Meteorol. Z. (Contrib. Atm. Sci.), Vol. **28**, No. 5, 417–428 (published online September 27, 2019)
© 2019 The authors

On skillful decadal predictions of the subpolar North Atlantic

Ines Höschel*, Sebastian Illing, Jens Grieger, Uwe Ulbrich and Ulrich Cubasch

Institut für Meteorologie, FU Berlin, Berlin, Germany

(Manuscript received December 20, 2018; in revised form March 25, 2019; accepted April 9, 2019)

Abstract

The North Atlantic is a crucial region for the prediction of weather and climate of North America and Europe and is the focus of this analysis. A skillful decadal prediction of the surface temperature was shown for several Earth system models, with the North Atlantic standing out as one region with higher predictive skill. This skill assessment concentrates on the rapid increase of the annual mean sea surface temperature of the North Atlantic subpolar gyre by about 1 K in the mid-1990s and the adjacent years. This event-oriented analysis adds creditability to the decadal predictions and reveals the potential for improvements. The ability to simulate the observed sea surface temperature in the North Atlantic is quantified by using four versions of decadal predictions, which differ in model resolution, initialization technique, and the reanalysis data used in the assimilation run. While all four versions can reproduce the mid-1990s warming of the subpolar North Atlantic, the characteristics differ with lead time and version. The higher vertical resolution in the atmosphere and the higher horizontal resolution in the ocean improve the decadal prediction for longer lead times, and the anomaly initialization outperforms the full-field initialization for short lead times. The effect from the two different ocean reanalysis products on the predictive skill is strongest in the first two prediction years; a substantial cooling instead of the warming in the central North Atlantic reduces the skill score for the North Atlantic sea surface temperature in one version, whereas a too large interannual variability, compared with observations, lowers the skill score in the other version. The cooling patches are critical since the resulting gradients in sea surface temperature and their effect on atmospheric dynamics deviate from observations, and, moreover, hinder the skillful prediction of atmospheric variables.

Keywords: decadal prediction, North Atlantic, sea surface temperature, model resolution, skill variations

1 Introduction

Ongoing efforts to establish climate prediction for a time range of up to a decade have led to a growing interest from decision makers in politics, finance, and industry on this topic. The North Atlantic (NA) is a key region for the prediction of the weather and climate in North America and Europe (Sutton and Hodson, 2005; Peings and Magnusdottir, 2014). It has been shown that decadal-scale fluctuations of sea surface temperature (SST) of the NA influence the number of Hurricanes (Trenberth and Shea, 2006), the Sahel precipitation (Martin and Thorncroft, 2014; Mohino et al., 2011), as well as the Mediterranean climate (Mariotti and DellAquila, 2012). Therefore, the understanding and the skillful prediction of the NA SST variability is essential for the skillful prediction of the Northern Hemisphere climate on a decadal timescale.

Decadal predictability arises from a response to changes in the concentrations of greenhouse gases (GHG) and influences from solar variability, as well as from low-frequency internal variability inherent to the Earth's climate system (Meehl et al., 2009). Decadal prediction systems use global climate models, which can simulate low-frequency internal variability (e.g., Giorgetta et al., 2013). These models are initialized from an observed state, and then freely evolve for the next 10 years (e.g., Bellucci et al., 2013; Marotzke et al., 2016; Robson et al., 2018). The skill of decadal prediction by these models can be influenced by the assimilation technique (Kröger et al., 2017), the type of reanalysis data (Kröger et al., 2012), and the model resolution (Prodhomme et al., 2016), among other factors. For the NA, Kröger et al. (2017) demonstrate that initializing from full oceanic fields triggers initial shocks in the ocean heat transport, which influences the predictive skill, while Pohlmann et al. (2013) show skill improvements due to the higher model resolution in the SST field for 2–5 lead years. Moreover, the predictive skill for the averaged NA SST is also influenced by the reanalysis data used during the assimilation procedure (Kröger et al., 2012).

A variety of decadal prediction systems show predictive skill for NA SST of up to a decade (e.g., Müller et al., 2012; Bellucci et al., 2013; Msadek et al., 2014; Marotzke et al., 2016; Robson et al., 2018). For example, Bellucci et al. (2013) used the climate model CMCC-CM to show that the skill of the decadal retrospective predictions (hindcasts) of SST in the extratropical NA is improved through initialization with near-observational states. In the NA, the subpolar gyre stands

*Corresponding author: Ines Höschel, Institut für Meteorologie, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6–10, 12165 Berlin, Germany, e-mail: ines.hoeschel@met.fu-berlin.de

© 2019 The authors

out as a region with a high predictive skill (Yeager and Robson, 2017).

In the mid-1990s, the subpolar gyre warmed rapidly by 1 K (Robson et al., 2012a), mainly because of an increased northward ocean heat transport related to a strengthening of the Atlantic circulation (Robson et al., 2012a). Robson et al. (2012b) used the UK Met Office decadal prediction system (DePreSys) to show that the rapid warming could have been predicted because of, primarily, the initialization with a strong Atlantic Meridional Overturning Circulation (AMOC).

Robson et al. (2012b) stated that the 1990s rapid warming is an excellent case study for assessing decadal predictions. Msadek et al. (2014) demonstrated a predictive skill of the mid-1990s warming in the subpolar gyre using the GFDL CM2.1 forecast system. Furthermore, the four ensemble members of the decadal hindcasts with the HiGEM climate model capture the warming (Robson et al., 2018). Different versions of the German decadal prediction system MiKlip (mittelfristige Klimaprognosen – midterm climate forecast; Marotzke et al., 2016) show a predictive skill for SST. For two versions of the MiKlip decadal prediction system, it was shown that a simulated SST index captures the warming of the subpolar gyre for short lead times (Marotzke et al., 2016; Müller et al., 2014).

Here, we examine the subpolar NA SST in decadal hindcasts from four versions of the German decadal prediction system MiKlip, with differences caused by the model resolution, the assimilation technique, and the type of reanalysis data used in the assimilation run. We focus on the rapid warming of the subpolar NA in the mid-1990s and the adjacent periods by considering the years from 1982 to 2010, and address the following two aspects in detail. 1) Which MiKlip versions capture the warming, and for which lead times, and does the higher model resolution increase the predictive skill in the subpolar NA? 2) How well do the four versions capture the spatial pattern associated with the mid-1990s warming in the subpolar NA, and how is this reflected in the skill score?

The next section describes the MiKlip model, the decadal hindcasts, and the evaluation method. Section 3 first presents the predicted time series of the SST index, with the spatial pattern associated with the warming of the subpolar NA in the hindcasts then compared with the observations. Next, the differences in the skill in predicting the subpolar NA SST are addressed by the use of the mean-squared-error skill score (MSESS) against different benchmarks. Finally, the discussion of these results is presented in Section 4.

2 Data and method

Global circulation models (GCM) are used in decadal prediction, and are run for approximately one thousand years with a constant external forcing of the conditions for the year 1850 in so-called equilibrium simulations. Transient historical simulations, which use the initial conditions from this equilibrium simulation, are forced with the observed solar variability, GHG, stratospheric ozone, and aerosol concentrations, and are combined with reanalysis data of the atmosphere and the ocean to produce near-observational initial conditions for decadal predictions. Decadal simulations with the GCM start every year (from 1961 until today) using these near-observational initialization data, and run freely for 10 years.

