

Geophysical Research Letters



RESEARCH LETTER

10.1029/2020GL091307

Key Points:

- North Atlantic sea surface temperature (SST) is better predicted in CMIP6 than in CMIP5 in both initialized hindcasts and noninitialized historical simulations
- High correlation of North Atlantic SST in CMIP6 historical simulations with observations indicates a prominent role for forcing after 1980.
- Fifty-five percent of the total observed post-1980 annual North Atlantic SST variance is explained by simulations with natural forcing only

Supporting Information:

- Supporting Information S1

Correspondence to:






L. F. Borchert,
leonard.borchert@locean.ipsl.fr

Citation:

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., & Mignot, J. (2021). Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophysical Research Letters*, *48*, e2020GL091307. <https://doi.org/10.1029/2020GL091307>

Received 3 JUN 2020
 Accepted 25 NOV 2020

Improved Decadal Predictions of North Atlantic Subpolar Gyre SST in CMIP6

Leonard F. Borchert¹ , Matthew B. Menary¹ , Didier Swingedouw² , Giovanni Sgubin², Leon Hermanson³ , and Juliette Mignot¹ 

¹LOcean Laboratory, Institut Pierre Simon Laplace (IPSL), Sorbonne Universités (SU/CNRS/IRD/MNHN), Paris, France, ²EPOC, Université de Bordeaux, Pessac, France, ³Met Office Hadley Centre, Exeter, UK

Abstract Due to its wide-ranging impacts, predicting decadal variations of sea surface temperature (SST) in the subpolar North Atlantic remains a key goal of climate science. Here, we compare the representation of observed subpolar SST variations since 1960 in initialized and uninitialized historical simulations from the 5th and 6th phases of the Coupled Model Intercomparison Project (CMIP5/6). Initialized decadal hindcasts from CMIP6 explain 88% of observed SST variance post-1980 in the subpolar gyre at lead years 5–7 (77% in uninitialized simulations) compared to 42% (8%) in CMIP5, indicating a more prominent role for forcing in driving observed subpolar SST changes than previously thought. Analysis of single-forcing experiments suggests much of this correlation is due to natural forcing, explaining ~55% of the observed variance. The amplitude of observed subpolar SST variations is underestimated in historical simulations and improved by initialization in CMIP6, indicating continued value of initialization for predicting North Atlantic SST.

Plain Language Summary Sea surface temperature (SST) fluctuations in the North Atlantic region are known to influence climate around the globe. Comparing retrospective predictions of North Atlantic SST with observations, we show that the most state-of-the-art climate models have improved in predicting North Atlantic SST for up to 10 years ahead compared to the previous generation of climate models. This recent improvement can be traced back to particularly well-predicted variations of North Atlantic SST after 1980. During this time, reactions to large volcanic eruptions and changes in solar activity, as well as inherent unforced variations, play an important role for the predictability of North Atlantic SST. Here, not only direct radiative forcing changes play a role, but there is also a dynamical response of the ocean that influences the final climate response. This study inspires hope that current climate models will show improved capability in predicting North Atlantic SST changes up to a decade ahead, particularly following large volcanic eruptions, but also otherwise.

1. Introduction

Sea surface temperature (SST) in the North Atlantic region has been shown to influence climate both locally and remotely, particularly temperature over Europe (Gastineau & Frankignoul, 2015; Sutton & Hodson, 2005) and northeast Asia (Monerie et al., 2018), the West African Monsoon (Dunstone et al., 2011), or the probability of occurrence of extremely warm summers in the Northern Hemisphere (Borchert et al., 2019a). Understanding and predicting North Atlantic SST variations is therefore an interesting and important scientific challenge. In this paper, we explore the advances that have been made in the prediction of decadal variations of North Atlantic SST from the Coupled Model Intercomparison Project (CMIP) phase 5 (Taylor et al., 2012) to 6 (Eyring et al., 2016).

Climate prediction several years ahead, so-called *decadal* climate prediction, has been a prominent topic in climate research for more than a decade (Marotzke et al., 2016; Pohlmann et al., 2005; D. Smith et al., 2007). Studies on such predictions commonly utilize coupled global climate models, simulating predictions of known past climate (*re-forecasts* or *hindcasts*) to examine the capability of these predictions to reproduce observed climate variations (their so-called *skill*). *Initialized* hindcasts (HCs) (which include information of the observed past climate) were mostly found to show improved decadal HC skill compared to uninitialized historical simulations (that only rely on external forcing of the climate system) (e.g., Befort et al., 2020; Boer et al., 2016; Brune & Baehr, 2020; Marotzke et al., 2016; Mignot et al., 2016; D. M. Smith et al., 2019).

© 2020. The Authors.
 This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

In CMIP5, skill improvement through initialization was particularly high in the North Atlantic subpolar gyre (SPG) region (Matei et al., 2012; S. G. Yeager & Robson, 2017). Predictions of decadal SPG SST variations were found to improve because initialization synchronized modeled fluctuations in the Atlantic Meridional Overturning Circulation (AMOC) and associated ocean heat transport with observations (e.g., Borchert et al., 2018; Robson et al., 2014; Swingedouw et al., 2013; S. Yeager et al., 2012; J. Zhang & Zhang, 2015). Moreover, North Atlantic SST variations were recently found to be caused by anthropogenic aerosol forcing except in the subpolar region (M. Watanabe & Tatebe, 2019). Such findings were commonly interpreted to illustrate that decadal SPG SST variations were primarily a product of internal climate variability. This notion, however, has been challenged by recent work arguing that multiannual SPG SST changes can also arise from external forcing (Haustein et al., 2019; Swingedouw et al., 2015). As such, it appears apposite to revisit the prediction skill of North Atlantic SPG SSTs in updated simulations for CMIP6.

Here, we examine decadal predictions of North Atlantic SST in a multimodel ensemble of 30 CMIP5 and 28 CMIP6 models across initialized and noninitialized simulations over their common time period since 1960.

2. Models and Methods

Our analysis is based on simulations from a total of 58 global climate models (Table S1). We use a set of 6 initialized decadal prediction ensembles from CMIP5 (henceforth HC5), and 7 initialized decadal prediction ensembles from CMIP6 (henceforth HC6). In addition, historical ensembles with 30 models from CMIP5 (henceforth HIST5) and 28 models from CMIP6 (henceforth HIST6) are considered. Finally, to study the contribution of individual forcings to observed climate variations, we also examine simulations from nine models contributing to the Detection and Attribution Model Intercomparison Project (DAMIP; Gillett et al., 2016) of CMIP6 (similar to a method employed by Bellucci et al. [2017] using CMIP5 models). Unless otherwise noted, the multimodel mean of the annual mean individual model ensemble means is considered (one-model-one-vote). An alternative choice whereby all ensemble members for all models are weighted equally (i.e., one-member-one-vote) has little impact on the results (Figure 2b).

