

Geophysical Research Letters



RESEARCH LETTER

10.1029/2021GL093258

Key Points:

- Model structural differences cause 2/3 of spread in North Atlantic Oscillation (NAO) projections for 2080–2099 and internal variability 1/3
- The NAO explains a large part of the spread (>40% locally) in North Atlantic mean sea level pressure projections due to internal variability
- At least 15 ensemble members are needed to detect a forced NAO response of 1 hPa, typical of the changes modeled by 2080–2099 under RCP8.5

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. M. McKenna,
C.McKenna1@leeds.ac.uk

Citation:

McKenna, C. M., & Maycock, A. C. (2021). Sources of uncertainty in Multimodel Large Ensemble projections of the winter North Atlantic Oscillation. *Geophysical Research Letters*, 48, e2021GL093258. <https://doi.org/10.1029/2021GL093258>

Received 5 MAR 2021

Accepted 3 JUL 2021

Sources of Uncertainty in Multimodel Large Ensemble Projections of the Winter North Atlantic Oscillation

C. M. McKenna¹  and A. C. Maycock¹ 

¹School of Earth and Environment, University of Leeds, Leeds, UK

Abstract Projections of the winter North Atlantic circulation exhibit large spread. Coupled Model Intercomparison Project archives typically provide only a few ensemble members per model, rendering it difficult to quantify reducible model structural uncertainty and irreducible internal variability (IV) in projections. We estimate using the Multimodel Large Ensemble Archive that model structural differences explain two-thirds of the spread in late 21st century (2080–2099) projections of the winter North Atlantic Oscillation (NAO). This estimate is biased by systematic model errors in the forced NAO response and IV. Across the North Atlantic, the NAO explains a substantial fraction of the spread in mean sea level pressure (MSLP) projections due to IV, except in the central North Atlantic. Conversely, the spread in North Atlantic MSLP projections associated with model differences is largely unexplained by the NAO. Therefore, improving understanding of the NAO alone may not constrain the reducible uncertainty in North Atlantic MSLP projections.

Plain Language Summary Variations in atmospheric circulation over the North Atlantic in winter are dominated by the North Atlantic Oscillation (NAO) pattern, which has a strong influence on regional climate and is often associated with severe weather events. It is uncertain how the NAO will respond to future changes in climate driven by human activity. This uncertainty in future projections has two main sources, which are yet to be fully quantified: first, there are large natural variations in the NAO on the timescale of many decades, which can mask the effect of long-term climate change on the NAO; second, different climate models have different representations of physical processes, which can lead to differences in the future climates they simulate. Here we estimate, using an unprecedented number of simulations from different climate models, that model structural differences explain the majority of uncertainty in late 21st century NAO projections. This result is important because it suggests that uncertainty in NAO projections could be reduced with improved knowledge of the physical processes involved. However, the NAO itself does not explain much of the model structural uncertainty in regional sea level pressure projections in and around the North Atlantic basin, suggesting other dynamical processes must be understood.

1. Introduction

The North Atlantic atmospheric circulation has a strong influence on Northern Hemisphere regional climate and is often associated with severe weather events (Buehler et al., 2011; Hurrell et al., 2003). For a given future greenhouse gas and aerosol forcing scenario, previous studies have found substantial spread in projections of late 21st century North Atlantic circulation change across models from the Coupled Model Intercomparison Project Phases 5 and 6 (CMIP5 and CMIP6, Collins et al., 2013; Oudar et al., 2020; Shepherd, 2014; Zappa et al., 2018). The model spread partly arises from competing large-scale drivers, such as upper and lower tropospheric temperature gradient changes (Harvey et al., 2014) and stratospheric circulation (Manzini et al., 2014; Simpson, Hitchcock, et al., 2018), with the relative dominance of each factor differing across models (Zappa & Shepherd, 2017).

The extent to which the spread in multimodel projections of the North Atlantic circulation is from model structural differences versus internal climate variability (IV) remains an open question. This is partly because models contributing to CMIP5/6 typically only provide a small number of realizations with different initial conditions and the same external forcing to sample IV. This makes it difficult to quantify the contributions of model structural uncertainty and IV to the spread in projections without making assumptions (e.g., approximating 21st century IV using IV in a stationary preindustrial climate; Hawkins & Sutton, 2009).

© 2021. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

This study aims to advance understanding of the roles of model structural error and IV in North Atlantic circulation projections. To achieve this, we use the recently available Multimodel Large Ensemble Archive (MMLEA, Deser et al., 2020) and data from CMIP5/6. We focus on the leading mode of variability in the North Atlantic circulation, the North Atlantic Oscillation (NAO), which is associated with changes in the strength and latitude of the eddy-driven jet (Woollings et al., 2010). To guide our investigation, we address the following questions:

1. What are the relative contributions of IV and model structural uncertainty to spread in NAO projections?
2. When do the forced NAO response and model differences in this response emerge from IV in the 21st century?
3. What is the minimum number of ensemble members required to separate the forced NAO response, and model differences in this response, from IV?
4. To what extent is spread in North Atlantic circulation projections explained by the NAO?

Addressing these questions will aid the interpretation of North Atlantic circulation projections, improving their utility, and provide guidance for designing future model experiments.

2. Methods

2.1. Datasets

The MMLEA contains large (16–100 members) initial-condition ensembles for seven comprehensive climate models (Table S1, Hazeleger et al., 2010; Jeffrey et al., 2013; Kay et al., 2015; Kirchmeier-Young et al., 2017; Maher et al., 2019; Rodgers et al., 2015; Schlunegger et al., 2019; Sun et al., 2018). We use historical and Representative Concentration Pathway (RCP) 8.5 simulations from the MMLEA models for the common period 1950–2099. RCP8.5 was chosen because only a small subset of the models is available for other RCPs. Since GFDL-ESM2M and GFDL-CM3 have similar atmosphere, ocean, sea-ice, and land components (Maher et al., 2021), and give similar results, we discard the smaller GFDL-CM3 ensemble from the MMLEA analysis. The winter North Atlantic circulation is described using monthly mean sea level pressure (MSLP) data averaged over December to February (DJF). Following Collins et al. (2013), the long-term climate response is computed as the 20-year epoch difference between a future period and a near-present-day period (updated to 1995–2014; year is for January).

