Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL084758

Forecast-Oriented Assessment of Decadal Hindcast Skill for North Atlantic SST

Leonard F. Borchert[1,2,3,4], **André Düsterhus**[2,5], **Sebastian Brune**[2], **Wolfgang A. Müller**[1], **and Johanna Baehr**[2]

[1]Max Planck Institute for Meteorology, Hamburg, Germany, [2]Institute for Oceanography, CEN, Universität Hamburg, Hamburg, Germany, [3]International Max Planck Research School on Earth System Modelling, Hamburg, Germany, [4]Now at Sorbonne Universités (SU/CNRS/IRD/MNHN), LOCEAN Laboratory, Institut Pierre Simon Laplace (IPSL), Paris, France, [5]Now at ICARUS, National University of Ireland Maynooth, Maynooth, Ireland

Abstract We demonstrate in this paper that conventional time-averaged decadal hindcast skill estimates can overestimate or underestimate the credibility of an individual decadal climate forecast. We show that hindcast skill in a long period can be higher or lower than skill in its subperiods. Instead of using time-averaged hindcast skill measures, we propose to use the physical state of the climate system at the beginning of the forecast to judge its credibility. We analyze hindcasts of North Atlantic sea surface temperature (SST) in an initialized prediction system based on the MPI-ESM-LR for the period 1901–2010. Subpolar North Atlantic Ocean heat transport (OHT) strength at hindcast initialization largely determines the skill of these hindcasts: We find high skill after anomalously strong or weak OHT, but low skill after average OHT. This knowledge can be used to constrain conventional hindcast skill estimates to improve the assessment of credibility for a decadal forecast.

Plain Language Summary Credible predictions of climate up to 10 years into the future, so-called decadal climate predictions, can be a potent tool for decision makers. However, previous work indicated that such predictions are sometimes credible and sometimes not. This study illustrates that knowing the physical state of the climate system at the start of a decadal climate prediction helps assessing the credibility of that prediction. Analyzing sea surface temperature in the North Atlantic as a case study, we show that northward heat transport in the ocean provides a good indicator of the credibility of decadal predictions in the North Atlantic. Unlike previous studies, we do not only analyze time-averaged prediction credibility, but look at individual predictions in the past. This makes our findings particularly relevant for individual forecasts and decision makers.

Key Points:
- Decadal hindcast skill estimates can misrepresent forecast skill as skill depends on the length of the time window used for calculation
- North Atlantic SST hindcast skill is high when subpolar ocean heat transport is strong or weak at initialization and low after average OHT
- Analyzing individual hindcasts shows that accounting for ocean heat transport helps to better constrain credibility of individual forecasts

Correspondence to:
L. F. Borchert,
leonard.borchert@locean-ipsl.upmc.fr

Citation:
Borchert, L. F., Düsterhus, A., Brune, S., Müller, W. A., & Baehr, J. (2019). Forecast-oriented assessment of decadal hindcast skill for North Atlantic SST. *Geophysical Research Letters*, *46*, 11,444–11,454. https://doi.org/10.1029/2019GL084758

Received 30 JUL 2019
Accepted 22 SEP 2019
Accepted article online 21 OCT 2019
Published online 30 OCT 2019

1. Introduction

Climate prediction for up to 10 years into the future, so-called decadal climate prediction, has received increasing scientific and public attention in recent years (e.g., Yeager & Robson, 2017). The strong interest in decadal climate predictions arises from the relevance of this time scale for decision makers who are interested in reliable forecasts of climate variability up to 10 years ahead (e.g., Boer et al., 2016). For any actual decadal climate forecast, it is paramount to know whether to trust the forecast—to understand its *credibility*—when it is issued. This credibility is commonly assessed for a fixed period of the past as *hindcast skill*. In this study we point out the importance to differentiate between hindcast skill and forecast credibility and suggest an approach to more accurately estimate forecast credibility from hindcast studies.

Conventional studies of decadal climate hindcasts result in one estimate of hindcast skill for the period in the past that is analyzed, often some 60 years (e.g., Boer et al., 2016; Yeager & Robson, 2017). This skill estimate is currently the best available estimate for the credibility of any individual forecast of the future that would be conducted using the same prediction system. While it was shown that the statistical significance of hindcast skill increases with the length of the time period used for its calculation (e.g., Müller et al., 2014), a recent study illustrated that hindcast skill in decadal predictions of North Atlantic sea surface temperature (SST) changes over time (Brune et al., 2018). A similar effect was also shown for seasonal hindcast skill of the North Atlantic Oscillation (O'Reilly et al., 2017; Weisheimer et al., 2017). Time-dependent hindcast skill indicates that a hindcast skill estimate derived for one fixed period in the past could not appropriately represent the

skill found in all time periods within that period. This might limit the applicability of conventional hindcast skill estimates to individual decadal climate forecasts. Because North Atlantic SST has a strong influence on climate in Europe (e.g., Årthun et al., 2017; Borchert et al., 2018; Gastineau & Frankignoul, 2015), the finding by Brune et al. (2018) also affects forecasts of European climate.

Changes in hindcast skill were hypothesized to be caused by physical processes. For decadal predictions of North Atlantic SSTs, Brune et al. (2018) hypothesized that phases of high hindcast skill were linked to very strong or very weak phases of the Atlantic Meridional Overturning Circulation (AMOC) and associated meridional ocean heat transport (OHT), while phases of low hindcast skill were linked to phases of average AMOC and OHT. In addition, case studies showed for the 1960s cooling and the 1990s warming of North Atlantic SST that weak and strong phases of AMOC preceding the respective events led to high SST hindcast skill in the North Atlantic (Yeager et al., 2012; Robson et al., 2012, 2014).

Borchert et al. (2018) illustrated that the skill of decadal North Atlantic SST hindcasts in the entire twentieth century was systematically influenced by a physical process proposed by Zhang (2008) and Zhang and Zhang (2015). This process involves a slowly southward propagating OHT anomaly in the North Atlantic, modulating ocean heat convergence and divergence in the subpolar North Atlantic on the decadal time scale. Decadal North Atlantic SST hindcast skill was shown to be on average conditioned by slowly southward propagating phases of OHT in the subpolar North Atlantic and the physical mechanism derived by Zhang and Zhang (2015) for up to a decade into the future (Borchert et al., 2018). While Borchert et al. (2018) showed an asymmetric influence of strong and weak phases of subpolar OHT on North Atlantic SST prediction skill, the influence of average OHT phases on decadal SST hindcast skill was not shown so far.

