Multi-Model Large Ensemble projections of the North Atlantic Oscillation during the 21st century

Christine M. McKenna[1,1] and Amanda Maycock[1,1]

[1]University of Leeds

November 30, 2022

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 Multi-Model 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.

Sources of uncertainty in multi-model large ensemble projections of the winter North Atlantic Oscillation

C. M. McKenna[1] and A. C. Maycock[1]

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

Corresponding author: Christine McKenna (C.McKenna1@leeds.ac.uk)

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 modelled by 2080-2099 under RCP8.5

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 Multi-Model 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 multi-model 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 realisations 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 pre-industrial climate; Hawkins & Sutton, 2009).

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 Multi-Model 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 member) initial-condition ensembles for 7 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 pre-industrial 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 multi-decadal 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). An 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 minimise 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 (90W-60E, 20N-55N) and a northern box (90W-60E, 55N-90N) in the North Atlantic. This index is less sensitive to differences in centres 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 EOF-based index (Ambaum et al., 2001; Dommenget & 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; using a future period gives similar results. 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 multi-model mean (MMM) historical NAO patterns; largely, the modelled 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 ($\%U_{\mathrm{IV}}$) and of model structural differences ($\%U_{\mathrm{MD}}$) 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 inter-member 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 re-arrange a two-sided Student's t-test for a difference of means (Text S2):

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

t_c is for p=0.025 and $2N_{\mathrm{min}}$−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 modelled range of IV, most modelled 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, grey 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 inter-model differences in the forced response of up to 5 hPa (Figure 1, coloured circles). Using Maher et al. (2021)'s uncertainty decomposition, we find that model 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 versus grey lines; Tables S1-S2). The model projections may therefore be overconfident: i.e., 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 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 non-robust (Figure 3a-b). 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 inter-model differences in the forced response are only detectable by 2080-2099, when $\%U_{MD}$ first dominates over $\%U_{IV}$ (Figure 3a-d). 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 (Figure 3e-f). This is consistent with inter-model 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 (Figure 3a-b) – requires N_{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 multi-model ensemble. For context, the interannual variability (standard deviation) in the DJF NAO index 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 (Figure 3c-d) – 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 3 members to detect. N_{min} is first minimised 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 non-robust in CMIP5/6 models with only 1 ensemble member.

Finally, we ask to what extent inter-model 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 inter-model spread in the forced MSLP response (e.g., see Greenland, eastern North America and central Europe; Figure 4, far-right column).

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 inter-model 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), 2 members are required for detection – meaning the

majority of model responses and differences are non-robust in CMIP5/6 models with only 1 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 analysed the NAO contribution to spread in projections from IV.

Regarding inter-model 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, inter-model 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 et al., 2014; Scaife & Smith, 2018) and by a factor of 10 on decadal timescales (Smith et al., 2020). This issue may also affect multi-decadal 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 inter-model 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 inter-model 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 utilise MMLEA to investigate the dynamical mechanisms further.

Figure 1. **Projections of the DJF NAO index for [2080-2099]−[1995-2014] in the CMIP6, CMIP5 and 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 IV (Text S3). Darker cyan/blue bars indicate the 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. Grey lines show the 2.5-97.5% range of 10^5 differences in 20-year epoch means of different observation-based records (Section 2.1), selected by randomly resampling with replacement. Grey 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.

Figure 2. **Inter-member variance in projections of DJF MSLP for [2080-2099]−[1995-2014] for each MMLEA model.** [Top row] Total (σ^2_{tot}); [Second row] 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 inter-member 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.

Figure 3. **Detecting a forced response in DJF NAO index and inter-model differences in this response. a-d**, NAO anomalies in 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-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 grey plumes using Table S2.

Figure 4. **Projections of ensemble mean DJF MSLP for [2080-2099]−[1995-2014] for each MMLEA model, and their inter-model variance.** [Top] Total; [Middle] NAO-congruent part; [Bottom] Residual. r^2 is the squared area-weighted pattern correlation between the total response and the NAO-congruent part.

Acknowledgments

CMM and ACM were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 820829 (CONSTRAIN project). ACM was supported by a NERC Independent Research Fellowship (NE/M018199/1) and The Leverhulme Trust (PLP-2018-278). We are grateful to Isla Simpson, Clara Deser, Flavio Lehner and John Fyfe for useful discussions about the MMLEA dataset and North Atlantic variability. We thank the CONSTRAIN project community for useful comments on this work and two anonymous reviewers for their constructive comments which improved the manuscript. We acknowledge the U.S. CLIVAR Working Group on Large Ensembles for providing the Multi-Model Large Ensemble Archive and Observational Large Ensemble data. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5/6. We thank the climate modelling groups for producing and making available their model output; the Earth System Grid Federation (ESGF) and UK Centre 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.

