-Serrano J, Guemas V, Doblas-Reyes FJ (2014) Added-value from initialization in predictions of Atlantic multi-decadal variability. Climate Dynamics DOI 10.1007/s00382-014-2370-7

Germe A, Chevallier M, Salas y Mélia D, Sanchez-Gomez E, Cassou C (2014) Interannual predictability of Arctic sea ice in a global climate model: regional contrasts and temporal evolution. Climate Dynamics 43(9-10):2519–2538, DOI 10.1007/s00382-014-2071-2

Giese BS, Ray S (2011) El Niño variability in simple ocean data assimilation (SODA), 1871 2008. Journal of Geophysical Research 116(C2):C02,024, DOI 10.1029/2010JC006695

Goddard L, Kumar a, Solomon a, Smith D, Boer G, Gonzalez P, Kharin V, Merryfield W, Deser C, Mason SJ, Kirtman BP, Msadek R, Sutton R, Hawkins E, Fricker T, Hegerl G, Ferro CaT, Stephenson DB, Meehl Ga, Stockdale T, Burgman R, Greene aM, Kushnir Y, Newman M, Carton J, Fukumori I, Delworth T (2012) A verification framework for interannual-to-decadal predictions experiments. Climate Dynamics DOI 10.1007/s00382-012-1481-2

Guemas V, Doblas-Reyes FJ, Lienert F, Soufflet Y, Du H (2012) Identifying the causes of the poor decadal climate prediction skill over the North Pacific. Journal of Geophysical Research: Atmospheres 117(D20), DOI 10.1029/2012JD018004

Hare SR, Mantua NJ (2000) Empirical evidence for North Pacific regime shifts in 1977 and 1989. Progress in Oceanography 47(2-4):103–145, DOI 10.1016/S0079-6611(00)00033-1

Hazeleger W, Guemas V, Wouters B, Corti S, Andreu-Burillo I, Doblas-Reyes FJ, Wyser K, Caian M (2013a) Multiyear climate predictions using two initialization strategies. Geophysical Research Letters 40(9):1794–1798, DOI 10.1002/grl.50355

Hazeleger W, Wouters B, van Oldenborgh GJ, Corti S, Palmer T, Smith D, Dunstone N, Kröger J, Pohlmann H, von Storch JS (2013b) Predicting multiyear North Atlantic Ocean variability. Journal of Geophysical Research: Oceans 118(3):1087–1098, DOI 10.1002/jgrc.20117

Ho CK, Hawkins E, Shaffrey L, Underwood FM (2012) Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions. Climate Dynamics 41(3-4):917–935, DOI 10.1007/s00382-012-1531-9

Ho CK, Hawkins E, Shaffrey L, Bröcker J, Hermanson L, Murphy JM, Smith DM, Eade R (2013) Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. Geophysical Research Letters 40(21):5770–5775, DOI 10.1002/2013GL057630

Hourdin F, Foujols MA, Codron F (2013) Impact of the LMDZ atmospheric grid configuration on the climate and sensitivity of the IPSL-CM5A coupled model. Climate Dynamics 40(9-10):2167–2192, DOI 10.1007/s00382-012-1411-3

Ingleby B, Huddleston M (2007) Quality control of ocean temperature and salinity profiles Historical and real-time data. Journal of Marine Systems 65(1-4):158–175, DOI 10.1016/j.jmarsys.2005.11.019

Karspeck A, Yeager S, Danabasoglu G, Teng H (2014) An evaluation of experimental decadal predictions using CCSM4. Climate Dynamics pp 1–17, DOI 10.1007/s00382-014-2212-7

Keenlyside N, Latif M, Jungclaus J (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453(May):1–5, DOI 10.1038/nature06921

Kim HM, Webster PJ, Curry JA (2012) Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. Geophysical Research Letters 39(10), DOI 10.1029/2012GL051644

Kim HM, Ham YG, Scaife Aa (2014) Improvement of Initialized Decadal Predictions over the North Pacific Ocean by Systematic Anomaly Pattern Correction. Journal of Climate p 140416111812004, DOI 10.1175/JCLI-D-13-00519.1

Kirtman B, Power S, Adedoyin J, Boer G, Bojariu R, Camilloni I, Doblas-Reyes F, Fiore A, Kimoto M, Meehl G, Prather M, Sarr A, Schär C, Sutton R, van Oldenborgh G, Vecchi G, Wang H, Sch\"ar C, van Oldenborgh G (2013) Near-term Climate Change: Projections and Predictability. In: Stocker T, D Qin GK, Plattner M, Tignor S, Allen J, Boschung A, Nauels Y, Xia Y, Bex P, Midgley V (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental

Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, chap 11, pp 953–1028

Kleeman R, Moore AM (1997) A Theory for the Limitation of ENSO Predictability Due to Stochastic Atmospheric Transients. Journal of the Atmospheric Sciences 54(6):753–767, DOI 10.1175/1520-0469(1997)054¡0753:ATFTLO¿2.0.CO;2

Knight JR, Allan RJ, Folland CK, Vellinga M, Mann ME (2005) A signature of persistent natural thermohaline circulation cycles in observed climate. Geophysical Research Letters 32(20):L20,708, DOI 10.1029/2005GL024233

Kumar A, Wang H, Xue Y, Wang W (2014) How Much of Monthly Subsurface Temperature Variability in the Equatorial Pacific Can Be Recovered by the Specification of Sea Surface Temperatures? Journal of Climate 27(4):1559–1577, DOI 10.1175/JCLI-D-13-00258.1

Latif M, Böning C, Willebrand J (2006) Is the thermohaline circulation changing? Journal of Climate 19:4631–4637, DOI 10.1175/JCLI3876.1

Lozier MS, Leadbetter S, Williams RG, Roussenov V, Reed MSC, Moore NJ (2008) The spatial pattern and mechanisms of heat-content change in the North Atlantic. Science 319(5864):800–3, DOI 10.1126/science.1146436