The German decadal prediction system MiKlip is based on the Max-Planck-Institute's Earth system model (MPI-ESM; Giorgetta et al., 2013), which is also used for the centennial-scale climate projection in the climate model intercomparison project phase 5 (CMIP5). The atmospheric part (ECHAM6; Stevens et al., 2013) has a horizontal resolution of T63 with 47 vertical levels in the low resolution (LR) version, and 95 levels in the mixed-resolution (MR) version, below an upper level at 1 Pa in both versions. The ocean component (MPIOM; Jungclaus et al., 2013) is a 40-layer ocean model with a horizontal resolution of 1.5° in the LR version and a 0.4° eddy-permitting resolution in the MR version, and includes a dynamic–thermodynamic sea-ice model. The ocean model is coupled via OASIS (Valcke, 2013) with the atmosphere model.

The MiKlip decadal prediction system uses a transient historical simulation (Giorgetta et al., 2013), and reanalysis data of the atmosphere and the ocean to create a near-observational assimilation run from which the decadal hindcasts are initialized. In the atmosphere model, full fields of vorticity, divergence, temperature, and sea-level pressure are nudged to ERA40 (1960–1989; Uppala et al., 2005) and ERA-interim (1990–2017; Dee et al., 2011) reanalyses. In the ocean model, full fields or anomalies of temperature and salinity are nudged to ORAS4 (Balmaseda et al., 2013) and GECCO2 (Köhl, 2015) reanalyses, respectively.

Beginning in the year 1961, and for every subsequent year on 1 January, an ensemble of free-running simulations is generated from the assimilation run using a lagged-day initialization method. Therefore, one specific version of decadal hindcasts is a collection of annually initiated ensembles of free-running simulations. The first (second, ..., ninth) year of the free-running simulation is called lead year 1 (2, ..., 9) and is abbreviated by ly1 (ly2, ..., ly9). That we only use the first 9 years of the experiments while skipping the last year is related to the consistency with the recommendations of Goddard et al. (2013), who restricted their analysis to years one to nine.

The natural and anthropogenic forcings of all simulations follow the observations until 2005, and subsequently the CMIP5 RCP 4.5 scenario. The anthropogenic forcing includes well-mixed GHG, ozone, and anthropogenic sulfate aerosols, while the natural forcing includes variations in the Earth's orbit, variability in the spectral solar irradiance, seasonally varying natural tropospheric aerosols, and stratospheric aerosols from volcanic eruptions (Giorgetta et al., 2013).

Table 1: Overview of datasets; [1]Pohlmann et al. (2013), [2]Marotzke et al. (2016), [3]Giorgetta et al. (2013), [4]Smith et al. (2008).

Dataset	Model resolution	Ocean initial.	Assimilated data	Ensemble size
b1-MR[1]	MR	anomaly	ORAS4	5
b1-LR[1]	LR	anomaly	ORAS4	10
pr-ora[2]	LR	full field	ORAS4	15
pr-gecco[2]	LR	full field	GECCO2	15
noAssim[3]	LR	no	no	10
ERSST[4]	observation	–	–	–

Four versions of decadal hindcasts from the two MiKlip generations, baseline1 and prototype, are analyzed (Table 1), differing in their horizontal and vertical model resolution (MR, LR), the assimilation method (anomaly or full field), and the ocean reanalysis data (ORAS4 or GECCO2).

In the baseline1 version (b1; Pohlmann et al., 2013), there are two versions of decadal hindcast ensembles, one five-member ensemble of the mixed-resolution model version (b1-MR), and one ten-member ensemble of the low-resolution model (b1-LR). Moreover, the full-field initialization is used in the atmosphere, and anomaly initialization with ORAS4-reanalysis is used in the ocean.

For the prototype prediction system (pr; Marotzke et al., 2016), a full-field initialization in the atmosphere and the ocean of the MPI-ESM-LR model is applied, with 15 ensemble members initialized from the ORAS4-assimilation run (pr-ora), and 15 ensemble members initialized from the GECCO2-assimilation run (pr-gecco). The only difference between the pr-ora and pr-gecco datasets is the reanalysis product used in the assimilation run.

The ensemble of transient simulations without any assimilation of reanalysis data (noAssim) is equivalent to the historical simulations (Giorgetta et al., 2013) as provided for CMIP5, albeit with an increased ensemble size of 10 members. The ensemble members of noAssim start from the equilibrium run with a time shift of 50 years. The ensemble mean of noAssim represents the climate adaptation due to changes in the concentration of GHG, aerosols, ozone and solar insolation, while the internal variability of the climate system is mainly canceled out through the 50-year time shift in the start dates of the ensemble members. Nevertheless, deviations from the above climate change signal in the noAssim ensemble mean can remain due to the limited number of ensemble members. The full ensemble sizes are used for analysis unless other ensemble sizes are mentioned.

The analysis is carried out for the same period 1982–2010 for all lead years as suggested in Boer et al. (2016), meaning the experiments from 1974 to 2002 are selected to analyze the ly9 simulation, and the experiments with the start years from 1975 to 2003 for the ly8 simulation, and so on. Before the skill of the hindcasts is assessed, the ensemble means of the annual mean anomalies are corrected for model drift and bias for each lead year following Goddard et al. (2013) and Boer et al. (2016) with the period from 1982 to 2010 as the reference. All model SST data are interpolated to the 2°×2° grid size of the observational data ERSST (Smith et al., 2008) before the comparison.

The domain of the NA SST index (NAI) is motivated by the difference between observations and the noAssim dataset (the simulation without any assimilation) within the box indicated in Figure 1. The NAI is defined as the area-weighted average of the annual anomalies of the SST to the mean from 1982 to 2010 of the NA from 65° W to 10° W and from 40° N to 65° N.

The mean-squared-error skill score (MSESS) is used to assess the accuracy of the hindcasts, and is a skill score based on the mean squared error (MSE). The MSE between the hindcast ensemble mean H_j and the observations O_j over $j = 1, \ldots, n$ initialization times is defined as

$$MSE_H = \frac{1}{n}\sum_{j=1}^{n}(H_j - O_j)^2. \tag{2.1}$$

The skill score compared to a reference prediction, such as the climatological forecast $\bar{O}$, can be written as

$$MSESS(H, \bar{O}, O) = 1 - \frac{MSE_H}{MSE_{\bar{O}}}, (-\infty, 1]. \tag{2.2}$$

The MSESS is a non-symmetric function with values in the interval $(-\infty, 1]$, with a positive value indicating an increased accuracy compared with the reference predictions, and a negative value a reduced accuracy. A significance denotes a skill difference distinct from zero exceeding the 95 % confidence level obtained by bootstrapping and resampling 500 times with replacement. For a more detailed explanation and the derivation, see Illing et al. (2014), Kadow et al. (2015) or Murphy (1988).

3 Results

The representation of the NA SST during the rapid warming in the mid-1990s and the adjacent periods in decadal hindcasts from four versions of the German decadal prediction system MiKlip is analyzed. The observed and the simulated NAI is investigated, and the spatial pattern associated with the warming is analyzed by comparing the observations with the hindcasts. Finally, the skill score MSESS is used to assess the improvements due to initialization and to quantify the differences between the versions regarding their skill in predicting the NA SST.