The historical (HIST5/6) simulations rely on external forcing to simulate observed climate (Eyring et al., 2016). DAMIP simulations are run with individual sets of external forcings over the historical period (1850–2014) while all others are kept constant, enabling a separate analysis of the impact of different forcings on climatic variability. The individual forcings considered in this study are greenhouse gas emissions (hist-GHG), anthropogenic aerosols (hist-aer), and natural forcings, that is volcanic and solar forcing (hist-nat). HC simulations are initialized from initial conditions including observations to improve their capability to reproduce internal climate variability (e.g., Brune & Baehr, 2020; Doblas-Reyes et al., 2013). All HC simulations considered here are run for 10 years, initialized every year from 1960 to 2005 (CMIP5) or 2014 (CMIP6). As in the historical simulations, all HC simulations include the real-world evolution of external forcings whereas a true future forecast would only be able to include projected estimates of these forcings. This approach allows for a more precise isolation of the effect of initialization but may also result in an overestimation of the decadal HC skill, in particular due to short-term volcanic forcing (Hermanson et al., 2020).

SST observations are obtained from the Hadley Centre Ice and Sea Surface Temperature (HadISST) data set (Rayner et al., 2003). SST from all models and observations were remapped to a regular $1^\circ \times 1^\circ$ grid prior to analysis. We define the North Atlantic SPG region as 50°W – 10°W , 45°N – 60°N (cf. Figures 1e–1h). Area averages are performed as area-weighted averages over the remapped data. We apply no detrending to the considered time series. We focus here on the fifth to seventh year after initialization (lead years 5–7) which reflects a balance between decadal-scale lead-time and moderate time-averaging, but our results are not particularly sensitive to the choice of this lead-time or the length of the averaging window considered (not shown). To match the 3-year mean applied in the HCs, all other multimodel ensemble means and observations are smoothed with a 3-year running mean. As the first decadal HCs in DCP are started in 1960 (i.e., the first year described by the HCs is 1961) and we utilize lead years 5–7, our analysis begins in 1965–1967. The end points for our analyses are defined by the end of reconstructed forcing in CMIP experiments (2005 for CMIP5 and 2014 for CMIP6) so as to not contaminate our analysis with a projected forcing. We thus analyze the time frame 1965–2014 for CMIP6 and 1965–2005 for CMIP5, assessed as anomalies against the

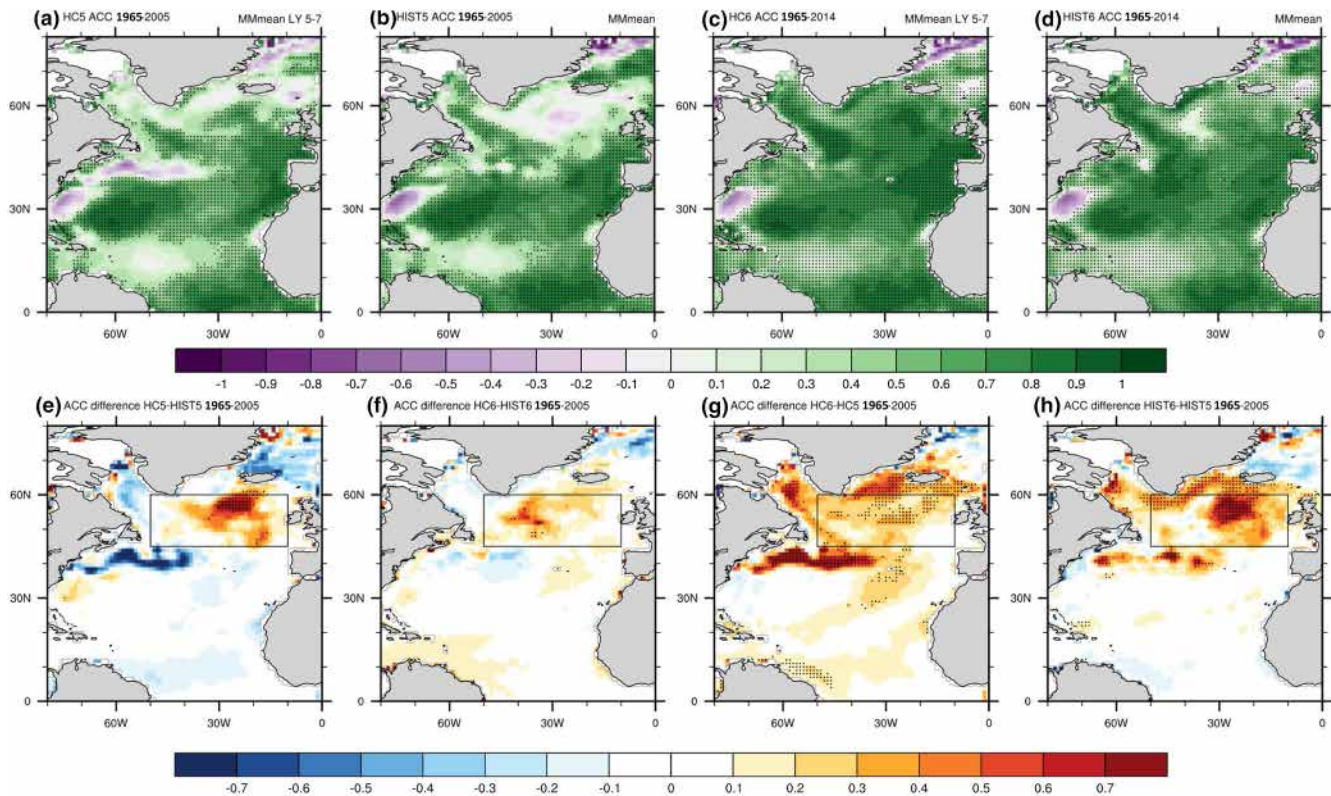


Figure 1. Multimodel ensemble mean decadal prediction skill (anomaly correlation coefficient; ACC) for annual mean nondetrended SST. (a) Skill of CMIP5 initialized decadal hindcasts at lead year 5–7 for the period 1965–2005, compared to skill in the multimodel ensemble mean of (b) CMIP5 historical simulations (1965–2005), (c) CMIP6 initialized hindcasts (1965–2014), and (d) CMIP6 historical simulations (1965–2014). The historical ensemble means are based on the same model subset as the HC5 and HC6 means, which were selected based on availability of simulations in HC5 (6 models) and HC6 (7 models) (Table S1). Skill differences are shown for the common period 1965–2005 between (e) HC5 and HIST5, (f) HC6 and HIST6, (g) HC6 and HC5, and (h) HIST6 and HIST5. Stippling shows where correlation or correlation differences are significantly different from zero (95% confidence, see Section 2). The box outlined in black in (e)–(h) shows the area used to calculate the SPG index.

average over the CMIP5 period in the individual model ensemble means (which equates to a lead-time-dependent mean bias correction).

