Clim Dyn (2018) 51:1947–1970
DOI 10.1007/s00382-017-3991-4

Time dependency of the prediction skill for the North Atlantic subpolar gyre in initialized decadal hindcasts

Sebastian Brune[1] · **André Düsterhus**[1] · **Holger Pohlmann**[2] · **Wolfgang A. Müller**[2] · **Johanna Baehr**[1]

Received: 12 January 2017 / Accepted: 22 October 2017 / Published online: 9 November 2017

Abstract We analyze the time dependency of decadal hindcast skill in the North Atlantic subpolar gyre within the time period 1961–2013. We compare anomaly correlation coefficients and temporal interquartile ranges of total upper ocean heat content and sea surface temperature for three differently initialized sets of hindcast simulations with the global coupled model MPI-ESM. All initializations use weakly coupled assimilation with the same full value nudging in the atmospheric component and different assimilation techniques for oceanic temperature and salinity: (1) ensemble Kalman filter assimilating EN4 observations and HadISST data, (2) nudging of anomalies to ORAS4 reanalysis, (3) nudging of full values to ORAS4 reanalysis. We find that hindcast skill depends strongly on the evaluation time period, with higher hindcast skill during strong multi-year trends, especially during the warming in the 1990s and lower hindcast skill in the absence of such trends. Differences between the prediction systems are more pronounced when investigating any 20-year subperiod within the entire hindcast period. In the ensemble Kalman filter initialized hindcasts, we find significant correlation skill for up to 5–8 lead years, albeit along with an overestimation of the temporal interquartile range. In the hindcasts initialized by anomaly nudging, significant correlation skill for lead years greater than two is only found in the 1980s and 1990s. In the hindcasts initialized by full value nudging, correlation skill is consistently lower than in the hindcasts initialized by anomaly nudging in the first lead years with re-emerging skill thereafter. The Atlantic meridional overturning circulation reacts on the density changes introduced by oceanic nudging, this limits the predictability in the subpolar gyre in the first lead years. Overall, we find that a model-consistent assimilation technique can improve hindcast skill. Further, the evaluation of 20 year subperiods within the full hindcast period provides essential insights to judge the success of both the assimilation and the subsequent hindcast quality.

1 Introduction

In decadal climate prediction, the skill depends on both the boundary conditions and the initial conditions (e.g., Cox and Stephenson 2007; Branstator and Teng 2012). Boundary conditions are addressed by using comprehensive globally coupled earth system models (ESM) in conjunction with prescribing the main external forcings such as solar irradiance, past volcanic activity, and atmospheric greenhouse gas concentrations. Initial conditions can be produced by assimilating past observations of the state of the climate system, such as atmospheric temperature, velocity and pressure, and oceanic temperature and salinity into the ESM (e.g., Smith et al. 2007; Pohlmann et al. 2009, among many others). The German initiative MiKlip (Marotzke 2016) develops a decadal prediction system based on the Max Planck Institute for Meteorology Earth System Model (MPI-ESM). This system has been widely used within MiKlip to test initialization strategies and subsequent hindcast simulations in the context of interannual to decadal forecasts (e.g., Müller et al. 2012; Pohlmann et al. 2013; Romanova and Hense 2015; Marini et al. 2016). The standard assimilation technique currently used in MiKlip is based on nudging to external reanalysis data, either anomaly or full value (Marotzke 2016).

✉ Sebastian Brune
sebastian.brune@uni-hamburg.de

[1] Institute of Oceanography, CEN, Universität Hamburg, Bundesstr.53, 20146 Hamburg, Germany

[2] Max Planck Institute for Meteorology, Bundesstr.53, 20146 Hamburg, Germany

Advantages and disadvantages of both anomaly nudging and full value nudging in initializing decadal predictions are controversially discussed (Smith et al. 2013; Volpi et al. 2016). Within the MiKlip prediction systems, however, both methods may severely hamper the ocean heat budget during assimilation and lead to strong imbalances in any subsequent forecasts, as recently investigated by Kröger et al. (in revision for Climate Dynamics) for the North Atlantic.

In recent years the use of climate predictions systems relying on the ensemble Kalman filter in assimilation has become more popular, e.g., Chang et al. (2013); Msadek et al. (2014) with the GFDL Climate Model (weakly coupled atmospheric and oceanic assimilation), Counillon et al. (2014) with the Norwegian climate model (ocean only assimilation), Karspeck et al. (2013, 2015b) with the Community Climate System Model (ocean assimilation forced by independently generated atmospheric assimilation). Brune et al. (2015) implemented an oceanic assimilation scheme in MPI-ESM based on the singular evolutive ensemble Kalman Filter (Pham et al. 1998; Pham 2001), a variant of the ensemble Kalman Filter (EnKF, Evensen 1994) using the Parallel Data Assimilation Framework (Nerger and Hiller 2013). In their setup they use a free atmosphere within MPI-ESM and the EnKF assimilates only oceanic temperature and salinity profiles from EN3, the predecessor of EN4 (Good et al. 2013), and sea surface temperature from HadISST (Rayner et al. 2003). An advantage of their EnKF assimilation is the consistent and direct assimilation of observations into the prediction system, without the need of a different ocean model or ESM, e.g., an oceanic re-analysis, facilitating the incorporation of observations. In their experiments Brune et al. (2015) detect significantly high correlation with reference data in the tropical and subtropical oceans. For our study the oceanic assimilation has been complemented by the standard MiKlip full value nudging to atmospheric re-analyses ERA40 (Uppala 2005) and ERAInterim (Dee 2011).

Typically, the skill of a decadal prediction system is assessed by simulating a large enough ensemble of initialized retrospective forecasts (hindcasts) with enough start dates over a time period in the past, where observations exist and against which we can test the hindcasts. For instance, in their decadal prediction system, Mignot et al. (2016) proposed for the time period 1961–2013 yearly start dates for the hindcasts and a reasonably large ensemble size of 9 or larger. The hindcast evaluation over the past is used to assess the skill of any real-time decadal forecast (Smith 2013). Although we have to acknowledge know that the external forcing, especially the volcanic forcing, is not exactly known for any real-time forecast, we use the known external forcing for our hindcasts. Therefore the skill in hindcast mode could be larger than the one expected in forecast mode (Timmreck et al. 2016). Nonetheless, the difference in skill between the prediction systems under consideration should remain the same.

Common methods to evaluate decadal hindcasts are often based on the phase (e.g., correlation) or distance (e.g., the root-mean-square error) compared to observational or near-observational reference data over a certain time period, and most often the maximum period common to both hindcast and observation is used. For subsurface oceanic temperatures, which are thought to make an important contribution to the variability of the climate system, variability can be dominated by time scales longer than 10 years. A prominent example in the North Atlantic region is the Atlantic multidecadal variability (AMV, Trenberth and Shea 2006; Smith et al. 2013). The hindcast skill of the model may depend on the specific phases of this variability. As such the skill of the initialization of the hindcasts and subsequently the skill of the hindcast simulations can be dependent on time.

Previous studies have shown that the North Atlantic is a key region to assess decadal predictions. Pronounced decadal prediction skill in the North Atlantic subpolar gyre (SPG) sea surface temperature and upper ocean heat content has been found in studies using initialized hindcasts with earlier versions of MPI-ESM (Pohlmann et al. 2009; Kröger et al. 2012; Matei et al. 2012; Polkova et al. 2015). The cooling of the SPG in the 1960s (Robson et al. 2014) and the warming of the SPG in the 1990s (Yeager et al. 2012; Robson et al. 2012; Msadek et al. 2014; Müller et al. 2014) could have been predicted using initialized forecasts of a global coupled ESM. Variability of the climate in the North Atlantic region and variability of the large scale oceanic circulation are thought to be deeply intertwined (Delworth et al. 1993). A probable link between atmosphere and ocean may be represented by the North Atlantic oscillation influencing near surface temperatures in the Labrador and Irminger Seas in the Western SPG (Yeager and Robson 2017, and references therein). Here, deep convection and densities at the western boundary of the North Atlantic may be modified, which in turn influence the the Atlantic meridional overturning circulation (AMOC, Hermanson et al. 2014; Buckley and Marshall 2016; Menary et al. 2016). Vice versa the variability in the AMOC and its associated northward heat transport is influencing temperatures in the upper ocean in the whole North Atlantic region (Pohlmann et al. 2006; Keenlyside et al. 2008; Zhang and Zhang 2015). In a recent study Delworth et al. (2017) established that on decadal and longer time scales the AMV responds to prolonged NAO phases, facilitated by large scale ocean dynamics in the North Atlantic Ocean, especially the AMOC.

In this study, we will compare the skill of MPI-ESM for upper ocean heat content (0-700m, HCT700) and sea surface temperature (SST) in the North Atlantic SPG in three differently initialized retrospective decadal forecasts (hindcasts) and an uninitialized simulation between 1961 and 2013. We use the integrated quantity HCT700 to assess the potential impact of deeper layers of the ocean on our analysis, compared to SST, which is more influenced by the atmosphere than HCT700. All three initialization methods use full value nudging to ERA40 and ERAInterim re-analysis in the atmosphere. Following methods are applied to incorporate information of oceanic temperature and salinity and complement the atmospheric assimilation: (1) EnKF assimilation of observational profiles, (2) anomaly nudging to ORAS4 re-analysis, (3) full value nudging to ORAS4 re-analysis. Because of the strong connections to oceanic temperatures in the SPG we will interpret our results in context with the simulated AMOC. Also, as outlined earlier, the AMOC is thought to represent an important link in the coupling of atmosphere and ocean on multidecadal time scales. As a highly integrated quantity the AMOC is sensitive to many parameters and therefore well suited to detect problems induced by any of the three initialization methods. However, the lack of observational data for the AMOC before the 2000s prohibits a full scale analysis of the AMOC similar to HCT700 and SST.

We specifically would like to establish an analysis scheme to identify differences in the MPI-ESM decadal hindcasts induced by the three initialization methods depending on the evaluation period. With that we lay the foundation for and try to start with an explanation why such differences exist and what they may implicate for the future development and interpretation of decadal prediction systems based upon MPI-ESM or other ESMs. For that purpose we evaluate hindcast correlations and interquartile ranges dependent on lead time up to 10 years and reference time periods of 20 years within 1961–2013. For the first time we would like to qualitatively assess the dependence of hindcast skill in MPI-ESM on the evaluation period. From our perspective this knowledge is needed in interpreting any real-time forecast skill.

The variability of our simulated decadal time series depends on three time scales: the long term trend with a time scale larger than 60 years, the intra- to multidecadal variability with a time scale of 2–30 years, the interannual variability with a time scale of 1–2 years. Our hindcasts draw skill from all three of them, removing any of them will therefore hamper a comprehensive analysis of the hindcast skill. The long term trend in all our simulations is mainly caused by the external forcing. The uninitialized historical simulation serves as a reference for the influence of the long term trend, we use the comparison of skill between the initialized hindcasts and the historical simulation to discuss the added value of the initialized hindcasts. That difference we assume to come mostly from the intra- to multidecadal time scale.

The remainder of this paper is structured as follows: we describe our initialized hindcasts, the uninitialized historical model run, and the observations as well as our analysis methods in Sect. 2. Results of our experiments for total upper ocean heat content and sea surface temperature are shown in Sect. 3, AMOC results are shown in Sect. 4. We summarize results and discussion and conclude our study with the main findings in Sect. 5.

2 Data and methods

2.1 Model simulations

We use the Max Planck Institute Earth system model with low resolution (MPI-ESM-LR, Giorgetta 2013), version 1.0.02, which consists of the atmospheric component ECHAM6 (Stevens 2013, ≈ 2.5° horizontal resolution, 47 levels up to 1 hPa), and the oceanic component MPIOM (Jungclaus et al. 2013, ≈ 1.5° horizontal resolution, 40 depth levels), both coupled once a day by OASIS3 (Valcke 2013). For all simulations we use observed solar irradiation, volcanic eruptions, and atmospheric greenhouse gas concentrations (RCP4.5 concentrations from 2006 onward) as boundary conditions, taken from CMIP5 (Taylor et al. 2012).

We analyze three different sets of hindcast simulations initialized by three different weakly coupled assimilation experiments and one uninitialized historical ensemble of simulations (see Table 1 for details). In all three assimilation experiments full values of atmospheric vorticity, divergence and temperature above the boundary layer, and sea level pressure are nudged to ECMWF reanalysis data (ERA40/ERAInterim, Uppala 2005; Dee 2011) at every atmospheric model timestep. Relaxation times are 6 h for vorticity, 48 h for divergence, and 0.25 h for temperature and sea level pressure. Assimilation information can only be exchanged between ECHAM6 and MPIOM during coupling times: once a day, hence all three assimilation simulations are weakly coupled.

The three assimilation experiments differ only in the way oceanic temperature and salinity is incorporated. In the EnKF assimilation experiment we use an 8-member global ensemble Kalman filter without inflation and localization (via the Parallel Data Assimilation Framework, PDAF, Nerger and Hiller 2013) to assimilate observed surface temperatures (HadISST, Rayner et al. 2003) and subsurface temperatures and salinities (EN4, Good et al. 2013) on a monthly basis into MPIOM (Brune et al. 2015). In their experiments assimilation of sub-50 m observations only leads to a better result compared to assimilation of all observations, when used together with a free atmosphere. Here, we include near-surface observations above 50 m depth to strengthen the influence of the oceanic component in context with the nudging we apply to the atmospheric component. For the EnKF analysis step we set the representativeness

Table 1 Overview of the assimilation experiments and the hindcast simulations, characteristics of the uninitialized simulation HIST are listed under hindcasts as well

	EnKF	NDGa	NDGf	HIST
Model	MPI-ESM-LR 1.0, ECHAM6 T63L47, MPIOM GR15L40			
Assim. data ocean	EN4, HadISST	ORAS4		–
Assim. method ocean	EnKF global variant	anomaly nudging	full nudging	–
	T, S, SST	T, S	T,S	
	obs. error 1 K, 1 psu	$\tau = 12$ days, 12 days	$\tau = 12$ days, 12 days	
	no inflation			
Assim. data atmosphere	ERA40/ERAInterim			–
Assim. method atmosphere	full nudging			–
	vorticity, divergence, and T above boundary layer, SLP			
	$\tau = 6$ h, 48 h, 0.25 h, 0.25 h			
Assim. ensenmble size	8	1	1	–
Hindcast initialization	direct	1-day lagged		pre-industrial
	from assim.	from January after assim.		control
Hindcast start	yearly at 1st of January 1961–2013			1850
Hindcast length	10 years			1850–2014
Hindcast ensemble size	8	8	8	8

error of the observations to 1 K for temperature and 1 psu for salinity.

In the anomaly nudged experiment baseline-1 (henceforth NDGa) model temperatures and salinities are nudged to ECMWF oceanic reanalysis (ORAS4, Balmaseda et al. 2013) anomalies added to the model climatology with respect to the 1958–2005 mean at every oceanic model timestep (Pohlmann et al. 2013, Kröger et al., in revision for Climate Dynamics). In the full value nudging experiment prototype-ORA (henceforth NDGf) the full values of temperatures and salinities are nudged to ORAS4 at every oceanic model timestep (Kröger et al., in revision for Climate Dynamics). In both NDGa and NDGf the nudging operates on the entire water column. The relaxation times are 12 days for temperature and 12 days for salinity. Thus the nudging is on the strong side of what has been recommended by DCPP-C (2016): 12–60 days, which is based among other works on Servonnat et al. (2015). Still, the nudging is much weaker than the 6 h Smith et al. (2013) used in their experiments. Due to the strong constraint the nudging technique imposes on the model, NDGa and NDGf assimilation experiments consist of only one member each.