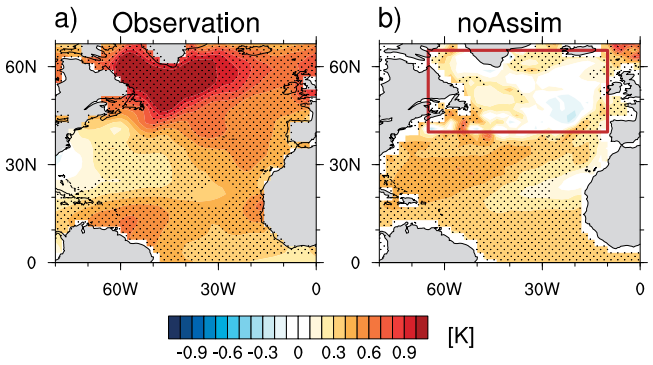

Figure 1: The difference in SST [K] of the 1998–2010 mean minus the 1982–1994 mean according to the observations (a) and the ensemble mean of 10 simulations without assimilation (noAssim) (b); the stippled area indicates statistical significance tested by a *t*-test with $\alpha = 0.01$; the brown box marks the region of more detailed investigations.

3.1 The North Atlantic index

The temporal development of the SST from the years 1982–2010 is analyzed by using the NAI, that is the area-weighted average over the domain of interest from the region without warming in the noAssim dataset indicated in Figure 1b. The temporal evolution of the index in the four versions is described by the ensemble means concatenated to one time series for each single lead year for each version.

The striking feature of the observed NAI is the rapid increase of about 0.8 K between 1994 and 1998 (Figure 2). In the observations, the mean of the index from the years 1982–1994 is −0.38 K, while the mean from the years 1998–2010, after the warming, is 0.38 K (Figure 3). During the colder conditions, the observed NAI vary between −0.62 K and −0.17 K, while the index shows anomalies between 0.13 K and 0.62 K in the warm period. Note that the observed annual temperature anomalies of the warm period do not overlap with those of the cold period. The noAssim dataset does not capture the observed change in the annual temperature anomalies between 1994 and 1998, showing rather a gradual increase in the NAI over the whole period (Figure 2).

In contrast, all initialized versions simulate a rapid increase in the mean temperature in the mid-1990s (Figure 2). While warming is, remarkably, simulated for all lead years, even for ly9 (see Figures 2 and 3a–d), there are, nonetheless, differences between the versions in the magnitude of the warming and in the interannual variability in the adjacent periods.

The magnitude of the temperature change in the b1-MR, pr-ora and pr-gecco datasets is closer to the observations for long lead times than for short lead times (Figure 3a–d). The variability is too small during the cold period in the anomaly-initialized versions for ly1, and too large especially for ly2 to ly4 of the b1-LR, pr-ora, and pr-gecco datasets. Overall, the full-field initialized versions show too large variability during the cold period. In contrast to the observations, the NAI values from the two periods overlap, especially in cases with too large variability.

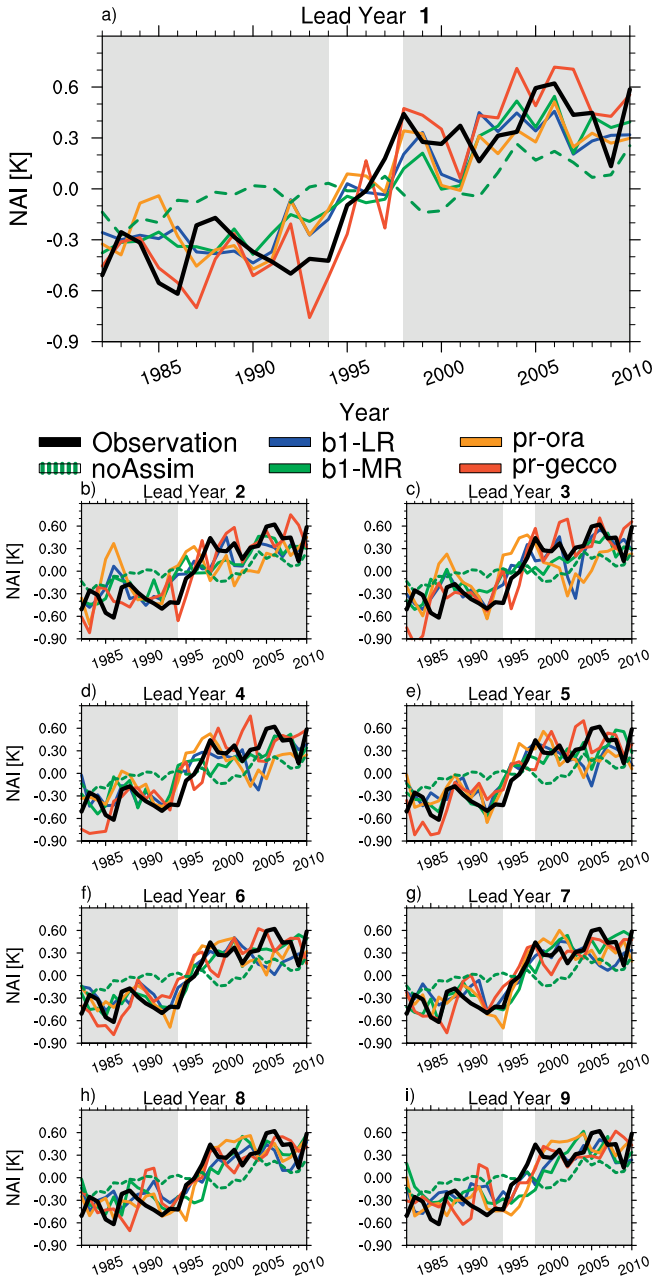

Figure 2: Time series of the NAI [K] for the observations (black), and the datasets noAssim (dashed, dark green), b1-LR (blue), b1-MR (green), pr-ora (yellow), pr-gecco (orange) for ly1 to ly9 (a)–(i).

3.2 The spatial pattern of the warming

The spatial pattern associated with the rapid warming in the subpolar NA is described by the difference of the mean temperature of the warm period from 1998 to 2010 and the cold period from 1982 to 1994 (Figures 1 and 4). The observations show significant warming over the NA, except for the Gulf Stream, with the observed warming pattern the most vigorous in the western part of the subpolar NA (Figure 1a). In contrast, the noAssim dataset reveals an entirely different spatial pattern (Figure 1b),

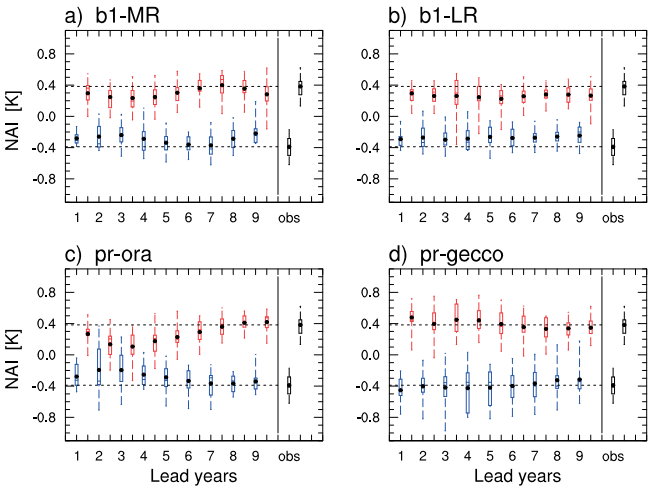

Figure 3: Boxplots (minimum, first quartile, median, third quartile, maximum) of NAI from ensemble means of the warm period 1998–2010 (red, observation black) and the cold period 1982–1994 (blue, observation black) of ly1 to ly9 of the b1-MR (a), b1-LR (b), pr-ora (c) and pr-gecco (d) datasets.

disagreeing with the observations in the missing warming in the whole subpolar NA (Figure 1b). There is significant warming in the tropics and the subtropics, but no warming north of 40° N in the noAssim ensemble mean. The following investigation concentrates on that region of the missing warming in the subpolar NA (the brown box in Figure 1b).