The present study provides a framework to harvest previously unused potential in conventional decadal hindcast skill estimates. We show that accounting for the strength of subpolar OHT at the beginning of decadal North Atlantic SST predictions can help in constraining hindcast skill estimates to make them more applicable for individual forecasts. Unlike previous studies, we do not only show the effect of the climate state at hindcast initialization for time averages, but connect the credibility of individual decadal hindcasts to the physical state of the climate system at their initialization. This study shows that accounting for the physical state of the climate system at the start of hindcasts enables a better constrained estimate of forecast credibility than common time-averaged hindcast skill estimates.

2. Data and Methods

2.1. Model

We analyze SST from simulations with the fully coupled Max Planck Institute Earth System Model in its low-resolution setup (MPI-ESM-LR) for the period 1901–2010 (data published: Modali, 2017). The simulations are based on the CMIP5 version of MPI-ESM-LR, consisting on the ocean model MPIOM (Jungclaus et al., 2013) at an average horizontal resolution of 1.5° and 40 vertical levels, and the atmospheric model ECHAM6 (Stevens et al., 2013) at the horizontal resolution T63 with 47 vertical levels. Specifically, we use three assimilation experiments with the MPI-ESM-LR, in which the coupled model was for the period 1901–2010 nudged daily toward three-dimensional ocean temperature and salinity fields (Müller et al., 2014). These temperature and salinity fields were taken from an ensemble of MPIOM simulations forced every 6 hr by fluxes of heat, freshwater, and momentum through bulk formulas at the ocean surface (Müller et al., 2015). These fluxes were calculated from the twentieth century reanalysis (Compo et al., 2011).

In January of every year of the assimilation run, a 10-year-long hindcast simulation with the free fully coupled MPI-ESM-LR was started from each of the three assimilation runs (Müller et al., 2014). We correct the mean bias of these individual hindcast simulations against the corresponding realization of the assimilation run, and form an ensemble mean of the hindcast simulations. Both the assimilation run and the hindcast simulations were previously shown to produce, despite the limited ensemble size, reasonable climate variability in the North Atlantic region (Borchert et al., 2018; Müller et al., 2015).

2.2. Postprocessing

Annual mean values of North Atlantic SSTs are analyzed either on an individual grid point basis or as an integrated value, as the Atlantic Multidecadal Variability (AMV; e.g., Delworth et al., 2017). We define the AMV as integrated SST anomalies in the North Atlantic between 0 and 80°N (as in, e.g., Ba et al., 2014; Sutton & Hodson, 2005). This index was in the past shown to be connected to ocean overturning variability (e.g., Delworth et al., 2017; Dijkstra et al., 2006). While its principal spatial characteristics can be reproduced

by a slab ocean model forced by stochastic heat flux variability at the ocean surface (Clement et al., 2015), active ocean overturning was shown to be important to set the time scale of varibaility for this AMV index (Zhang et al., 2016). Note that unlike in previous studies, we do not low-pass filter the time series of North Atlantic SST to obtain the AMV index. This ensures that AMV predictions are based on individual years and not decade-long averages.

We calculate total depth-integrated North Atlantic ocean heat transport (OHT) at each latitude using the formula presented in Jayne and Marotzke (2001). For the analysis presented in this paper, we focus on subpolar OHT, using OHT at 50°N (henceforth OHT_{50N}). To illustrate the influence of OHT_{50N} on hindcast skill, we contrast average and anomalous phases of OHT_{50N}. Anomalous phases comprise strong and weak phases of OHT_{50N}. To analyze time series of similar lengths, we separate the OHT_{50N} time series using a criterion of a two-third standard deviation around the mean. More strict criteria for the selection of anomalous phases of OHT_{50N}, for example, 2 standard deviations, amplify the results presented in this paper, but reduce the length of analyzed time series dramatically, reducing the robustness of the result. The OHT_{50N} phases separated by a two-third standard deviation consist of 54 years for average OHT_{50N}, and 55 years for anomalous OHT_{50N}.

2.3. Hindcast Skill Assessment

We examine decadal hindcast skill of annual mean North Atlantic SST in this study, focusing on predictions 5–7 years into the future (on *lead years* 5–7). Skill is assessed using the anomaly correlation coefficient (ACC; e.g., Jolliffe & Stephenson, 2012), root-mean-square error (RMSE; e.g., Jolliffe & Stephenson, 2012), and using a contigency table approach. We use SST from the Hadley Center Sea Ice and Sea Surface Temperature reanalysis data set (HadISST; Rayner et al., 2003) and the ensemble mean of the assimilation model experiments (Müller et al., 2015) as reference for skill estimation. We show hindcast skill against both reference data sets throughout the study to ensure that our findings are not contaminated by physical inconsistencies between the model and observations.

Statistical significance of hindcasts is in this study evaluated by bootstrapping hindcast start years 1,000 times with replacement. The significance level is, unless stated otherwise, 1%. When applying a significance test to a map, we control for the False Discovery Rate using the procedure outlined in Wilks (2016), with $\alpha_{FDR} = 0.1$.

We calculate hindcast skill for the entire time series (i.e. 1901-2010) as well as for all possible 30, 40, 50, 60, and 70-year-long sub-periods of the entire time series. We correct the mean bias of the hindcast ensemble against the assimilation run and subtract the linear trend from both the hindcast ensemble mean as well as from the reference time series prior to calculating ACC, such that the global warming trend does not dominate our findings.

3. Time-Dependent Decadal SST Hindcast Skill

AMV hindcast skill is highly significant at lead years 5–7, independent of the reference data set that is used for skill calculation (thick black dot in Figures 1a and 1b). However, breaking down hindcast skill estimates to 40-year periods, we show that the ACC for AMV changes over time, taking values between $\sim$ 0 and 0.8 for a given 40-year period (Figures 1a and 1b). As a consequence, AMV hindcast skill is significant in some periods and not significant in other periods. The hindcast skill estimate for North Atlantic SST therefore strongly depends on the period that it is calculated for (as also shown by Brune et al. (2018), albeit for Atlantic subpolar gyre SST). This finding illustrates the complexity of the problem of estimating the credibility of an individual AMV forecast.

3.1. The Effect of Time Averaging on Decadal SST Hindcast Skill

Going beyond previous studies, we analyze the dependence of AMV hindcast skill at lead years 5–7 on the length of the time period for which skill is evaluated (Figure 1c). This analysis indicates the potential size of the error induced by assuming time-averaged hindcast skill to reflect forecast credibility.