For all three experiments hindcasts are initialized January, 1st every year from 1961 to 2013 from their respective assimilation experiment and are running freely with CMIP5 external forcing for 10 years. The three sets of hindcasts are initialized as follows: each member of the 8-member EnKF hindcast is directly initialized from the respective member of the EnKF assimilation experiment at the beginning of each January. For the 8-member NDGa hindcast and the 8-member NDGf hindcast we use 1-day-lagged initialization from the respective assimilation experiments, thus member 1 is initialized by the assimilation as of January 1st, member 2 is initialized by the assimilation as of January 2nd, and so forth, then model dates are reset and all members start at January 1st.

The uninitialized model simulation (HIST) is represented by an 8-member ensemble started from different states of a pre-industrial control run (part of the contribution to CMIP5), start dates have been set to January 1850 and the simulations run until the end of 2013 using the CMIP5 external forcings.

2.2 Reference data and heat content calculation

We assess the hindcast skill for 0–700 m total upper ocean heat content (HCT700) and sea surface temperature (SST), using NOAA's Ocean Climate Laboratory heat content data (NOCL, Levitus et al. 2012), and the Met Office Hadley Centre's sea ice and sea surface temperature data set (HadISST, Rayner et al. 2003) as reference data. Both data sets are created from raw observations using varieties of the objective analysis approach. Since observations of subsurface oceanic quantities are sparse compared to those of ocean surface data, we assume the NOCL heat content data to be less accurate than the HadISST data.

For the AMOC, reference data is very limited and only available at specific latitudes such as 26°N where RAPID-MOCHA (Smeed et al. 2014) became operational in 2004. This time series is not long enough for our analysis. We therefore compare our simulated AMOC against multi-model results published in the studies of Pohlmann et al. (2013), and (Karspeck et al. 2015a).

We define the total upper ocean heat content (0–700m) HCT700 for each horizontal grid cell i as:

$$HCT700(i) = c_p \rho_{ref} A(i) \sum_{z(k)=0m}^{z(k)=700m} \theta(k)(z(k) - z(k-1)), \quad (1)$$

with the heat capacity of water $c_p = 4000\, J\, kg^{-1}\, K^{-1}$, the reference density of water $\rho_{ref} = 1025\, kg\, m^{-3}$, the area $A(i)$ of the grid cell i, the depth $z(k)$ in m and the potential temperature $\theta(k)$ in °C of layer k. Our definition of the upper ocean heat content differs from the one used by Levitus et al. (2012) in that we do not calculate the anomaly with reference to the 1961-1990 mean. Thus there is a time independent bias between simulation and NOCL reference data, which our analysis, consisting of correlations and interquartile ranges, is not depending on.

2.3 Analysis method

We base our study on yearly mean quantities over the period 1961 to 2013 in the North Atlantic SPG (60°W–10°W, 50°N–65°N, Robson et al. 2012, see black rectangle in Fig. 1), all data is interpolated to a 1° grid. For HCT700 we sum over all grid cells within the SPG and down to 700 m depth, for SST we average over all grid cells within the SPG.

Prior to the analysis of HCT700 and SST we correct the bias of the initialized hindcasts with respect to the assimilation experiment used for initialization analogously to the bias correction scheme proposed in ICPO (2011) and Pohlmann et al. (2013). For any lead year t of any 10-year hindcast simulation x initialized in year s we subtract the mean difference between the same lead year in all other 10-year hindcast simulations and the respective assimilation experiment a. The corrected hindcast value $y_s(t)$ is therefore calculated as:

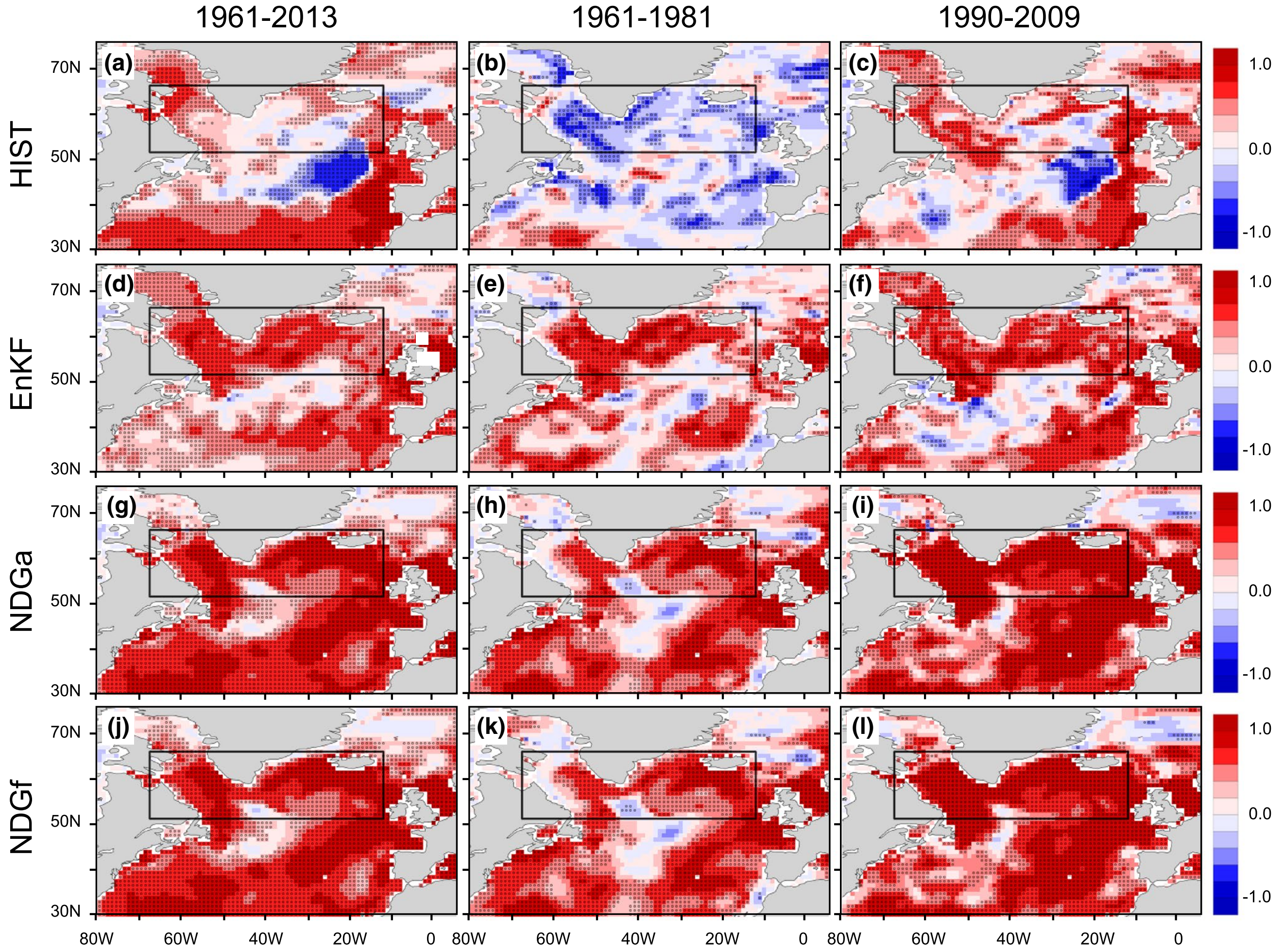


Fig. 1 Correlation of total upper ocean heat content (0–700m) with NOCL (Levitus et al. 2012) during assimilation for 1961–2013 (left column), 1961–1981 (middle column), and 1990–2009 (right column) for **a–c** HIST, **d–f** EnKF, **g–i** NDGa, **j–l** NDGf. Stippling indicates significance at the 95% level. The geographic definition of the North Atlantic subpolar gyre is indicated by the black rectangle

$$y_s(t) = x_s(t) - \frac{1}{52} \sum_{\substack{i = 1961 \\ i \neq s}}^{2013} x_i(t) - a_i \tag{2}$$

Please note again that we use as the reference for each hindcast system the respective assimilation experiment instead of observations.

For our analysis we construct lead year time series from our 10 year long individual hindcasts. Thus, for instance, the lead year 1 time series running from 1961 to 2013 is a result of concatenating the first lead years of every individual hindcast.

For our analysis we do not remove any trend because from the physical point of view we treat trends as an integral part of the multidecadal variability. Also, from the statistical point of view it is difficult to remove trends over all evaluation periods and in all simulations in a consistent way.

We measure the hindcast skill by assessing the temporal variability of simulations and reference data in two ways. We use the anomaly correlation coefficient between simulations and reference data to assess the phase of the variability, and we compare the interquartile range (IQR) of the temporal probability distribution of each simulation and reference data to assess the amplitude of the variability.

The correlation between hindcast and reference data for any evaluation period (start time t_s, end time t_e) is computed from the respective SPG lead year time series averaged over the whole ensemble (Pearson or ordinary correlation, see Wilks 2011):

$$\mathrm{Correlation}_{t_e,t_s}(x,y) = \frac{\sum_{i=t_s}^{t_e}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=t_s}^{t_e}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=t_s}^{t_e}(y_i - \bar{y})^2}}, \tag{3}$$

with the hindcasts ensemble mean x, the reference ensemble mean y, and their time means $\bar{x}$ and $\bar{y}$, all calculated over the respective evaluation period (t_s,t_e).

In this study we use the temporal IQR over any evaluation period of the lead year time series. We define it as the difference between the 25th and 75th percentile (Wilks 2011) of the respective SPG lead year time series of all individual ensemble members combined:

$$\mathrm{IQR}_{t_0,t_n}(z,ly) = q_{0.75}(z^1_{t_0} \ldots z^N_{t_n}, ly) - q_{0.25}(z^1_{t_0} \ldots z^N_{t_n}, ly), \tag{4}$$

with the lead year ly, z any of the hindcast or reference data sets, N the corresponding ensemble size, and start time t_0 and end time t_n of the evaluation period. In our study, the temporal IQR represents a measure of the distribution of only the internal amplitudes of all time series within an ensemble, not only the ensemble mean. A wide (narrow) distribution leads to a large (small) IQR. If the distribution is other than normal, e.g., bimodal, the IQR is more robust than the standard deviation. The temporal IQR shall not be confused with the ensemble spread. However, the larger the ensemble spread when compared to the temporal amplitude, the wider the distribution and the larger the IQR. By definition and unlike all error based measures the temporal IQR is independent from any bias in the mean of the time series or any phase shifts within the time series. The computation of the temporal IQR does not require any reference data. However, for interpretation the simulated temporal IQR needs to be compared against the temporal IQR of reference data as a measure of the similarity in the distribution of the time series over all resolved time scales. In the remainder of this manuscript we will abbreviate the term "temporal IQR" to "IQR".

Both correlation and IQR are calculated for the entire hindcast period 1961–2013, and for all 20 year time periods within 1961–2013, i.e., 1961–1980, 1962–1981, and so on until 1994–2013. For periods incorporating 1961–1969, the number of data points may be reduced by 1–10, because the first hindcast of lead year 2 is available only for 1962, for lead year 3 for 1963, for lead 10 for 1970. Consequently, full data coverage for the 20-year periods for all lead years is maintained for periods starting after 1969. On the use of a 20-year subperiod for our analysis: we also tested a shorter subperiod of 15 years, compared to 20 years results were noisy and more uncertain. And we tested a longer subperiod of 30 years, compared to 20 years results were less noisy but also with less difference between the systems.

We compute the significance of correlation values at the 95% level by bootstrapping (500 realizations) both the simulated and reference ensemble mean time series and evaluating both tails of the resulting probability distribution with a p value of 0.025, respectively. In a similar manner we compute the uncertainty range in the ensemble mean values, for both correlation and IQR, where we bootstrap the total ensemble for 500 realizations of the ensemble mean. We evaluate the simulated IQR being similar to the reference IQR by computing the significance of the opposite event—simulated and reference IQR are significantly different at the 95% level—using the same bootstrapping approach as for correlation.

3 Total upper ocean heat content and sea surface temperature

3.1 Correlation maps of HCT700 and SST in the North Atlantic

Correlation of HCT700 for the uninitialized simulation HIST against the reference data NOCL (Fig. 1a–c) is for the entire time period 1961–2013 high in the Southwestern and Northern part of the Labrador Sea, the Northern part of the Irminger Sea and south of the Iceland-Scotland Ridge, the evolution of these parts of the SPG is relatively well represented by HIST. This is contrasted by lower correlations south of the Iceland Basin. These low correlations are presumably related to the displaced North Atlantic Current, which is consistently shifted to the South in all of our simulations. For 1961–1981, a cooling period for the SPG, the regions with high correlations described above turn negative in HIST, there is almost no positive correlation skill during that time period in the SPG. For 1990–2009, a warming period for the SPG, the correlations are higher than for the entire time period in the Labrador and Irminger Seas. Similar to 1961–2013 there are high correlations between Iceland and Scotland and negative correlations in the region south of the Iceland Basin.