All four versions of decadal hindcasts can simulate warming in the subpolar NA (Figure 4), and the warming is statistically significant in large parts of the analyzed region irrespective of lead time. The simulated overall spatial pattern closely mirrors the observed one, although there are considerable variations between the hindcasts regarding the amplitude and the location of the strongest warming and the occurrence of spots of fictitious cooling in the central NA and near the Gulf Stream. The spatial agreement of the predicted warming pattern with observations is quantified by Pearson's pattern correlation coefficient r computed over the area in Figure 4 (identical to the brown box in Figure 1b). Individual coefficients are printed in the title of each of the panels.

The b1-MR version, which has the highest horizontal resolution in the ocean and the highest vertical resolution in the atmosphere, shows a better corresponding pattern with observations at long lead times, with the highest pattern correlations for ly8 ($r = 0.48$) and ly9 ($r = 0.47$). In the eastern NA, the warming is stronger than that observed until ly8 (Figure 4, first column). There is a zone of moderate fictitious cooling in the central NA until ly4, with the maximum extent for ly3. Furthermore, the magnitude of a fictitious cooling in the Gulf Stream grows with the lead time.

The b1-LR version (the same set-up as the b1-MR version, but with a lower resolution) shows a different development with lead time than the b1-MR version with respect to the spatial agreement with observations (Figure 4, second column). In ly1, the spatial pattern of the b1-LR version is close to the b1-MR version, apart from the strength of the fictitious cooling. The cooling in the Gulf Stream is stronger in the LR version, but the pattern correlations of the b1-LR and b1-MR versions are similar in ly1. With increasing lead times, both versions initially show decreasing correlations with the observed anomalies, but the b1-MR version regains stronger agreement for both the NAI and the temperature pattern towards ly8 and ly9.

The full-field-initialized pr-ora dataset (Figure 4, third column) shows an area of strong fictitious cooling in the central NA until ly5, in contrast to the anomaly-initialized b1-LR version, with the problem strongest for ly2. This is also reflected in the NAI, where the values between the cold and warm period overlap, and the temperature changes are too small in the first five lead years (Figure 3c), but the pattern agreement with observations improves after ly5. In ly9 ($r = 0.34$), the pattern correlation coefficient is higher than in ly1 to ly5. Notably, in the pr-ora dataset, the highest pattern correlation coefficient in ly7 ($r = 0.53$) exceeds the highest values in the b1-LR ($r = 0.34$) and b1-MR ($r = 0.48$) versions.

The pr-gecco version has the highest spatial correlation with observations among all versions for all lead times, although it overestimates the observed strength of the warming in the east (Figure 4, right column); this overestimation vanishes with lead time, and the warming pattern becomes more realistic. While there is a minimum of warming in the observations in the south-west corner of the area considered, the pr-gecco version produces a cooling patch at the southern border of the area, which increases in intensity from ly5 to ly9. Nevertheless, the warming pattern of the pr-gecco version in ly9 is closest to the observations ($r = 0.66$) among all versions and all lead times (Figure 4 xxxvi).

In summary, the spatial agreement with observations is mainly impaired through two aspects (Figure 4). Firstly, the baseline1 and pr-gecco datasets have too strong warming compared with the observations in the eastern part of the subpolar NA. Secondly, patches of fictitious cooling occur in almost all versions of the prediction system, but for different lead times, and at different specific locations and extensions.

3.3 Skill assessment

The predictive skill due to the initialization of the simulations from the assimilation run is presented in terms of the MSESS of the decadal hindcasts with respect to the reference hindcast noAssim. Figure 5 reveals that all the initialized decadal hindcasts show an improved skill with respect to the noAssim hindcast for the NAI in nearly all the cases for the period considered, the exception being no skill improvement for the pr-ora dataset for ly2 and ly3. The differences between the versions are small in ly1, ly7, ly8, ly9, while the full-field initialization ranks clearly below the anomaly initialization

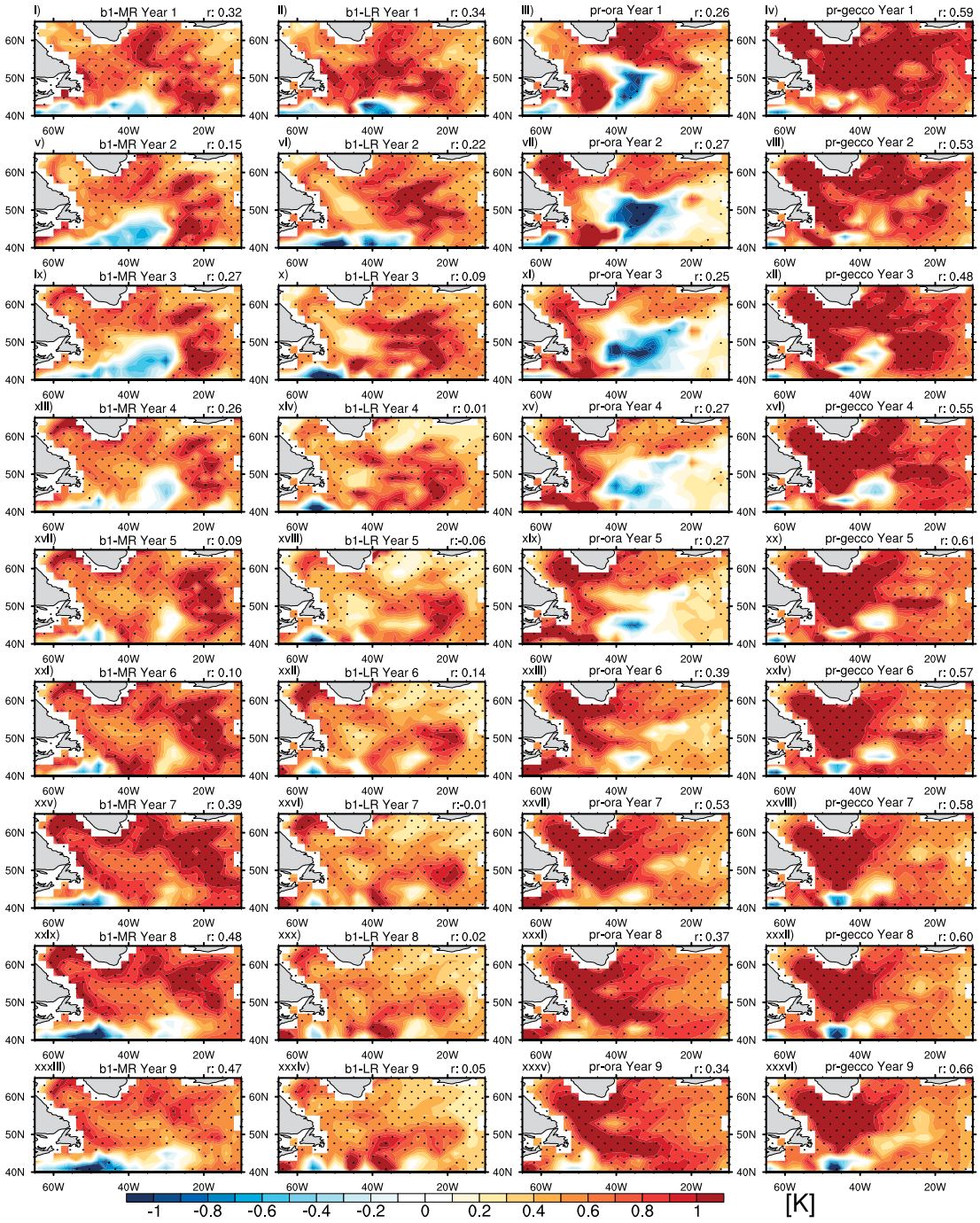

Figure 4: Difference in mean SST from 1998–2010 and 1982–1994; ly1 to ly9 from b1-MR – first column, b1-LR – second column, pr-ora – third column, pr-gecco – right column; r – Pearson pattern correlation coefficient: correlation with observation; the stippled areas indicate statistical significance tested by a t-test with $\alpha = 0.01$.