We also use historical and RCP8.5 simulations from 39 CMIP5 models (Taylor et al., 2012), and historical and Shared Socioeconomic Pathway (SSP) 5–8.5 simulations from 36 CMIP6 models (Eyring et al., 2016, Table S2). The forcing scenarios changed in CMIP6, where SSP5-8.5 has the most similar total end-of-century radiative forcing to RCP8.5 (Meinshausen et al., 2020). However, there are differences in the mix of forcings between the RCP and SSP scenarios (Meinshausen et al., 2011, 2020) to be borne in mind when comparing results.

Generally, only a few ensemble members are available for the CMIP5/6 simulations, so we estimate IV using the preindustrial control (piControl) runs. Model drift is eliminated by subtracting each run's long-term linear trend. Various observation-based datasets are used to evaluate the spread in model projections against observed IV. Since multidecadal timescales are our focus, we use two centennial-scale reanalysis datasets: the NOAA-CIRES-DOE 20th Century Reanalysis version 3 (20CRv3; Compo et al., 2011; Slivinski et al., 2019) and the ECMWF 20th Century Reanalysis (ERA20C, Poli et al., 2016). A 1000 member “Observational Large Ensemble” (Obs LE; McKinnon & Deser, 2018) is also used, which contains synthetic historical trajectories produced by a statistical model based on observed climate statistics. We use the full extent of Obs LE (1921–2014), and the longer common period of 1900–2010 for 20CRv3 and ERA20C to minimize sampling issues. Forced trends in 20CRv3 and ERA20C are estimated and removed using linear least squares regression; Obs LE by construction has no forced MSLP trend (McKinnon & Deser, 2018).

All model and observation-based data were bilinearly interpolated onto a common 2° horizontal grid; this procedure does not alter our results.

2.2. NAO Definition

Following Stephenson et al. (2006) and Baker et al. (2018), the NAO index is defined as the difference in area-averaged MSLP between a southern box (90°W–60°E, 20°N–55°N) and a northern box (90°W–60°E, 55°N–90°N) in the North Atlantic. This index is less sensitive to differences in centers of action between observations and models than the station-based index (Hurrell et al., 2003; Stephenson et al., 2006) and is also less variable enabling easier detection of a forced NAO response from IV. Furthermore, it is less affected by issues of interpretability that occur when using a mathematically constructed empirical orthogonal function (EOF)-based index (Ambaum et al., 2001; Dommenges & Latif, 2002; Stephenson et al., 2006).

Each MMLEA model's historical NAO pattern (Figure S1) is constructed from the regression slopes obtained by regressing historical (1951–2014) timeseries of DJF MSLP at each grid-point onto the NAO index timeseries. Note that using a future period gives similar results and that all timeseries are first linearly detrended. The pattern is defined separately for each ensemble member and then the ensemble mean is calculated (Simpson et al., 2020). The NAO-congruent part of an MSLP anomaly map is obtained by multiplying the historical NAO pattern by the NAO index anomaly. Figure S1 also shows observation-based and CMIP5/6 multimodel mean (MMM) historical NAO patterns; largely, the modeled and observation-based patterns are highly correlated.

2.3. Statistical Methods

In each MMLEA model, uncertainty due to IV is mainly estimated as the standard deviation across ensemble members (Deser et al., 2012). The externally forced response is estimated using the ensemble mean. The percentage variance contribution of IV ($\%U_{IV}$) and of model structural differences ($\%U_{MD}$) to the total uncertainty in MMLEA projections is quantified following Maher et al. (2021; Text S1).

A forced response is described as “robust” if it is statistically detectable from IV at the 95% confidence level. Two-sided confidence intervals for a forced response (μ) are calculated as $\mu \pm t\sigma / \sqrt{N}$ (von Storch & Zwiers, 1999). t is the Student's t -distribution value for $p = 0.025$ and $N - 1$ degrees of freedom, σ is the intermember standard deviation of the epoch difference, and N is the ensemble size.

To estimate the minimum ensemble size (N_{\min}) required to detect a robust forced NAO index response of a given magnitude (X) between any two 20-year epochs, we follow Screen et al. (2014) and rearrange a two-sided Student's t -test for a difference of means (Text S2):

$$N_{\min} = 2t_c^2 \times (\sigma / X)^2.$$

t_c is for $p = 0.025$ and $2N_{\min} - 2$ degrees of freedom, and σ is the standard deviation of 20-year epoch means due to IV. N_{\min} is calculated for a difference in forced response (X) where σ is for differences in 20-year means.

3. Results

Figure 1 shows winter NAO index anomalies between 2080–2099 and 1995–2014 in the CMIP6, CMIP5, and MMLEA models. For both CMIP5/6 ensembles, the MMM response in the NAO index is ~ 1.5 hPa. However, the MMM responses are generally small compared to the spread across the individual models. While some models have large positive NAO anomalies exceeding their modeled range of IV, most modeled anomalies are smaller than IV. The range of NAO anomalies is 6 hPa in CMIP6 and 7 hPa in CMIP5, comparable to the observed range of NAO variability (Figure 1, gray box), where 86% and 79% of models agree on sign respectively.

Given many CMIP5/6 models only have one ensemble member available, it is impossible to separate the spread in projections into parts due to structural model differences and IV. Despite this limitation, uncertainty in projections is often examined using these models (e.g., Hawkins & Sutton, 2009). The MMLEA models suggest there are indeed substantial intermodel differences in the forced response of up to 5 hPa (Figure 1, colored circles). Using Maher et al. (2021)'s uncertainty decomposition, we find that model