Luo J, Masson S, Behera S (2005) Seasonal climate predictability in a coupled OAGCM using a different approach for ensemble forecasts. Journal of climate 18:4474–4497, DOI 10.1175/JCLI3526.1

Luo JJ, Masson S, Behera SK, Yamagata T (2008) Extended ENSO Predictions Using a Fully Coupled Ocean-Atmosphere Model. Journal of Climate 21(1):84–93, DOI 10.1175/2007JCLI1412.1

Madec G (2008) NEMO ocean engine. Tech. Rep. 27, Institut Pierre Simon Laplace

Magnusson L, Alonso-Balmaseda M, Corti S, Molteni F, Stockdale T (2012) Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. Climate Dynamics 41(9-10):2393–2409, DOI 10.1007/s00382-012-1599-2

Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (1997) A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. Bulletin of the American Meteorological Society 78(6):1069–1079, DOI 10.1175/1520-0477(1997)078¡1069:APICOW¿2.0.CO;2

Marini C, Frankignoul C (2013) An attempt to deconstruct the Atlantic Multidecadal Oscillation. Climate Dynamics 2013(3-4):607–625, DOI 10.1007/s00382-013-1852-3

Matei D, Pohlmann H, Jungclaus J, Müller W, Haak H, Marotzke J (2012) Two Tales of Initializing Decadal Climate Prediction Experiments with the ECHAM5/MPI-OM Model. Journal of Climate 25(24):8502–8523, DOI 10.1175/JCLI-D-11-00633.1

Meehl G, Teng H (2012) Case studies for initialized decadal hindcasts and predictions for the Pacific region. Geophysical Research Letters 39(22), DOI 10.1029/2012GL053423

Meehl G, Hu A, Tebaldi C (2010) Decadal Prediction in the Pacific Region. Journal of Climate 23:2259–2973, DOI 10.1175/2010JCLI3296.1

Meehl GA, Goddard L, Boer G, Burgman R, Branstator G, Cassou C, Corti S, Danabasoglu G, Doblas-Reyes F, Hawkins E, Karspeck A, Kimoto M, Kumar A, Matei D, Mignot J, Msadek R, Navarra A, Pohlmann H, Rienecker M, Rosati T, Schneider E, Smith D, Sutton R, Teng H, van Oldenborgh GJ, Vecchi G, Yeager S (2014) Decadal Climate Prediction: An Update from the Trenches. Bulletin of the American Meteorological Society 95(2):243–267, DOI 10.1175/BAMS-D-12-00241.1

Mehta VM, Wang H, Mendoza K (2013) Decadal predictability of tropical basin average and global average sea surface temperatures in CMIP5 experiments with the HadCM3, GFDL-CM2.1, NCAR-CCSM4, and MIROC5 global Earth System Models. Geophysical Research Letters 40(11):2807–2812, DOI 10.1002/grl.50236

Merryfield WJ, Lee W, Boer GJ, Kharin VV, Pal B, Scinocca JF, Flato GM (2010) The first coupled historical forecasting project (CHFP1). Atmosphere-Ocean 48(4):263–283, DOI 10.3137/AO1008.2010

Mignot J, Frankignoul C (2003) On the interannual variability of surface salinity in the Atlantic. Clim Dyn 20:555–565, DOI 10.1007/s00382-002-0294-0

Mignot J, Swingedouw D, Deshayes J, Marti O, Talandier C, Séférian R, Lengaigne M, Madec G (2013) On the evolution of the oceanic component of the IPSL climate models from CMIP3 to CMIP5: A mean state comparison. Ocean Modelling 72:167–184, DOI 10.1016/j.ocemod.2013.09.001

Minobe S (2000) Spatio-temporal structure of the pentadecadal variability over the North Pacific. Progress in Oceanography 47(2-4):381–408, DOI 10.1016/S0079-6611(00)00042-2

Mochizuki T, Ishii M, Kimoto M, Chikamoto Y, Watanabe M, Nozawa T, Sakamoto TT, Shiogama H, Awaji T, Sugiura N, Toyoda T, Yasunaka S, Tatebe H, Mori M (2010) Pacific decadal oscillation hindcasts relevant to near-term climate prediction. Proceedings of the National Academy of Sciences of the United States of America 107(5):1833–7, DOI 10.1073/pnas.0906531107

Neelin JD, Battisti DS, Hirst AC, Jin FF, Wakata Y, Yamagata T, Zebiak SE (1998) ENSO theory. Journal of Geophysical Research 103(C7):14,261, DOI 10.1029/97JC03424

Newman M (2007) Interannual to Decadal Predictability of Tropical and North Pacific Sea Surface Temperatures. Journal of Climate 20:2333–2356, DOI 10.1175/JCLI4165.1

Newman M (2013) An empirical benchmark for decadal forecasts of global surface temperature anomalies. Journal of Climate 26(14):5260–5269, DOI 10.1175/JCLI-D-12-00590.1

van Oldenborgh GGJ, Doblas-Reyes FJF, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Climate Dynamics 38(7-8):1263–1280, DOI 10.1007/s00382-012-1313-4

Ortega P, Lehner F, Swingedouw D, Masson-Delmotte, Valerie Raible C, Casado M, Yiou P (2015a) A model-tested North Atlantic Oscillation reconstruction for the past millennium. Nature in press

Ortega P, Mignot J, Swingedouw D, Sévellec F, Guilyardi E (2015b) Reconciling two alternative mechanisms behindbidecadal AMOC variability. Progress in Oceanography 137(A):237–249, DOI 10.1016/j.pocean.2015.06.009

Perigaud CM, Cassou C (2000) Importance of oceanic decadal trends and westerly wind bursts for forecasting El Niño. Geophysical Research Letters 27(3):389–392, DOI 10.1029/1999GL010781

Persechino A, Mignot J, Swingedouw D (2013) Decadal predictability of the Atlantic meridional overturning circulation and climate in the IPSL-CM5A-LR model. Climate dynamics 40(9-10):2359–2380, DOI 10.1007/s00382-012-1466-1