In the EnKF assimilation (Fig. 1d–f) high correlations extend to the entire Labrador and Irminger Seas, the Iceland-Scotland Ridge and parts of the North Atlantic Drift to the West of Ireland. South of the Iceland Basin low correlations persist. The resulting correlation pattern is similar to a horseshoe of high correlations and encompasses the whole North Atlantic. For 1961–1981, this horseshoe pattern is thinner than in 1961–2013 and 1990–2009. In the NDGa assimilation (Fig. 1g–i) the area of high correlation is further enlarged. However, a distinct area of low correlations persists between the Labrador and Irminger Seas, especially in 1961–1981. The NDGf assimilation (Fig. 1j–l) shows similar correlation patterns as NDGa. The difference in the nudging methodology (anomaly vs. full value) does not significantly impact the correlation during assimilation. In all three systems the incorporation of sub-surface temperatures during assimilation improves the correlation of HCT700 when compared to HIST. The

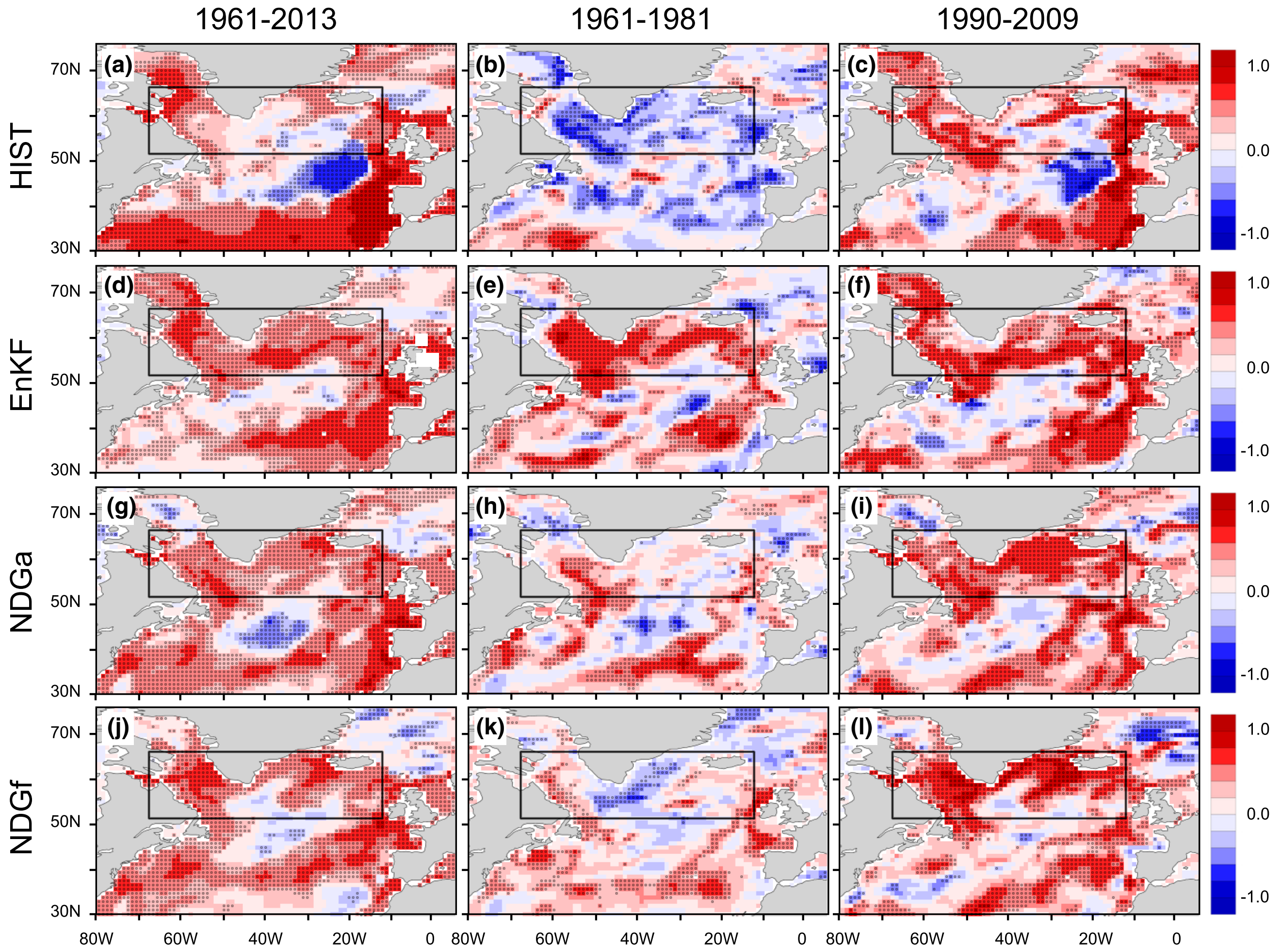


Fig. 2 Correlation of simulated lead year 2 upper ocean heat content (0–700m) with NOCL (Levitus et al. 2012) 1961–2013 (left column), 1961–1981 (middle column), and 1990–2009 (right column) for **a–c** HIST, **d–f** EnKF, **g–i** NDGa, **j–l** NDGf. Stippling indicates significance at the 95% level. The geographic definition of the North Atlantic subpolar gyre is indicated by the black rectangle

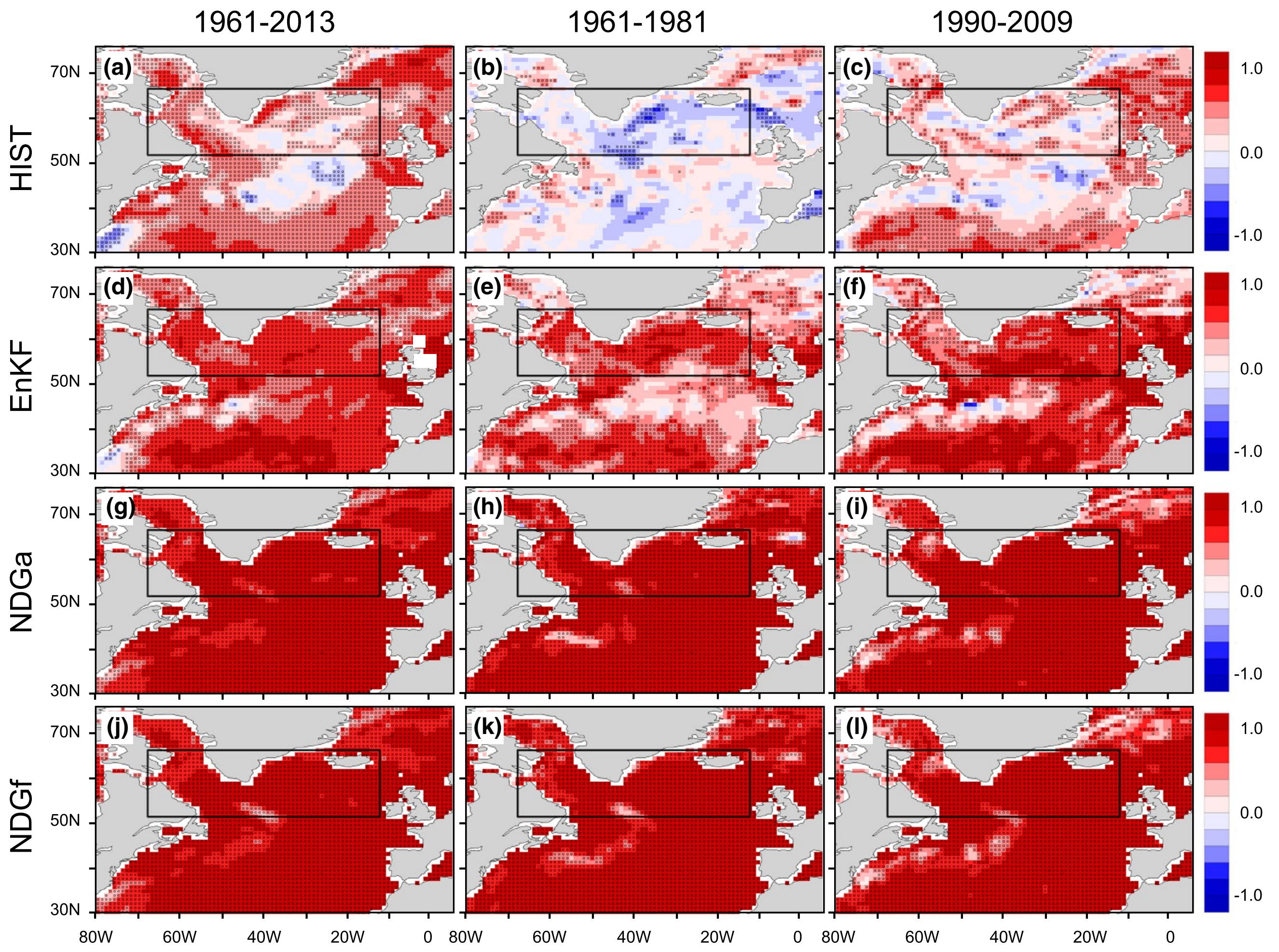


Fig. 3 Correlation of sea surface temperature with HadISST (Rayner et al. 2003) during assimilation for 1961–2013 (left column), 1961–1981 (middle column), and 1990–2009 (right column) for **a**–**c** HIST, **d**–**f** EnKF, **g**–**i** NDGa, **j**–**l** NDGf. Stippling indicates significance at the 95% level. The geographic definition of the North Atlantic subpolar gyre is indicated by the black rectangle

improvement is larger during the cooling of the SPG (time period 1961–1981) and smaller during the warming of the SPG (time period 1990–2009).

Lead year 2 HCT700 correlation skill for all three initialized systems is high at similar locations but covers overall smaller regions than in the assimilation (Fig. 2). The EnKF system shows a horseshoe pattern of high correlation as in the assimilation, albeit thinned out for all three time periods. In the NDGa system areas of high correlation are similar to the EnKF system for the entire time period 1961–2013 and for 1990–2009. But for 1961–1981 correlations are low in the Irminger Sea and the Northern Iceland Basin, the horseshoe pattern is broken. The NDGf assimilation retains a horseshoe of high correlations for 1961–2013 and 1990–2009. But for 1961–1981 correlations are lower in the Irminger Sea than in NDGa, they even turn significantly negative.

For the entire time period 1961–2013 all three prediction systems lead to a similar HCT700 correlation pattern in the SPG which is improved over HIST. However, for 1961–1981 both NDGa and NDGf correlations are significantly lower than EnKF (a correlation difference of 0.5 is deemed significant at the 95% level after our bootstrapping analysis). During this time period, hindcast correlation skill for both nudging systems is constrained to parts of the Labrador Sea in the West and the Rockall Plateau in the East. For lead year 2 hindcasts, the EnKF system is the only of the three initialized systems to retain for all three time periods the horseshoe pattern of high correlation, which is evident in all three assimilations.

Correlations of SST against the reference data HadISST are in HIST and in all assimilations generally higher than correlations of HCT700 (Fig. 3). In HIST and EnKF, low correlations persist south of the Iceland Basin in all three time periods. Both the NDGa and NDGf assimilations

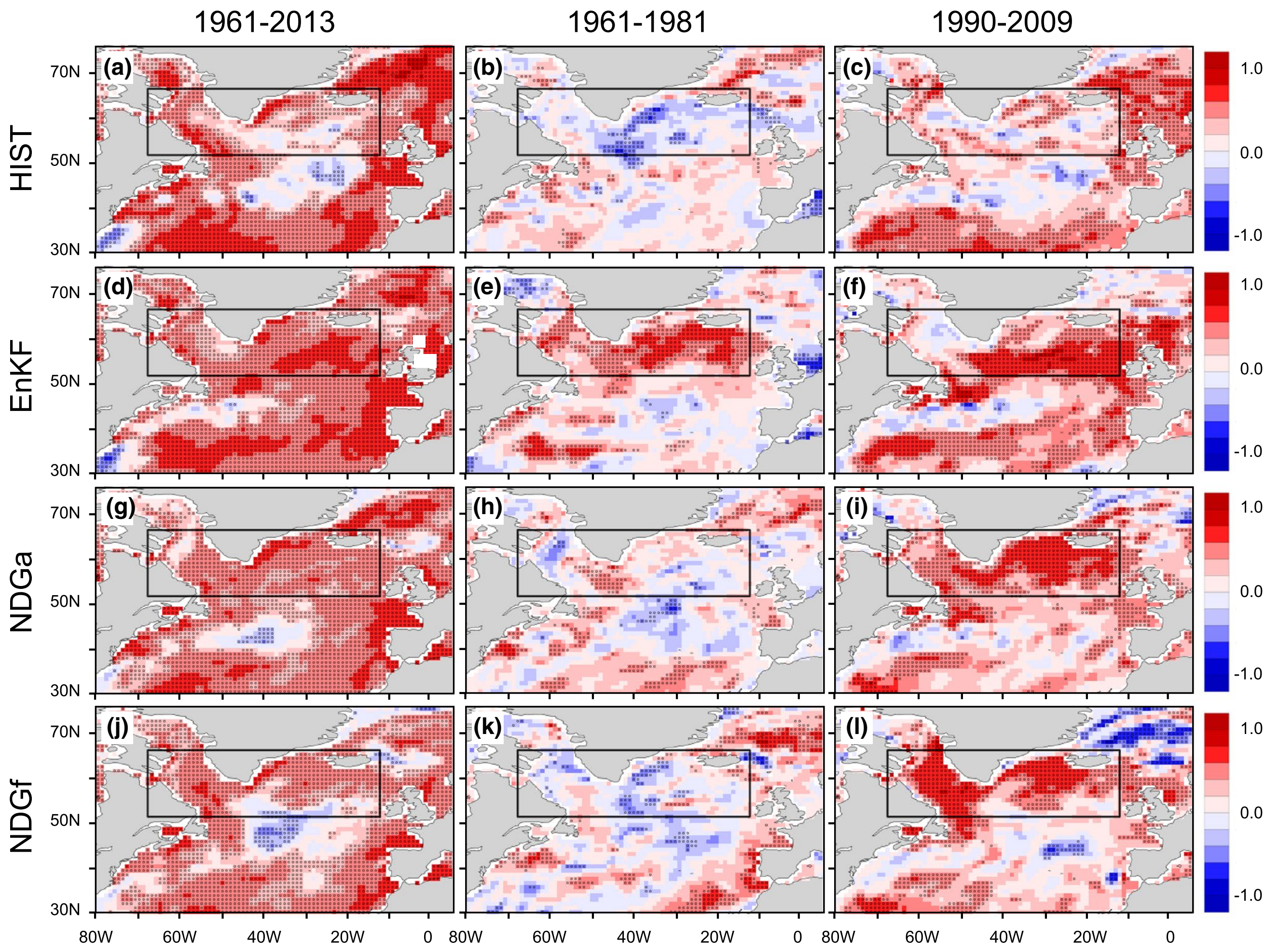


Fig. 4 Correlation of simulated lead year 2 sea surface temperature with HadISST (Rayner et al. 2003) 1961–2013 (left column), 1961–1981 (middle column), and 1990–2009 (right column) for **a**–**c** HIST, **d**–**f** EnKF, **g**–**i** NDGa, **j**–**l** NDGf. Stippling indicates significance at the 95% level. The geographic definition of the North Atlantic subpolar gyre is indicated by the black rectangle

closely meet the reference SST, as can be expected due to the nudging methodology.

For lead year 2 correlations of SST against HadISST are smaller than during assimilation (Fig. 4). For the EnKF system the area of high correlations in 1961–2013 is similar to the assimilation, despite the decrease in correlation strength, and extends over nearly the whole SPG. For 1961–1981 the area of high correlation is decreased when compared to the entire time period. Only the northern part of the horseshoe pattern of high correlation evident in HCT700 persists in SST. For 1990–2009 this horseshoe pattern emerges again in SST. For the NDGa system high correlations persist in the whole SPG over the entire time period, whereas for 1961–1981 correlations are low almost everywhere. For 1990–2009 high correlations re-emerge in the SPG. For the NDGf system the relative decrease in correlation between assimilation and lead year 2 for the entire time period and 1961–1981 is larger compared to the other systems. The NDGf system shows the lowest correlation skill of all three initialized systems over these time periods, both in strength and area. For 1990–2009 correlation for NDGf is higher than for the other systems in the Labrador Sea, and slightly lower in the rest of the SPG.