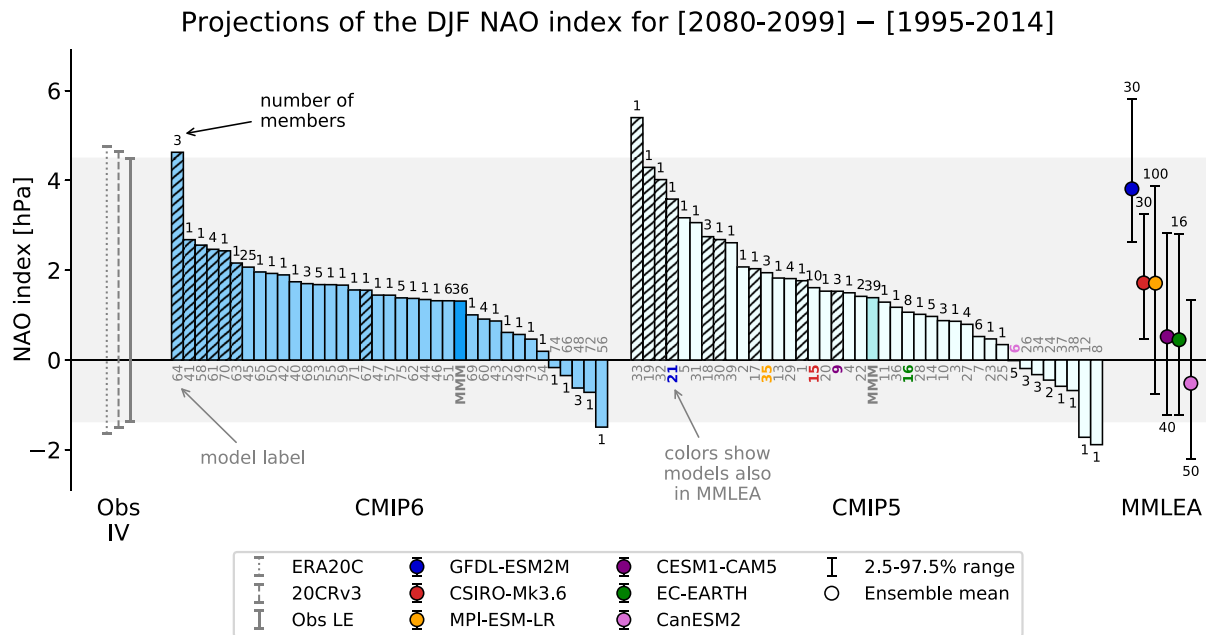


Figure 1. Projections of the December to February (DJF) North Atlantic Oscillation (NAO) index for (2080–2099) – (1995–2014) in the CMIP6, CMIP5, and Multimodel Large Ensemble Archive (MMLEA) models. For CMIP5/6 models, ensemble means are shown if more than one ensemble member is available. Hatching indicates a CMIP5/6 model anomaly that is larger than 2 standard deviations of model-specific internal variability (IV) (Text S3). Darker cyan/blue bars indicate the multimodel mean (MMM). Whiskers for MMLEA models indicate the 2.5%–97.5% range of responses across the ensemble members. Section 2.1 describes the model forcing scenarios. Gray lines show the 2.5%–97.5% range of 10^3 differences in 20-year epoch means of different observation-based records (Section 2.1), selected by randomly resampling with replacement. Gray shaded box shows this range for Obs LE. The observation-based IV estimates are shifted by the CMIP6 MMM anomaly for comparison with the inter-model spread.

structural differences and IV contribute to 66% and 34%, respectively, of the total uncertainty in MMLEA NAO projections. The following sections examine each source of uncertainty in detail.

3.1. Uncertainty From Internal Variability

In several MMLEA models, the forced winter NAO response is smaller than IV as measured by the ensemble spread (Figure 1). Using the ensemble spread to assess the range of possible futures assumes that the models adequately represent observed NAO variability. However, as in previous studies (Bracegirdle et al., 2018; Kim et al., 2018; Kravtsov, 2017; Simpson, Deser, et al., 2018; Wang et al., 2017), we find that most CMIP5/6 and MMLEA models underestimate low frequency NAO variability compared to observation-based datasets (Figure 1, black whiskers vs. gray lines; Tables S1 and S2). The model projections may therefore be overconfident: that is, a larger part of the uncertainty in the future real-world NAO response may be from IV. When model-based estimates of IV are adjusted to an observation-based estimate (Text S1), IV and model structural differences each contribute to half of the total uncertainty in the adjusted MMLEA projections. These estimates also depend on the models simulating a realistic forced NAO response; Section 4 discusses this further.

Now, we ask to what extent the NAO explains uncertainty in North Atlantic circulation projections due to IV. Figure 2 presents for each MMLEA model a decomposition of the total ensemble spread in MSLP (top row) into an NAO-congruent part (second row) and a residual (third row). The total uncertainty from IV is generally largest at high northern latitudes, extending from Greenland to Northern Europe, as well as in the central North Atlantic. There is also larger uncertainty from IV in north-eastern North America and continental Europe. The NAO contributes to a large proportion (>50%; Figure 2, bottom row) of the uncertainty in MSLP projections at high latitudes, and a substantial proportion (up to 50%) of the uncertainty around the Mediterranean region. The large residual uncertainty in projections in the central Atlantic and western Europe is largely associated with the East Atlantic (EA) pattern (Figure S2), the second dominant mode of