Pohlmann H, Smith DM, Balmaseda Ma, Keenlyside NS, Masina S, Matei D, Müller Wa, Rogel P (2013) Predictability of the mid-latitude Atlantic meridional overturning circulation in a multi-model system. Climate Dynamics 41(3-4):775–785, DOI 10.1007/s00382-013-1663-6

Ray S, Giese BS (2012) Historical changes in El Niño and La Niña characteristics in an ocean reanalysis. Journal of Geophysical Research 117(C11):C11,007, DOI 10.1029/2012JC008031

Ray S, Swingedouw D, Mignot J, Guilyardi E (2015) Effect of surface restoring on subsurface variability in a climate model during 1949-2005. Climate Dynamics 44(9-10):2333–2349, DOI 10.1007/s00382-014-2358-3

Rayner NA (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. Journal of Geophysical Research 108(D14):4407, DOI 10.1029/2002JD002670

Reichler T, Kim J, Manzini E, Kröger J (2012) A stratospheric connection to Atlantic climate variability. Nature Geoscience 5(September):783–787, DOI 10.1038/NGEO1586

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily High-Resolution-Blended Analyses for Sea Surface Temperature. Journal of Climate 20(22):5473–5496, DOI 10.1175/2007JCLI1824.1

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

Séférian R, Bopp L, Gehlen M, Swingedouw D, Mignot J, Guilyardi E, Servonnat J (2014) Multi-year prediction of Tropical Pacific Marine Productivity. PNAS 111(32):11,646–11,651, DOI 10.1073/pnas.1315855111

Servonnat J, Mignot J, Guilyardi E, Swingedouw D, Séférian R, Labetoulle S (2014) Reconstructing the subsurface ocean decadal variability using surface nudging in a perfect model framework. Climate Dynamics DOI 10.1007/s00382-014-2184-7

Smith DM, Scaife Aa, Boer GJ, Caian M, Doblas-Reyes FJ, Guemas V, Hawkins E, Hazeleger W, Hermanson L, Ho CK, Ishii M, Kharin V, Kimoto M, Kirtman B, Lean J, Matei D, Merryfield WJ, Müller Wa, Pohlmann H, Rosati A, Wouters B, Wyser K (2012) Real-time multi-model decadal climate predictions. Climate Dynamics 41(11-12):2875–2888, DOI 10.1007/s00382-012-1600-0

Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. Climate Dynamics 41(11-12):3325–3338, DOI 10.1007/s00382-013-1683-2

Sugiura N, Awaji T, Masuda S, Toyoda T, Igarashi H, Ishikawa Y, Ishii M, Kimoto M (2009) Potential for decadal predictability in the North Pacific region. Geophysical Research Letters 36(20):L20,701, DOI 10.1029/2009GL039787

Sutton RT, Hodson DLR (2005) Atlantic Ocean Forcing of North American and European Summer Climate. Science 309. no. 5(2005):115–118, DOI 10.1126/science.110949616

Swingedouw D, Mignot J, Labetoulle S, Guilyardi E, Madec G (2013) Initialisation and predictability of the AMOC over the last 50 years in a climate model. Climate Dynamics 40(9-10):2381–2399, DOI 10.1007/s00382-012-1516-8

Swingedouw D, Ortega P, Mignot J, Guilyardi E, Masson-Delmotte V, Butler PG, Khodri M, Séférian R (2015) Bidecadal North Atlantic ocean circulation variability controlled by timing of volcanic eruptions. Nature communications 6:6545, DOI 10.1038/ncomms7545

Taylor KE, Stouffer RJ, Meehl GA (2012) An Overview of CMIP5 and the Experiment Design. Bulletin of the American Meteorological Society 93(4):485–498, DOI 10.1175/BAMS-D-11-00094.1

Trenberth KE, Hurrell JW (1994) Decadal atmosphere-ocean variations in the Pacific. Clim Dyn 9(6):303–319, DOI 10.1007/BF00204745

Vial J, Dufresne JL, Bony S (2013) On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. Climate Dynamics 41(11-12):3339–3362, DOI 10.1007/s00382-013-1725-9

Voldoire A, Claudon M, Caniaux G, Giordani H, Roehrig R (2014) Are atmospheric biases responsible for the tropical Atlantic SST biases in the CNRM-CM5 coupled model? Climate Dynamics DOI 10.1007/s00382-013-2036-x

Volpi D, Doblas-Reyes FJ, García-Serrano J, Guemas V (2013) Dependence of the climate prediction skill on spatiotemporal scales: Internal versus radiatively-forced contribution. Geophysical Research Letters 40(12):3213–3219, DOI 10.1002/grl.50557

Weisheimer A, Palmer TN, Doblas-Reyes FJ (2011) Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. Geophysical Research Letters 38(16), DOI 10.1029/2011GL048123

Yeager S, Karspeck A, Danabasoglu G, Tribbia J, Teng H (2012) A Decadal Prediction Case Study: Late Twentieth-Century North Atlantic Ocean Heat Content. Journal of Climate

25(15):5173–5189, DOI 10.1175/JCLI-D-11-00595.1

Yeh SW, Kang YJ, Noh Y, Miller AJ (2011) The North Pacific Climate Transitions of the Winters of 1976/77 and 1988/89. Journal of Climate 24(4):1170–1183, DOI 10.1175/2010JCLI3325.1

Zhang R (2007) Anticorrelated multidecadal variations between surface and subsurface tropical North Atlantic. Geophysical Research Letters 34(12):L12,713, DOI 10.1029/2007GL030225

Zhang S, Rosati A, Delworth T (2010) The Adequacy of Observing Systems in Monitoring the Atlantic Meridional Overturning Circulation and North Atlantic Climate. Journal of Climate 23(19):5311–5324, DOI 10.1175/2010JCLI3677.1