We assess the skill of the model simulations against HadISST using a suite of skill metrics, namely the Pearson correlation coefficient (or anomaly correlation coefficient, ACC; Jolliffe & Stephenson, 2003), residual ACC score (D. M. Smith et al., 2019) and the Mean Square Skill Score (MSSS; e.g., D. M. Smith et al., 2020). Formulas used to calculate these skill scores are given in the supporting information. The ACC measures whether a linear relationship exists between modeled and observed anomalies, while MSSS measures the absolute difference between modeled and observed anomalies. We use observed climatology over the period 1965–2005 as a benchmark for our MSSS calculation. The residual ACC aims to measure skill in the internally generated signal. This is achieved by removing the forced component, estimated as the historical multimodel ensemble mean, from both HC and observations prior to calculation of the correlation (D. M. Smith et al., 2019). All indices indicate perfect agreement between model and observations when they take the value 1, and lower skill at lower values. We use ACC^2 to analyze the observed variance explained by the different model simulations. Statistical significance of prediction skill estimates is assessed using a Monte-Carlo process of 1,000 bootstraps with replacement on the time-dimension using blocks of 3 years to take the low-pass filtering into account (Jolliffe & Stephenson, 2003).

We also assess the possible effect of larger ensemble sizes in CMIP6 than in CMIP5 on our results. We resample the CMIP6 ensembles based on individual ensemble members 10,000 times with replacement, decreasing the CMIP6 ensemble size to the respective CMIP5 ensemble size, and calculate ACC during each iteration. This results in distributions of possible skill estimates from CMIP6 models using CMIP5