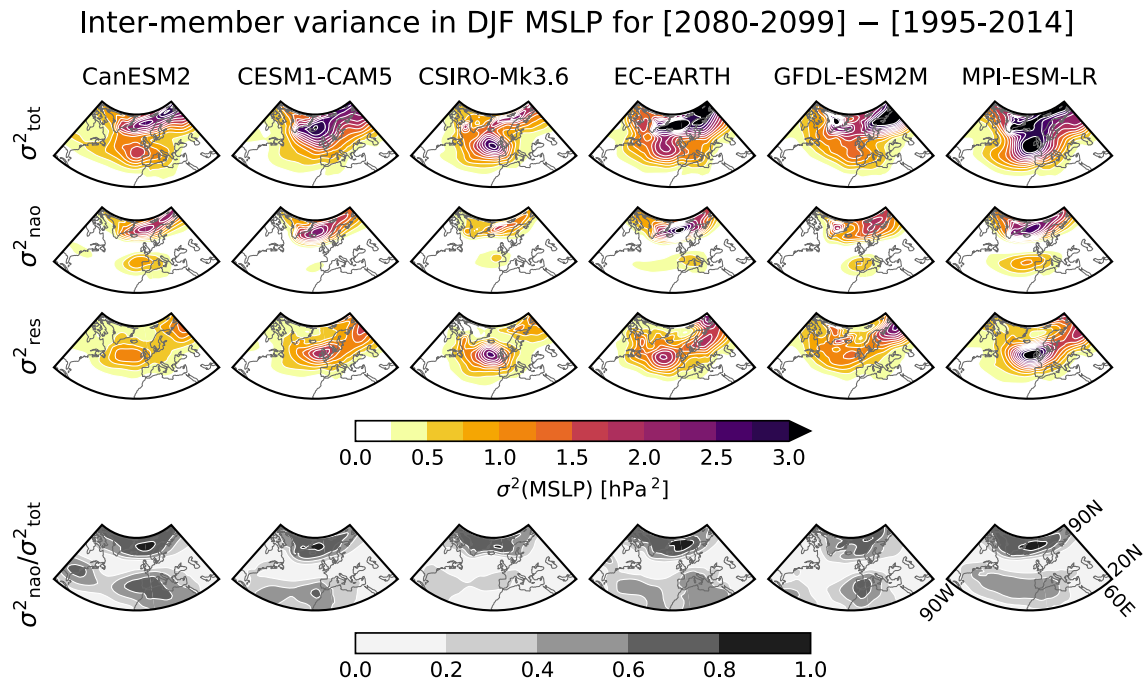


Figure 2. Inter-member variance in projections of December to February (DJF) mean sea level pressure (MSLP) for (2080–2099) – (1995–2014) for each Multimodel Large Ensemble Archive (MMLEA) model. (Top row) Total (σ^2_{tot}); (Second row) North Atlantic Oscillation (NAO)-congruent part (σ^2_{nao}); (Third row) Residual (σ^2_{res}); (Bottom row) Proportion of total variance explained by NAO. σ^2_{nao} is obtained by regressing the total intermember spread in MSLP on the spread in NAO-congruent MSLP at each grid-point. σ^2_{res} is the variance in the residuals of this regression.

circulation variability in the North Atlantic sector (Barnston & Livezey, 1987; Moore et al., 2011; Wallace & Gutzler, 1981).

3.2. Uncertainty in the Forced Response

Figure 1 shows structural differences in the late 21st century forced NAO response across the MMLEA models. Here, we ask: When do the forced NAO response and model structural differences in the response become detectable from IV? In the early-to-mid 21st century, most individual model responses are small and nonrobust (Figures 3a and 3b). GFDL-ESM2M is one exception, having a relatively large and robust positive NAO response by 2020–2039. By 2060–2079, most of the model responses become large enough to be detected from IV, except for EC-EARTH due to its smaller response and ensemble size (Figure 3c). Regarding detection of model differences in response, in the mid-21st century only GFDL-ESM2M is robustly distinguishable from the other models (Figure 3b). By 2060–2079, the only model with a negative NAO response (CanESM2; Böhnisch et al., 2020) becomes distinct from other models (Figure 3c). By 2080–2099, CSIRO-Mk3.6 and MPI-ESM-LR develop stronger positive responses and become distinct from CESM1-CAM5 and EC-EARTH (Figure 3d). In short, most of the models simulate a robust forced NAO response by 2060–2079. However, most intermodel differences in the forced response are only detectable by 2080–2099, when $\%U_{\text{MD}}$ first dominates over $\%U_{\text{IV}}$ (Figures 3a–3d). This largely holds when the model-based IV estimates are adjusted to an observation-based estimate (Figure S3).

We now calculate the minimum ensemble size (N_{min}) required to robustly detect a forced NAO index response, and model differences in this response, given a certain magnitude of IV. First, note that N_{min} is larger when identifying differences in forced response between models than when identifying a response of equivalent magnitude in one model (Figures 3e and 3f). This is consistent with intermodel differences in forced response emerging from IV later. An NAO index response of 0.5 hPa, typical of early-to-mid 21st century MMLEA responses (Figures 3a and 3b), requires $N_{\text{min}} = 10, 20,$ or 40 to detect in a model with low (2.5th percentile), median, or high (97.5th percentile) IV, basing the IV magnitude on the CMIP5/6 multimodel ensemble. For context, the interannual variability (standard deviation) in the DJF NAO index