Using ACC, we show AMV hindcast skill for long time periods and their shorter subperiods by systematically shortening the time window for which skill is evaluated from 70 to 30 years. Skill estimates derived for short time windows are sometimes as high as skill estimates for long time windows and sometimes lower. This shows that assessing hindcast skill for long time windows of, for example, 50 years likely results in a high skill estimate but that this skill estimate does not necessarily reflect the skill within every subperiod.

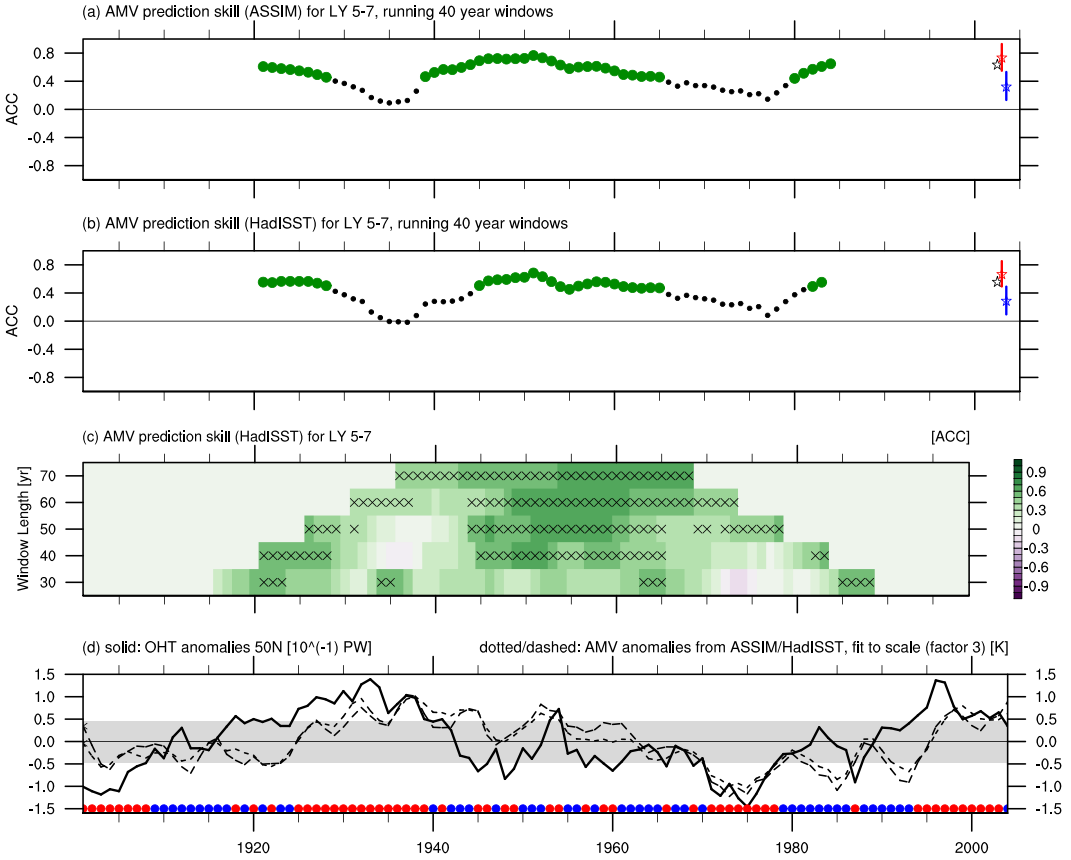

Figure 1. (a)–(c) Atlantic Multidecadal Variability (AMV) hindcast skill at lead years 5–7. (a) Change of hindcast skill estimates over time for a running 40-year time window, evaluated against the assimilation run. Each dot is centered on the time period that is used for its calculation. Green dots indicate statistical significance at the 1% level. The red, blue, and black asteriks indicate AMV hindcast skill for hindcasts started in years of anomalous and average OHT_{50N} and for the entire time series, respectively. The vertical red and blue lines through the asterisks mark the tenth to ninetieth percentile of 1,000 bootstrap samples of the respective underlying distribution. (b) As in (a) but with HadISST observations as a reference. (c) Time-dependent AMV hindcast skill against HadISST varies with the length of the skill evaluation time window (*y* axis). Here, crosses denote statistical significance at the 1% level. (d) Time series for OHT_{50N} (solid line), AMV from ASSIM (dotted line), and HadISST (dashed line). AMV time series were multiplied by 3 to fit the scale of the *y* axis. The gray area denotes the selection criterion for anomalous and average phases of OHT_{50N} used in this study: two thirds of a standard deviation. Colored dots at the bottom indicate whether a year falls into an anomalous (red) or average (blue) OHT_{50N} phase. SST = sea surface temperature; ACC = anomaly correlation coefficient; HadISST = Hadley Center Sea Ice and Sea Surface Temperature reanalysis.

Figure 1c also shows the effect of low-frequency AMV variability on hindcast skill. For very long evaluation time windows that cover a full oscillation of the AMV (i.e., 70 year windows), skill is significant across all time periods. For shorter time windows that include at least one-half AMV oscillation (40- to 60-year windows), hindcast skill is sometimes significant and sometimes not. Even shorter evaluation time windows (30 years) show hardly any skill. This highlights the importance of the low-frequency AMV oscillation in setting SST hindcast skill in the North Atlantic region.

Figure 1 thus provides a strong indication that a hindcast skill estimate derived for a period of several years is unlikely to reflect the hindcast skill of every individual hindcast within that period. Conventional measures of hindcast skill can thus overestimate or underestimate the credibility of an individual forecast.