For all three assimilations and all time periods, correlations in the SPG are higher for SST than for HCT700. In lead year 2 of the initialized hindcasts, correlation skill may be higher in SST (1961–2013), higher in HCT700 (1961–1981, 1990–2009 only for NDGa and NDGf), or differently distributed (1990–2009 for EnKF). Differences in between the three initialized systems tend to be better detectable in SST than in HCT700, for example, the difference between EnKF and NDGa and NDGf in 1990-2009. Correlation skill of SST may be higher than in HCT700 right after the assimilation but decreases faster than in HCT700 over the lead years because SST is more influenced by the atmosphere and HCT700 is thermally more inert. However, the difference

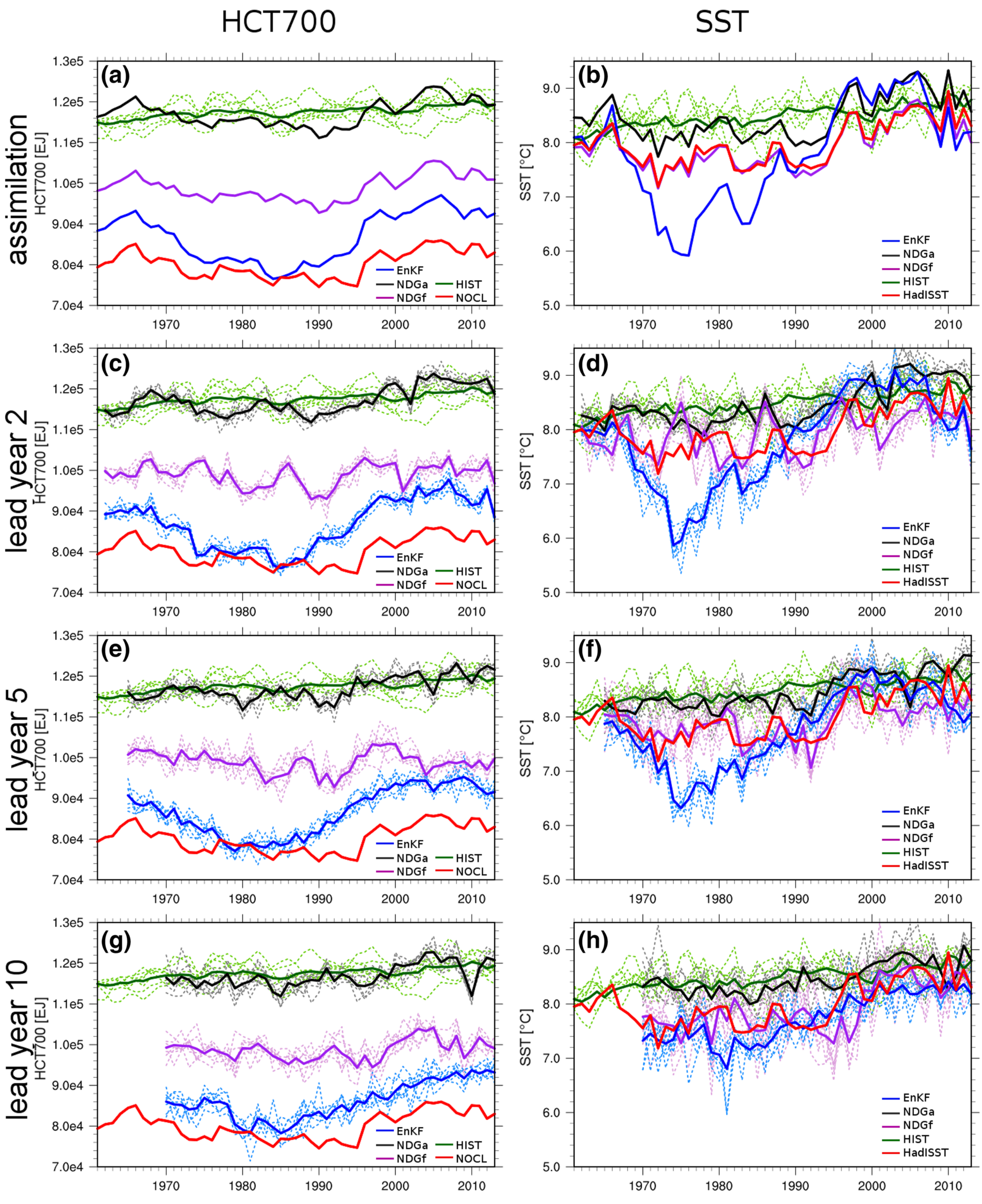


Fig. 5 Time series of SPG integrated total upper ocean heat content (left column, in EJ) and SPG mean SST (right column, in °C) of NOCL (shifted by 8000 EJ) and HadISST, respectively (red), HIST (green), EnKF (blue), NDGa (black), and NDGf (purple) for assimilation (**a**, **b**), and lead years 2 (**c**, **d**), 5 (**e**, **f**), and 10 (**g**, **h**), solid lines: ensemble mean, dashed lines: individual ensemble members

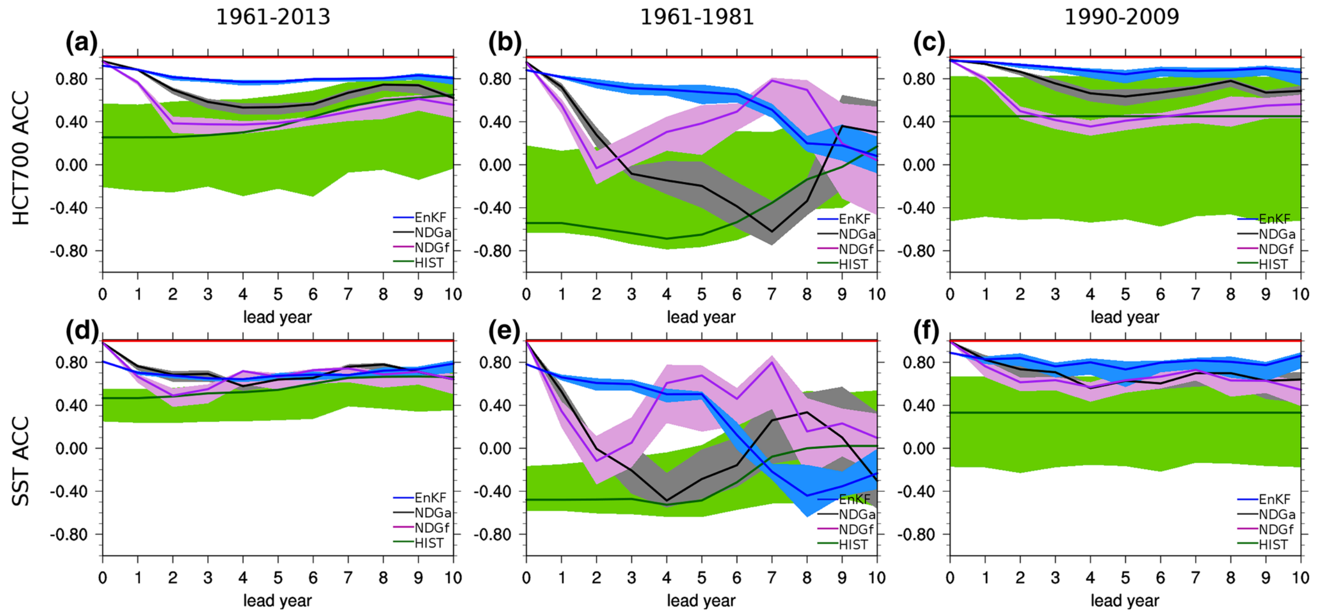


Fig. 6 Correlation of SPG upper ocean heat content (0–700 m) with NOCL (above) and of SPG mean sea surface temperature with HadISST (below) for HIST (green), EnKF (blue), NDGa (black), and NDGf (purple) for lead years 0 (assimilation) to 10: **a**, **d** for the time period 1961–2013, **b**, **e** for the time period 1961–1981, **c**, **f** for the time period 1990–2009, solid lines: ensemble mean, shaded areas indicate spread based on 95% of the bootstrapped ensemble means. Two correlations are distinctly different, when their respective shaded areas do not overlap

between the individual prediction systems and HIST is in all three time periods more pronounced in HCT700. In our view this already indicates the advantage in using both variables, HCT700 and SST, in the analysis.

The nudging methodologies in the ocean lead to high correlation during assimilation but may introduce problems in retaining the high correlation skill in the hindcasts, as we see in the time period 1961–1981. This problem does not seem to exist to the same extent in the EnKF system, where the model is not fit as strongly to observed temperatures. A strong fit to the oceanic re-analysis ORAS4 before the initialization may introduce serious problems, such as

Table 2 HCT700 correlation with NOCL (Levitus et al. 2012) and IQR, in bold: highest correlations and IQR nearest to NOCL, across all systems and with respect to lead times

	EnKF	NDGa	NDGf	HIST	EnKF	NDGa	NDGf	HIST	NOCL
1961–2013	Ensemble mean correlation				IQR of combined PDF [EJ]				
assim.	0.92	**0.96**	**0.96**	0.25	11100	4700	**4500**	3900	5700
ly 1	**0.88**	**0.88**	0.77	0.25	11100	**5400**	4600	3900	5700
ly 2	**0.81**	0.70	0.38	0.26	10700	**5600**	4900	3900	6000
ly 5	**0.77**	0.54	0.39	0.36	10800	**4400**	4300	3700	5700
ly 10	**0.80**	0.62	0.56	0.65	8200	**4600**	4400	3700	5900
1961–1981	Ensemble mean correlation				IQR of combined PDF [EJ]				
assim.	0.88	**0.95**	**0.95**	– 0.54	8300	**2600**	2700	3400	2500
ly 1	**0.82**	0.72	0.55	– 0.54	8900	3400	**2300**	3400	2500
ly 2	**0.76**	0.28	– 0.03	– 0.59	8300	3500	4200	**3400**	3300
ly 5	**0.68**	– 0.20	0.38	– 0.65	7000	**2800**	3400	3300	2800
ly 10	0.08	**0.30**	0.04	0.17	5400	3600	**3500**	3900	2400
1990–2009	Ensemble mean correlation				IQR of combined PDF [EJ]				
assim.	0.97	**0.98**	0.97	0.45	10900	**7100**	7000	3500	7100
ly 1	**0.96**	0.94	0.80	0.45	9900	**6700**	5600	3500	7100
ly 2	**0.93**	0.87	0.50	0.45	7900	**6700**	5000	3500	7100
ly 5	**0.84**	0.64	0.41	0.45	**6400**	5000	4900	3500	7100
ly 10	**0.86**	0.69	0.56	0.45	**6300**	4800	5100	3500	7100

Table 3 SST correlation with HadISST (Rayner et al. 2003) and IQR, in bold: highest correlations, and IQR nearest to HadISST, across all systems and with respect to lead times

	EnKF	NDGa	NDGf	HIST	EnKF	NDGa	NDGf	HIST	NOCL
1961–2013	Ensemble mean correlation				IQR of combined PDF [°C]				
assim.	0.81	**0.98**	**0.98**	0.47	1.53	0.56	**0.58**	0.41	0.62
ly 1	0.70	**0.76**	0.67	0.47	1.43	0.56	**0.60**	0.41	0.62
ly 2	0.68	**0.69**	0.49	0.48	1.33	0.61	**0.67**	0.41	0.67
ly 5	**0.67**	0.64	0.66	0.54	1.21	0.59	**0.60**	0.39	0.67
ly 10	**0.79**	0.64	0.64	0.66	**0.79**	0.53	0.87	0.39	0.73
1961–1981	Ensemble mean correlation				IQR of combined PDF [°C]				
assim.	0.78	**0.99**	**0.99**	− 0.48	1.34	**0.26**	**0.26**	0.39	0.24
ly 1	**0.66**	0.53	0.34	− 0.48	1.58	**0.35**	**0.35**	0.39	0.24
ly 2	**0.61**	− 0.01	− 0.12	− 0.48	1.25	0.31	0.57	**0.38**	0.37
ly 5	0.50	− 0.29	**0.68**	− 0.49	0.86	0.29	0.54	**0.39**	0.35
ly 10	− 0.23	− 0.30	**0.10**	0.02	0.49	0.47	0.70	**0.38**	0.38
1990–2009	Ensemble mean correlation				IQR of combined PDF [°C]				
assim.	0.89	**0.99**	**0.99**	0.33	1.15	**0.98**	**0.98**	0.29	0.91
ly 1	**0.83**	0.82	0.76	0.33	**0.93**	0.66	0.61	0.29	0.91
ly 2	**0.84**	0.74	0.61	0.33	**0.74**	0.69	0.60	0.29	0.91
ly 5	**0.73**	0.63	0.63	0.33	0.52	0.48	**0.57**	0.29	0.91
ly 10	**0.86**	0.64	0.54	0.33	0.50	0.49	**1.00**	0.29	0.91

a strong initialization shock, in the subsequent hindcasts. In fact, the good fit to observed oceanic temperature and salinity imposed by the nudging methodology in NDGa and NDGf assimilations does not guarantee that dynamic features such as the North Atlantic Current as part of the ocean circulation are well met, even if the imperfections of the re-analysis were small. In a recent study, Kröger et al. (in revision for Climate Dynamics) investigate this problem further. They establish a link between anomalous heat fluxes introduced by nudging to re-analysis data inconsistent with the model climatology and the decrease in hindcast skill of ocean temperature in the SPG. Monitoring and mitigating those anomalous heat fluxes would certainly improve initialization.

3.2 Time series of integrated HCT700 and mean SST over the SPG

The simulated time series of HCT700 integrated and SST averaged over the SPG (Fig. 5) represent the basis of our further analysis of the multidecadal variability in the SPG region. For HCT700 (left column in Fig. 5) HIST shows a warming over the entire time period 1961–2013. The absence of multidecadal variability in HIST can be expected due to the lack of initialization. The reference data NOCL and the three assimilations are dominated by multidecadal variations with warm phases in the 1960s and 2000s and a cold phase in the 1980s/early 1990s. The amplitude between warm and cold phases is about 1000 EJ for NOCL, NDGa and NDGf, but twice as large for the EnKF assimilation. There is a phase shift regarding the coldest years between NOCL (1989–1995), NDGa (1991) and NDGf (1991) on one side and the EnKF assimilation (1984–1986) on the other side. Such a shift cannot be seen in the warm phases.

Lead year 2 and five of the hindcasts show a similar multidecadal variability as the assimilation time series. In the time series for hindcast lead years 10 the warm phase in the 1960s cannot be resolved any more, but the cold phase in the 1980s/early 1990s and the warm phase in the 2000s can still be predicted, albeit with a decreased amplitude in the EnKF system when compared to the assimilation.

In principle, all three prediction systems are able to simulate the multidecadal variability found in NOCL reference HCT700 integrated over the whole SPG, with both the NDGa and NDGf systems matching the amplitude of the NOCL reference time series better than the EnKF system. The EnKF system is able to reproduce cooling and warming in the SPG over several lead years, but with the disadvantage of overestimating the amplitudes. A similar overestimation is apparent in a MPI-ESM assimilation experiment with only the atmospheric nudging applied (not shown). The EnKF assimilation does not damp this overestimation. However, the influence of the atmospheric nudging is stronger for SST than for HCT700. We further discuss this in the context of coupled data assimilation in Sect. 5.