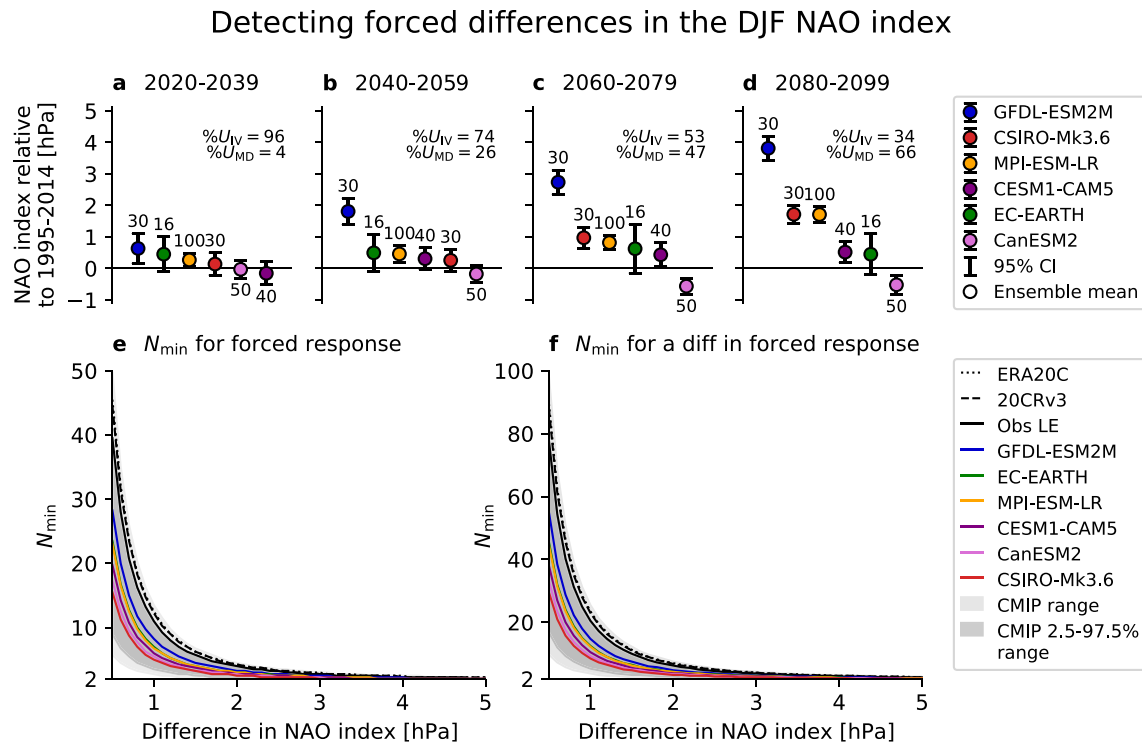


Figure 3. Detecting a forced response in December to February (DJF) North Atlantic Oscillation (NAO) index and inter-model differences in this response. (a–d), NAO anomalies in Multimodel Large Ensemble Archive (MMLEA) models for future 20-year epochs (1995–2014 baseline). Whiskers are 95% confidence intervals and numbers indicate ensemble size. Section 2.3 defines $\%U_{IV}$ and $\%U_{MD}$. (e), N_{\min} required to detect a forced NAO response of a given magnitude at the 95% confidence level based on IV estimates from MMLEA models, CMIP5/6 models, and observation-based datasets (Text S2 and S3). (f), As in (e) but for detecting a difference in forced response; note different y-axis scale. Single CMIP5/6 models can be located within the gray plumes using Table S2.

is ~ 4 hPa in the observation-based datasets. N_{\min} is doubled to 20, 40, or 80 to detect a difference in NAO index response of 0.5 hPa between two models. N_{\min} for a high IV model is similar to N_{\min} calculated using observation-based IV estimates. All subsequent results use the high IV estimate and thus provide an upper bound on N_{\min} . To detect larger NAO responses of 1 hPa and 2 hPa, typical of late 21st century MMLEA responses (Figures 3c and 3d), at least 15 or 5 members are required, respectively. This becomes 30 or 10 members for a difference in response. The largest MMLEA model response, and difference in response, of ~ 4 hPa in 2080–2099 (Figure 3d) requires only three members to detect. N_{\min} is first minimized at 2 for a response of 5 hPa or a difference in response of 7 hPa. Therefore, when considering more realistic IV estimates, most NAO anomalies and model differences in Figure 1 are nonrobust in CMIP5/6 models with only one ensemble member.

Finally, we ask to what extent intermodel spread in the forced response of North Atlantic circulation projects onto the NAO structure and, therefore, reflects differences in the response of the NAO to external forcing. The forced MSLP response is rather different across the MMLEA models (Figure 4, top row). For example, in CSIRO-Mk3.6, GFDL-ESM2M, and MPI-ESM-LR there is a north-south dipole in pressure anomalies, which is not present in CanESM2, CESM1-CAM5, and EC-EARTH. This is associated with inter-model spread in the NAO-congruent MSLP response (Figure 4, middle row). However, while a substantial portion of the forced North Atlantic MSLP response is NAO-congruent in some models (e.g., 80% in GFDL-ESM2M), this is not true of other models (e.g., EC-EARTH), and there are large residuals in all models (Figure 4, bottom row). Besides limited regions at high latitudes and in Southern Europe, the MSLP residuals contribute to the majority of the intermodel spread in the forced MSLP response (e.g., see Greenland, eastern North America and central Europe; Figure 4, far-right column).

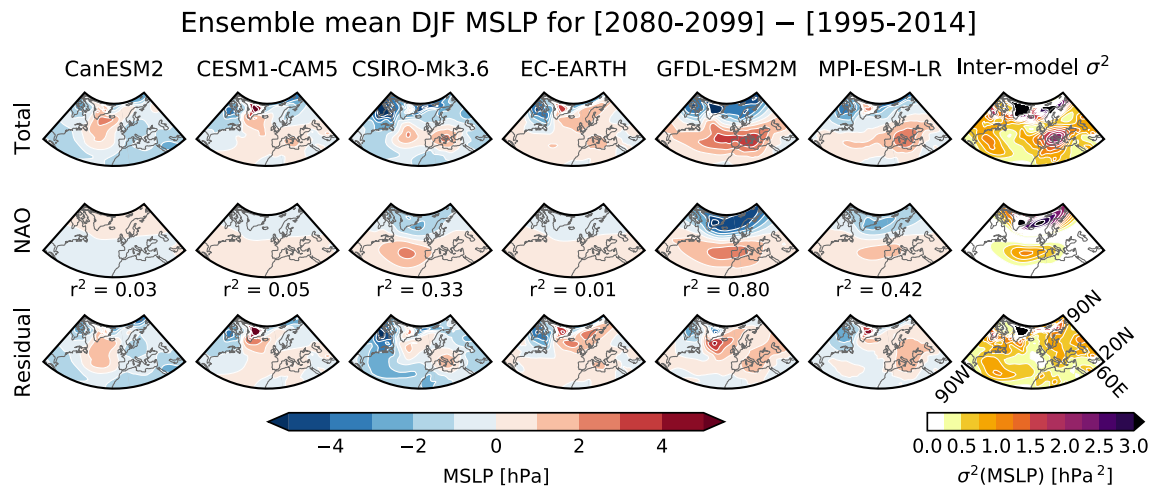


Figure 4. Projections of ensemble mean December to February (DJF) mean sea level pressure (MSLP) for (2080–2099) – (1995–2014) for each Multimodel Large Ensemble Archive (MMLEA) model, and their inter-model variance. (Top) Total; (Middle) North Atlantic Oscillation (NAO)-congruent part; (Bottom) Residual. r^2 is the squared area-weighted pattern correlation between the total response and the NAO-congruent part.

4. Discussion and Conclusions

The results presented here have improved our understanding of North Atlantic circulation projections in various ways.

First, while the CMIP5/6 models under RCP8.5/SSP5-8.5 show a mean response in the winter NAO index of ~ 1.5 hPa during the late 21st century (2080–2099) compared to near-present-day (1995–2014), the individual model responses span 6–7 hPa and less than 90% of models agree on the sign of response. The MMLEA models suggest that approximately two-thirds of the large intermodel spread in CMIP5/6 could be explained by potentially reducible model structural differences and one-third by irreducible uncertainty from IV. While previous studies have noted the large spread in North Atlantic circulation projections (Section 1), this study is the first to quantify these sources of uncertainty using large initial-condition ensembles performed by a subset of CMIP5 models. The real-world relevance of this separation relies on models correctly reproducing the observed magnitude of low frequency IV and forced NAO response. We find the former is generally underestimated in models as in previous studies, but note the latter may also be underestimated (see below).