Data availability statement

The Multi-Model 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.sarahschlunegger.com/large-ensemble-archive). The CMIP5 and CMIP6 datasets were downloaded from CEDA/JASMIN (timestamps of 21-23 September 2020 and 4 December 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

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

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

Böhnisch, 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., Fichefet, T., Friedlingstein, P., et al. (2013). Long-term Climate Change: Projections, Commitments and Irreversibility. In T. F. Stocker, 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*

406 *Climate Change* (pp. 1029–1136). Cambridge, UK, and New York: Cambridge
407 University Press. https://doi.org/10.1017/CBO9781107415324.024

408 Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., et al.
409 (2011). The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal*
410 *Meteorological Society*, *137*, 1–28. http://dx.doi.org/10.1002/qj.776

411 Deser, C., Hurrell, J. W., & Phillips, A. S. (2017). The role of the North Atlantic Oscillation in
412 European climate projections. *Climate Dynamics*, *49*, 3141–3157.
413 https://doi.org/10.1007/s00382-016-3502-z

414 Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020).
415 Insights from Earth system model initial-condition large ensembles and future prospects.
416 *Nature Climate Change*, *10*, 277–286. https://doi.org/10.1038/s41558-020-0731-2

417 Deser, C., Magnusdottir, G., Saravanan, R., & Phillips, A. (2004). The Effects of North Atlantic
418 SST and Sea Ice Anomalies on the Winter Circulation in CCM3. Part II: Direct and
419 Indirect Components of the Response. *Journal of Climate*, *17*(5), 877–889.
420 https://doi.org/10.1175/1520-0442(2004)017<0877:TEONAS>2.0.CO;2

421 Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change
422 projections: the role of internal variability. *Climate Dynamics*, *38*, 527–546.
423 https://doi.org/10.1007/s00382-010-0977-x

424 Dommenget, D., & Latif, M. (2002). A Cautionary Note on the Interpretation of EOFs. *Journal*
425 *of Climate*, *15*(2), 216–225.
426 https://doi.org/10.1175/1520-0442(2002)015%3c0216:ACNOTI%3e2.0.CO;2

427 Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., et al. (2016). Skilful
428 predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*,
429 *9*, 809–814. https://doi.org/10.1038/ngeo2824

430 Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N.
431 (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 Intercomparison 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., Ştefănescu, 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. Kushner, G. Ottersen, & M. Visbeck (Eds.), *The North Atlantic Oscillation: Climate Significance and Environmental Impact, Geophysical Monograph Series* (Vol. 134, 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*

458 *Meteorological and Oceanographic Journal*, *63*(1), 1–13.
459 https://doi.org/10.22499/2.6301.001

460 Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The Community
461 Earth System Model (CESM) Large Ensemble Project: A Community Resource for
462 Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the*
463 *American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-
464 13-00255.1

465 Kim, W. M., Yeager, S., Chang, P., & Danabasoglu, G. (2018). Low-Frequency North Atlantic
466 Climate Variability in the Community Earth System Model Large Ensemble. *Journal of*
467 *Climate*, *31*(2), 787–813. https://doi.org/10.1175/JCLI-D-17-0193.1

468 Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2016). Attribution of extreme events
469 in Arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571.
470 https://doi.org/10.1175/JCLI-D-16-0412.1

471 Kravtsov, S. (2017). Pronounced differences between observed and CMIP5-simulated
472 multidecadal climate variability in the twentieth century. *Geophysical Research Letters*,
473 *44*, 5749–5757. https://doi.org/10.1002/2017GL074016

474 Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., &
475 Marotzke, J. (2019). The Max Planck Institute grand ensemble: Enabling the exploration
476 of climate system variability. *Journal of Advances in Modeling Earth Systems*, *11*, 2050–
477 2069. https://doi.org/10.1029/2019MS001639

478 Maher, N., Power, S. B. & Marotzke, J. (2021). More accurate quantification of model-to-model
479 agreement in externally forced climatic responses over the coming century. *Nature*
480 *Communications*, *12*, 788. https://doi.org/10.1038/s41467-020-20635-w