List of Tables

List of Figures

Initialization strategy	ens. size	Start dates	length(yrs)	Name	Remark
Non-initialized	3	yearly (1961-2013)	10	HIST	independent long-term historical simulations: HIST1, HIST2, HIST3
continuous surface nudging	3	yearly (1961-2013)	10	NUDG	independent long-term nudged simulations: NUDG1, NUDG2, NUDG3
surface nudging	3	Every 5 years (1961-2006) (CMIP5)	10	DEC1	launched from NUDG1
surface nudging	3	Every 5 years (1961-2006) (CMIP5)	10	DEC2	launched from NUDG2
surface nudging	3	Yearly (1961-2013)	10	**DEC3**	launched from NUDG3
surface nudging	9	Every 5 years (1961-2006)	10	**DEC9**	from DEC1+DEC2+DEC3

Table 1 table summarizing the hind cast simulations used in this study. We specify in particular the initialization strategy, the number of members of the ensemble, the start dates frequency, the length (in years) of each hindcasts. The final columns gives some additional remarks for clarity.

Atlantic - [30°N-60°N]	std (psu)	EN3	ORAS4	SODA	HIST	NUDG	DEC3	SST
EN3 / ERSST	0.025	1	0.05	**0.77**	0.23	0.08	-0.27	0.35
ORAS4	0.028	-	1	0.17	0.14	0.10	0.17	-0.34
SODA	0.032	-	-	1	0.42	-0.20	-0.47	0.19
HIST	0.065	-	-	-	1	-0.57	-0.66	**0.80**
NUDG	0.094	-	-	-	-	1	**0.79**	**0.64**
DEC3	0.081	-	-	-	-	-	1	**0.69**
Pacific - [30°N-45°N]	std	EN3	ORAS4	SODA	HIST	NUDG	DEC3	SST
EN3 / ERSST	0.016	1	**0.72**	**0.60**	0.32	0.27	0.29	-0.10
ORAS4	0.027	-	1	**0.86**	0.36	0.26	0.32	0.06
SODA	0.019	-	-	1	0.23	0.14	0.14	0.44
HIST	0.016	-	-	-	1	0.24	-0.12	-0.02
NUDG	0.022	-	-	-	-	1	0.51	0.06
DEC3	0.022	-	-	-	-	-	1	0.42

Table 2 correlation between SSS time series in different regions in the reanalysis (ORAS4 and SODA respectively), and the HIST, NUDG and DEC3 time series computed from the model simulations as described in the text at the forecast range 2-5 years. The last column gives the correlation between the SSS and the SST time series for dataset separately. Significant correlation at the 90% level with a two-sided student test have been highlighted in bold

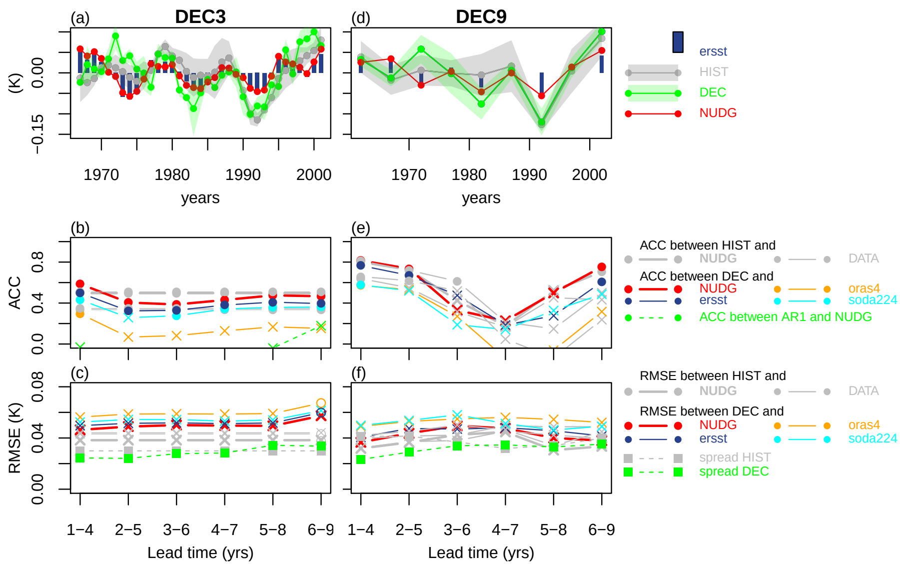

Fig. 1 (a) and (d): Time series of the detrended ensemble mean forecast anomalies averaged over the forecast years 2-5 (green, DEC3 (a), DEC9 (b)) and the accompanying non-initialized (grey) experiments of the global-mean sea surface temperature (SST). The green and grey shadings respectively show the spread of the forecasts. The red line shows the time series from the nudged experiment. The observational time series from the ERSST dataset are represented with dark blue vertical bars, where a 4-year running mean has been applied for consistency with the time averaging of the predictions. The time axis corresponds to the first year of the forecast period (i.e. year 2 of each forecast). (b) and (e): Correlation of the ensemble mean with the NUDG reference (thick red and grey lines respectively, for the DEC and HIST forecast ensembles), along the forecast time for 4-year averages. The figure also shows the correlation of DEC with ERSST (dark blue), ORAS4 (orange) and SODA (light blue) in thin lines, together with their counterparts for the HIST ensemble (grey thin lines', different data sets not identified with colors). Significant correlations according to a one-sided 90% confidence level with a t-distribution are represented with a circle, non significant ones with a cross. The number of degrees of freedom has been computed taking into account the autocorrelation of the time series, which are different for each forecast time. A filled circle indicates significant correlations but not passing a two-sided t-test for the differences between the DEC and HIST correlations. (c) and (f): RMSE of the ensemble mean along the forecast time for 4-year forecast averages are plotted with solid lines. Circles are used where the DEC skill is significantly better than the HIST skill with 90% confidence using a two-sided F-test. Dashed lines represent the ensemble spread estimated as the standard deviation of the anomalies around the multi-model ensemble mean. Green line is for the spread of the initialized hindcasts (DEC3 (c), DEC9 (e)), grey dashed lines for the non-initialized ones.