Second, as expected from the relatively large IV of the winter NAO, we find a relatively long time horizon for detecting a forced NAO response. The MMLEA models suggest that the forced NAO response is only detectable from IV by 2060–2079 and that model structural uncertainty in the forced response is detectable by 2080–2099. Uncertainty in NAO projections is therefore largely irreducible for most of the 21st century. While individual MMLEA models have larger NAO responses that are distinct from IV and other models earlier, this is generally not the case. This highlights a benefit of using the new MMLEA archive here, whereas previous studies have been limited to using a single-model large initial-condition ensemble to quantify the time of emergence of a forced circulation response (Deser et al., 2012, 2017).

Third, we show that a relatively large ensemble size is required to robustly separate the forced NAO response, and model differences in this response, from IV. A typical response (or model difference) of 1–2 hPa over the 21st century requires at least 15–5 (30–10) ensemble members to detect based on realistic estimates of IV. Even for very large responses (model differences) of around 5 hPa (7 hPa), two members are required for detection, meaning the majority of model responses and differences are nonrobust in CMIP5/6 models with only one ensemble member. This result is relevant to the growing application of emergent constraint techniques for narrowing uncertainty in future projections, as this relies on knowledge of the forced response and differences in forced responses across ensembles of models. Future model intercomparison experiment designs should consider the required ensemble sizes for examining regional climate signals (e.g., Milinski et al., 2020).

Finally, we have examined the extent to which the spread in North Atlantic MSLP projections is NAO-congruent. Regarding spread from IV, this is large in most North Atlantic regions and surrounding land areas, where the NAO explains over 50% of the inter-member spread in individual MMLEA models at higher latitudes and up to 50% around the Mediterranean region. The residual spread in the central Atlantic and western Europe is largely explained by the EA pattern. That the spread in projections from IV is largely explained by dominant modes of atmospheric variability agrees with Deser et al. (2012). These results build on those of Deser et al. (2017), who only analyzed the NAO contribution to spread in projections from IV.

Regarding intermodel spread in the forced North Atlantic MSLP response, while this is largely NAO-congruent at high latitudes and in Southern Europe, the majority of the spread is not NAO-congruent. Therefore, improving understanding of the NAO alone may not constrain the reducible uncertainty in North Atlantic MSLP projections. This is surprising considering previous work demonstrating the resemblance of externally forced model responses to the dominant modes of IV (Deser et al., 2004, 2012). The large residual uncertainty in the forced MSLP response over Greenland may be associated with local near-surface temperature changes over orography and/or the extrapolation of pressure to mean sea level.

These results have some limitations. First, MSLP only provides one perspective of the circulation. When using the zonal wind at 850 hPa (U850), which is related to the meridional pressure gradient, we find a shift in the regions with large uncertainty from IV (Figure S4). Furthermore, intermodel spread in the forced U850 response appears more NAO-congruent over Europe than for MSLP (Figure S5).

Second, models appear to underestimate predictable forced NAO variations by a factor of 2 on seasonal timescales (Baker et al., 2018; Dunstone et al., 2016; Eade et al., 2014; Scaife & Smith, 2018; Scaife et al., 2014) and by a factor of 10 on decadal timescales (Smith et al., 2020). This issue may also affect multidecadal NAO projections, though given the limited temporal extent of the observational record this is difficult to assess. If it does, this implies an underestimation of model differences in the forced NAO response and therefore the contribution of the NAO to intermodel spread in the forced circulation response, as well as an overestimation of the time horizon and “true” ensemble size required to detect a forced NAO response from IV. A further limitation of our analysis is that the MMLEA models may not span the full range of forced NAO responses in the CMIP5/6 models. However, it is difficult to assess this given the small ensemble sizes for most CMIP5/6 models.

The dynamical mechanisms responsible for intermodel spread in the forced North Atlantic circulation response need to be understood to identify potential physical constraints on the spread. Oudar et al. (2020) identified various mechanisms within CMIP5/6 projections, but could not determine which are relevant for spread from IV and/or model differences. Harvey et al. (2020)'s results suggest that mean state biases in the North Atlantic jet do not provide a useful constraint. Future studies could utilize MMLEA to investigate the dynamical mechanisms further.

Data Availability Statement

The Multimodel Large Ensemble Archive and Observational Large Ensemble data can be accessed at <http://www.cesm.ucar.edu/projects/community-projects/MMLEA/>. The GFDL-ESM2M large ensemble data used here can be accessed from the Princeton Large Ensemble Archive through Globus (<https://www.sarah-schlunegger.com/large-ensemble-archive>). The CMIP5 and CMIP6 datasets were downloaded from CEDA/JASMIN (timestamps of September 21–23, 2020 and December 4, 2020, respectively); these are publicly available through the Earth System Grid Federation at <https://esgf-index1.ceda.ac.uk/projects/esgf-ceda/>. The observational datasets can be downloaded from https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html (20CRv3) and <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-20c> (ERA20C).

References

- Ambaum, M. H. P., Hoskins, B. J., & Stephenson, D. B. (2001). Arctic Oscillation or North Atlantic Oscillation? *Journal of Climate*, *14*(16), 3495–3507. [https://doi.org/10.1175/1520-0442\(2001\)014<3495:AOONAO>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<3495:AOONAO>2.0.CO;2)
- Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A., & Scaife, A. A. (2018). An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, *45*, 7808–7817. <https://doi.org/10.1029/2018GL078838>

Acknowledgments

C. M. McKenna and A. C. Maycock were supported by the European Union's Horizon 2020 Research And Innovation Program under grant agreement No 820829 (CONSTRAIN project). A. C. Maycock was supported by a NERC Independent Research Fellowship (NE/M018199/1) and The Leverhulme Trust (PLP-2018-278). The authors are grateful to Isla Simpson, Clara Deser, Flavio Lehner, and John Fyfe for useful discussions about the MMLEA dataset and North Atlantic variability. They thank the CONSTRAIN project community for useful comments on this work and two anonymous reviewers for their constructive comments which improved the manuscript. They acknowledge the U.S. CLIVAR Working Group on Large Ensembles for providing the Multimodel Large Ensemble Archive and Observational Large Ensemble data. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP5/6. They thank the climate modeling groups for producing and making available their model output; the Earth System Grid Federation (ESGF) and UK Center for Environmental Data Analytics (CEDA) JASMIN cluster for archiving the data and providing access; and the multiple funding agencies who support CMIP5/6, ESGF, and CEDA/JASMIN.