3.2. OHT and AMV Hindcast Skill

A first indication of whether the skill of a hindcast will be higher or lower than the time-averaged hindcast skill estimate might be gained from the strength of subpolar OHT at hindcast initialization (as was suggested by, e.g., Borchert et al., 2018; Robson et al., 2012). In fact, OHT_{50N} variability in the assimilation run consistently leads AMV in both assimilation and HadISST by 5–10 years as indicated by a lag correlation

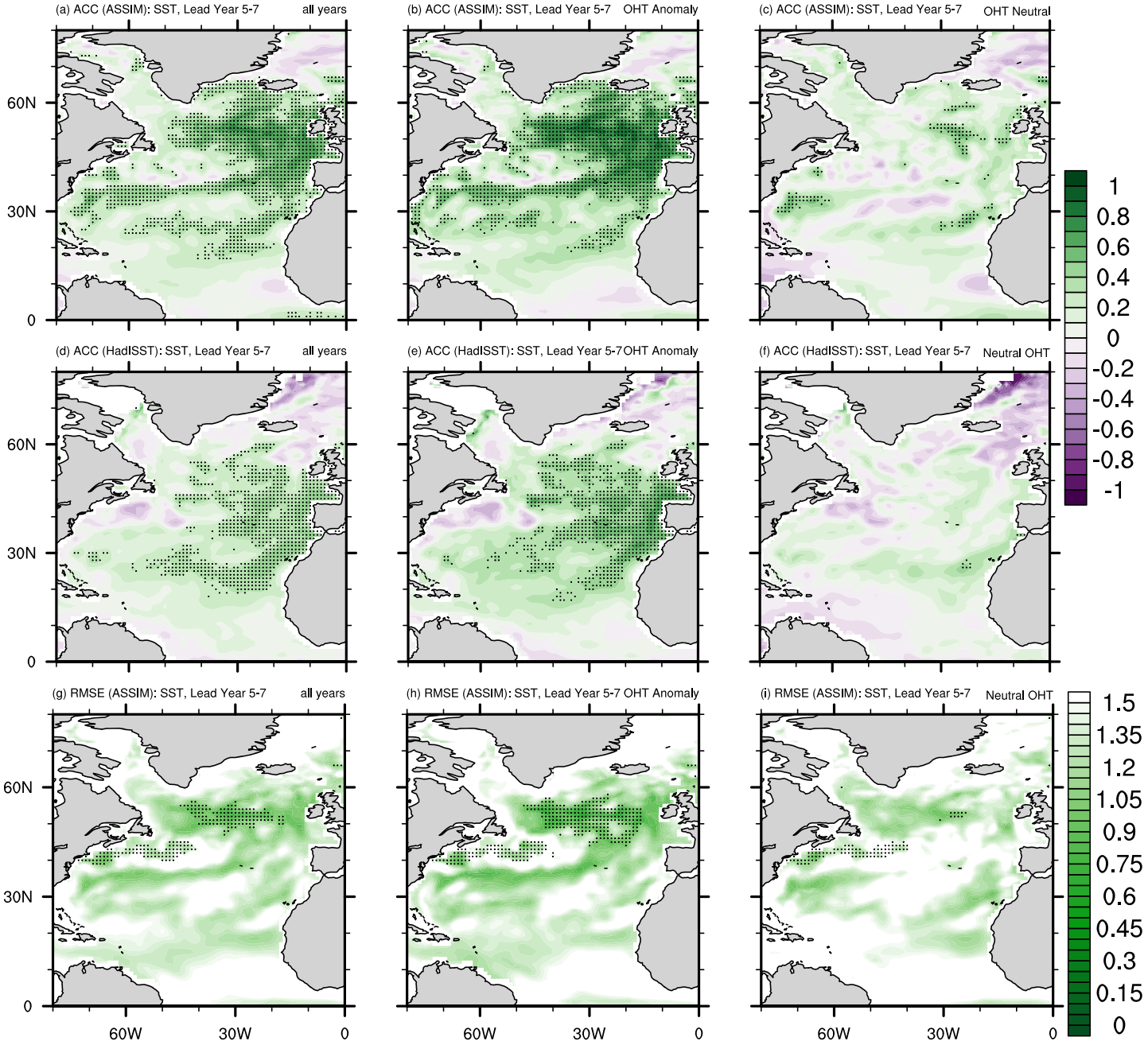

Figure 2. Anomaly correlation coefficient (ACC) maps [corr.] for sea surface temperature at lead years 5–7, using (a)–(c) the assimilation run, and (d)–(f) HadISST observations as a reference. Maps of RMSE [K] at lead years 5–7 using the assimilation run for reference (g)–(i). (a, d, g) Calculated for all hindcasts. (b, e, h) Calculated for hindcasts that were initialized in years of anomalous OHT_{50N} in the assimilation run. (c, f, i) Calculated for average OHT_{50N} phases. Stippling shows statistical significance at the $\alpha_{FDR} = 0.1$ level. HadISST = Hadley Center Sea Ice and Sea Surface Temperature reanalysis; RMSE = root-mean-square error; OHT = ocean heat transport.

analysis. Highest correlations occur when OHT_{50N} leads by 5–10 years at values >0.6 for both the assimilation run and HadISST (not shown). This indicates a physical relationship between OHT and SST as was also described in Zhang (2008), Zhang and Zhang (2015), and Borchert et al. (2018).

We evaluate hindcast skill separately for anomalous and average OHT_{50N} at hindcast initialization to assess the influence of OHT_{50N} on decadal AMV hindcast skill. When OHT_{50N} is anomalous at hindcast initialization, the hindcast skill estimate is among the highest estimates we find (assimilation run: 0.78, HadISST: 0.69; red asterisk in Figures 1a and 1b). Conversely, average phases of OHT_{50N} yield comparatively low AMV

hindcast skill when evaluated as ACC (assimilation run: 0.32, HadISST: 0.28; blue asterisk in Figures 1a and 1b). These skill values are clearly different: Their uncertainty bars do not overlap in Figures 1a and 1b.

4. Using Ocean Heat Transport Strength to Constrain Decadal SST Hindcast Skill Estimates

This section illustrates the specific influence of OHT_{50N} on decadal AMV hindcast skill. North Atlantic SST is significantly predictable at lead years 5–7 against the assimilation run (Figure 2a). The SST pattern that is predictable at this time scale closely resembles the horseshoe pattern that characterizes the AMV (not shown) as well as its temporal evolution (Figure 1d; also, see Borchert et al., 2018). The highest decadal hindcast skill is found in the subpolar gyre region.

We find strong time dependence of ACC hindcast skill for SST fields (Figures 2b and 2c). As skill maps differ strongly between anomalous and average OHT_{50N} phases, hindcast skill for North Atlantic SST fields shows the same time dependence as AMV hindcast skill (cf. Figure 1.)

There is a strong influence of OHT_{50N} variability on decadal SST hindcast skill with high hindcast skill in the entire subpolar gyre region after anomalous OHT_{50N} phases (Figure 2b). The complete absence of significant subpolar SST hindcast skill after average OHT_{50N} phases (Figure 2c) illustrates the strong influence of subpolar OHT variability on decadal SST hindcast skill in the North Atlantic region.