481 Manzini, E., Karpechko, A. Y., Anstey, J., Baldwin, M. P., Black, R. X., Cagnazzo, C., et al.
482 (2014). Northern winter climate change: Assessment of uncertainty in CMIP5 projections
483 related to stratosphere-troposphere coupling. *Journal of Geophysical Research:*
484 *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. 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., et al. (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., et al. (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., et al. (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

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

565 Zappa, G., & Shepherd, T. G. (2017). Storylines of Atmospheric Circulation Change for
566 European Regional Climate Impact Assessment. *Journal of Climate, 30*(16), 6561–6577.
567 https://doi.org/10.1175/JCLI-D-16-0807.1

Figure 1.

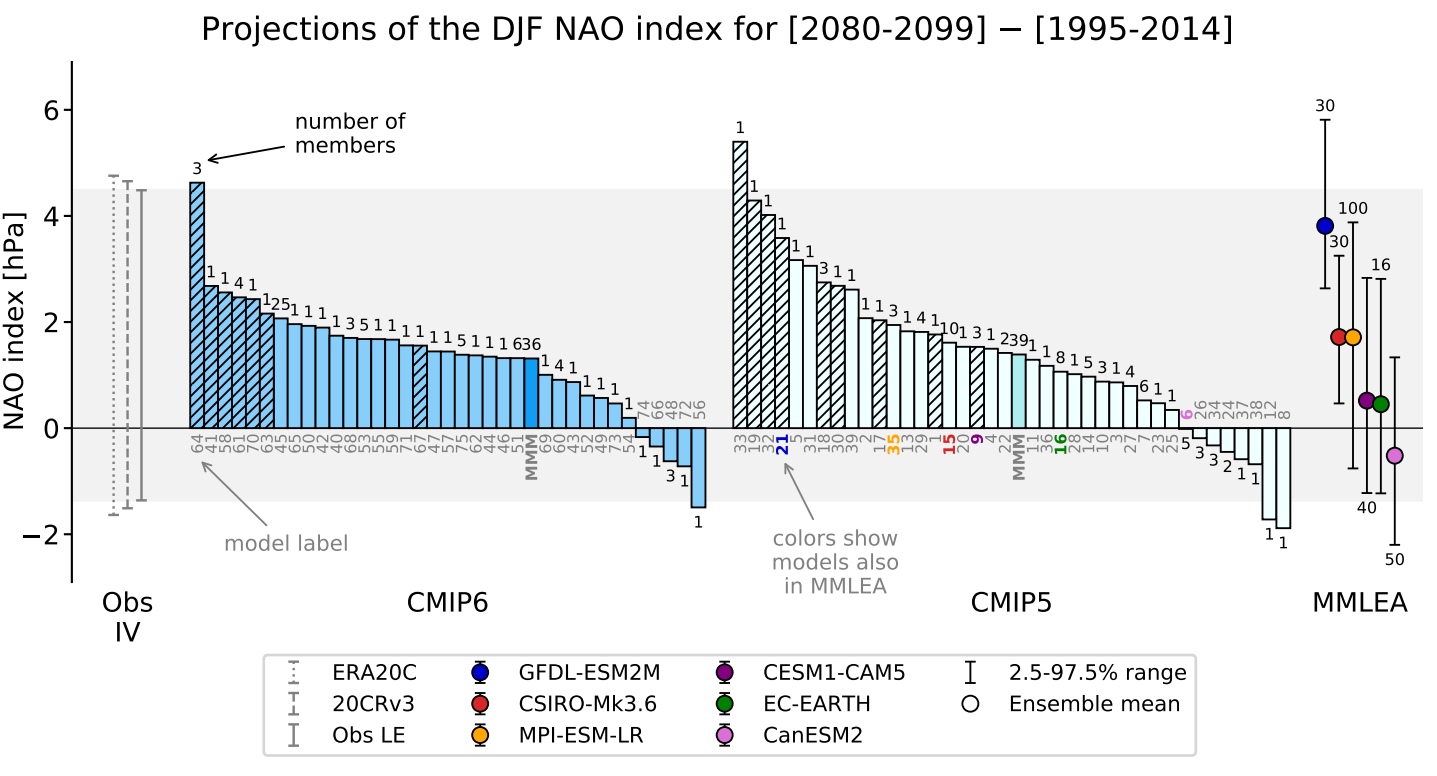

Projections of the DJF NAO index for [2080-2099] − [1995-2014]
NAO index [hPa]
number of members
model label
colors show models also in MMLEA
Obs IV
CMIP6
CMIP5
MMLEA
ERA20C
20CRv3
Obs LE
GFDL-ESM2M
CSIRO-Mk3.6
MPI-ESM-LR
CESM1-CAM5
EC-EARTH
CanESM2
2.5-97.5% range
Ensemble mean

Figure 2.