The SPG mean SST time series (right column in Fig. 5) show relatively similar multidecadal variations as in HCT700 and a similar overestimation by the EnKF system. Amplitudes between the warm and cold phases are 1.8 K for the reference time series from HadISST, 1.2 K for NDGa and NDGf assimilations and 3 K for the EnKF assimilation, decreasing throughout the hindcasts. The warm phases in

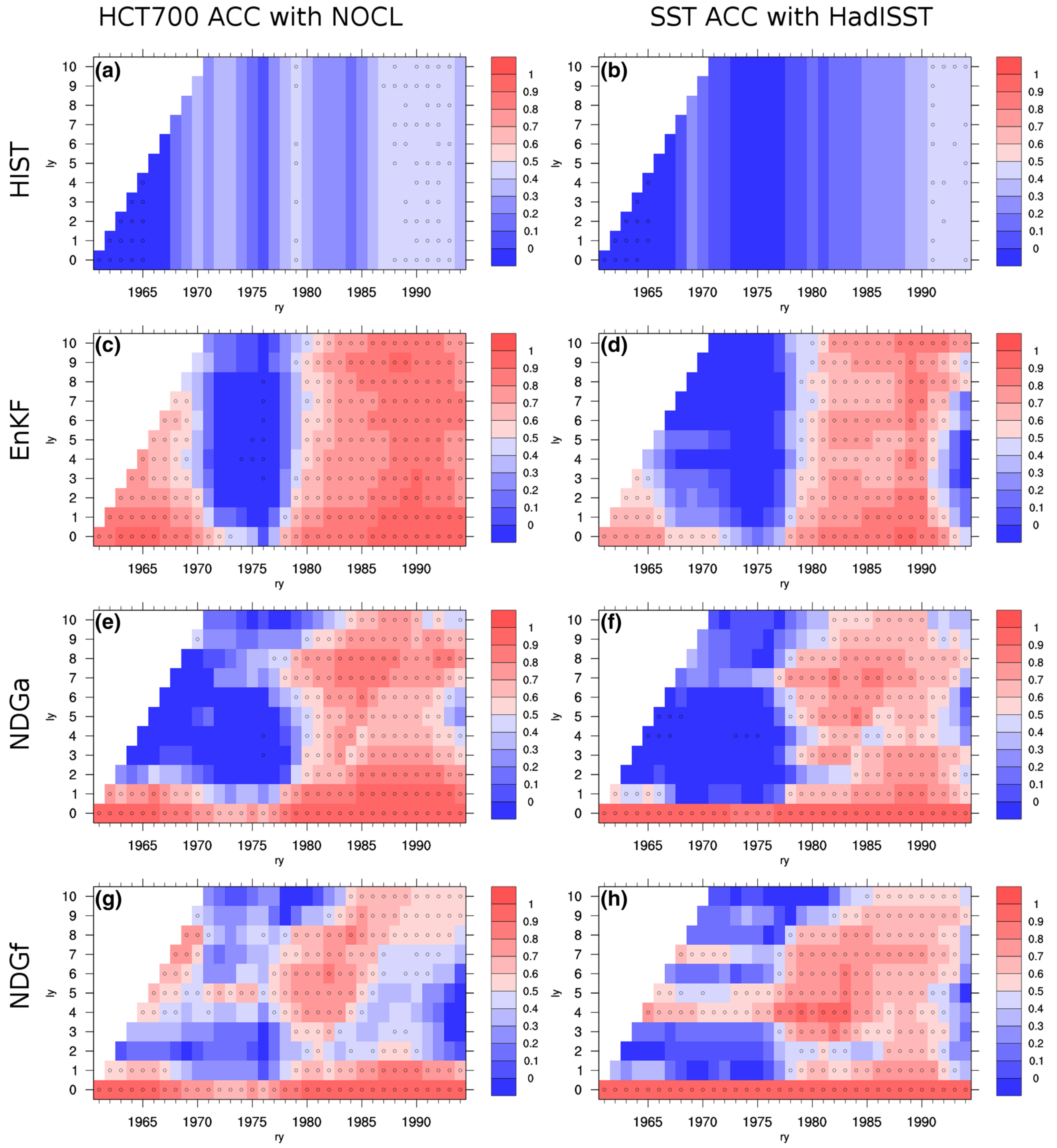


Fig. 7 Correlation of SPG upper ocean heat content (0–700 m) with NOCL (left column) and of SPG mean sea surface temperature with HadISST (right column) of **a**, **b**) HIST, **c**, **d**) EnKF, **e**, **f** NDGa, **g**, **h** NDGf for lead years 0 (assimilation) to 10, calculated for a 20 year moving window between 1961 and 2013, the start year of this window is indicated at the x-axis. The curves for time period 1961–1981 in Fig. 6b,e are for each simulation represented in in color code as the first diagonal column starting at reference year 1961. The curves for the time period 1990–2009 in Fig. 6c,f are represented in color code as the vertical column indicated in 1990. Stippling indicates significance at the 95% level. Please note the color scale from 0 to 1

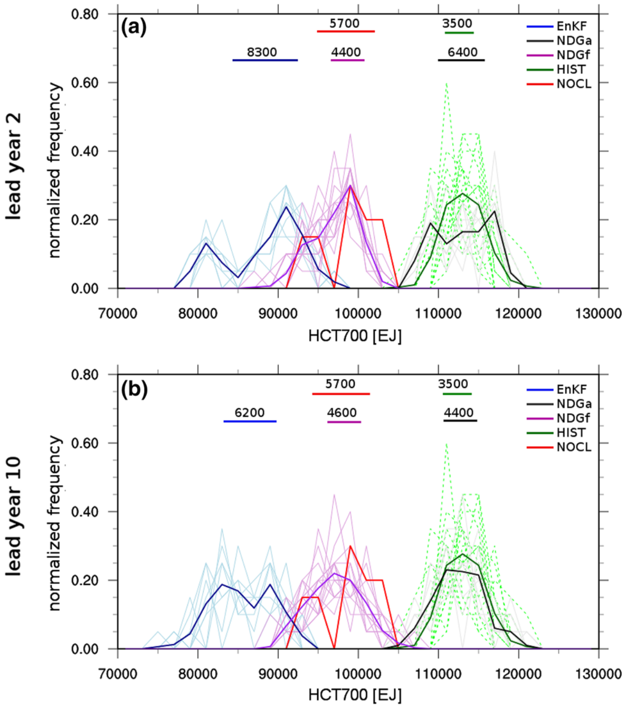


Fig. 8 Temporal PDF of SPG upper ocean heat content for the period 1990–2009 for NOCL (red, shifted by 10000 EJ), HIST (green), EnKF (blue), NDGa (black), and NDGf (purple) for lead years 2 (**a**) and 10 (**b**), solid lines: PDF of the combined ensemble, dashed lines: PDF of individual ensemble members. IQRs of the combined PDF are indicated as numbers and horizontal bars

the 1960s and 2000s are simulated by all prediction systems, moreover the warm phases agree between SST and HCT700. For HadISST, NDGa and NDGf that is also the case for the cold phase in the 1980s/early 1990s. In the EnKF SST time series the cold phase is in the 1970s, much earlier than in the SST time series of the other systems and also earlier than in the EnKF HCT700 time series. It is shifted to the early 1980s in lead year 10. The NDGf system shows in lead year 2 three anomalous warming events in 1968, 1975 and 1986, which are hard to detect during assimilation or in lead years 5 or 10. These short-term warmings are not simulated by any of the other systems, except for the 1986 event in NDGa lead year 2, and they are stronger in SST than in HCT700. At least two of these events (1975, 1986) seem to be related to anomalously strong AMOC events at 45°N simulated by NDGf, see Sect. 4 and the discussion therein. It is worth noting that the variability of the ensemble mean is damped in the EnKF and NDGa systems over the lead years, but not in the NDGf system (see subsection on IQR below).

3.3 Dependence of correlation on lead year and evaluation period

The lead year dependent correlation between the simulated HCT700 integrated over the SPG and NOCL for the entire hindcast period 1961-2013 (Fig. 6a; Table 2) are positive for HIST (0.25-0.65) due to the long-term warming by the

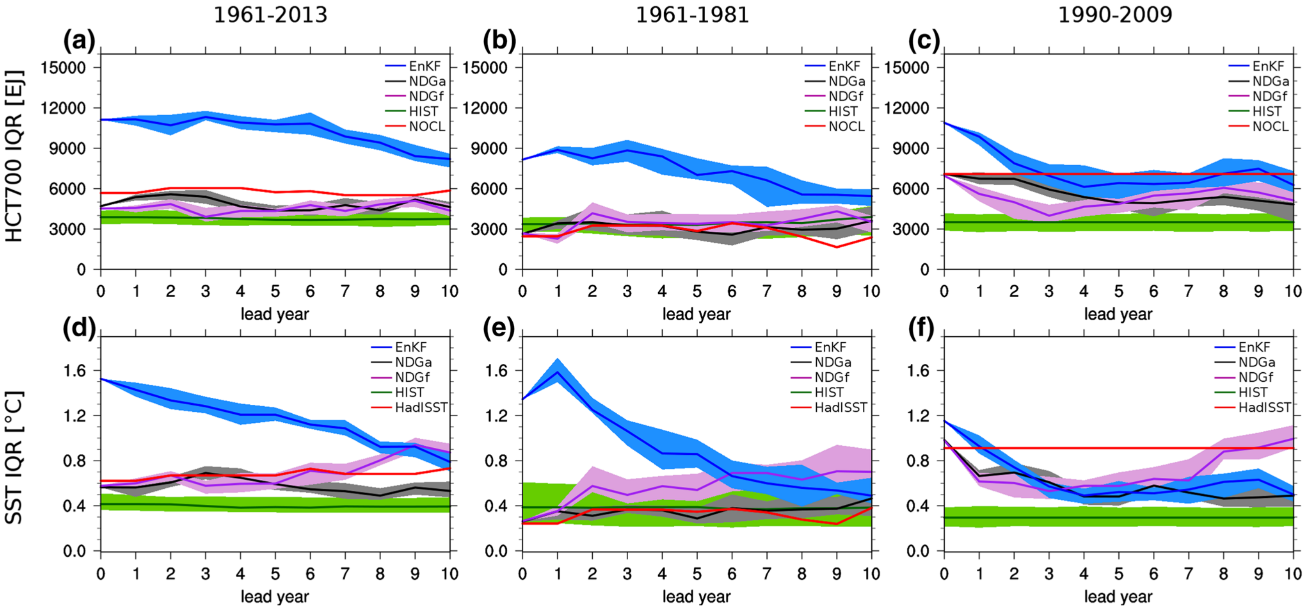


Fig. 9 IQR of SPG upper ocean heat content (0–700 m) (above) and SPG mean sea surface temperature with HadISST (below) for the reference data (red), HIST (green), EnKF (blue), NDGa (black), and NDGf (purple) for lead years 0 (assimilation) to 10: **a**, **d** for the time period 1961–2013, **b**, **e** for the time period 1961–1981, **c**, **f** for the time period 1990-2009, solid lines: combined ensemble PDF, shaded areas indicate spread based on 95% of the bootstrapped combined ensemble PDF. Two IQRs are distinctly different, when their respective shaded areas do not overlap

external forcing. The differences in high frequency content in any individual member of HIST leads to the large uncertainty range. For the EnKF system correlation skill is close to 0.9 in assimilation mode, slightly dropping off until lead year 4 to 0.8 and remaining there until lead year 10. For the first 6 lead years this is distinctly higher than correlation in HIST, i.e. uncertainty ranges do not overlap. Up to lead year 6 the hindcast skill benefits from the EnKF initialization. The NDGa system shows a decrease in correlation from 0.95 in assimilation mode down to 0.5 in lead year 5. It is distinctly higher than HIST up to lead year 2. The NDGf system shows the same correlation as NDGa in assimilation mode, but decreases to 0.4 in lead years 2 to 5, the drop in lead year 2 is related to the warming events simulated by NDGf in lead year 2 for 1968, 1975 and 1986. NDGf system correlations are distinctly higher than HIST only in the first lead year. Both nudged systems show a re-emergence of correlation towards lead year 9 but so does HIST. The time series for lead year 9 and 10 do not include the cooling in the 1960s since they start at 1969 and 1970, respectively. This may lead to a higher skill in these lead years when compared to the skill of the smaller lead years.

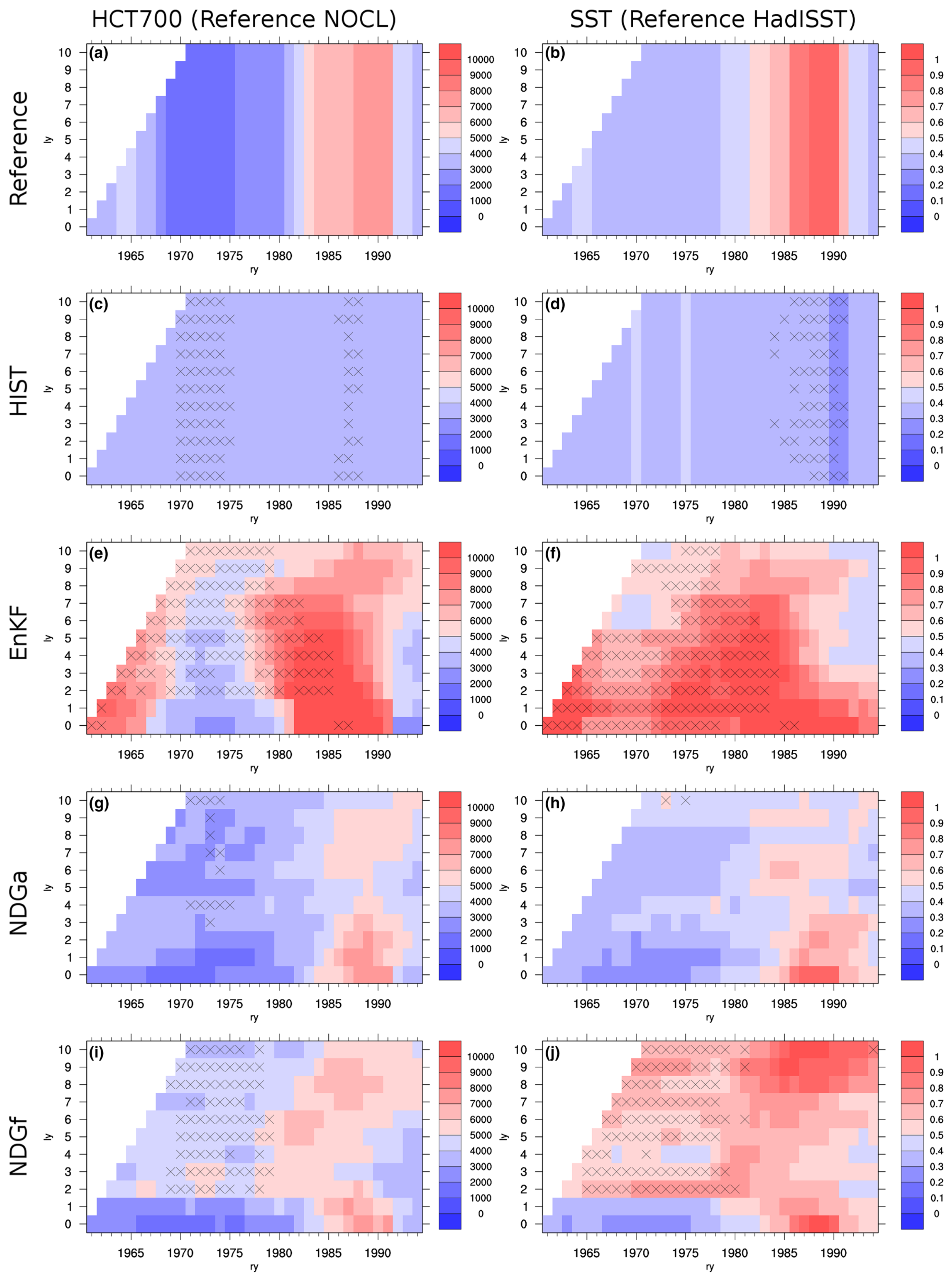
Considering only the cooling period 1961–1981 (Fig. 6b), correlations for HIST are negative. HIST cannot reproduce the cooling of the SPG. Correlation skill for EnKF is between 0.6 and 0.8 for lead years up to 5, as such distinctly better than HIST, but dropping off in later lead years. As for the entire time period there is a sharp decrease in correlation skill for NDGa and NDGf after assimilation. However, the nudging systems differ distinctly in lead years 7 and 8: NDGa with a low correlation (−0.7), NDGf with a re-emerging high correlation (0.8).