Using HadISST as a reference for hindcast skill assessment yields qualitatively similar results as using the assimilation run (Figures 2d–2f). The predictable SST pattern resembles the AMV-horseshoe (Figure 2d), and we find a similar dependence of decadal SST hindcast skill on the initial phase of OHT_{50N} when evaluating skill against HadISST observations as when evaluating skill against the assimilation run (Figures 2e and 2f).

We use RMSE hindcast skill to illustrate the robustness of the ACC-based findings. While RMSE maps (Figures 2g–2i) do not show a prominent horseshoe pattern, there is significant decadal hindcast skill in the subpolar gyre region. Furthermore, hindcasts are more skillful after anomalous than after average phases of OHT_{50N}. This indicates that the effect of the initial OHT_{50N} phase on SST hindcast skill is robust.

5. Assessing the Credibility of Individual Hindcasts to Judge Forecast Credibility

5.1. The Skill of Individual AMV Hindcasts

While having demonstrated that hindcast skill estimates should be broken down into physical states to harvest their full potential, the analysis so far operates on time averages. In the following, we assess the skill of individual hindcasts to demonstrate how knowing the physical state of the climate system at the start of an individual forecast can indicate its credibility.

We assess the performance of individual hindcasts as a blueprint to the performance of an individual forecast. We compare the predicted AMV value to the corresponding observed value for every individual hindcast. Specifically, we use a contingency table approach to map the performance of individual hindcasts into four categories: anomalous AMV predicted and observed (*correct anomalous*); anomalous AMV predicted and average AMV observed (*false anomalous*); average AMV predicted and anomalous AMV observed (*false average*); and average AMV predicted and observed (*correct average*). A contingency table can be visualized as a scatter plot with hindcast AMV value on the x axis and observed AMV value on the y axis.

Between 1901 and 2010, 57.4% of AMV hindcasts show correct anomalous or correct average AMV values against the assimilation run at lead years 5–7 (for HadISST this number is 49.5%). These values qualitatively correspond well to those found using ACC for the same time series (0.68 and 0.59, respectively; cf. Figures 1a and 1b), indicating that this simple method of assessing hindcast performance leads to reasonable results. This approach does not only provide an intuitive measure of hindcast credibility, but it also enables the assessment of individual hindcasts as well as averages. We will now examine the different skills between hindcasts that were started in years of anomalous and average OHT_{50N} more closely, using the contingency table approach.

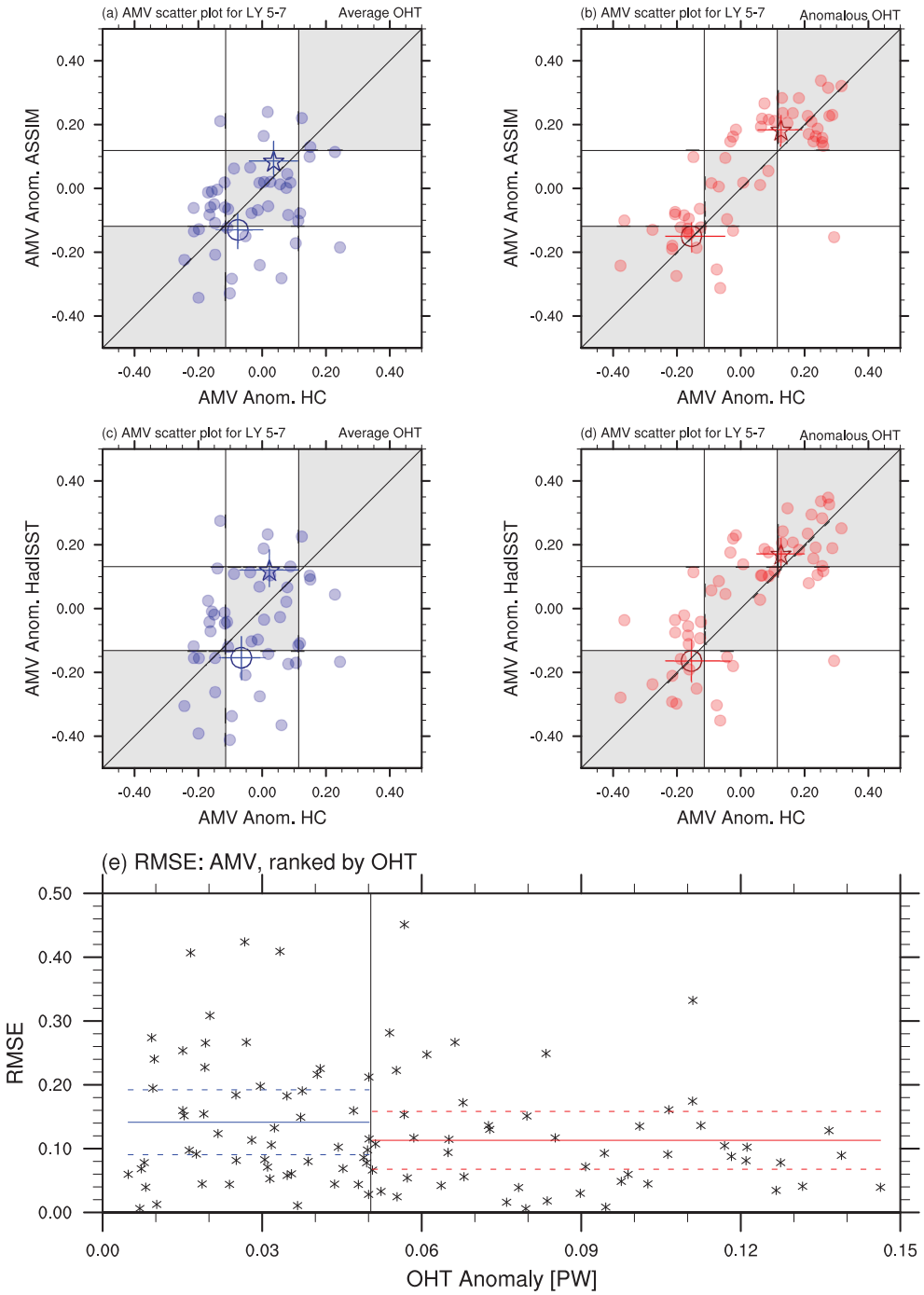

Figure 3. (a) A scatter plot comparing predicted Atlantic Multidecadal Variability (AMV) anomalies (*x* axis) at lead years 5–7 to the corresponding AMV anomalies observed in the assimilation run (*y* axis). Blue dots denote AMV anomalies corresponding to average OHT_{50N} phases at hindcast initialization. The asterisk/circle denote the average positive/negative observed AMV anomaly and the corresponding average predicted AMV anomaly. Colored lines show the uncertainty of the position of the asterisk and circle based on 10,000 bootstrap samples (0.1% confidence). Quandrants on the scatter plot in which hindcasts predict the AMV phase defined as anomalous (more than 1 standard deviation from the mean) or average (within 1 standard deviation from the mean) are shown in gray. (b) As in (a) but showing anomalous OHT_{50N} phases in red. (c) and (d) As in (a) and (b) but using observed AMV phases from HadISST as a reference. Panel (e) shows the skill of individual hindcasts using RMSE evaluated against HadISST as a function of absolute OHT_{50N} anomaly at hindcast initialization. The black vertical line represents the criterion that we use to identify anomalous and average OHT_{50N}. Solid blue and red lines show the average RMSE for average and anomalous OHT_{50N} respectively, while the dashed lines illustrate a standard deviation around the respective mean. RMSE = root-mean-square error; OHT = ocean heat transport.