- Barnston, A. G., & Livezey, R. E. (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, *115*(6), 1083–1126. [https://doi.org/10.1175/1520-0493\(1987\)115<1083:CSAPOL>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2)
- Böhmsch, A., Ludwig, R., & Leduc, M. (2020). Using a nested single-model large ensemble to assess the internal variability of the North Atlantic Oscillation and its climatic implications for central Europe. *Earth System Dynamics*, *11*, 617–640. <https://doi.org/10.5194/esd-11-617-2020>
- Bracegirdle, T. J., Lu, H., Eade, R., & Woollings, T. (2018). Do CMIP5 models reproduce observed low-frequency North Atlantic jet variability? *Geophysical Research Letters*, *45*, 7204–7212. <https://doi.org/10.1029/2018GL078965>
- Buehler, T., Raible, C. C., & Stocker, T. F. (2011). The relationship of winter season North Atlantic blocking frequencies to extreme cold or dry spells in the ERA-40. *Tellus A: Dynamic Meteorology and Oceanography*, *63*(2), 174–187. <https://doi.org/10.1111/j.1600-0870.2010.00492.x>
- Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichet, T., Friedlingstein, P., et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In T. F. Stocker, D. Qin, G.-K. Plattner, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (pp. 1029–1136). Cambridge, UK, and New York: Cambridge University Press. <https://doi.org/10.1017/CBO9781107415324.024>
- 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>
- Deser, C., Hurrell, J. W., & Phillips, A. S. (2017). The role of the North Atlantic Oscillation in European climate projections. *Climate Dynamics*, *49*, 3141–3157. <https://doi.org/10.1007/s00382-016-3502-z>
- Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*, 277–286. <https://doi.org/10.1038/s41558-020-0731-2>
- Deser, C., Magnusdottir, G., Saravanan, R., & Phillips, A. (2004). The effects of North Atlantic SST and sea ice anomalies on the winter circulation in CCM3. Part II: Direct and indirect components of the response. *Journal of Climate*, *17*(5), 877–889. [https://doi.org/10.1175/1520-0442\(2004\)017<0877:TEONAS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2004)017<0877:TEONAS>2.0.CO;2)
- Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*, 527–546. <https://doi.org/10.1007/s00382-010-0977-x>
- Dommenget, D., & Latif, M. (2002). A cautionary note on the interpretation of EOFs. *Journal of Climate*, *15*(2), 216–225. [https://doi.org/10.1175/1520-0442\(2002\)015<0216:ACNOTI%3e2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<0216:ACNOTI%3e2.0.CO;2)
- Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., et al. (2016). Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, *9*, 809–814. <https://doi.org/10.1038/ngeo2824>
- Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*, 5620–5628. <https://doi.org/10.1002/2014GL061146>
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model inter-comparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*, 1937–1958. <https://doi.org/10.5194/gmd-9-1937-2016>
- Harvey, B. J., Cook, P., Shaffrey, L. C., & Schiemann, R. (2020). The response of the northern hemisphere storm tracks and jet streams to climate change in the CMIP3, CMIP5, and CMIP6 climate models. *Journal of Geophysical Research: Atmospheres*, *125*, e2020JD032701. <https://doi.org/10.1029/2020JD032701>
- Harvey, B. J., Shaffrey, L. C., & Woollings, T. J. (2014). Equator-to-pole temperature differences and the extra-tropical storm track responses of the CMIP5 climate models. *Climate Dynamics*, *43*, 1171–1182. <https://doi.org/10.1007/s00382-013-1883-9>
- Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. <https://doi.org/10.1175/2009BAMS2607.1>
- Hazeleger, W., Severijns, C., Semmler, T., Ștefănescu, S., Yang, S., Wang, X., et al. (2010). EC-Earth. *Bulletin of the American Meteorological Society*, *91*(10), 1357–1364. <https://doi.org/10.1175/2010BAMS2877.1>
- Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (2003). An overview of the North Atlantic Oscillation. In J. W. Hurrell, Y. Kushnir, G. Ottersen, et al. (Eds.), *The North Atlantic Oscillation: Climate significance and environmental impact*. Geophysical monograph series (pp. 1–35). Washington, DC: American Geophysical Union. <https://doi.org/10.1029/134GM01>
- Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., et al. (2013). Australia's CMIP5 submission using the CSIRO-Mk3.6 model. *Australian Meteorological and Oceanographic Journal*, *63*(1), 1–14. <https://doi.org/10.22499/2.6301.001>
- Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. <https://doi.org/10.1175/BAMS-D-13-00255.1>
- Kim, W. M., Yeager, S., Chang, P., & Danabasoglu, G. (2018). Low-frequency North Atlantic climate variability in the community earth system model large ensemble. *Journal of Climate*, *31*(2), 787–813. <https://doi.org/10.1175/JCLI-D-17-0193.1>
- Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017). Attribution of extreme events in Arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571. <https://doi.org/10.1175/JCLI-D-16-0412.1>
- Kravtsov, S. (2017). Pronounced differences between observed and CMIP5-simulated multidecadal climate variability in the twentieth century. *Geophysical Research Letters*, *44*, 5749–5757. <https://doi.org/10.1002/2017GL074016>
- Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., et al. (2019). The Max Planck Institute grand ensemble: Enabling the exploration of climate system variability. *Journal of Advances in Modeling Earth Systems*, *11*, 2050–2069. <https://doi.org/10.1029/2019MS001639>
- Maher, N., Power, S. B., & Marotzke, J. (2021). More accurate quantification of model-to-model agreement in externally forced climatic responses over the coming century. *Nature Communications*, *12*, 788. <https://doi.org/10.1038/s41467-020-20635-w>
- Manzini, E., Karpechko, A. Y., Anstey, J., Baldwin, M. P., Black, R. X., Cagnazzo, C., et al. (2014). Northern winter climate change: Assessment of uncertainty in CMIP5 projections related to stratosphere-troposphere coupling. *Journal of Geophysical Research: Atmospheres*, *119*, 7979–7998. <https://doi.org/10.1002/2013JD021403>
- McKinnon, K. A., & Deser, C. (2018). Internal variability and regional climate trends in an observational large ensemble. *Journal of Climate*, *31*(17), 6783–6802. <https://doi.org/10.1175/JCLI-D-17-0901.1>
- Meinshausen, M., Nicholls, Z. R. J., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., et al. (2020). The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development*, *13*, 3571–3605. <https://doi.org/10.5194/gmd-13-3571-2020>
- Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., et al. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, *109*, 213–241. <https://doi.org/10.1007/s10584-011-0156-z>