For the warming period 1990–2009 (Fig. 6c), the correlation of HIST with NOCL is 0.45 with a large uncertainty ranging between −0.5 and 0.8. The EnKF system correlations are above 0.8 for all lead years, they are distinctly better than HIST in lead years up to 3. Correlations of both the NDGa and NDGf systems are not distinctly different from HIST after lead year 2 (NDGa) and 1 (NDGf), and distinctly lower when compared to the EnKF system for lead years 1–9.

It is more difficult for the initialized systems to improve correlations over HIST in the warming period 1990–2009 than in the cooling period 1961–1981, although the absolute correlation skill, at least for the EnKF system, may be better for all lead years in 1990–2009 than in 1961–1981. The contribution of external forcing alone to the high HCT700 correlation skill in the warming period 1990–2009 is already so large that additional improvement due to initialization of the hindcast is hard to attain.

There are several differences when comparing correlation skill of SST (Fig. 6d–f; Table 3) with HCT700. In all three time periods the uncertainty range in HIST is smaller in SST than in HCT700. Also, correlation skill for the EnKF system is generally slightly lower in SST than in HCT700. For the entire time period, correlation skill for HIST is higher for SST than for HCT700 for the first lead years, possibly because the long-term warming trend has a larger impact on SST than on HCT700. Correlation skill for NDGa and NDGf is higher in SST from lead year 2 on, for the NDGa system skill re-emerges 1–2 years earlier in SST than in HCT700. As a consequence SST correlation skill after lead year 3 is not distinctly different between all three initialized systems. Moreover, none of the initialized systems shows distinctly higher skill than HIST. For 1961–1981 correlation skill for NDGa and NDGf also re-emerges earlier in SST than in HCT700. Therefore in SST the EnKF system shows distinctly higher correlation than the other systems only in lead years 2–4, and the NDGf system has higher skill, but not distinctly higher, than the other systems in lead years 5–7. For 1990–2009 the EnKF system has the highest SST correlation skill for lead years 2–10, albeit it is decreased when compared to HCT700. EnKF skill is distinctly better than NDGa and NDGf only in lead years 4 and 10. For the three time periods under consideration, correlation skills differ more distinctly between the three prediction systems in HCT700 than in SST. Although the uncertainty range of HIST is always larger in HCT700 than in SST, improvements in correlation skill over HIST are easier detectable in HCT700, possibly related to the larger thermal inertia in the sub-surface ocean.

When we analyze all 20-year time periods within 1961–2013 (Fig. 7), correlation skill for HIST never exceeds 0.5, it is similar in SST and HCT700 for time periods starting in the 1960s and 1990s and higher in HCT700 than in SST for time periods starting in the 1970s and 1980s. For the EnKF system correlation skill is for the assimilation and all hindcast lead years high during the cooling of the SPG in the 1960s, high again during the warming of the SPG in the 1990s, but low in between. Skill is higher for EnKF in HCT700 than in SST in most time periods. The lead year dependence of the EnKF skill is low for both variables. For the NDGa system correlation skill is generally high during assimilation. In the cooling in the 1960s correlation drops off after lead year 1, in the warming during the 1990s correlation decreases until lead years 2-4, but increases again in later lead years. Hindcast correlation skill is higher for HCT700 than for SST. For the NDGf system correlation skill decreases from generally high values during assimilation to lowest values in lead year 2. Skill re-emerges in lead year 4 in almost all time periods and stays high for HCT700 in the time periods starting in the 1960s and 1980s, for SST only in the time periods starting after 1980. In contrast to the EnKF and NDGa systems, correlation skill is higher in HCT700 only in the early time periods, whereas it is higher in SST in the later time periods. For the three initialized systems, the


HCT700 (Reference NOCL)
SST (Reference HadISST)
Reference
HIST
EnKF
NDGa
NDGf
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
ly
ry
0
1
2
3
4
5
6
7
8
9
10
1965
1970
1975
1980
1985
1990
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

◂**Fig. 10** IQR of SPG upper ocean heat content (0–700 m) (left column, in EJ) and of SPG mean sea surface temperature (right column, in °C) of **a** NOCL HCT700 and **b** HadISST SST, of **c**, **d** HIST, **e**, **f** EnKF, **g**, **h** NDGa, **i**, **k** NDGf for lead years 0 (assimilation) to 10, calculated for a 20 year moving window between 1961 and 2013, the start year of this window is indicated at the x-axis. The curves for time period 1961-1981 in Fig. 9b,e are for each simulation represented in in color code as the first diagonal column starting at reference year 1961. The curves for the time period 1990–2009 in Fig. 9c,f are represented in color code as the vertical column indicated in 1990. Stippling indicates significance at the 95% level. Crosses indicate significant difference between simulated and reference IQR at the 95% level, thus these IQRs cannot be similar here

20 year periods incorporating either the cooling in the 1960s or the warming in the 1990s show higher correlation skill than the 20 year periods starting in the 1970s in between the cooling and warming.

Our results show that in both HCT700 and SST the correlation of all three prediction systems can be high in time periods covering pronounced multiyear variability, in particular trends on time scales of the order of 10 years, such as the warming in the 1990s. The increase in correlation skill is only partly captured by the uninitialized experiment, which due to simulating a monotonous increase in temperature over the entire hindcast period has got higher skill during warming periods. Our initialized hindcasts outperform the uninitialized simulations over the time period 1990–2009, up to which lead year depends on the prediction system: up to lead year 1 for NDGf, up to lead years 4–6 for EnKF. As such we can corroborate the findings of Robson et al. (2012) using DePreSys and Yeager et al. (2012) using CCSM4 that the warming of the SPG in the 1990s could have been predicted with today's prediction systems. In time periods with a multiyear decrease of oceanic temperature in the SPG, such as the cooling in the 1960s, only the EnKF system provides high hindcast correlation skill. For the EnKF initialized hindcasts we agree with the results from Robson et al. (2014), they could have predicted the cooling of SPG in the 1960s with DePreSys. In the 1970s and 1980s, a period with small multiyear trends in between the cooling and warming of the SPG, all our prediction systems show low correlation skill.

3.4 Dependence of IQR on lead year and evaluation period

We complement our assessment of the correlation skill with the interquartile range (IQR, Eq. 4) of the temporal probability distribution of the simulated SPG time series and its dependence on lead year and evaluation period. Ideally, the longer the IQR of the initialized hindcasts stay close to the IQR of reference data the higher the hindcast IQR skill. However, the faster the initialized hindcasts drift to the climatological state of the model the earlier the IQR of the initialized hindcasts is close to the IQR of the uninitialized simulation. In theory, the evolution of the IQR of an initialized prediction system from reference data IQR to uninitialized simulation IQR can be used to measure the transition time scales onto which such systems may improve the predictability over the uninitialized simulation. In the following we test the feasibility of these theoretical considerations with our prediction systems.

The IQRs of all three prediction systems change from lead year 2 to lead year 10 as do the shapes of the PDFs (Fig. 8). Here, the most visible changes are for the EnKF and NDGa systems. In lead year 2, their PDFs show a bimodal structure similar to the reference data, a result of distinctly cold waters at the western boundary and distinctly warm waters in the eastern part of the SPG. As a consequence both IQRs are relatively large and more similar to the reference NOCL than to the uninitialized simulation HIST. In lead year 10, their PDFs resemble a narrower shape than in lead year 2, closer to a one peak structure. The IQRs are smaller and thus more similar to HIST than to NOCL.

The lead year dependent IQRs of reference data and HIST are moderately separated for the entire time period 1961–2013, they are similar for 1961–1981 and well separated in 1990–2009 (Fig. 9). A transition of hindcast IQRs from reference data IQR towards the IQR of HIST could be detected for 1961–2013 and 1990–2009, we therefore focus on these time periods. In general the EnKF system shows for both HCT700 and SST an overestimation of IQRs compared to reference data and HIST. The reason for the large IQRs is the overestimation of amplitudes in both the HCT700 and SST time series by the EnKF system. The IQRs are largest in the early lead years and decrease in the later lead years. Only in the warming period 1990–2009 the EnKF systems IQRs are between reference data and HIST. And only for SST during 1990–2009 a transition of IQRs from close to reference to close to HIST can be identified. The IQRs of the NDGa system are generally closer to reference data than HIST after assimilation and in the earlier lead years and closer to HIST than reference data in later lead years (except for HCT700 1961–2013). For the NDGa system we find the evolution from an IQR close to reference data towards an IQR close to HIST for both HCT700 and SST in 1990–2009 and partly in SST in 1961–2013. The IQRs of NDGf are as close to reference data as NDGa after assimilation. For the time periods 1961–2013 and 1990–2009 they decrease until lead year 3 and increase again in the later lead years. There is no evidence of a consistent evolution towards HIST over 10 lead years in the NDGf system.

For all 20-year periods within 1961–2013, the IQR of the reference data for both HCT700 and SST show a minor maximum in the 1960s and a major maximum in the 1990s with a minimum in the 1970s/1980s inbetween (Fig. 10). None of these are evident in the uninitialized simulations IQRs,

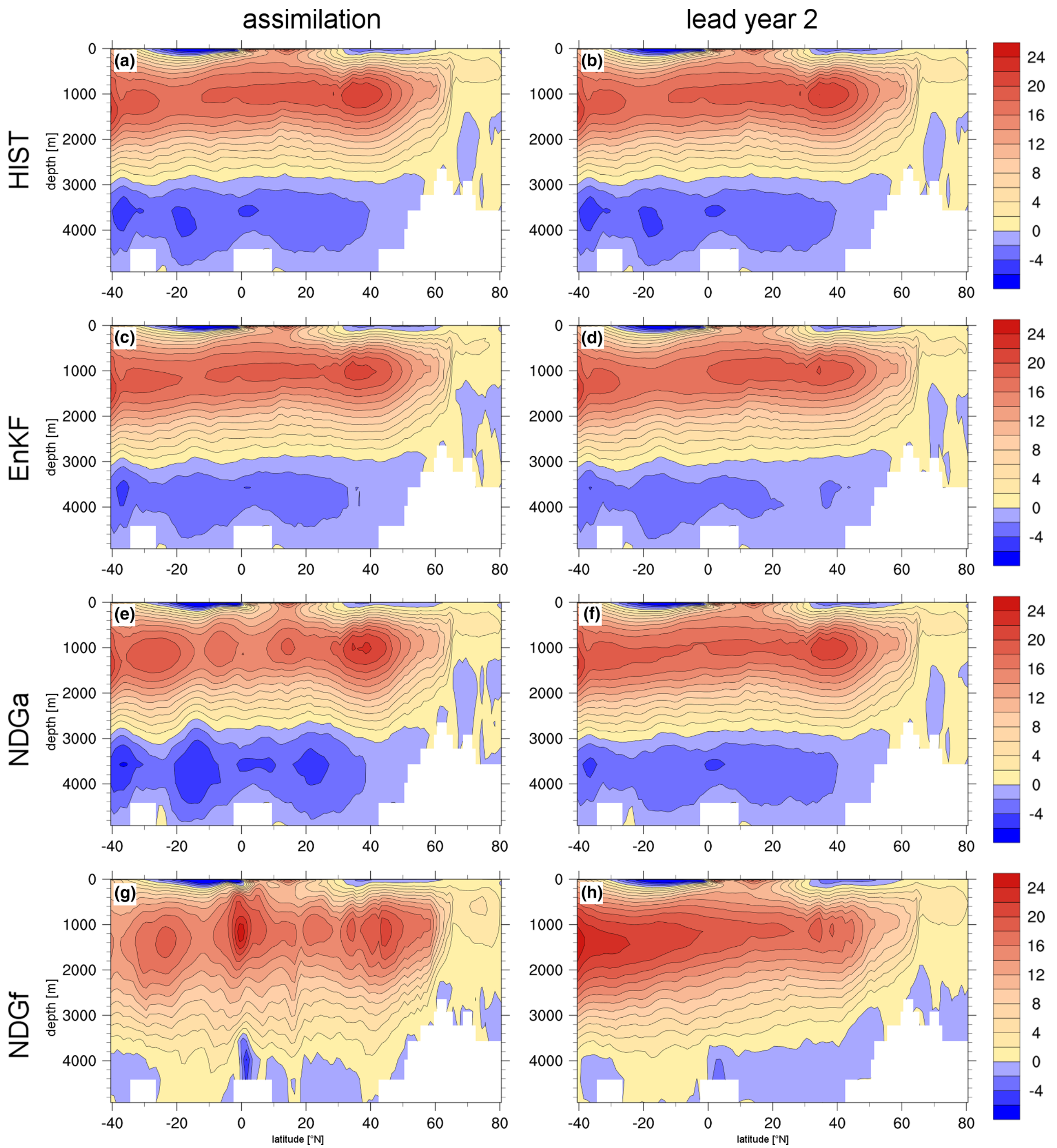


Fig. 11 Section of the 1961–2013 time mean AMOC in Sv as a function of latitude and depth, during assimilation (left column) and in lead year 2 (right column) for **a**, **b** HIST, **c**, **d** EnKF, **e**, **f** NDGa, **h**, **i** NDGf. Positive (negative) values indicate a clockwise (counter-clockwise) direction

which remain almost unchanged for all time periods. In contrast, the EnKF system's IQR for HCT700 show these two maximums and the minimum, but overestimates the IQR for almost all time periods in HCT700 and even stronger in SST. For HCT700 EnKF IQRs are close to reference data during assimilation only in the periods starting in 1972–1975. For SST the IQR of the EnKF system is close to reference data in hindcasts only in periods starting in the late 1980s, including 1990-2009, as mentioned above. Here, the IQR decreases over lead time towards the one of HIST. It is only in these