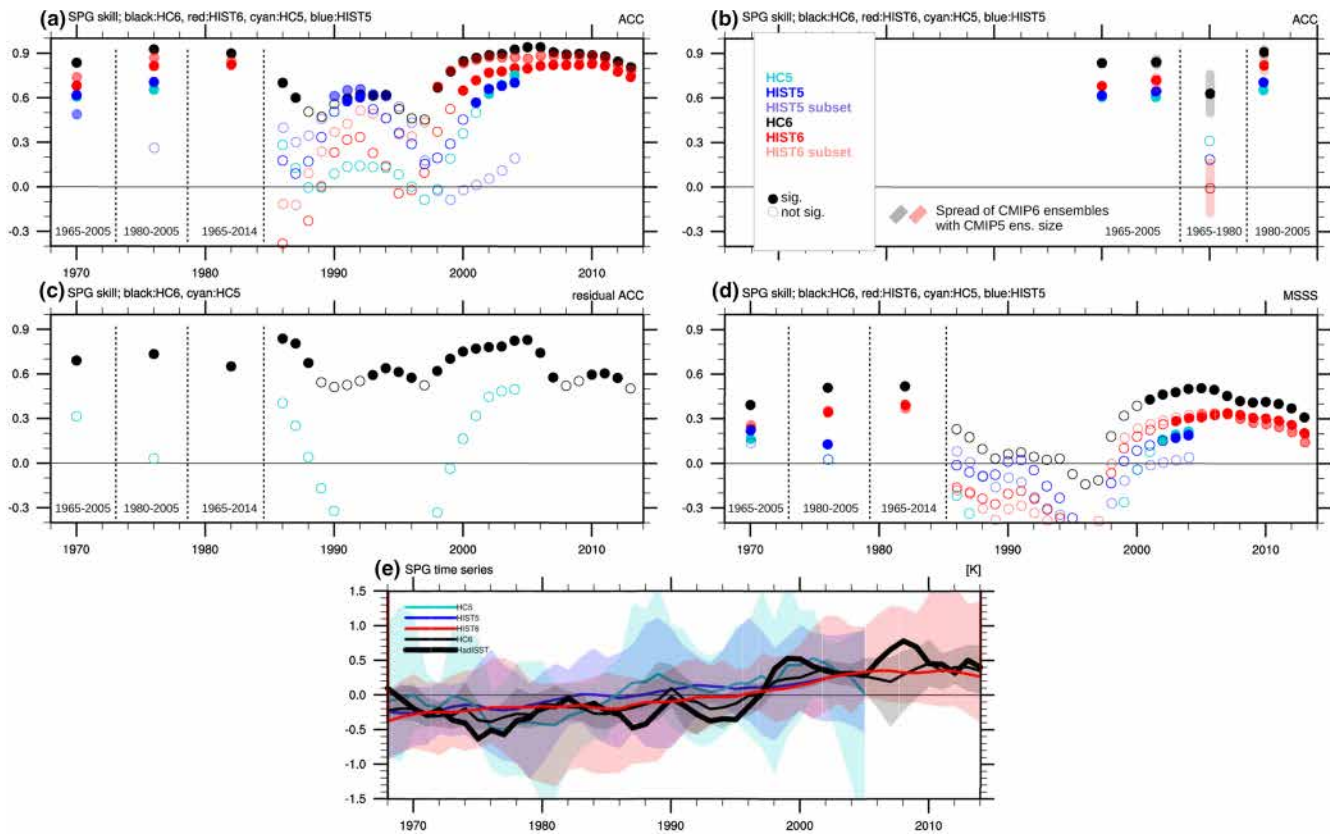


Figure 2. Multimodel mean time-dependent prediction skill at lead years 5–7 and nondetrended time series of the SPG SST index. (a) ACC for a rolling 20-year window (full/empty circles) in HC5 (cyan), HIST5 (blue), HC6 (black), and HIST6 (red), positioned over the last year of the respective 20-year period. Circles/dots in lighter shades show skill for the model subset used in Figure 1. Circles/dots on the left show skill for the full CMIP5 and CMIP6 periods as well as the period since 1980, specified above. Full circles indicate significant skill, empty circles indicate that skill is not significantly different from 0 (95% confidence, see Section 2). (b) ACC for the (from left to right) CMIP5 period using the one-model-one-vote hierarchy and (twice) the one-member-one-vote hierarchy, as well as the pre-1980 and post-1980 period using the one-member-one-vote hierarchy. Shading illustrates the likely range around CMIP6 skill if the corresponding CMIP5 ensemble sizes were used. (c) and (d) same as (a), but evaluating skill using the residual ACC and MSSS skill metrics, respectively (see Section 2). Note that calculating residual ACC involves subtracting the historical simulation time series from the initialized one, eliminating the need to illustrate historical skill which equals 0 by design. (e) Time series of SPG SST anomalies in observations (thick black), HC5 (cyan), HIST5 (blue), HC6 (black), and HIST6 (red). Shading shows the spread of the single model ensemble means represented in the multimodel ensemble mean.

ensemble sizes, while allowing for individual models to be excluded altogether. Two standard deviations around the mean of these distributions are defined as the likely range of skill found in CMIP6 models with CMIP5 ensemble sizes. If the actual skill diagnosed in CMIP5 models lies outside that range, it is defined as significantly different from the skill diagnosed in CMIP6.

We use time-dependent HC skill (Borchert et al., 2019b; Brune et al., 2018) over a rolling 20-year time window across the time series to highlight time periods of particularly high or low skill (so-called *windows of opportunity*; Mariotti et al., 2020), and to attribute skill in these periods to particular forcings.

3. Decadal Prediction Skill for SST in the North Atlantic Region

We first analyze ACC skill for North Atlantic SST in multimodel ensembles based on the 6 (7) models for which yearly initialized simulations for HC5 (HC6) were available, along with the corresponding historical simulations of the same models in HIST5 (HIST6). North Atlantic SST simulated by CMIP5 models generally correlates less well with observations than that simulated by CMIP6 models (Figures 1a–1d). While the HC5 ensemble mean shows generally higher correlation with the observations than the corresponding multimodel mean of HIST5 (Figures 1a and b), the same comparison of the ensemble means of HC6 and HIST6 yields less clear results (Figures 1c and 1d). This indicates reduced improvement through initialization

in CMIP6 compared to CMIP5 for the ACC metric. The North Atlantic SPG region is emblematic for this result: for the CMIP5 period until 2005, CMIP5 models show particularly strong improvement of ACC through model initialization for SST there (Figure 1e), while CMIP6 models display much smaller and only marginally significant SPG SST ACC improvement (Figure 1f).

These findings do not mean a degradation of skill in the initialized predictions from CMIP5 to CMIP6. In fact, HC6 predicts SST variations in the Labrador Sea, Gulf Stream region, and around the northern edge of the SPG significantly better than HC5 (Figure 1g). In the SPG region, both HC5 and HC6 show high skill, with improvements toward CMIP6. The decreased improvement of SPG SST skill through initialization in CMIP6 compared to CMIP5 mainly originates from an improved capability of HIST6 to reproduce SPG SST changes compared to HIST5 (Figure 1h). Consequently, new models indicate that forced SST changes (represented by the HIST experiments) might have played a larger role in North Atlantic SST variations since the 1960s than previously thought.

4. SST Skill in the Subpolar Gyre Region

We now focus on the skill increase from CMIP5 to CMIP6 in North Atlantic SPG SST. For predictions of SPG SST for the CMIP5 period (1965–2005, Figure 2a), the HC5 (6 models; 37% observed variance explained), HIST5 (30 models; 37%), HC6 (7 models; 72%), and HIST6 (28 models; 47%) multimodel ensembles, as well as the HIST5 and HIST6 model subset used in the previous analysis (6 and 7 models, respectively) (Table S1), show significant ACC skill. Available simulations show higher ACC for the CMIP6 period (1965–2014), with larger amounts of observed variance explained (HIST6: 65%, HC6: 81%).

Initialized and historical simulations from both model generations capture climate variations particularly well after 1980 (Figure 2a). This is also reflected in the explained variances at 50% for HIST5, 42% for HC5, 65% for HIST6, and 88% for HC6. While the 6-model HIST5 subset shows much lower skill after 1980 than the corresponding HC5, the full available 30-model HIST5 ensemble shows high correlation to observations. The strong increase of correlation from the 6-model HIST5 subset to the full 28-model HIST5 ensemble points to a relevant effect of ensemble size in our findings.

The increase in ensemble size from CMIP5 to CMIP6 partly explains the skill increase from HIST5 to HIST6 in the CMIP5 period (Figure 2b), indicating that only part of the skill increase is attributable to forcing-related improvements from CMIP5 to CMIP6 over the entire CMIP5 period. Both HC6 and HIST6 simulations show high ensemble size-related uncertainty in skill estimates pre-1980. ACC for HIST5 falls within the spread of HIST6 ACC pre-1980, indicating insignificant difference between the two. Conversely, ACC in both CMIP6 ensembles stands out significantly from that in CMIP5 post-1980 (Figure 2b). The significantly improved correlation in HIST6 over HIST5 in the post-1980 period is therefore robust for equally sized ensembles and thus cannot only be explained by ensemble size increase. Meanwhile, model initialization robustly improves ACC for SPG SST in the 1960s and 1970s and after 1980, as well as for the full CMIP5 and CMIP6 periods (Figure 2a). This skill improvement is highest in CMIP5 post-1980, taking the average skill from statistical insignificance to significance. As noted above, lower skill improvement through initialization in CMIP6 than in CMIP5 originates from the generally very high skill in the HIST6 ensemble, which is found to be particularly high after 1980. Applying a Fisher transform to the correlation values, we find that the skill increase from initialization is similarly high between CMIP5 and CMIP6 for the CMIP5 period, while the correlation increase through initialization in the post-1980 period is much lower in CMIP6 than in CMIP5. This emphasizes the particularly high correlation of simulated SPG SST in HIST6 to observations.