- Milinski, S., Maher, N., & Olonscheck, D. (2020). How large does a large ensemble need to be? *Earth System Dynamics*, *11*, 885–901. <https://doi.org/10.5194/esd-11-885-2020>
- Moore, G. W. K., Pickart, R. S., & Renfrew, I. A. (2011). Complexities in the climate of the subpolar North Atlantic: a case study from the winter of 2007. *Quarterly Journal of the Royal Meteorological Society*, *137*, 757–767. <https://doi.org/10.1002/qj.778>
- Oudar, T., Cattiaux, J., & Douville, H. (2020). Drivers of the northern extratropical eddy-driven jet change in CMIP5 and CMIP6 models. *Geophysical Research Letters*, *47*, e2019GL086695. <https://doi.org/10.1029/2019GL086695>
- Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., et al. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, *29*(11), 4083–4097. <https://doi.org/10.1175/JCLI-D-15-0556.1>
- Rodgers, K. B., Lin, J., & Frölicher, T. L. (2015). Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, *12*(11), 3301–3320. <https://doi.org/10.5194/bg-12-3301-2015>
- Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*, 2514–2519. <https://doi.org/10.1002/2014GL059637>
- Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, *1*, 28. <https://doi.org/10.1038/s41612-018-0038-4>
- Schlunegger, S., Rodgers, K. B., Sarmiento, J. L., Frölicher, T. L., Dunne, J. P., Ishii, M., & Slater, R. (2019). Emergence of anthropogenic signals in the ocean carbon cycle. *Nature Climate Change*, *9*, 719–725. <https://doi.org/10.1038/s41558-019-0553-2>
- Screen, J. A., Deser, C., Simmonds, I., & Tomas, R. (2014). Atmospheric impacts of Arctic sea-ice loss, 1979–2009: Separating forced change from atmospheric internal variability. *Climate Dynamics*, *43*, 333–344. <https://doi.org/10.1007/s00382-013-1830-9>
- Shepherd, T. (2014). Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geoscience*, *7*, 703–708. <https://doi.org/10.1038/ngeo2253>
- Simpson, I. R., Bacmeister, J., Neale, R. B., Hannay, C., Gettelman, A., Garcia, R. R., (2020). An evaluation of the large-scale atmospheric circulation and its variability in CESM2 and other CMIP models. *Journal of Geophysical Research: Atmospheres*, *125*, e2020JD032835. <https://doi.org/10.1029/2020JD032835>
- Simpson, I. R., Deser, C., McKinnon, K. A., & Barnes, E. A. (2018). Modeled and observed multidecadal variability in the North Atlantic jet stream and its connection to sea surface temperatures. *Journal of Climate*, *31*(20), 8313–8338. <https://doi.org/10.1175/JCLI-D-18-0168.1>
- Simpson, I. R., Hitchcock, P., Seager, R., Wu, Y., & Callaghan, P. (2018). The downward influence of uncertainty in the Northern Hemisphere stratospheric polar vortex response to climate change. *Journal of Climate*, *31*(16), 6371–6391. <https://doi.org/10.1175/JCLI-D-18-0041.1>
- Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., et al. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*, 2876–2908. <https://doi.org/10.1002/qj.3598>
- 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*, 796–800. <https://doi.org/10.1038/s41586-020-2525-0>
- Stephenson, D., Pavan, V., Collins, M., Junge, M., Quadrelli, R., & Participating CMIP2 Modelling Groups. (2006). North Atlantic Oscillation response to transient greenhouse gas forcing and the impact on European winter climate: A CMIP2 multi-model assessment. *Climate Dynamics*, *27*(4), 401–420. <https://doi.org/10.1007/s00382-006-0140-x>
- Sun, L., Alexander, M., & Deser, C. (2018). Evolution of the global coupled climate response to Arctic Sea ice loss during 1990–2090 and its contribution to climate change. *Journal of Climate*, *31*(19), 7823–7843. <https://doi.org/10.1175/JCLI-D-18-0134.1>
- Taylor, K. E., Stouffer, R. J., & 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>
- von Storch, H., & Zwiers, F. W. (1999). *Statistical analysis in climate research*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511612336>
- Wallace, J. M., & Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, *109*(4), 784–812. [https://doi.org/10.1175/1520-0493\(1981\)109<0784:TITGHF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2)
- Wang, X., Li, J., Sun, C., & Liu, T. (2017). NAO and its relationship with the Northern Hemisphere mean surface temperature in CMIP5 simulations. *Journal of Geophysical Research: Atmospheres*, *122*, 4202–4227. <https://doi.org/10.1002/2016JD025979>
- Woollings, T., Hannachi, A., & Hoskins, B. (2010). Variability of the North Atlantic eddy-driven jet stream. *Quarterly Journal of the Royal Meteorological Society*, *136*, 856–868. <https://doi.org/10.1002/qj.625>
- Zappa, G., Pithan, F., & Shepherd, T. G. (2018). Multimodel evidence for an atmospheric circulation response to Arctic sea ice loss in the CMIP5 future projections. *Geophysical Research Letters*, *45*, 1011–1019. <https://doi.org/10.1002/2017GL076096>
- Zappa, G., & Shepherd, T. G. (2017). Storylines of atmospheric circulation change for European regional climate impact assessment. *Journal of Climate*, *30*(16), 6561–6577. <https://doi.org/10.1175/JCLI-D-16-0807.1>