5.2. The Skill of Individual AMV Hindcasts Depends on the Initial Climate State

The scatter plots show that in general, AMV hindcasts perform better after anomalous AMV phases than after average ones (Figures 3a–3d). Of the AMV hindcasts corresponding to anomalous OHT_{50N}, 67.3% predict the correct phase of the AMV as observed in the assimilation run, while only 45.7% of AMV hindcasts corresponding to average OHT_{50N} predict the correct AMV phase (for HadISST, these numbers are 57.4% and 43.5%, respectively).

In the assimilation run and observations, anomalous phases of OHT_{50N} are generally followed by strong positive or negative AMV phases (in 74.5% and 63.6% of cases in the assimilation run and HadISST, respectively; Figures 3b and 3d). As the hindcasts mostly reproduce this behavior, AMV predictions of anomalous AMV phases are well constrained after anomalous OHT_{50N} (54.6% and 43.7% correct AMV phases predicted against the assimilation run and HadISST, respectively). On average, the predicted anomalous phase of AMV that is predicted by the hindcast corresponds well to that observed, indicating significant credibility of predictions of anomalous AMV following anomalous OHT_{50N}.

Average phases of OHT_{50N} are in the assimilation run and HadISST less indicative of the following AMV anomalies than anomalous ones. After average OHT_{50N}, average and anomalous AMV phases are similarly likely to occur (average in 60.8% and 56.5% of cases in the assimilation run and HadISST, respectively; Figures 3a and 3c). This behavior is represented by the hindcasts. Hindcasts predict correct average AMV phases against the assimilation run and HadISST in 30.4% and 26.1% of cases, respectively. However, hindcasts are just as likely to wrongly predict anomalous AMV phases from average OHT_{50N} (30.8% and 30.4% against the assimilation run and HadISST, respectively). This indicates insignificant hindcast skill after average OHT_{50N}.

The effect of subpolar OHT on AMV hindcast skill seems less pronounced for negative AMV anomalies than for positive AMV anomalies (Figures 3a–3d). In other words, knowing the phase of OHT_{50N} at the beginning of a decadal AMV prediction is likely to have a more beneficial influence on the skill estimate when predicting an anomalously positive AMV phase than when predicting a negative one. This was already noted in Borchert et al. (2018) and is likely related to a quick and relatively unpredictable compensation of negative oceanic heat anomalies in the North Atlantic from the atmosphere.

5.3. RMSE for Individual Hindcasts

Another method to estimate the credibility of an individual hindcast besides judging whether the correct AMV phase is predicted is quantifying the absolute difference between observed and predicted AMV anomaly. This can be accomplished using a root-mean-square error (RMSE), for example. In Figure 3e, we show the RMSEs of individual hindcasts as a function of the strength of OHT_{50N} at their initialization. We use absolute values for OHT_{50N}, so we do not differenciate between particularly strong and particularly weak OHT_{50N}.

Anomalous OHT_{50N} lead to generally lower RMSE than average OHT_{50N}. Unlike in previous analyses, however, the difference between anomalous and average OHT_{50N} is in this case not significant. This indicates that there is still room for improvement in the models, or that the link between OHT_{50N} and North Atlantic SST only acts on the sign of AMV, not on absolute values. RMSE does, however, show a tendency to lower skill in average compared to anomalous phases of OHT_{50N}. As this tendency is in line with the argumentation we bring forward in this paper, we consider this finding encouraging for future research.

6. Discussion

We illustrate in this study that decadal hindcast skill of North Atlantic SST is strongly influenced by subpolar ocean heat transport in the North Atlantic. As a result, there is unused potential in time-averaged hindcast skill estimates. We also show that assessing the credibility of individual hindcasts shows the same dependence on the initial conditions as integrated hindcast skill, illustrating that our findings have a high importance for the estimation of the credibility of individual decadal climate forecasts. Some of our findings, however, warrant discussion.

In this study we analyze low-frequency variability in the North Atlantic ocean. The time series of 110 years of length we use here covers approximately two full oscillations of AMV. This may not be sufficient to draw general conclusions from this study. On the other hand, the length of time series in this study is about twice that of current studies with initialized climate models (e.g., Marotzke et al., 2016; Yeager & Robson, 2017),

and it covers the majority of observational record that is currently available at monthly temporal resolution (e.g., Müller et al., 2015).Thus, the results presented here are based on the current state of the art. However, we recommend a repetition of our experiments with longer time series when they become available.

In the absence of long observational time series of OHT_{50N}, it cannot be guaranteed that the connection between OHT_{50N} phases and SST variability in the HadISST is physical. This physical connection can, however, be shown for the assimilation run (Borchert et al., 2018). Imperfect representations of North Atlantic physics in the model therefore explain generally lower hindcast skill when evaluating against HadISST than when using the assimilation run as a reference. It is in principle possible that the phenomenon we describe and explain in this paper stems from a systematic bias of the model simulation from observations in the North Atlantic. Moreover, Kröger et al. (2017) showed that nudging techniques similar to the one applied in this study can lead to unrealistic representation of North Atlantic climate, particularly ocean overturning which might influence OHT. We approach this potential caveat by showing results from the assimilation run alongside those from observations, as the assimilation run represents a hybrid between observations and the free model and would also be impacted by possible problems with the initialization procedure. The issues outlined in Kröger et al. (2017) may not be as prominently present in this study because we use a model-consistent ocean estimate for nudging (as also discussed in Borchert et al., 2018). Because we find qualitatively indistinguishable evolutions of SST patterns in the assimilation run and the observation data set, we conclude that the results presented in this paper result from the described physical mechanism and not from a systematic model bias.