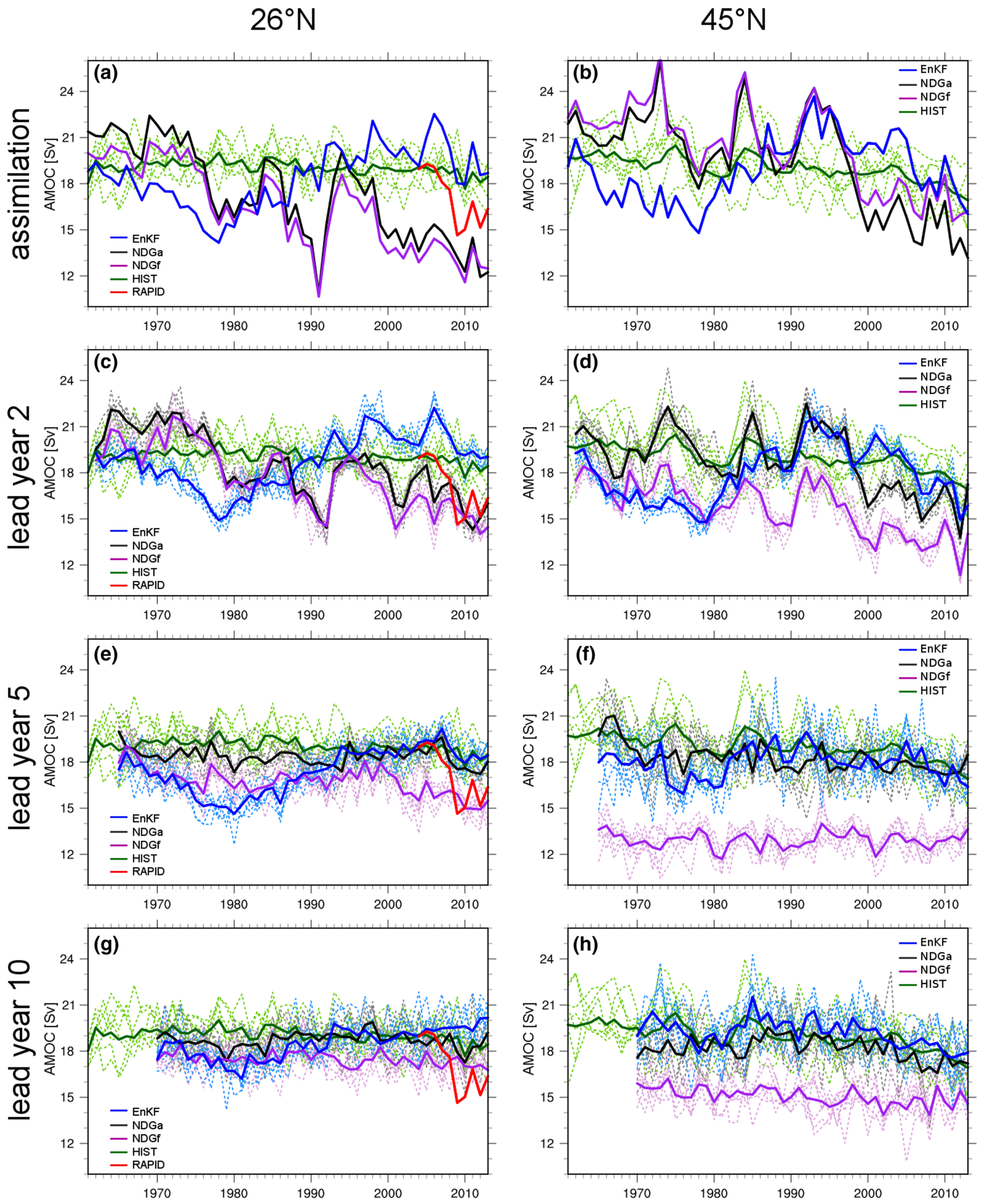


Fig. 12 Time series of AMOC in Sv at 26°N (left column) and at 45° N (right column) of HIST (green), EnKF (blue), NDGa (black), and NDGf (purple) for assimilation (**a**, **b**), and lead years 2 (**c**, **d**), 5 (**e**, **f**), and 10 (**g**, **h**), solid lines: ensemble mean, dashed lines: individual ensemble members. For 26°N observations for 2004–2015 from RAPID-MOCHA are shown in red

reference periods, where a IQR transition from reference data NOCL to HIST can be detected in the EnKF system. The NDGa system IQRs represent the two maximums and the minimum of the reference data relatively well in the assimilation and in lead years up to 2. For the reference periods starting in 1970–1979 the IQRs are increasing towards the IQR of HIST over lead time. For the reference periods starting in 1986–1990 the IQRs are decreasing towards HIST over lead time. A transition from reference data IQR to the one of HIST occurs more often in NDGa than in EnKF. In the NDGf system the IQRs of reference data are well met during assimilation and lead year one. IQRs distinctly increase in lead year 2 over most time periods, especially in SST. In contrast to NDGa the IQRs of NDGf do not evolve towards HIST in the later lead years. For SST in particular, they increase again in the later lead years towards the IQR of the HadISST. An IQR transition to HIST cannot be established in the NDGf system.

Our IQR analysis thus complements the correlation analysis in the following ways. For the EnKF hindcasts time periods of high correlations are accompanied by large, mostly overestimated, IQRs, a result of the multiyear cooling/warming in these periods. The PDF of reference data and of the EnKF system may even display a double peak during these times, a result of the presence of warmer than normal or colder than normal water in the SPG. In comparison to reference data the EnKF hindcasts are thus able to successfully reproduce the phase of these multiyear to multidecadal variations (high correlation), albeit discounted by an increased amplitude (IQR too large). This represents a strong contrast to the nudged hindcasts. Their IQRs are on average closer to the reference IQR than the EnKF hindcasts, but they are not able to reproduce the full variability in the phase of the SPG time series as measured by the correlation. Instead, anomalous events may be simulated as in NDGf. This leads to lower correlations for the nudging systems than for EnKF. An extreme case is represented by the uninitialized simulation, which shows IQRs almost constant over time. It is not able to reproduce any of the multiyear changes evident in the reference data, and high correlations are only occurring when the reference temperature is increasing.

4 Atlantic meridional overturning circulation

4.1 Time mean AMOC sections

Against the background of North Atlantic atmosphere-ocean interconnections we analyze the Atlantic meridional overturning circulation as an integrated dynamic quantity influenced by, among other parameters, the SPG temperatures.

The difference between the time mean AMOC section of HIST and the EnKF assimilation is small (Fig. 11) and they both agree with the reconstructed 20th century time mean AMOC within MPI-ESM in Müller et al. (2015), their Fig. 7b. This is also true in the EnKF hindcasts for lead year 2 (and lead years 3–5, not shown). There are two main circulation cells, a positive cell above and a negative cell below 3000 m depth. In contrast, the AMOC in the NDGa system is changed when compared to HIST. Both upper and lower cell are divided into several sub-cells during assimilation. However, this division is hardly detectable in lead year 2 (and lead years 3-5, not shown) of the NDGa hindcasts, where the AMOC is again close to the AMOC of HIST. In the NDGf system the AMOC is heavily disturbed during assimilation with sharp changes from latitude to latitude and almost no lower cell anymore. The disturbances are less pronounced in lead year 2, however, the AMOC is still considerably away from HIST. This does not change for lead years 3-5 (not shown). It is worth noting that the time mean AMOC for both NDGa and NDGf assimilations does not only differ between them, but also differs considerably from the one found in the ORAS4 re-analysis, compare Fig. 11 with Karspeck et al. (2015a), their Fig. 1. The nudging methodology transfers the temperatures and salinities from ORAS4 to MPI-ESM, but not the time mean state of the AMOC.

The change in AMOC from assimilation to hindcast in both NDGa and NDGf and the "wrong" AMOC state through all lead years in NDGf may be responsible for the drop in correlation skill for HCT700 and SST in the SPG. While the SST and HCT700 at very short lead times may be primarily driven by a persistence of the assimilated anomalies, a wrong AMOC structure may result in a wrong heat transport that may hamper the prediction of SST and HCT700 beyond the interannual time scale. In NDGa and NDGf this my be caused by an inconsistent energy budget in the North Atlantic introduced by the nudging methodology, see Kröger et al. (in revision for Climate Dynamics). Such a violation of the energy budget could be balanced on time scales up to several years. This balancing in turn may lead to a different thermodynamic state of the ocean with both a different AMOC and a different temperature field.

4.2 Time series AMOC at 26°N and 45°N

The AMOC at 26°N as simulated by HIST does not show any multidecadal variability (Fig. 12). In contrast, the EnKF system in lead years up to 5 shows strong AMOC phases in the 1960s and 2000s similar to the warm HCT700 and SST phases. A weak AMOC is simulated around 1980, just between the corresponding cold phases in SST (around 1975) and HCT700 (around 1985). In the NDGa and NDGf systems the multidecadal variability of the AMOC at 26°N is similar to each other but different from the EnKF system or any of the variability seen in

HCT700 and SST. Both the NDGa and NDGf assimilation simulate a sudden drop in AMOC strength in 1991, which is also simulated in lead year 2, but not in lead years 5 nor 10. This weak event is also visible in the ORAS4 re-analysis data (Karspeck et al. 2015a, their Fig. 7)

At 45°N the EnKF systems multidecadal variability is as high as at 26°N, maximum phases occur in the early 1960s and the mid-1990s. The EnKF systems increase in AMOC strength until 1995 and drop thereafter corresponds to what was found by Pohlmann et al. (2013) to be the multidecadal variability all models in their study agree upon, compare Fig. 12 with Pohlmann et al. (2013), their Figs. 2b and 3b. In the NDGa and NDGf systems the multidecadal variability is different to the EnKF system, anomalously strong AMOC events are simulated by both NDGa and NDGf assimilations in 1973 and 1984. A third strong AMOC phase in 1992-1995 coincides with the maximum AMOC phase in the EnKF system. These events are also simulated in the hindcast lead year 2, but not in lead years 5 or 10. All three events are known to be present in the ORAS4 re-analysis (Karspeck et al. 2015a, their Fig. 7). They relate these events to anomalously deep convection and anomalously cold temperatures from the surface down to below 3000 m depth in the Labrador Sea as simulated by ORAS4 (Karspeck et al. 2015a, their Fig. 18).

In NDGa and NDGf the nudging methodology directly transfers anomalous temperatures and their associated densities from the ORAS4 re-analysis to MPI-ESM during assimilation. Already during assimilation MPI-ESM reacts with an anomalously strong AMOC during these events. However, the time mean AMOC in ORAS4 and NDGa and NDGf is different, in NDGa and NDGf it seems to be a mix of the models climatological time mean AMOC and the density variations picked up during the nudging toward temperatures and salinities of the ORAS4 re-analysis. In the hindcasts initialized by either nudging the simulated AMOC responds to these events in lead years 1 and 2.

5 Discussion and conclusion

In this study we established differences in hindcast skill in the SPG for HCT700 and SST within the time period 1961–2013 for three decadal prediction systems based on MPI-ESM. Skill not only differs between the systems but also depends on the evaluation period. This two-fold dependence of decadal hindcast skill impacts the way how we need to evaluate and compare decadal prediction systems in the future: Evaluation of one time period alone is not enough to fully apprehend the systems hindcast skill and the resulting forecast skill we could expect from that system. The difference between the prediction systems depend on the time period as well, e.g., correlation skill of hindcasts initialized by the EnKF system is higher than the one for hindcasts initialized by NDGa in the earlier 20-year time periods, whereas this relation is reversed in the later 20-year time periods within 1961–2013.

Considering both correlation and IQR together for lead years 1 to 5, the EnKF system often shows the highest correlation skill but the least fit with the reference data IQR. The NDGa system often has the highest fit in IQR but lower hindcast correlation skill than EnKF. The NDGf system often has even lower correlation skill than NDGa and exhibits its problems in the match of reference data IQR.

The EnKF system and the nudged systems differ in the way assimilation skill is mirrored in hindcast skill. In general, we expect hindcast skill to be lower than assimilation skill. This is the case for all three prediction systems. However, the decrease in correlation skill and the difference in IQR between assimilation and the first two hindcast years is lower in the EnKF system than in NDGa or in NDGf. That means for the correlation skill in particular, high assimilation skill due to nudging does not necessarily lead to high skill in the subsequent hindcasts. For the determination of a good model state for the initialization of decadal hindcasts it is not enough to assess the assimilation skill alone.

The skill of the hindcasts in the SPG depends on the method used for their initialization. Although all three prediction systems seem to be able to capture the multidecadal variability in the SPG to a certain extent, every initialization method comes with distinct problems. While the EnKF assimilation captures well the phase of this variability, it overestimates the amplitude when compared with reference data. A similar overestimation we also find in an assimilation with MPI-ESM where only atmospheric variables are nudged to re-analysis data. Further insight to the possible impact of the atmosphere on the oceanic variability during assimilation could be gained by carrying out a twin assimilation experiments with a free ocean. The EnKF assimilation changes the ocean variability where and when observations are available, but cannot overwrite the impact of the atmospheric nudging to the ocean. In the hindcasts initialized by EnKF the overestimation decays after lead year 5. An advantage of our EnKF assimilation setup in this study is that the AMOC is not disturbed. On the other hand the full value nudging in the ocean in NDGf causes a strong fit to re-analysis temperatures and salinities, due to the nudging the direct atmospheric impact is completely discarded. As a disadvantage the nudging also directly transfers the imperfections of the re-analysis to MPI-ESM. As a consequence the prediction system reacts with an anomalous interannual AMOC, and the resulting anomalous northward heat transport consistently destroys correlation skill in SPG temperatures in the first lead years after initialization (Kröger et al., in revision for Climate Dynamics). However, skill seems to re-emerge after lead year 3, a sign that at least the anomalous northward heat transport into the SPG may recover after this time. In comparison the anomaly nudging in the ocean in NDGa induces only moderate disturbances, the AMOC recovers faster and correlation skill in the first lead years is higher than in the full value nudging hindcasts. Some of the problems we can diagnose already during assimilation, e.g., overestimation of amplitudes in EnKF and the disturbed AMOC in NDGf. Other problems we can only detect in the hindcasts, e.g., correlation drop in SST in NDGa and NDGf.

In a broader sense the three initializations presented in this paper can be embedded in the ongoing discussion how to jointly assimilate atmospheric and oceanic variables for decadal prediction in a global coupled ESM. In the oceanic part special emphasis has to be paid to a balanced representation of the impact of both atmospheric and oceanic observations. For example, the initialization in the EnKF system could benefit from an assimilation setup with more ensemble members, localization and inflation, and possibly an assimilation interval shorter than one month, at least where observations exist frequently enough. At the same time disturbances to the AMOC during assimilation, for instance what happened in NDGf, should be kept under control as they may impact the hindcast skill in the SPG over several lead years. In view of a consistent coupled assimilation the extension of the EnKF to the assimilation of atmospheric parameters may have a positive impact.

From our analysis of the hindcast skill of three decadal prediction systems based on MPI-ESM we draw the following conclusions:

- In the North Atlantic subpolar gyre hindcast skill strongly depends on the evaluation period chosen, evaluating one, even long, hindcast period alone may be misleading. The analysis of 20-year subperiods helps to reveal strong differences between the prediction systems, these differences may almost be hidden if the analysis focused on the hindcast skill for the full time period 1961–2013 alone.
- Hindcast skill is generally better in periods of strong multiyear trends (the cooling in the 1960s and the warming in the 1990s) and weaker when multiyear trends are small (for example in the 1970s).
- Hindcasts initialized with an oceanic ensemble Kalman filter (EnKF) show higher correlation skill for lead years up to 5 and longer in both upper ocean heat content and sea surface temperature when compared to the systems based on anomaly and full value nudging. The EnKF initialized hindcasts are the only hindcasts under consideration, which reproduce the cooling in the 1960s AND the warming in the 1990s, albeit with an overestimated amplitude.
- Depending on the assimilation method, high correlation skill in the assimilation experiment does not necessarily transfer to high skill in the hindcast experiments.
- Supplementing the correlation analysis with the analysis of the interquartile range of the temporal probability density function improves the assessment of the similarity of initialized hindcasts against reference data and uninitialized simulations.