Analyzing other skill metrics for SPG SST highlights the continued importance of model initialization for decadal prediction skill in CMIP6, even after 1980. Residual ACC for HC6 is comparable to the full ACC, indicating a strong signal beyond the forced component that can be reproduced in initialized decadal predictions (Figure 2c), which is far better than in HC5, indicating a potential improvement in the initialization schemes of HC6, based on this metric. Initialization also appears to be important in capturing the full amplitude of the SPG SST signal, as highlighted by the MSSS improvement for both HC5 and HC6 (compared to the corresponding subset of HIST5 and HIST6; Figure 2d), albeit a stronger improvement is found in CMIP6. Both initialized and noninitialized CMIP6 simulations show improved capability to capture the full amplitude of decadal SPG SST changes over CMIP5 simulations (Figure 2d), so the MSSS is generally

significant in CMIP6 simulations. Finally, in contrast with CMIP5, initialization strongly decreases the intermodel spread in SPG SST in CMIP6 (HC6: 0.7 K vs. HIST6: 1.4 K; HC5: 1.5 K vs. HIST5: 1.6 K), making predictions more robust (i.e., less sensitive to the effects of individual models, Figure 2e). In HC6, intermodel spread is particularly low, which indicates exceptionally robust decadal North Atlantic SST predictions in CMIP6 among different models. This effect is particularly strong in, but not entirely dominated by, the prominent shift in North Atlantic SPG SST between 1995 and 1999 (observed: 1 K, HC6: 0.5 K, HC5: 0.2 K, HIST6: 0.2 K, HIST5: 0.1 K) (e.g., S. Yeager et al., 2012).

5. Contributions of Different Forcings to Decadal SPG SST Variations

The improved capability of HIST6 (compared to HIST5) to capture SPG SST variations since 1980 in the full ensembles indicates an improved response of CMIP6 models to forcing during that time. Using a 9-model ensemble of the DAMIP simulations (see Section 2), we attempt to disentangle the contributions of different forcings to the SPG SST signal, investigating whether the high skill in HIST6 after 1980 can be attributed to the response to an individual forcing. For consistency, we here use a subset of HIST6 that only contains those 9 models that provide DAMIP simulations (Table S1).

A decomposition of the different forcing contributions to the total ACC in HIST6 can only explain the full historical signal if the different forcings combined result in similar skill as HIST6. The 9-model HIST6 ensemble subset we use in this analysis shows no pronounced ACC differences to the full CMIP6 historical ensemble (Figure 3a; compare to Figure 2a). A linear sum of the multimodel ensemble mean anomalies for the individual forcing simulations (hist-aer, hist-GHG, and hist-nat) shows both comparable full-period and time-dependent skill as HIST6 (Figure 3a). The time series of the combined forcings also resembles the HIST6 multimodel ensemble mean (Figure 3b). The forcing decomposition of SPG SST variations using DAMIP experiments is thus approximately linear and appropriate for our purpose.

Correlations of the simulations with isolated hist-aer, hist-GHG, and natural (hist-nat) forcings to observations indicate the degree to which the individual forcings explain observed variability. We find a number of somewhat consistent signals among models. Several of the single models agree that anthropogenic aerosols only explain an insignificant amount of observed SPG SST annual variability until very recently (Figure 3c, 3d); greenhouse gas forcing explains a significant amount of SPG SST variations during a very short time window around the 1990s (Figure 3e, 3f); and natural forcing explains much of the SPG SST variations observed since the 1980s, but not before that (Figure 3g, 3h). Due to the substantial amount of model spread found in this analysis, however, these findings are sensitive to the set of models used in the analysis. After 1980, hist-aer, hist-GHG, and hist-nat explain 0% (two standard deviation spread of resampling the individual models 10,000 times with replacement: -6 to 10%), 16% (-1 to 35%), and 55% (34%–59%) of the observed SST variance, respectively. These findings implicate natural forcing over the other examined forcings as an important driver of the high skill in HIST6 after 1980.

The results from our analysis of DAMIP simulations illustrate that recently observed variations of SPG SST do not exclusively originate from internal variability, but are also strongly related to response to natural forcing. In the past, changes in SPG SST have been associated with AMOC variations, leading to prediction skill (e.g., Borchert et al., 2018; S. G. Yeager & Robson, 2017; R. Zhang, 2008). For the 6-model set of hist-nat simulations used by Menary et al. (2020), we find a strong lagged relationship between AMOC at 35°N and SPG SST (Figure S1) when AMOC leads by about 4 years (details on this analysis can be found in Text S2, supporting information). The conceptual model of harmonic AMOC response to major volcanic eruptions developed in Swingedouw et al. (2015) well represents decadal AMOC variability in hist-nat (Figure S1), indicating that volcanic eruptions contribute strongly to the simulated natural forcing-related SPG SST variations from the hist-nat simulations (as also suggested by Hermanson et al. [2020]) through a dynamical mechanism involving ocean circulation. It is therefore possible that a mechanism of major volcanic eruptions leading to a strengthening of the AMOC that subsequently warms the SPG (Swingedouw et al., 2015) explains a relevant portion of the observed post-1980 SPG SST variations as a response to the eruptions of Mt. Agung in 1963, El Chichon in 1982, and Mt. Pinatubo in 1991 (Santer et al., 2016). Note that the dynamical adjustment timescale of Atlantic circulation can be up to 15 years according to Mignot et al. (2011) and Swingedouw et al. (2015), thereby emphasizing the possible role of early eruptions such as Mt. Agung,