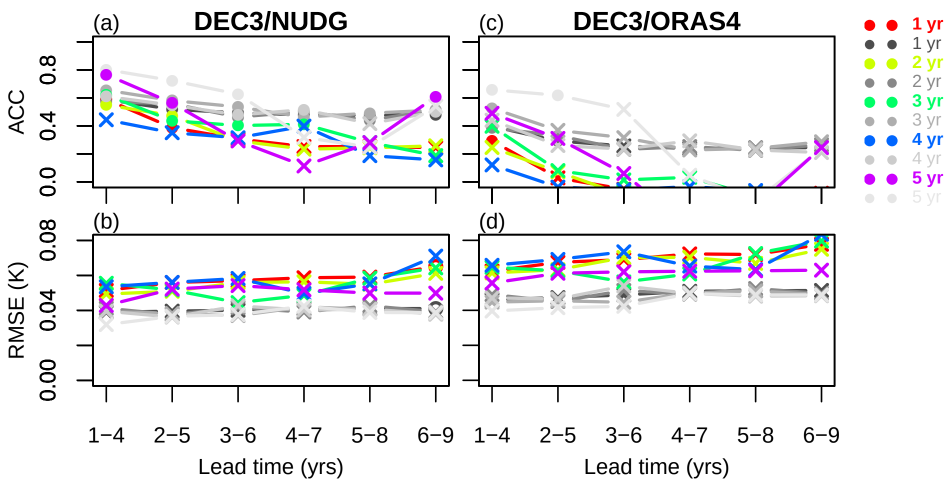

Fig. 2 (a) Potential ACC skill score of global mean SST with start dates taken with an interval of 1 to 5 years from 1961 to 2005 in DEC3. Grey lines show the corresponding skill for the HIST ensemble. (b) as (a) for the RMSE. (c) and (d) Same as (a) and (b) for the skill scores computed against ORAS4. Hindcasts launched between 1961 and 2005 were used here, but anomalies were not computed against a common verification period since this would be too restrictive for the longest start date intervals (see section 2.4 for details).

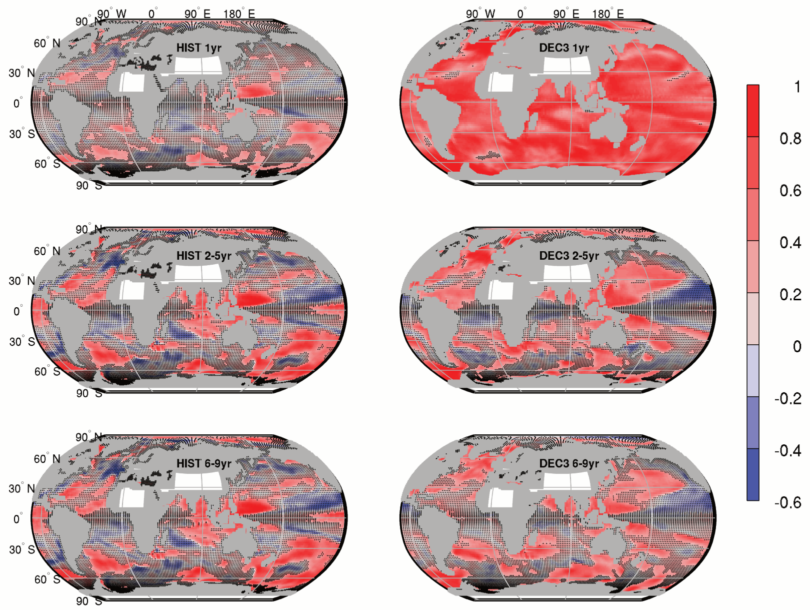

Fig. 3 ensemble mean ACC of detrended SST in the HIST (left) and DEC3 (right) hindcasts against the NUDG simulation, for a lead time of 1 year (top), 2-5 years (middle) and 6-9 years (bottom). Non-significant correlations at the 90% confidence level are marked with black dots.

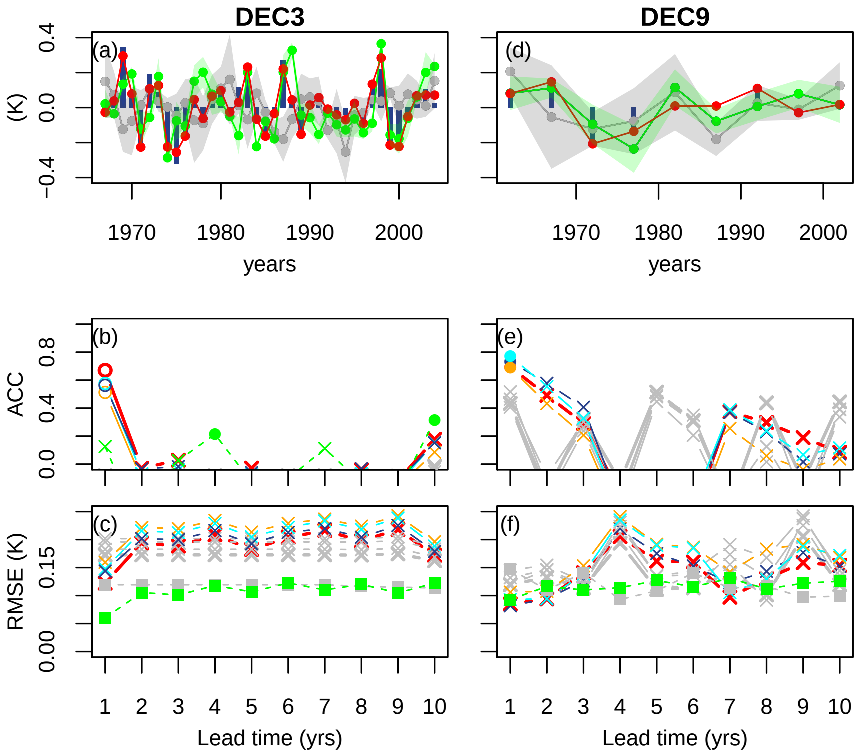

Fig. 4 Same as Fig. 1 for SST averaged over the region [20°S-20°N]. In the upper panels, HIST and DEC time series are considered for a lead time of 1 year. In the middle and bottom panels, note that the forecast ranges are not 4-year averaged.