Further, some methodological constraints need to be discussed. Although previous studies suggested that the prediction system used produces reasonable climate variability (e.g., Borchert et al., 2018; Müller et al., 2014), the findings presented here are based on only one climate model and will have to be replicated with other predictions systems to ultimately assess their robustness. The high hindcast skill we find against observations prior to 1930 (cf. Figure 1) suggests that the relative sparsity of SST observations pre-1960 does not systematically influence the skill of hindcasts. The time-dependent ACC estimates presented in this paper are based on individually detrended time periods of different lengths. From the analyses presented here, it cannot be ruled out that detrending relatively short time periods removes some predictable signal that is found due to autocorrelation in longer evaluation time periods. Without detrending, however, the time-dependent ACC estimates are not substantially different from the ACC estimates presented here (not shown). This indicates that the predictable signal in AMV variability was in the twentieth century larger than the predictable signal from the warming trend. The same is valid for using subpolar AMOC as a criterion to divide hindcast skill estimates: The results are qualitatively indistinguishable from those using OHT. This indicates that both can be used interchangeably and warrants an applicability of the results presented here to similar previous findings using AMOC (e.g., Matei et al., 2012; Persechino et al., 2013; Robson et al., 2012, 2014).

Previous work showed that using AMOC or OHT variability at a subpolar latitude as a precursor to North Atlantic SST variability maximizes the time lag between the OHT and SST anomaly because of the slow southward propagation of OHT phases in the North Atlantic (Zhang & Zhang, 2015; Borchert et al., 2018). As a consequence, our results are insensitive to the exact choice of latitude for OHT variability, as long as subpolar OHT is chosen. Possible latitudes range between 45° and 55°N (see Figure 2b in Borchert et al., 2018, using the same model simulation). Our results are also insensitive to the precise definition of the OHT threshold for anomalous phases, and to the chosen lead year. The latter is because subpolar OHT influences AMV at all subdecadal lead times in the assimilation run (not shown). We therefore argue that, despite some methodological shortcomings of this study, the results and conlusions presented in this paper are robust.

Our findings also address the common assumption in hindcast studies that the robustness of hindcast skill estimates increases with increasing the time window for which skill is evaluated (e.g., Müller et al., 2014). Our findings show that this assumption generally holds for the statistical robustness of hindcast skill. However, results presented here also illustrate that a hindcast skill estimate calculated for a long time period can overestimate or underestimate the skill of individual hindcasts within that period. With increasing length of evaluation window of hindcasts, this error increases. As a result, short hindcast time series can outperform long hindcast time series when estimating the credibility of an individual climate forecast. To harvest the full potential of short time series, however, they need to be combined with a physical mechanism.

7. Conclusion

We provide in this study several lines of evidence that subpolar OHT variability is an important source of decadal SST prediction skill in the North Atlantic region. Hindcast skill for North Atlantic SST can change significantly over time, with ACC taking values between 0 and 0.8. We show that the difference in SST hindcast skill is connected to average and anomalous phases of subpolar OHT and highlight the importance of our findings analyzing the recently shown time dependency of decadal SST hindcast skill. Using individual hindcasts, we then demonstrate how our findings can be used to constrain hindcast skill estimates to judge the credibility of a single decadal climate forecast.

This study demonstrates that conventional hindcast skill estimates can fail to provide a suitable estimate of the credibility of a single forecast when the forecast is started. Instead, these estimates overestimate or underestimate skill when being applied to estimate forecast credibility. We show that by taking the physical state of the climate system at the start of a climate forecast into account, conventional hindcast skill estimates can be constrained such that the credibility of a forecast is reflected more appropriately. For decadal North Atlantic SST predictions, a good physical variable to indicate the skill of a forecast at its start is subpolar ocean heat transport. In other regions, for other target variables, and on other time scales, other physical mechanisms are likely to apply (see, e.g., Neddermann et al., 2019; O'Reilly et al., 2017; Weisheimer et al., 2017).

We highlight that skill estimates derived from hindcast analysis need to be based on individual years in order to be attached to individual forecasts. New skill measures going beyond the very simple approach we present in this paper will therefore have to be developed in the future and combined with conventional hindcast skill estimates and physical mechanisms to truly estimate the credibility of individual forecasts.

Acknowledgments
This research was supported by the International Max Planck Research School on Earth System Modelling, Hamburg (L. F. B.); the German Ministry of Education and Research under MiKlip FlexForDec (Grant 01LP1519A; L. F. B., W. A. M.) and MiKlip AODA-PENG (Grant 01LP1516A; S. B., J. B.); and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2037 "Climate, Climatic Change, and Society"—Project: 390683824, contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg (J. B., A. D.). The model simulations were performed at the German Climate Computing Centre. The model output used in this study can be accessed here (http://catalogue.ceda.ac.uk/uuid/72fcd8b56e6d4e468a80cfa01d645d20). Data and scripts used in this study are available from *publications@mpimet.mpg.de*. The authors thank three anonymous reviewers as well as Jürgen Kröger and Dirk Olonscheck for helpful comments on this paper.

References

Årthun, M., Eldevik, T., Viste, E., Drange, H., Furevik, T., Johnson, H. L., & Keenlyside, N. S. (2017). Skillful prediction of northern climate provided by the ocean. *Nature Communications*, *8*, 15875. https://doi.org/10.1038/ncomms15875

Ba, J., Keenlyside, N. S., Latif, M., Park, W., Ding, H., Lohmann, K., et al. (2014). A multi-model comparison of Atlantic multidecadal variability. *Climate Dynamics*, *43*, 2333–2348. https://doi.org/10.1007/s00382-014-2056-1

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., et al. (2016). The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*, 3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Borchert, L. F., Müller, W. A., & Baehr, J. (2018). Atlantic Ocean heat transport influences interannual–to–decadal surface temperature predictability in the North Atlantic Region. *Journal of Climate*, *31*, 6763–6782. https://doi.org/10.1175/JCLI-D-17-0734.1

Brune, S., Düsterhus, A., Pohlmann, H., Müller, W. A., & Baehr, J. (2018). Time dependency of the prediction skill for the North Atlantic subpolar gyre in initialized decadal hindcasts. *Climate Dynamics*, *51*, 1947–1970. https://doi.org/10.1007/s00382-017-3991-4

Clement, A., Bellomo, K., Murohy, L. N., Cane, M. A., Mauritsen, T., Rädel, G., & Stevens, B. (2015). The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science*, *16*, 320–324. https://doi.org/10.1126/science.aab3980