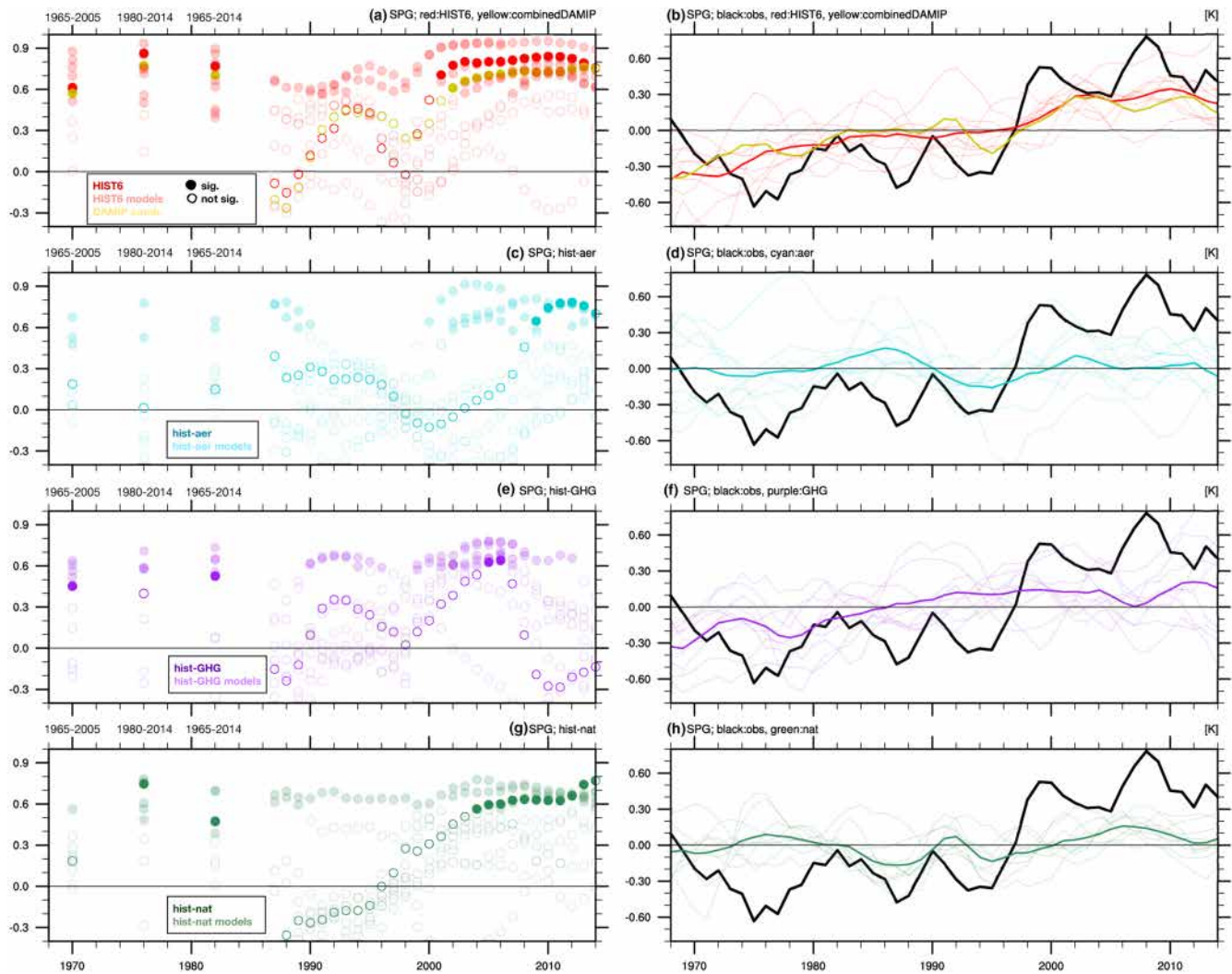


Figure 3. SPG SST ACC skill and time series in a 9-model multimodel ensemble for which DAMIP simulations are available (see Table S1) from (a), (b) HIST6 (red), and for simulations from the DAMIP scenarios (c), (d) hist-aer (cyan), (e), (f) hist-GHG (purple), and (g), (h) hist-nat (green). The linear sum of the multimodel ensemble means of hist-aer, hist-GHG, and hist-nat is shown in (a), (b) in yellow. (a, c, e, g) ACC for a rolling 20-year window (dots/circles), positioned over the last year of the respective 20-year period. ACC is examined for the multimodel mean (solid colors) and the individual models (shading). Circles/dots on the left show skill for the full CMIP5 and CMIP6 periods as well as the period since 1980, specified above. Full circles indicate significant skill, empty circles indicate that skill is not significantly different from 0 (95% confidence, see Section 2). (b, d, f, h) Time series of SPG SST anomalies in observations (thick black) as well as the multimodel ensemble mean (solid colors) and the individual models (shading).

as well as possible interferences, either constructive (El Chichon) or destructive (Pinatubo, cf. Figure S1) between the eruption responses as shown in Swingedouw et al. (2015; their Figure 9).

A strong combined influence of internal variability and, since 1980, volcanic forcing on SPG SST variations can also explain why we find only a surprisingly limited influence of GHG emissions on SPG SST in recent years; the GHG signal is masked by a strong influence of other factors on SPG SST variations. The limited impact of GHG emissions on SPG SST is in line with findings presented by Bellucci et al. (2017), based on a DAMIP-like partitioning of forcing contributions to North Atlantic SST in CMIP5 models.

6. Discussion

We demonstrate an improvement of decadal North Atlantic SPG SST HC skill across different metrics in initialized HCs in CMIP6 compared to CMIP5. For the commonly used correlation-based ACC metric we only find limited improvement of SPG SST prediction skill through initialization in CMIP6 after 1980. This

can be attributed to particularly skillful CMIP6 historical simulations. These HIST6 simulations show much higher ACC for SPG SST than the corresponding simulations from CMIP5 post-1980, which indicates a particularly strong role for forcing in modulating SPG SST during this time. Using DAMIP simulations, we show that the high skill after 1980 in the HIST6 simulations is likely a result of their accurate response to natural forcing.

The central point of this study is that a variety of improvements from CMIP5 to CMIP6 act together to enhance the representation of decadal-scale North Atlantic SPG SST variations in initialized and noninitialized simulations in CMIP6. Possible structural improvements in CMIP6 that could have led to this improved prediction of North Atlantic SST include larger ensembles, enhanced horizontal and vertical model resolution (e.g., Drews & Greatbatch, 2016), better parametrization schemes, interactive aerosols (Menary et al., 2020), more realistic forcing, or, for the initialized HCs, more sophisticated initialization methods in CMIP6. The latter is likely to be one cause for the improved SPG SST skill in HC6, as initialized HCs from CMIP5 predict SPG SST variations much less skillfully than those from CMIP6 (cf. Figure 2). The influence of improved forcing can be illustrated by looking at the CESM1.1 decadal HC simulations used in this study as part of HC6, as they are based on a CMIP6 model that was run with CMIP5 forcing (S. G. Yeager et al., 2018). Since these simulations show similarly high skill as other CMIP6 HCs (not shown), the increased skill in CMIP6 appears to originate from the updated models' improved capability to produce the correct response to forcing of SPG SST rather than improvement in the forcing terms themselves. This illustration is, however, only based on one model and thus lacks robustness. Ensemble size explains a part of the improvement of SPG SST representation in CMIP6 compared to CMIP5, but this effect is not particularly strong in the interesting period after 1980, where ensemble size differences do not explain a significant portion of the skill increase from CMIP5 to CMIP6 (cf. Figure 2b). The most promising candidate to explain the improvement in historical simulations from CMIP5 to CMIP6 in representing North Atlantic SST variations since the 1980s is therefore improved response to forcing in CMIP6, owing to model improvements. Future studies should certainly explore the contributions of individual improvements from CMIP5 to CMIP6 on SPG SST representation, as well as highlight models that perform particularly well.