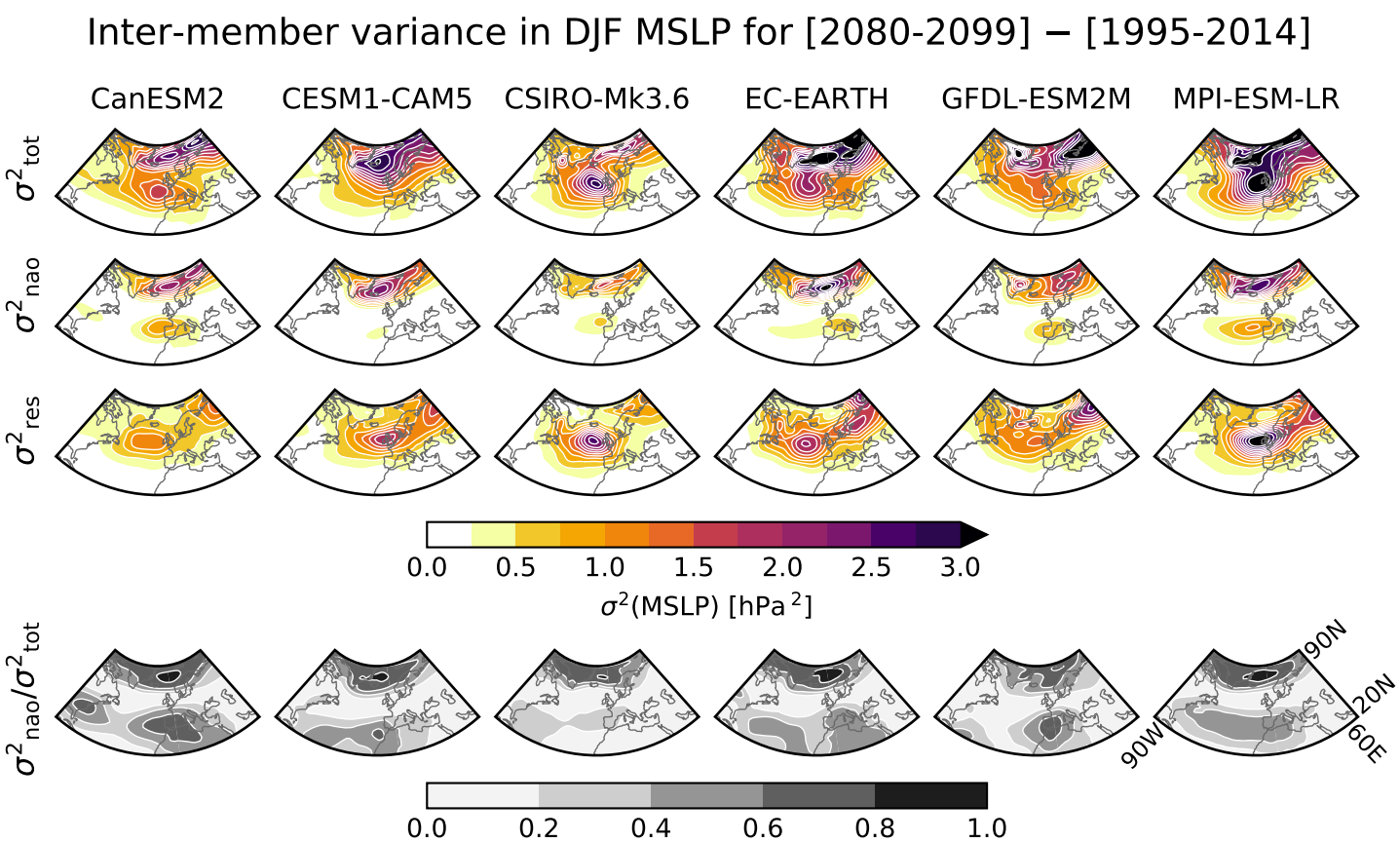

Inter-member variance in DJF MSLP for [2080-2099] − [1995-2014]
CanESM2
CESM1-CAM5
CSIRO-Mk3.6
EC-EARTH
GFDL-ESM2M
MPI-ESM-LR
σ2tot
σ2nao
σ2res
0.0 0.5 1.0 1.5 2.0 2.5 3.0
σ2(MSLP) [hPa 2]
σ2nao/σ2tot
90N
120N
60E
90W
0.0 0.2 0.4 0.6 0.8 1.0

Figure 3.