in ly3. The b1-MR version shows the highest MSESS for most lead years, except for ly3, ly8 and ly9. The evaluation per gridpoint of the skill of different versions with respect to the noAssim dataset (Figure 6) indicates significant skill improvement due to initialization in regions of the Icelandic, the Irminger, the West Greenland, and the Labrador Currents, as well as in the Labrador Sea and the Flemish Cap in different lead years. The b1-MR version shows the greatest improvements among the initialized hindcasts for nearly all lead times. Skill improvements are realized in the first prediction year and at longer lead times, while lead years two to four are dominated by significant skill loss in all versions. Cooling spots and areas of too strong warming (Figure 4) are reflected in the skill loss in the Gulf Stream region and the central NA, which are largest in the full-field initialized versions pr-ora and pr-gecco.

The effects from model resolution, initialization technique, and reanalysis data on the predictive skill can be evaluated by using the prediction of one of the respective versions of the decadal prediction system as a reference, rather than the noAssim dataset (Figure 7).

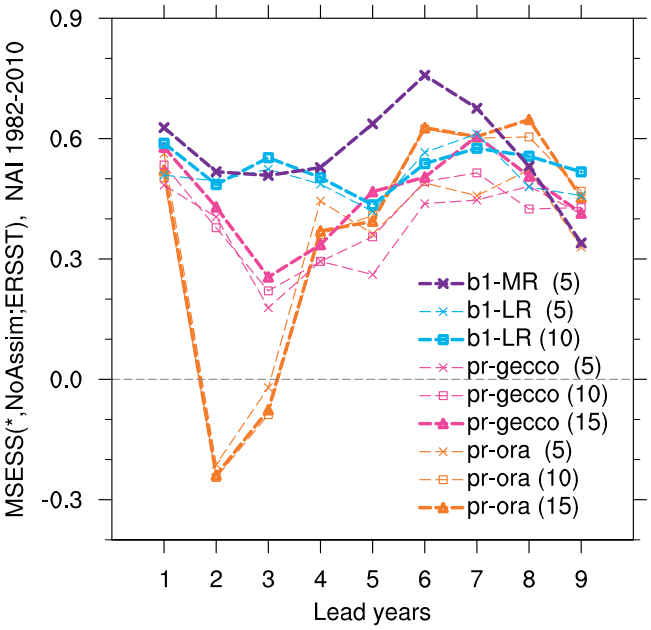

Figure 5: The MSESS of the NAI (1982–2010) for the initialized versions with respect to the noAssim version for individual lead years (the ensemble size is indicated in parenthesis).

First, the influence of the model resolution on the predictive skill is illustrated by computing the skill score of the b1-MR version with respect to b1-LR as the reference prediction. The increased model resolution improves the skill over large areas of the NAI region at all lead times (Figure 7, left column), with the most substantial improvements seen for ly4 to ly7. Significant differences are caused with respect to the initialization technique (Figure 7, middle column), where the anomaly-initialized version excels in the center of the mid-latitudes of the NA. These significant improvements in the b1-LR version compared with pr-ora are seen in all lead years, but with decreasing strength. In contrast, the full-field initialized pr-ora shows higher predictive skill around 40° N in all lead years and the skill increases with lead time north of 60° N (blue in Figure 7, middle column). The dependency of the predictive skill on the reanalysis dataset used in the assimilation is visible in Figure 7, right column, where the significant differences are most substantial in the first two lead years. In ly1, the pr-ora version excels, apart from the center of the mid-latitudes. In ly2, the pr-gecco version reveals better predictions than pr-ora in large parts of the NA (blue in Figure 7, right column). The discrepancies are reduced in later lead years, except in the NA Current, where the pr-gecco version yields the improved predictive skill.

4 Summary and discussion

We have shown that the ensemble means of the four different versions of the MiKlip decadal prediction system reproduce the rapid warming of the subpolar NA SST, irrespective of their differences and of the lead time. The uninitialized hindcast noAssim does not show the rapid increase in the NAI, but produces a steady increase of temperatures. The representation of this period of rapid rise must thus be due to the initialization and not the external forcing (i.e., the effects of the changing GHG concentrations are represented by the ensemble mean of the noAssim hindcast, while the internal variability is mainly canceled out through the 50-year-shifted initial condition from the pre-industrial equilibrium simulation).

With respect to the spatial pattern of the temperature change, the noAssim hindcast is very different from both the observations and the decadal forecasts. The individual patterns of the predicted SST change reveal a development of the predictions with the lead year, but the shortest lead times do not imply the highest correlations. This effect is visible in all versions of the prediction system, even if individual features of the pattern are specific for a part of the versions and lead times. Examples are the cooling spots in the south-west to southern border of the area investigated, which is present in the b1-MR version and, in particular, at short lead times in the b1-LR version, while a similar spot mainly occurs in the pr-gecco dataset at long lead times. Only the pr-ora version produces a large and strong cooling in the central parts of the area considered, and at lead times shorter than five years. Also, the areas of largest warming differ, and are found in the central to eastern part of the area in the two baseline1 versions, but more focused on the central to the western part in the two prototype versions.

An initialization with a strong AMOC is a key to the prediction of the NA warming (Robson et al., 2012a; Robson et al., 2012b), which is the case in the analyzed MiKlip prediction system (supplemental material), and serves as a possible explanation of the pattern in the anomaly-initialized versions. This is because the spatial change in the SST reveals a region with a cooling in the Gulf Stream / NA Current region and warming elsewhere, which is similar to the pattern described in Borchert et al. (2018), who show that a strong AMOC at 50° N leads to this kind of pattern 3 to 10 years later. Additionally, the AMOC in their assimilation run is strong between 1994 and 2001, which fits the time horizon considered here.

The full-field-initialized versions show significant differences due to the type of reanalysis, illustrating the remarkable uncertainties in the ocean reanalysis products. The cooling spot in the central NA reduces the average temperature change at short lead times in the pr-ora version, while the NAI shows a too large amplitude of the year-to-year variability in the pr-gecco version. In general, the influence of the reanalysis product on the predictive skill is in agreement with the findings reported in Kröger et al. (2012). We attribute the discrepancies to the time dependence of the skill in decadal prediction systems (Brune et al., 2018), and to the lead time averaging.