HC skill is particularly high in our analysis in CMIP6 initialized HC, related to large ensemble size, model improvements, and careful HC initialization. In particular, initialization is found to improve prediction skill whenever comparatively small and short signals are to be predicted (cf. Figure 2). Due to the signal-to-noise problem in initialized predictions (e.g., D. M. Smith et al., 2020), however, initialized HCs still underestimate the magnitude of decadal SPG SST variations, and therefore require careful treatment to extract the predictable signal.

In our analysis, we find similar time-dependent ACC as previous studies on decadal predictions of North Atlantic SST (Borchert et al., 2019b; Brune et al., 2018). This indicates that these previously found windows of opportunity for the period 1960-today are robust beyond the MPI-ESM-LR model suite that was employed in both previous studies. As windows of opportunity are here shown to largely coincide between CMIP6 initialized HCs and historical simulations (cf. Figure 2a), while windows of opportunity in residual ACC that explicitly excludes the forced component do not show strong variations (cf. Figure 2c), the reason for changes from low-to-high skill appears to lie in the forcing. Nonetheless, there are two components to the high skill post-1980: a predictable forced signal indicated by high skill in HIST6, and a predictable internal signal indicated by high skill in residual ACC. For times of opposing trends of forced component and internal variability like the late 1960s and early 1970s (although barely covered by our analysis because the first HCs are started in 1960), we find high skill in initialized HCs and residual ACC, but not in the historical runs (cf. Figure 2). Understanding the timing of different forcings and internal variability is therefore crucial to understand windows of opportunity for decadal HCs (as also argued in S. Yeager et al., 2012).

7. Conclusions

We analyze a multimodel ensemble of initialized decadal HCs from CMIP5 and CMIP6 to show the increased capability of CMIP6 models to predict decadal variations in SPG SST. CMIP6 HCs explain up to 88% of observed North Atlantic SST variance at lead years 5–7, whereas CMIP5 HCs only explain 42% after 1980. This improvement can be traced back to a good representation of SPG SST in CMIP6 historical simulations,

explaining up to 65% of observed variance after the 1980s, compared to their CMIP5 equivalents which explain only 50% of observed variance. This points to a strong role of forcing in SPG SST variations during that time period.

Fifty-five percent (two standard deviation spread of resampling the individual models 10,000 times with replacement: 34%–59%) of observed SPG SST variance after 1980 in CMIP6 can be attributed to natural forcing. Anthropogenic aerosol and greenhouse gas forcing only explain 0% (–6 to 10%) and 16% (–1 to 35%) of observed SST variance, respectively. The main cause for the good representation of observed SPG SST variations in CMIP6 historical simulations after 1980 is an accurate response to natural forcing. We suggest this response originates from an AMOC-related lagged response to volcanic eruptions, with a possible contribution of solar forcing. Both increased ensemble size in CMIP6 and high prediction skill after the end of the CMIP5 period are shown to be a factor for the improved representation of SPG SST in CMIP6 compared to CMIP5 as well. While CMIP6 historical simulations are shown to be sufficient to reproduce the sign of decadal SPG SST anomalies, model initialization remains important to predict their full amplitude and limit the uncertainty of predictions. The significant predictable signal beyond forcing (i.e., residual ACC) emphasizes that initialized model simulations continue to be powerful tools to explore climate predictability and perform predictions.

Data Availability Statement

Both CMIP5 and CMIP6 outputs were downloaded from the ESGF-IPSL node [https://esgf-node.ipsl.upmc.fr/projects/esgf-ipsl/](https://esgf-node.ipsl.upmc.fr/projects/esgf-ips/). HadISST data were obtained from <https://www.metoffice.gov.uk/hadobs/hadisst/> and are © British Crown Copyright, Met Office, 2020, provided under a Non-Commercial Government Licence <https://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/version/2/>.

Acknowledgments

The authors acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5 and CMIP6. The authors thank the climate modeling groups (listed in Table S1) for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. L. F. Borchert, D. Swingedouw, G. Sgubin, and J. Mignot were supported by the EUCP project funded by the European Union's Horizon 2020 programme, Grant Agreement number 776613. M. B. Menary was supported by the EPICE project funded by the European Union's Horizon 2020 programme, Grant Agreement number 789445. L. Hermanson was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. The authors are grateful toward two anonymous reviewers as well as Sebastian Brune, Brady Ferster and Reinel Sospedra-Alfonso for discussions on data and the manuscript.