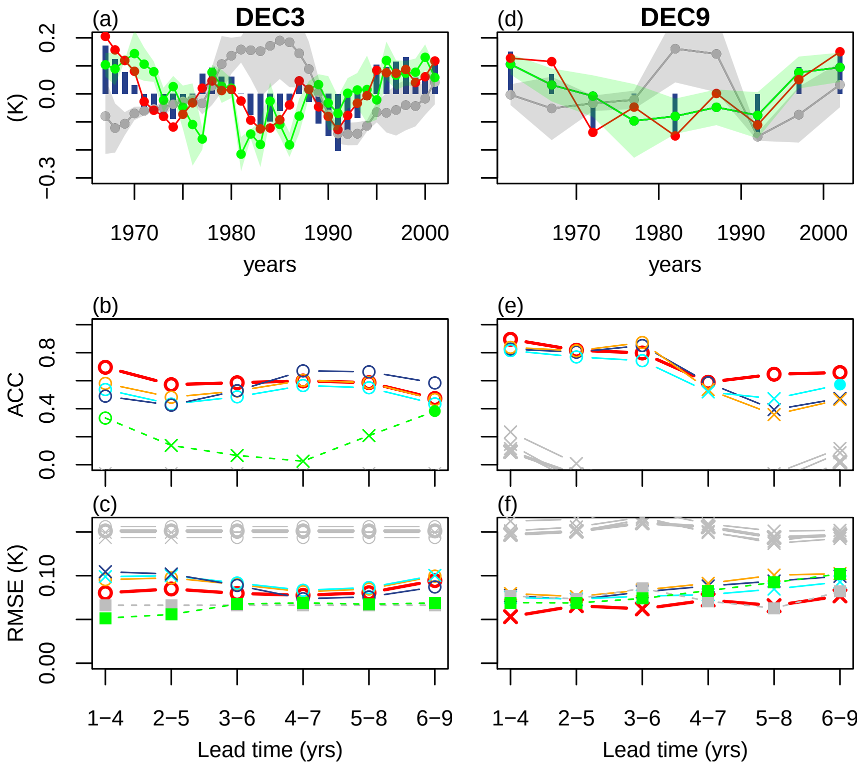

Fig. 5 Same as Fig. 1 for SST averaged over the region [0-60°N] in the Atlantic

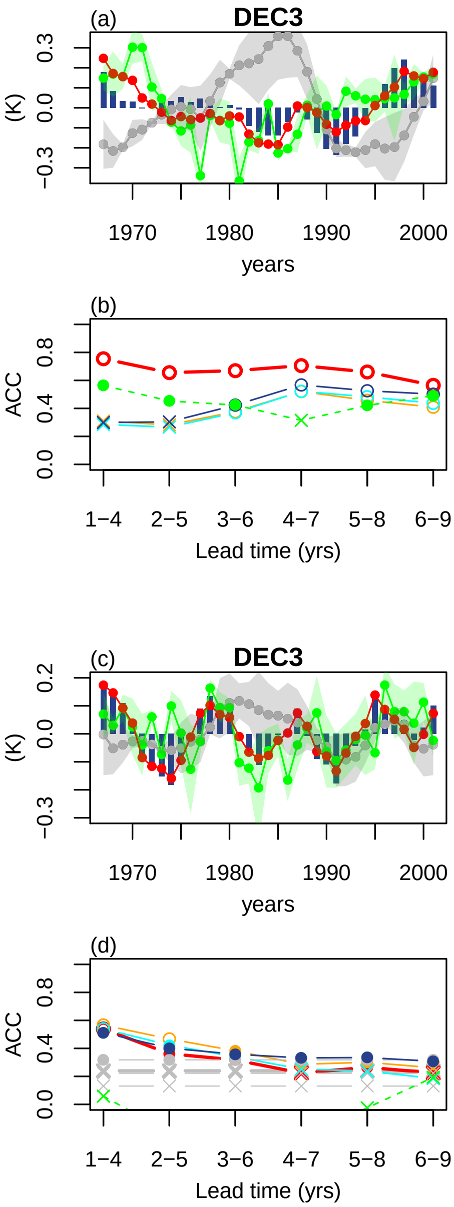

Fig. 6 Same as Fig. 1 (a) and (b) for SST averaged over the mid latitudes [30°N-60°N] (a and b) and low latitude [0-30°N] (c and d) in the Atlantic.

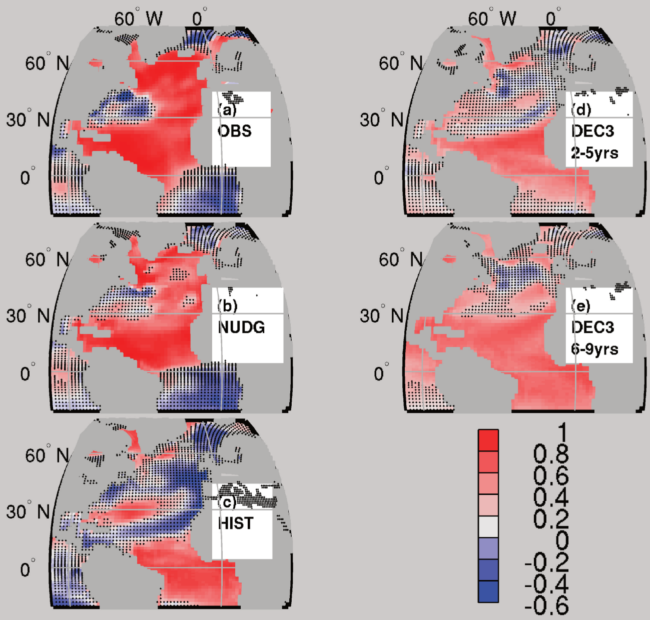

Fig. 7 Correlation of observed ERSST time series averaged between 0 and 60°N in the Atlantic against the SST field in (a) ERSST (b-c) NUDG and HIST respectively, (d-e) DEC3 at forecast range 2-5 years and 6-9 years respectively. All SST fields are linearly detrended and considered as averages over 4 consecutive years. Non-significant correlations at the 90% level are marked with the black dots.