The low skill at short lead times in the full-field hindcasts is consistent with results published by Kröger

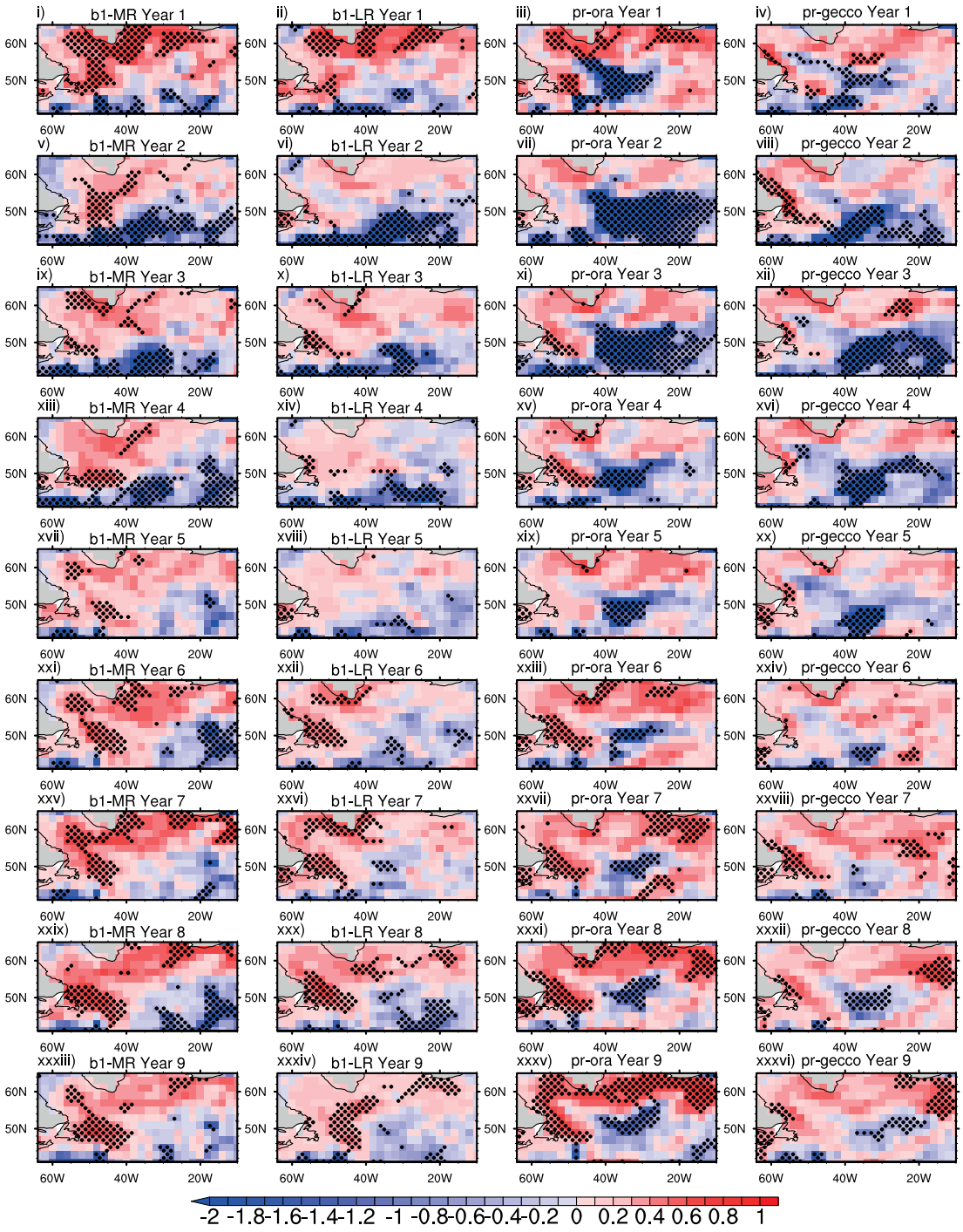

Figure 6: The MSESS of SST from the years 1982 to 2010 with respect to the noAssim dataset: ly1 to ly9 for b1-MR – first column, b1-LR – second column, pr-ora – third column, pr-gecco – right column; reference: ERSST data; the stippled areas indicate significance by bootstrapping.

et al. (2017), who ascribe it to initial shocks in the decadal hindcasts resulting from the poor initialization of the oceanic flow. In their analysis, the nudging of temperature and salinity induces changes in ocean heat and mass transport. These changes and the net heat exchange at the surface do not correspond to the tendencies of the ocean heat content in the subpolar gyre in the assimilation runs. The discrepancies are larger with the full field than with the anomaly nudging, which appears to be consistent with the skill of the full-field-initialized hindcasts being lower than in their anomaly counterparts during the first five lead years.

The comparison of the b1-LR and b1-MR versions shows the effect of the increased model resolution on the predictive skill of the subpolar NA SST. Overall, the b1-MR version is better alongside the subpolar gyre, but shows deficits in the eastern mid-latitudes. Nevertheless, the b1-MR version reveals a better pattern agreement than the b1-LR version at longer lead times, and, consequently, higher local skill scores than the b1-LR version

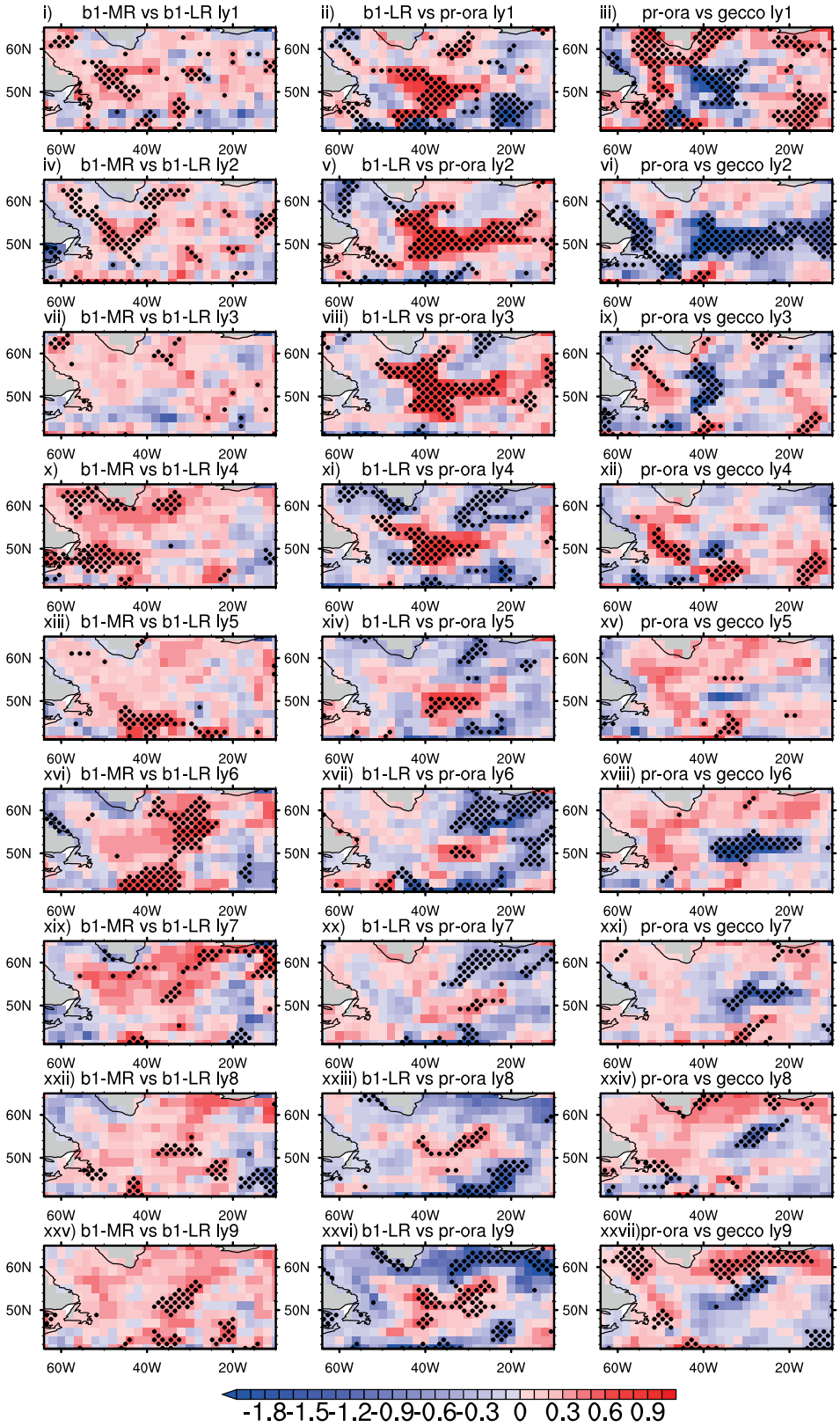

Figure 7: The MSESS of SST (1982–2010) from ly1 to ly9 accounting for resolution (five ensemble members): b1-MR against b1-LR (left column), for the initialization method (10 ensemble members): b1-LR against pr-ora (middle column), for the reanalysis used in assimilation run (15 ensemble members): pr-ora against pr-gecco (right column); reference: ERSST data; the stippled areas indicate significance by bootstrapping.