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., et al. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*, 1–28. https://doi.org/10.1002/qj.776

Delworth, T. L., Zeng, F., Zhang, L., Zhang, R., Vecchi, G. A., & Yang, X. (2017). The central role of ocean dynamics in connecting the North Atlantic Oscillation to the extratropical component of the Atlantic Multidecadal Oscillation. *Journal of Climate*, *30*, 3789–3805. https://doi.org/10.1175/JCLI-D-16-0358.1

Dijkstra, H. A., te Raa, L., Schmeits, M., & Gerrits, J. (2006). On the physics of the Atlantic Multidecadal Oscillation. *Ocean Dynamics*, *56*, 36–50. https://doi.org/10.1007/s10236-005-0043-0

Gastineau, G., & Frankignoul, C. (2015). Influence of the North Atlantic SST variability on the atmospheric Circulation during the Twentieth Century. *Journal of Climate*, *28*, 1396–1416. https://doi.org/10.1175/JCLI-D-14-00424.1

Jayne, S. R., & Marotzke, J. (2001). The dynamics of ocean heat transport variability. *Reviews of Geophysics*, *39*, 385–411. https://doi.org/10.1029/2000RG000084

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.) Oxford: Wiley Ltd.

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., et al. (2013). Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI–Earth System Model. *Journal of Advances in Modeling Earth Systems*, *5*, 422–446. https://doi.org/10.1002/jame.20023

Kröger, J., Pohlmann, H., Sienz, F., Marotzke, J., Baehr, J., Köhl, A., et al. (2017). Full-field initialized decadal predictions with the MPI earth system model: An initial shock in the North Atlantic. *Climate Dynamics*, *51*, 2593–2608. https://doi.org/10.1007/s00382-017-4030-1

Marotzke, J., Müller, W. A., Vamborg, F. S. E, Becker, P., Cubasch, U., Feldmann, H., et al. (2016). MiKlip: A national research project on decadal climate prediction. *Bulletin of the American Meteorological Society*, *97*(12), 2379–2394. https://doi.org/10.1175/BAMS-D-15-00184.1

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

Modali, K. (2017). SPECS - MPI-ESM-LR model output prepared for SPECS decadal (1901–2015). Centre for Environmental Data Analysis, 2017. http://catalogue.ceda.ac.uk/uuid/72fcd8b56e6d4e468a80cfa01d645d20

Müller, W. A., Matei, D., Bersch, M., Jungclaus, J. H., Haak, H., Lohmann, K., et al. (2015). A twentieth-century reanalysis forced ocean model to reconstruct the North Atlantic climate variation during the 1920s. *Climate Dynamics*, *44*, 1935–1955. https://doi.org/10.1007/s00382-014-2267-5

Müller, W. A., Pohlmann, H., Sienz, F., & Smith, D. (2014). Decadal climate predictions for the period 1901–2010 with a coupled climate model. *Geophysical Research Letters*, *41*, 2100–2107. https://doi.org/10.1002/2014GL059259

Neddermann, N.-C., Müller, W. A., Dobrynin, M., Düsterhus, A., & Baehr, J. (2019). Seasonal predictability of European summer climate re-assessed. *Climate Dynamics*, *53*, 3039–3056. https://doi.org/10.1007/s00382-019-04678-4

O'Reilly, C. H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T., Schaller, N., & Woollings, T. (2017). Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, *44*, 5729–5738. https://doi.org/10.1002/2017GL073736

Persechino, A., Mignot, J., Swingedouw, D., Labetoulle, S., & Guilyardi, E. (2013). Decadal predictability of the Atlantic Meridional Overturning Circulation and climate in the IPSL-CM5A-LR Model. *Climate Dynamics*, *40*, 2359–2380. https://doi.org/10.1007/s00382-012-1466-1

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Robson, J. I., Sutton, R. T., & Smith, D. M. (2012). Initialized decadal predictions of the rapid warming of the North Atlantic Ocean in the mid-1990s. *Geophysical Research Letters*, *39*, L19713. https://doi.org/10.1029/2012GL053370

Robson, J. I., Sutton, R. T., & Smith, D. M. (2014). Decadal predictions of the cooling and freshening of the North Atlantic in the 1960s and the role of ocean circulation. *Climate Dynamics*, *42*, 2353–2365. https://doi.org/10.1007/s00382-014-2115-7

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Cruger, T., Rast, S., et al. (2013). Atmospheric component of the MPI–M Earth System Model ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*, 146–172. https://doi.org/10.1002/jame.20015

Sutton, R. T., & Hodson, D. L. (2005). Atlantic Ocean forcing of North American and European summer climate. *Science*, *309*, 115–118. https://doi.org/10.1126/science.1109496

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A., & Palmer, T. (2017). Atmospheric seasonal forecasts of the twentieth century: Multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their Potential Value for Extreme Event Attribution. *Quarterly Journal of the Royal Meteorological Society*, *143*, 917–926. https://doi.org/10.1002/qj.2976

Wilks, D. (2016). The stippling shows statistically significant grid points—How research results are routinely overstated and overinterpreted and what to do about it. *Bulletin of the American Meteorological Society*, *97*, 2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1

Yeager, S., Karspeck, A., Danabasoglu, G., Tribbia, J., & Teng, H. (2012). A decadal predictions case study: Late twentieth-century North Atlantic Ocean Heat Content. *Journal of Climate*, *25*, 5173–5189. https://doi.org/10.1175/JCLI-D-11-00595.1

Yeager, S. G., & Robson, J. I. (2017). Recent progress in understanding and predicting Atlantic Decadal climate variability. *Current Climate Change Reports*, *3*, 112–127. https://doi.org/10.1007/s40641-017-0064-z

Zhang, R. (2008). Coherent surface–subsurface fingerprint of the Atlantic Meridional Overturning Circulation. *Geophysical Research Letters*, *35*, L20705. https://doi.org/10.1029/2008GL035463

Zhang, R., Sutton, R., Danabasoglu, G., Delworth, T. L., Kim, W. M., Robson, J. I., & Yeager, S. G. (2016). Comment on The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science*, *24*, 1527. https://doi.org/10.1126/science.aaf1660

Zhang, J., & Zhang, R. (2015). On the evolution of Atlantic Meridional Overturning fingerprint and implications for decadal predictability in the North Atlantic. *Geophysical Research Letters*, *42*, 5419–5426. https://doi.org/10.1002/2015GL064596