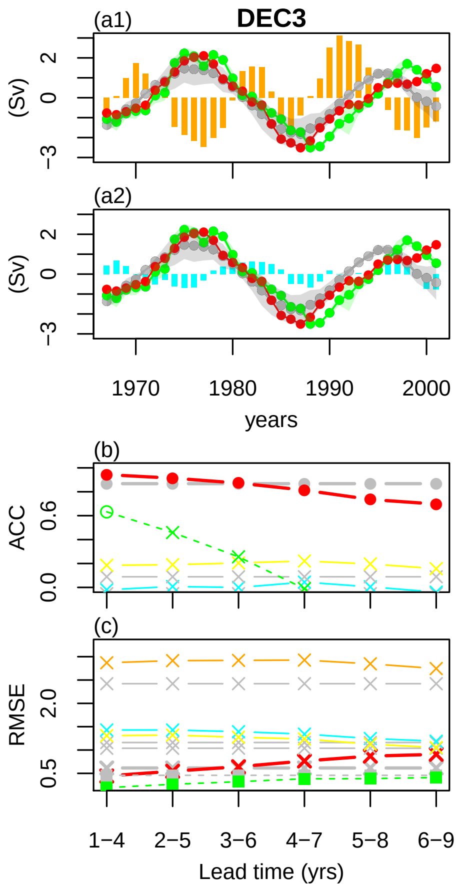

Fig. 8 Same as Fig. 1 for the AMOC maximum at 48°N verified against ORAS4 (a1) and SODA (a2). The yellow line on panel (b) and (c) shows the skill scores (ACC and RMSE) of the AMOC computed against the reconstruction proposed by Latif et al (2006), using a dipole of SST between the Northern and Southern Atlantic.

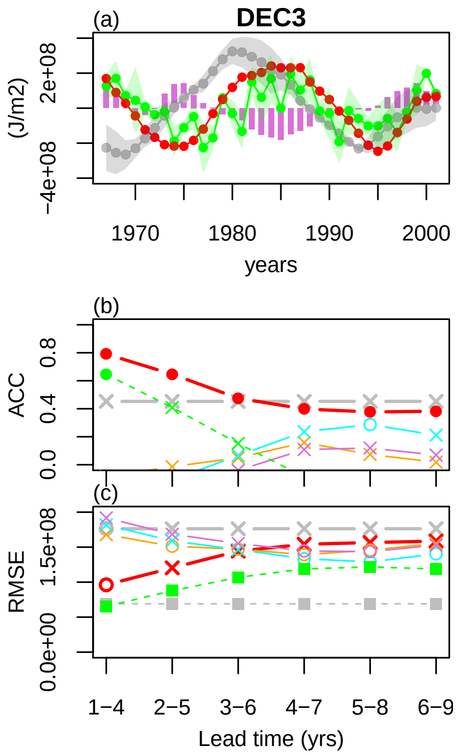

Fig. 9 Same as Fig. 1 for the oceanic heat content integrated down to 300m and averaged over the North Atlantic sub polar region [30°N-60°N]. The purple bars in panel (a) and purple lines in panel (b) and (c) correspond to the heat content computed from the EN3 dataset.

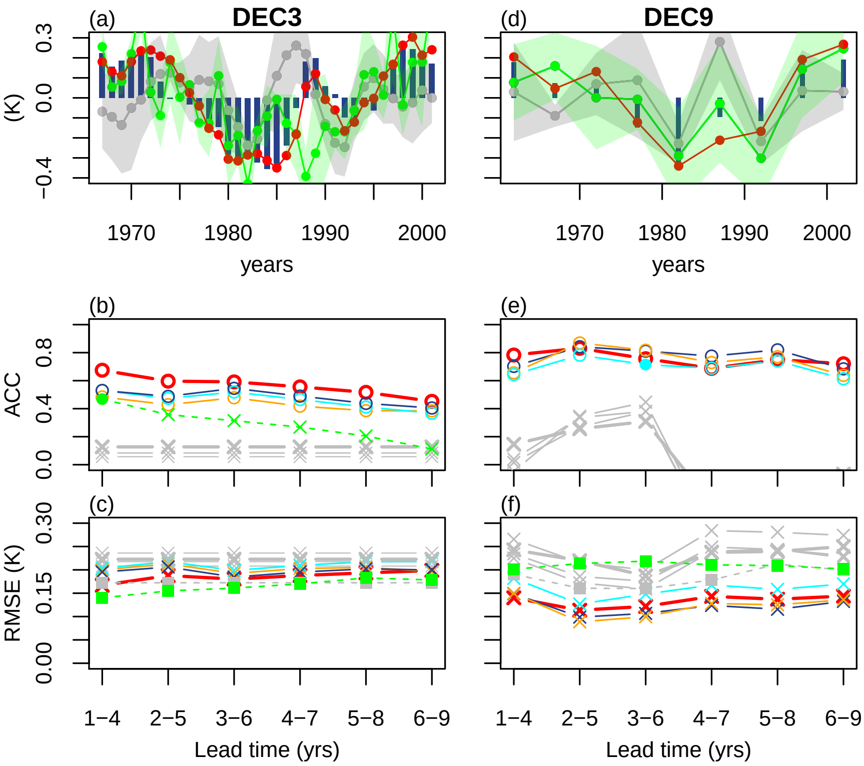

Fig. 10 Same as Fig. 1 for SST averaged over the region [30°N-45°N] in the Pacific