in large areas of the subpolar NA. There are two possible sources of the improved skill of the b1-MR version compared with the b1-LR version due to the difference in experimental design, since these versions differ in their horizontal resolution of the ocean model and their vertical resolution of the atmosphere model, especially in the stratosphere.

In the eddy-permitting ocean model used for the b1-MR version, Jungclaus et al. (2013) find a weaker NA deep-water cell, and a stronger Antarctic bottom-water cell than in the b1-LR version. Differences in the stream function may affect the changes in the SST on multi-year timescales, and serve as a possible explanation of the better pattern agreement of the b1-MR version at longer lead times and its improved skill.

The higher vertical stratospheric resolution of the b1-MR version enables a better representation of stratospheric processes. For example, Pohlmann et al. (2013) show that, in the b1-MR version, the predictive skill of the quasi-biennial oscillation in the stratosphere remains significant by up to four lead years. The higher resolution in the stratosphere may affect the NA SST prediction, since multidecadal variability in the stratospheric circulation can be generated within an ocean–troposphere–stratosphere model (e.g., Schimanke et al., 2011). Reichler et al. (2012) demonstrate that the variations in the sequences of stratospheric circulation anomalies, combined with the persistence of individual anomalies, significantly affect the NA. Thus, the higher resolved stratosphere possibly improves the representation of ocean–troposphere–stratosphere interaction on the decadal timescale, and, consequently enhances the predictive skill of the SST in the b1-MR version. Furthermore, the combination of external forcing and the better representation of processes potentially plays a role for the increased skill, as Ottera et al. (2010) show that changes in the external solar and volcanic forcing are important for multi-decadal fluctuations in NA SST.

To conclude the points made above, we see the following possible implications.

Firstly, there is no clear best version of the decadal prediction system regarding the NA SST from 1982 to 2010. The hindcast from the model with the increased resolution shows a few advantages, but the otherwise moderate improvements and the poor performance in the eastern part of the subpolar NA currently do not justify the higher computational costs.

Secondly, there is a dependence of the regional skill on the reanalysis product at all lead times. The significant regional differences in skill show the urgent need for a broadening of the efforts to improve the ocean reanalysis products in terms of evaluation, model development, and extension of ocean observations.

Thirdly, the fictitious cooling patches implicate changes in the SST gradients, which deviate from the observed gradients. Therefore, this appears as one possible source of the reduced predictive skill of atmospheric variables, such as the precipitation and wind speed, while impacting relevant events, such as heat waves and storms in the decadal hindcasts.

Finally, the question of how to make progress in decadal prediction remains. We recommend avoiding initial shocks by not using full-field initialization unless the model climatology is close enough to the observations, or by using an ensemble Kalman filter (Brune et al., 2018) to deliver an initial state closer to model climatology. Additionally, a higher resolution is beneficial, but it should be explored why there is no clear benefit in the first few lead years. For example, is the initial state a problem in combination with the slow transport of anomalies in the ocean? An enhanced ocean-model resolution would be worthwhile for key regions, for example, in the north-west Atlantic with its steep temperature gradients.

Acknowledgments

We acknowledge funding from the Federal Ministry of Education and Research in Germany (BMBF) through the research program *MiKlip II* (FKZ: 01LP1519B, 01LP1520A). The NOAA_ERSST_V3b data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/

References

Balmaseda, M.A., K. Mogensen, A.T. Weaver, 2013: Evaluation of the ecmwf ocean reanalysis system ORAS4. – Quart. J. Roy. Meteor. Soc. **139**, 1132–1161, DOI: 10.1002/qj.2063.

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

Boer, G.J., D.M. Smith, C. Cassou, F. Doblas-Reyes, G. Danabasoglu, B. Kirtman, Y. Kushnir, M. Kimoto, G.A. Meehl, R. Msadek, W.A. Mueller, K.E. Taylor, F. Zwiers, M. Rixen, Y. Ruprich-Robert, R. Eade, 2016: The decadal climate prediction project (dcpp) contribution to CMIP6. – Geosci. Model Develop. **9**, 3751–3777, DOI: 10.5194/gmd-9-3751-2016.

Borchert, L.F., W.A. Müller, J. Baehr, 2018: Atlantic ocean heat transport influences interannual-to-decadal surface temperature predictability in the north atlantic region. – J. Climate **31**, 6763–6782, DOI: 10.1175/JCLI-D-17-0734.1.

Brune, S., A. Düsterhus, H. Pohlmann, W.A. Müller, J. Baehr, 2018: Time dependency of the prediction skill for the north atlantic subpolar gyre in initialized decadal hindcasts. – Climate Dyn. **51**, 1947–1970, DOI: 10.1007/s00382-017-3991-4.

Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.J. Morcrette, B.K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.N. Thepaut, F. Vitart, 2011: The era-interim reanalysis: configuration and performance of the data assimilation system. – Quart. J. Roy. Meteor. Soc. **137**, 553–597, DOI: 10.1002/qj.828ls.

Giorgetta, M.A., J. Jungclaus, C.H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, T. Crueger, M. Esch, K. Fieg, K. Glushak, V. Gayler, H. Haak, H.D. Hollweg, T. Ilyina, S. Kinne, L. Kornblueh, D. Matei, T. Mauritsen, U. Mikolajewicz, W. Mueller, D. Notz, F. Pithan, T. Raddatz, S. Rast, R. Redler, E. Roeckner, H. Schmidt, R. Schnur, J. Segschneider, K.D. Six, M. Stockhause, C. Timmreck, J. Wegner, H. Widmann, K.H. Wieners, M. Claussen, J. Marotzke, B. Stevens, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the coupled model intercomparison project phase 5. – J. Adv. Model. Earth Sys. **5**, 572–597, DOI: 10.1002/jame.20038.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. Mason, B. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C. Ferro, D. Stephenson, G.A. Meehl, T. Stockdale, R. Burgman, A. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, T. Delworth, 2013: A verification framework for interannual-to-decadal predictions experiments. – Climate Dyn. **40**, 245–272, DOI: 10.1007/s00382-012-1481-2.

Illing, S., C. Kadow, O. Kunst, U. Cubasch, 2014: Murcss: A tool for standardized evaluation of decadal hindcast systems. – J. Open Res. Software **2**, e24, DOI: 10.5334/jors.bf.

Jungclaus, J.H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, J.S. von Storch, 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. Model. Earth Sys. **5**, 422–446, DOI: 10.1002/jame.20023.