References

- Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2020). Constraining projections using decadal predictions. *Geophysical Research Letters*, *47*, e2020GL087900. <https://doi.org/10.1029/2020GL087900>
- Bellucci, A., Mariotti, A., & Gualdi, S. (2017). The role of forcings in the twentieth-century North Atlantic multidecadal variability: The 1940–75 North Atlantic cooling case study. *Journal of Climate*, *30*, 7317–7337. <https://doi.org/10.1175/JCLI-D-16-0301.1>
- 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., Düsterhus, A., Brune, S., Müller, W. A., & Baehr, J. (2019b). Forecast-oriented assessment of decadal hindcast skill for North Atlantic SST. *Geophysical Research Letters*, *46*, 11444–11454. <https://doi.org/10.1029/2019GL084758>
- 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*(17), 6763–6782. <https://doi.org/10.1175/JCLI-D-17-0734.1>
- Borchert, L. F., Pohlmann, H., Baehr, J., Neddermann, N., Suarez-Gutierrez, L., & Müller, W. A. (2019a). Decadal predictions of the probability of occurrence for warm summer temperature extremes. *Geophysical Research Letters*, *46*, 14042–14051. <https://doi.org/10.1029/2019GL085385>
- Brune, S., & Baehr, J. (2020). Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions. *WIREs Climate Change*, *11*, e637. <https://doi.org/10.1002/wcc.637>
- 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*(5–6), 1947–1970. <http://doi.org/10.1007/s00382-017-3991-4>
- Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., Garcia-Serrano, J., Guemas, V., Kimoto, M., et al. (2013). Initialized near-term regional climate change prediction. *Nature Communications*, *4*(1). <http://dx.doi.org/10.1038/ncomms2704>
- Drews, A., & Greatbatch, R. J. (2016). Atlantic multidecadal variability in a model with an improved North Atlantic current. *Geophysical Research Letters*, *43*, 8199–8206. <https://doi.org/10.1002/2016GL069815>
- Dunstone, N. J., Smith, D. M., & Eade, R. (2011). Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. *Geophysical Research Letters*, *38*, L14701. <https://doi.org/10.1029/2011GL047949>
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*, 1937–1958. <https://doi.org/10.5194/gmd-9-1937-2016>
- Gastineau, G., & Frankignoul, C. (2015). Influence of the North Atlantic SST variability on the atmospheric circulation during the twentieth century. *Journal of Climate*, *28*(4), 1396–1416. <https://doi.org/10.1175/JCLI-D-14-00424.1>
- Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6. *Geoscientific Model Development*, *9*, 3685–3697. <https://doi.org/10.5194/gmd-9-3685-2016>
- Hausein, K., Otto, F. E. L., Venema, V., Jacobs, P., Cowtan, K., Hausfather, Z., et al. (2019). A Limited Role for Unforced Internal Variability in Twentieth-Century Warming. *Journal of Climate*, *32*(16), 4893–4917. <http://dx.doi.org/10.1175/jcli-d-18-0555.1>
- Hermanson, L., Bilbao, R., Dunstone, N., Ménégoz, M., Ortega, P., Pohlmann, H., et al. (2020). Robust multiyear climate impacts of volcanic eruptions in decadal prediction systems. *Journal of Geophysical Research: Atmosphere*, *125*, e2019JD031739. <https://doi.org/10.1029/2019JD031739>

- Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast verification. A practitioner's guide in atmospheric science*. Hoboken, NJ: John Wiley & Sons Ltd.
- Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., et al. (2020). Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. *Bulletin of the American Meteorological Society*, 101(5), E608–E625. <http://dx.doi.org/10.1175/bams-d-18-0326.1>
- Marotzke, J., Müller, W. A., Vamborg, F., 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(24), 8502–8523. <https://doi.org/10.1175/JCLI-D-11-00633.1>
- Menary, M. B., Robson, J., Allan, R. P., Booth, B. B. B., Cassou, C., Gastineau, G., et al. (2020). Aerosol-forced AMOC changes in CMIP6 historical simulations. *Geophysical Research Letters*, 47, e2020GL088166. <https://doi.org/10.1029/2020GL088166>
- Mignot, J., Garcia-Serrano, J., Swingedouw, D., Germe, A., Nguyen, S., Ortega, P., et al. (2016). Decadal prediction skill in the ocean with surface nudging in the IPSL-CM5A-LR climate model. *Climate Dynamics*, 47(3-4), 1225–1246. <http://dx.doi.org/10.1007/s00382-015-2898-1>
- Mignot, J., Khodri, M., Frankignoul, C., & Servonnat, J. (2011). Volcanic impact on the Atlantic Ocean over the last millennium. *Climate of the Past*, 7, 1439–1455. <https://doi.org/10.5194/cp-7-1439-2011>
- Monerie, P.-A., Robson, J., Dong, B., & Dunstone, N. (2018). A role of the Atlantic Ocean in predicting summer surface air temperature over North East Asia? *Climate Dynamics*, 51(1-2), 473–491. <http://doi.org/10.1007/s00382-017-3935-z>
- Pohlmann, H., Botzet, M., Latif, M., Roesch, A., Wild, M., & Tschuck, P. (2005). Estimating the decadal predictability of a coupled AOGCM. *Journal of Climate*, 17(22), 4463–4472.
- 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, 4407. <https://doi.org/10.1029/2002JD002670>
- Robson, J., Sutton, R., & Smith, D. (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>
- Santer, B., Solomon, S., Ridley, D., Fyfe, J., Beltran, F., Bonfils, C., et al. (2016). Volcanic effects on climate. *Nature Climate Change*, 6(1), 3–4. <http://dx.doi.org/10.1038/nclimate2859>
- Smith, D., CusackColman, S. A. W., Folland, C. K., Harris, G. R., & Murphy, J. M. (2007). Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317(5839), 796–799. <https://doi.org/10.1126/science.1139540>
- Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., et al. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, 2(1). <http://dx.doi.org/10.1038/s41612-019-0071-y>
- Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., et al. (2020). North Atlantic climate far more predictable than models imply. *Nature*, 583(7818), 796–800. <http://dx.doi.org/10.1038/s41586-020-2525-0>
- Sutton, R. T., & Hodson, D. L. R. (2005). Atlantic ocean forcing of North American and European summer climate. *Science*, 309(5731), 115–118. <https://doi.org/10.1126/science.1109496>
- Swingedouw, D., Mignot, J., Labetoule, S., Guilyardi, E., & Madec, G. (2013). Initialization and predictability of the AMOC over the last 50 years in a climate model. *Climate Dynamics*, 40, 2381–2399. <https://doi.org/10.1007/s00382-012-1516-8>
- Swingedouw, D., Ortega, P., Mignot, J., Guilyardi, E., Masson-Delmotte, V., Butler, P. G., et al. (2015). Bidecadal North Atlantic ocean circulation variability controlled by timing of volcanic eruptions. *Nature Communications*, 6(1). <http://dx.doi.org/10.1038/ncomms7545>
- Taylor, K. E., Stouffer, R., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4), 485–498. <https://doi.org/10.1175/BAMS-D-11-00094.1>
- Watanabe, M., & Tatebe, H. (2019). Reconciling roles of sulfate aerosol forcing and internal variability in Atlantic multidecadal climate changes. *Climate Dynamics*, 53, 4651–4665. <https://doi.org/10.1007/s00382-019-04811-3>
- Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (2018). Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the community Earth system model. *Bulletin of the American Meteorological Society*, 99, 1867–1886. <https://doi.org/10.1175/BAMS-D-17-0098.1>
- Yeager, S., Karspeck, A., Danabasoglu, G., Tribbia, J., & Teng, H. (2012). A decadal prediction case study: Late twentieth-century North Atlantic ocean heat content. *Journal of Climate*, 25, 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, J., & Zhang, R. (2015). On the evolution of Atlantic meridional overturning circulation fingerprint and implications for decadal predictability in the North Atlantic. *Geophysical Research Letters*, 42, 5419–5426. <https://doi.org/10.1002/2015GL064596>