With our evaluation of differences in the initialized hindcasts depending on initialization method and evaluation period we lay the base for both an improvement in the explanation of how MPI-ESM reacts on assimilation of atmospheric and oceanic quantities and an improvement in the methods used to assimilate these quantities on decadal time scales.

Acknowledgements We thank two anonymous reviewers for their constructive and helpful comments for improving the manuscript. We thank Kameswarrao Modali and Helmuth Haak for technical help with the model, and Lars Nerger, AWI Bremerhaven, for providing PDAF and support in its implementation. Sea surface temperature data from HadISST and EN4 oceanic profile data have been retrieved through www.metoffice.gov.uk/hadobs, and NOCL heat content data through http://www.nodc.noaa.gov. This research was supported by the German Ministry of Education and Research (BMBF) under the MiKlip projectcs AODA-PENG (Grants 01LP1157C, 01LP1516A; SB, JB) and FlexForDec (Grant 01LP1519A; HP, WM) and through the Cluster of Excellence CliSAP (EXC177), Universität Hamburg, funded through the German Science Foundation (DFG) (AD, JB). The model simulations were performed at the German Climate Computing Centre (DKRZ).

References

Balmaseda MA, Mogensen K, Weaver AT (2013) Evaluation of the ECMWF ocean reanalysis system ORAS4. Quart J Roy Meteor Soc 139(674):1132–1161. https://doi.org/10.1002/qj.2063

Branstator G, Teng H (2012) Potential impact of initialization on decadal predictions as assessed for CMIP5 models. Geophys Res Lett 39(12): https://doi.org/10.1029/2012GL051974

Brune S, Nerger L, Baehr J (2015) Assimilation of oceanic observations in a global coupled Earth system model with the SEIK filter. Ocean Model 96 (Part 2):254–264. https://doi.org/10.1016/j.ocemod.2015.09.011

Buckley MW, Marshall J (2016) Observations, inferences, and mechanisms of the atlantic meridional overturning circulation: a review. Rev Geophys 54(1):5–63. https://doi.org/10.1002/2015RG000493

Chang YS, Zhang S, Rosati A, Delworth TL, Stern WF (2013) An assessment of oceanic variability for 1960–2010 from the GFDL ensemble coupled data assimilation. Climate Dyn 40(3–4):775–803. https://doi.org/10.1007/s00382-012-1412-2

Counillon F, Bethke I, Keenlyside NS, Bentsen M, Bertino L, Zheng F (2014) Seasonal-to-decadal predictions with the ensemble Kalman filter and the Norwegian Earth System Model: a twin experiment. Tellus A 66. https://doi.org/10.3402/tellusa.v66.21074

Cox P, Stephenson D (2007) A changing climate for prediction. Science 317(5835):207–208. https://doi.org/10.1126/science.1145956

DCPP-C (2016) Technical Note for DCPP-Component C—II. Recommendations for ocean restoring and ensemble generation. Tech. rep., World Climate Research Programme. https://www.wcrp-climate.org/wgsip/documents/Tech-Note-2.pdf

Dee DP et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quart J Roy Meteor Soc 137(656):553–597. https://doi.org/10.1002/qj.828

Delworth TL, Manabe S, Stouffer RJ (1993) Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. J Clim 6(11):1993–2011. https://doi.org/10.1175/1520-0442(1993)006<1993:IVOTTC>2.0.CO;2

Delworth TL, Zeng F, Zhang L, Zhang R, Vecchi GA, Yang X (2017) The central role of ocean dynamics in connecting the North Atlantic oscillation to the extratropical component of the Atlantic multidecadal oscillation. J Clim 30(10):3789–3805. https://doi.org/10.1175/JCLI-D-16-0358.1

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99(C5):10,143–10,162. https://doi.org/10.1029/94JC00572

Giorgetta MA et al (2013) Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. J Adv Mod Earth Sys 5(3):572–597. https://doi.org/10.1002/jame.20038

Good SA, Martin MJ, Rayner NA (2013) EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. J Geophys Res 118(12):6704–6716. https://doi.org/10.1002/2013JC009067

Hermanson L, Eade R, Robinson NH, Dunstone NJ, Andrews MB, Knight JR, Scaife AA, Smith DM (2014) Forecast cooling of the Atlantic subpolar gyre and associated impacts. Geophys Res Lett 41(14):5167–5174. https://doi.org/10.1002/2014GL060420, 2014GL060420

ICPO (2011) Decadal and bias correction for decadal climate predictions. Tech. Rep. 150, International CLIVAR Project Office, http://www.clivar.org/sites/default/files/documents/ICPO150_Bias.pdf

Jungclaus JH, Fischer N, Haak H, Lohmann K, Marotzke J, Matei D, Mikolajewicz U, Notz D, von Storch JS (2013) Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model. J Adv Mod Earth Sys 5(2):422–446. https://doi.org/10.1002/jame.20023

Karspeck AR, Yeager SG, Danabasoglu G, Hoar T, Collins N, Raeder K, Anderson JL, Tribbia J (2013) An ensemble adjustment kalman filter for the CCSM4 ocean component. J Clim 26(19):7392–7413. https://doi.org/10.1175/JCLI-D-12-00402.1

Karspeck AR, Stammer D, Köhl A, Danabasoglu G, Balmaseda M, Smith DM, Fujii Y, Zhang S, Giese B, Tsujino H, Rosati A (2015a) Comparison of the atlantic meridional overturning circulation between 1960 and 2007 in six ocean reanalysis products. Clim Dyn :1–26, https://doi.org/10.1007/s00382-015-2787-7

Karspeck AR, Yeager SG, Danabasoglu G, Teng H (2015b) An evaluation of experimental decadal predictions using CCSM4. Clim Dyn 44:907–923. https://doi.org/10.1007/s00382-014-2212-7

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Roeckner E (2008) Advancing decadal-scale climate prediction in the north atlantic sector. Nature 453(7191):84–88

Kröger J, Müller WA, von Storch JS (2012) Impact of different ocean reanalyses on decadal climate prediction. Clim Dyn 39(3–4):795–810. https://doi.org/10.1007/s00382-012-1310-7

Levitus S, Antonov JI, Boyer TP, Baranova OK, Garcia HE, Locarnini RA, Mishonov AV, Reagan JR, Seidov D, Yarosh ES, Zweng MM (2012) World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. Geophys Res Lett 39(10). https://doi.org/10.1029/2012GL051106

Marini C, Polkova I, Köhl A, Stammer D (2016) A comparison of two ensemble generation methods using oceanic singular vectors and atmospheric lagged initialization for decadal climate prediction. Mon Wea Rev 144(7):2719–2738. https://doi.org/10.1175/MWR-D-15-0350.1

Marotzke J (2016) MiKlip: a national research project on decadal climate prediction. Bull Amer Meteor Soc 97(12):2379–2394. https://doi.org/10.1175/BAMS-D-15-00184.1

Matei D, Pohlmann H, Jungclaus JH, Müller WA, Haak H, Marotzke J (2012) Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. J Clim 25(24):8502–8523. https://doi.org/10.1175/JCLI-D-11-00633.1

Menary MB, Hermanson L, Dunstone NJ (2016) The impact of Labrador Sea temperature and salinity variability on density and the subpolar AMOC in a decadal prediction system. Geophys Res Lett 43(23):12,217–12,227. https://doi.org/10.1002/2016GL070906,2016GL070906

Mignot J, García-Serrano J, Swingedouw D, Germe A, Nguyen S, Ortega P, Guilyardi E, Ray S (2016) Decadal prediction skill in the ocean with surface nudging in the IPSL-CM5A-LR climate model. Climate Dyn 47(3):1225–1246. https://doi.org/10.1007/s00382-015-2898-1

Msadek R, Delworth TL, Rosati A, Anderson W, Vecchi G, Chang YS, Dixon K, Gudgel RG, Stern WF, Wittenberg A, Yang X, Zeng F, Zhang R, Zhang S (2014) Predicting a decadal shift in North Atlantic climate variability using the GFDL forecast system. J Clim 27(17):6472–6496. https://doi.org/10.1175/JCLI-D-13-00476.1

Müller W, Matei D, Bersch M, Jungclaus J, Haak H, Lohmann K, Compo G, Sardeshmukh P, Marotzke J (2015) A twentieth-century reanalysis forced ocean model to reconstruct the north atlantic climate variation during the 1920s. Clim Dyn 44(7–8):1935–1955. https://doi.org/10.1007/s00382-014-2267-5

Müller WA, Baehr J, Haak H, Jungclaus JH, Kröger J, Matei D, Notz D, Pohlmann H, von Storch JS, Marotzke J (2012) Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. Geophys Res Lett 39(22). https://doi.org/10.1029/2012GL053326

Müller WA, Pohlmann H, Sienz F, Smith DM (2014) Decadal climate predictions for the period 1901–2010 with a coupled climate model. Geophys Res Lett 41:2100–2107. https://doi.org/10.1002/2014GL059259

Nerger L, Hiller W (2013) Software for ensemble-based data assimilation systems—implementation strategies and scalability. Comput Geosci 55:110–118. https://doi.org/10.1016/j.cageo.2012.03.026

Pham DT (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon Wea Rev 129(5):1194–1207. https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA$>2.0.CO;2

Pham DT, Verron J, Gourdeau L (1998) Singular evolutive Kalman filters for data assimilation in oceanography. C R Acad Sci, Ser II 326(4):255–260. https://doi.org/10.1016/S1251-8050(97)86815-2

Pohlmann H, Sienz F, Latif M (2006) Influence of the multidecadal atlantic meridional overturning circulation variability on european climate. J Clim 19(23):6062–6067. https://doi.org/10.1175/JCLI3941.1

Pohlmann H, Jungclaus JH, Köhl A, Stammer D, Marotzke J (2009) Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic. J Climate 22(14):3926–3938. https://doi.org/10.1175/2009JCLI2535.1

Pohlmann H, Müller WA, Kulkarni K, Kameswarrao M, Matei D, Vamborg FSE, Kadow C, Illing S, Marotzke J (2013a) Improved forecast skill in the tropics in the new MiKlip decadal climate predictions. Geophys Res Lett 40(21):5798–5802. https://doi.org/10.1002/2013GL058051

Pohlmann H, Smith DM, Balmaseda MA, Keenlyside NS, Masina S, Matei D, Müller WA, Rogel P (2013b) Predictability of the mid-latitude atlantic meridional overturning circulation in a multi-model system. Climate Dyn 41(3):775–785. https://doi.org/10.1007/s00382-013-1663-6

Polkova I, Köhl A, Stammer D (2015) Predictive skill for regional interannual steric sea level and mechanisms for predictability. J Clim 28(18):7407–7419. https://doi.org/10.1175/JCLI-D-14-00811.1

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14). https://doi.org/10.1029/2002JD002670

Robson JI, Sutton RT, Lohmann K, Smith DM, Palmer MD (2012a) Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. J Clim 25(12):4116–4134. https://doi.org/10.1175/JCLI-D-11-00443.1

Robson JI, Sutton RT, Smith DM (2012b) Initialized decadal predictions of the rapid warming of the north Atlantic ocean in the mid 1990s. Geophys Res Lett 39(19): https://doi.org/10.1029/2012GL053370

Robson JI, Sutton RT, Smith DM (2014) Decadal predictions of the cooling and freshening of the North Atlantic in the 1960s and the role of ocean circulation. Climate Dyn 42(9):2353–2365. https://doi.org/10.1007/s00382-014-2115-7

Romanova V, Hense A (2015) Anomaly transform methods based on total energy and ocean heat content norms for generating ocean dynamic disturbances for ensemble climate forecasts. Clim Dyn :1–21. https://doi.org/10.1007/s00382-015-2567-4

Servonnat J, Mignot J, Guilyardi E, Swingedouw D, Séférian R, Labetoulle S (2015) Reconstructing the subsurface ocean decadal variability using surface nudging in a perfect model framework. Clim Dyn 44(1):315–338. https://doi.org/10.1007/s00382-014-2184-7

Smeed DA, McCarthy GD, Cunningham SA, Frajka-Williams E, Rayner D, Johns WE, Meinen CS, Baringer MO, Moat BI, Duchez A, Bryden HL (2014) Observed decline of the Atlantic meridional overturning circulation 2004–2012. Ocean Sci 10(1):29–38. https://doi.org/10.5194/os-10-29-2014

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317(5839):796–799. https://doi.org/10.1126/science.1139540

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

Smith DM et al (2013b) Real-time multi-model decadal climate predictions. Clim Dyn 41(11–12):2875–2888. https://doi.org/10.1007/s00382-012-1600-0

Stevens B et al (2013) Atmospheric component of the MPI-M earth system model: ECHAM6. J Adv Mod Earth Sys 5(2):146–172. https://doi.org/10.1002/jame.20015

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Amer Meteor Soc 93(4):485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Timmreck C, Pohlmann H, Illing S, Kadow C (2016) The impact of stratospheric volcanic aerosol on decadal-scale climate predictions. Geophys Res Lett 43(2):834–842. https://doi.org/10.1002/2015GL067431,2015GL067431

Trenberth KE, Shea DJ (2006) Atlantic hurricanes and natural variability in 2005. Geophys Res Lett 33(12). https://doi.org/10.1029/2006GL026894,l12704

Uppala SM et al (2005) The ERA-40 re-analysis. Quart J Roy Meteor Soc 131(612):2961–3012. https://doi.org/10.1256/qj.04.176

Valcke S (2013) The OASIS3 coupler: a European climate modelling community software. Geosci Model Dev 6(2):373–388. https://doi.org/10.5194/gmd-6-373-2013

Volpi D, Guemas V, Doblas-Reyes FJ (2016) Comparison of full field and anomaly initialisation for decadal climate prediction: towards an optimal consistency between the ocean and sea-ice anomaly initialisation state. Clim Dyn :1–15. https://doi.org/10.1007/s00382-016-3373-3

Wilks D (2011) Statistical methods in the atmospheric sciences, international geophysics series, vol 100. Academic Press, New York

Yeager SG, Robson JI (2017) Recent progress in understanding and predicting atlantic decadal climate variability. Curr Clim Chang Reports 3(2):112–127. https://doi.org/10.1007/s40641-017-0064-z

Yeager SG, Karspeck AR, Danabasoglu G, Tribbia J, Teng H (2012) A decadal prediction case study: late twentieth-century North Atlantic Ocean heat content. J Clim 25(15):5173–5189. https://doi.org/10.1175/JCLI-D-11-00595.1

Zhang J, Zhang R (2015) On the evolution of atlantic meridional overturning circulation fingerprint and implications for decadal predictability in the north atlantic. Geophys Res Lett 42(13):5419–5426. https://doi.org/10.1002/2015GL064596,2015GL064596