Kadow, C., S. Illing, O. Kunst, H. Rust, H. Pohlmann, W. Müller, U.Cubasch, 2015: Evaluation of forecasts by accuracy and spread in the MIKLIP decadal climate prediction system. – Meteorol. Z. **25**, 631–643, DOI: 10.1127/metz/2015/0639.

Köhl, A., 2015: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the atlantic. – Quart. J. Roy. Meteor. Soc. **141**, 166–181, DOI: 10.1002/qj.2347.

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

Kröger, J., H. Pohlmann, F. Sienz, J. Marotzke, J. Baehr, A. Köhl, K. Modali, I. Polkova, D. Stammer, F.S.E. Vamborg, W.A. Müller, 2017: Full-field initialized decadal predictions with the mpi earth system model: an initial shock in the north atlantic. – Climate Dyn. **51**, 2593–2608, DOI: 10.1007/s00382-017-4030-1.

Mariotti, A., A. DellAquila, 2012: Decadal climate variability in the mediterranean region: roles of large-scale forcings and regional processes. – Climate Dyn. **38**, 1129–1145, DOI: 10.1007/s00382-011-1056-7.

Marotzke, J., W.A. Müller, F.S.E. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H.W. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J.G. Pinto, H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, M. Ziese, 2016: MIKLIP – a national research project on decadal climate prediction. – Bull. Amer. Meteor. Soc. **97**, 2379–2394, DOI: 10.1175/BAMS-D-15-00184.1.

Martin, E.R., C.D. Thorncroft, 2014: The impact of the AMO on the West African monsoon annual cycle. – Quart. J. Roy. Meteor. Soc. **140**, 31–46, DOI: 10.1002/qj.2107.

Meehl, G.A., L. Goddard, J. Murphy, R.J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M.A. Giorgetta, A.M. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimot, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, T. StockdaleG, 2009: Decadal prediction can it be skillful?. – Bull. Amer. Meteor. Soc. **90**, 1467–1485, DOI: 10.1175/2009BAMS2778.1.

Mohino, E., S. Janicot, J. Bader, 2011: Sahel rainfall and decadal to multi-decadal sea surface temperature variability. – Climate Dyn. **37**, 419–440, DOI: 10.1007/s00382-010-0867-2.

Msadek, R., T.L. Delworth, A. Rosati, W. Anderson, G. Vecchi, Y.S. Chang, K. Dixon, R.G. Gudgel, W. Stern, A. Wittenberg, X. Yang, F. Zeng, R. Zhang, S. Zhang, 2014: Predicting a Decadal Shift in North Atlantic Climate Variability Using the GFDL Forecast System. – J. Climate **27**, 6472–6496, DOI: 10.1175/JCLI-D-13-00476.1.

Müller, W.A., J. Baehr, H. Haak, J.H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J.S. von Storch, J. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the max planck institute for meteorology. – Geophys. Res. Lett. **39**, L22707, DOI: 10.1029/2012GL053326.

Müller, W.A., H. Pohlmann, F. Sienz, D. Smith, 2014: Decadal climate predictions for the period 1901-2010 with a coupled climate model. – Geophys. Res. Lett. **41**, 2100–2107, DOI: 10.1002/2014GL059259.

Murphy, A.H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. – Mon. Wea. Rev. **116**, 2417–2424.

Ottera, O.H., M. Bentsen, H. Drange, L. Suo, 2010: External forcing as a metronome for atlantic multidecadal variability. – Nature Geosci. **3**, 688–694, DOI: 10.1038/NGEO955.

Peings, Y., G. Magnusdottir, 2014: Forcing of the wintertime atmospheric circulation by the multidecadal fluctuations of the north atlantic ocean. – Env. Res. Lett. **9**, 034018, DOI: 10.1088/1748-9326/9/3/034018.

Pohlmann, H., W.A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F.S.E. Vamborg, C. Kadow, S. Illing, J. Marotzke, 2013: Improved forecast skill in the tropics in the new miklip decadal climate predictions. – Geophys. Res. Lett. **40**, 5798–5802, DOI: 10.1002/2013GL058051.

Prodhomme, C., L. Batté, F. Massonnet, P. Davini, O. Bellprat, V. Guemas, F.J. Doblas-Reyes, 2016: Benefits of increasing the model resolution for the seasonal forecast quality in ec-earth. – J. Climate **29**, 9141–9162, DOI: 10.1175/JCLI-D-16-0117.1.

Reichler, T., J. Kim, E. Manzini, J. Kröger, 2012: A stratospheric connection to atlantic climate variability. – Nature Geosci. **5**, 783–787, DOI: 10.1038/NGEO1586.

Robson, J., R. Sutton, K. Lohmann, D. Smith, M.D. Palmer, 2012a: Causes of the rapid warming of the north atlantic ocean in the mid-1990s. – J. Climate **25**, 4116–4134, DOI: 10.1175/JCLI-D-11-00443.1.

Robson, J.I., R.T. Sutton, D.M. Smith, 2012b: Initialized decadal predictions of the rapid warming of the north atlantic ocean in the mid 1990s. – Geophys. Res. Lett. **39**, L19713, DOI: 10.1029/2012GL053370.

Robson, J., I. Polo, D.L.R. Hodson, D.P. Stevens, L.C. Shaffrey, 2018: Decadal prediction of the north atlantic subpolar gyre in the higem high-resolution climate model. – Climate Dyn. **50**, 921–937, DOI: 10.1007/s00382-017-3649-2.

Schimanke, S., J. Körper, T. Spangehl, U. Cubasch, 2011: Multi-decadal variability of sudden stratospheric warmings in an aogcm. – Geophys. Res. Lett. **38**, L01801, DOI: 10.1029/2010GL045756.

Smith, T.M., R.W. Reynolds, T.C. Peterson, J. Lawrimore, 2008: Improvements to noaa's historical merged land-ocean surface temperature analysis (1880-2006). – J. Climate **21**, 2283–2296.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler,, E. Roeckner, 2013: Atmospheric component of the mpi-m earth system model: Echam6. – J. Adv. Model. Earth Sys. **5**, 146–172, DOI: 10.1002/jame.20015.

Sutton, R.T., D.L.R. Hodson, 2005: Atlantic ocean forcing of north american and european summer climate. – Science **309**, 115–118.

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

Uppala, S.M., P.W. Kallberg, A.J. Simmons, U. Andrae, V.D.C. Bechtold, M. Fiorino, J.K. Gibson, J. Haseler, A. Herandez, G.A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R.P. Allan, E. Andersson, K. Arpe, M.A. Balmaseda, A.C.M. Beljaars, L. Vandeberg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Holm, B.J. Hoskins, L. Isaksen, P.A.E.M. Janssen, R. Jenne, A.P.McNally, J.F. Mahfouf, J.J. Morcrette, N.A. Rayner, R.W. Saunders, P. Simon, A. Sterl, K.E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, J. Woollen, 2005: The era-40 re-analysis. – Quart. J. Roy. Meteor. Soc. **131**, 2961–3012, DOI: 10.1256/qj.04.176.

Valcke, S., 2013: The OASIS3 coupler: a European climate modelling community software. – Geosci. Model Develop. **6**, 373–388, DOI: 10.5194/gmd-6-373-2013.

Yeager, S.G., J.I. Robson, 2017: Recent progress in understanding and predicting Atlantic decadal climate variability. – Curr. Climate Change Reports **3**, 112–127, DOI: 10.1007/s40641-017-0064-z.

The pdf version (Adobe Java Script must be enabled) of this paper includes an electronic supplement:
Table of content – Electronic Supplementary Material (ESM)

Figures S1, S2, S3, S4