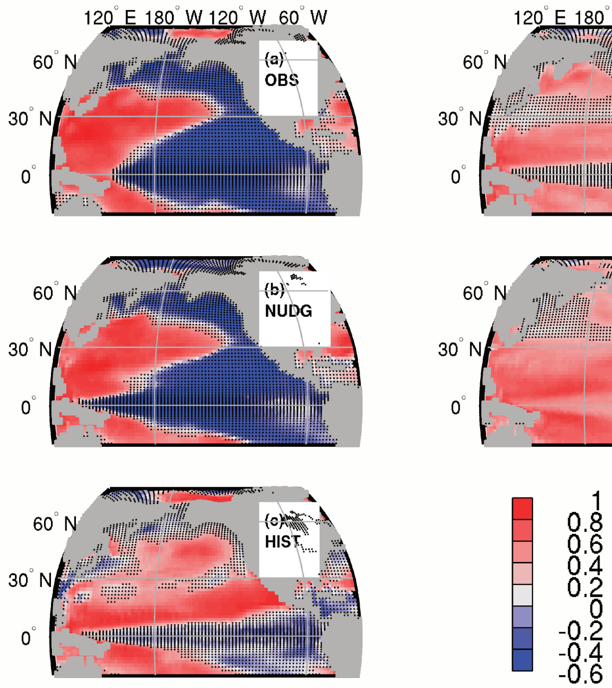

Fig. 11 Correlation of observed ERSST time series averaged between 30°N and 45°N in the Pacific against the SST field in (a) ERSST (b-c) NUDG and HIST respectively, (d-e) DEC3 at forecast range 2-5 years and 6-9 years respectively. All SST fields are linearly detrended and considered as averages over 4 consecutive years. Non-significant correlations at the 90% level are marked with the black dots.

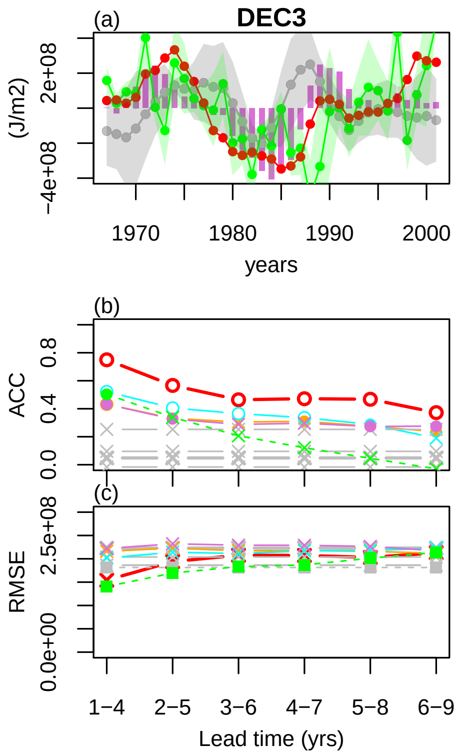

Fig. 12 Same as Fig. 9 averaged over the Pacific extratropical region [30°N-45°N].

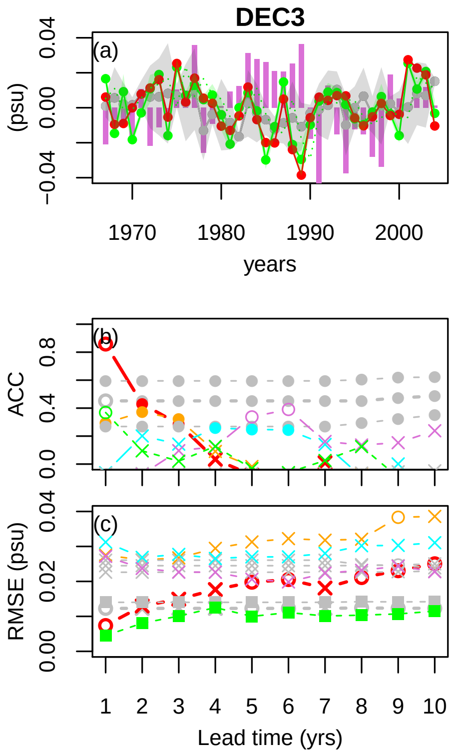

Fig. 13 Same as Fig. 4 (left), but for the SSS (average over the latitude band [20°S-20°N]). The purple bars in panel (a) and purple lines in panel (b) and (c) are from EN3 dataset.

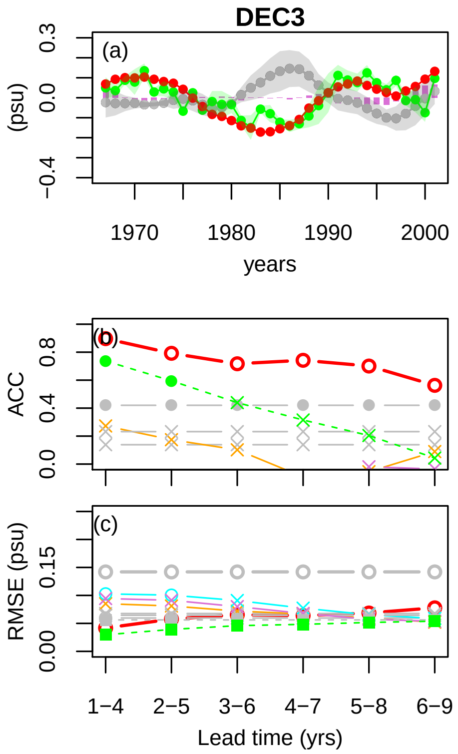

Fig. 14 Same as Fig. 1 for SSS averaged over the region [30°N-60°N] in the Atlantic