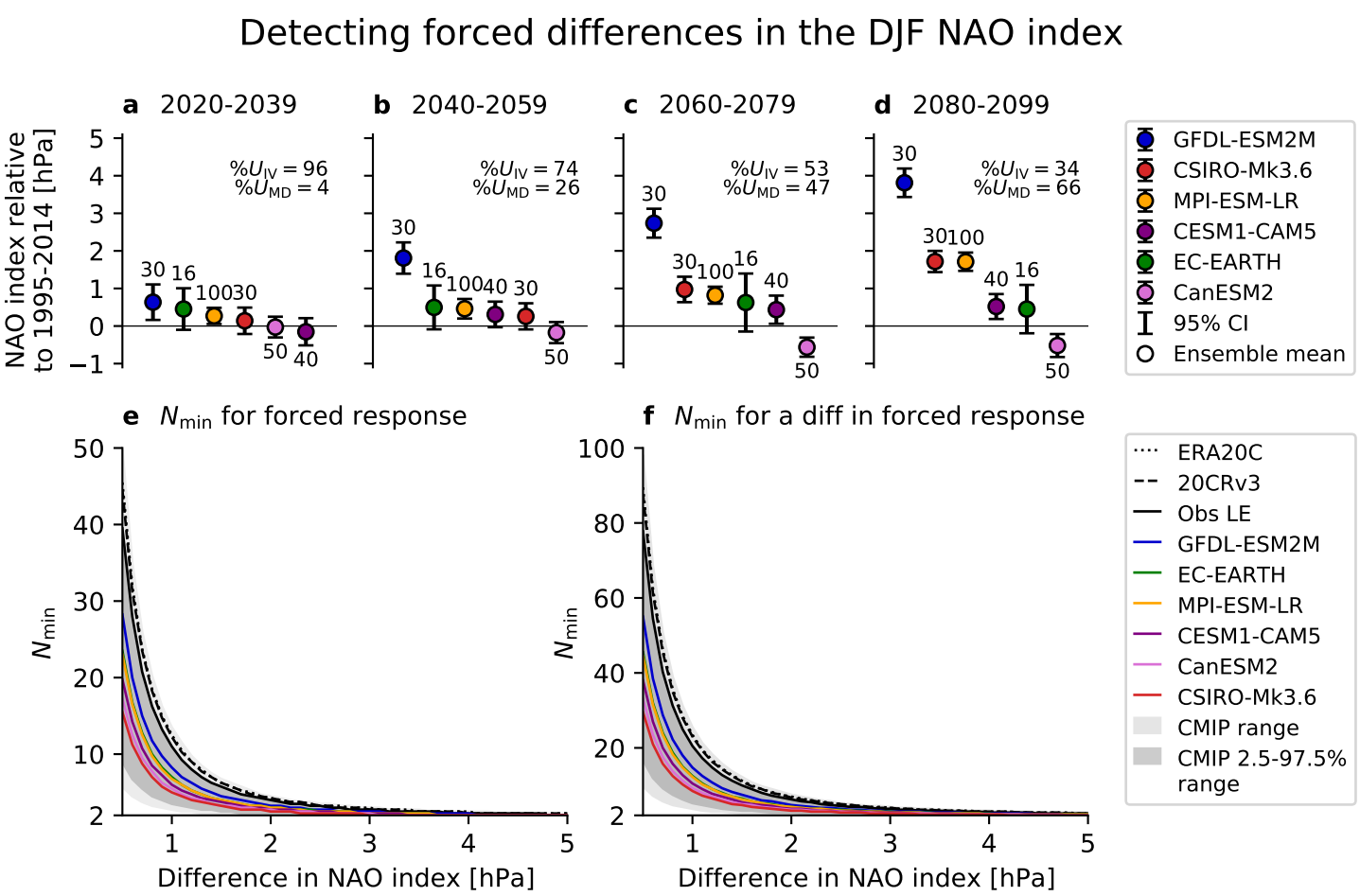
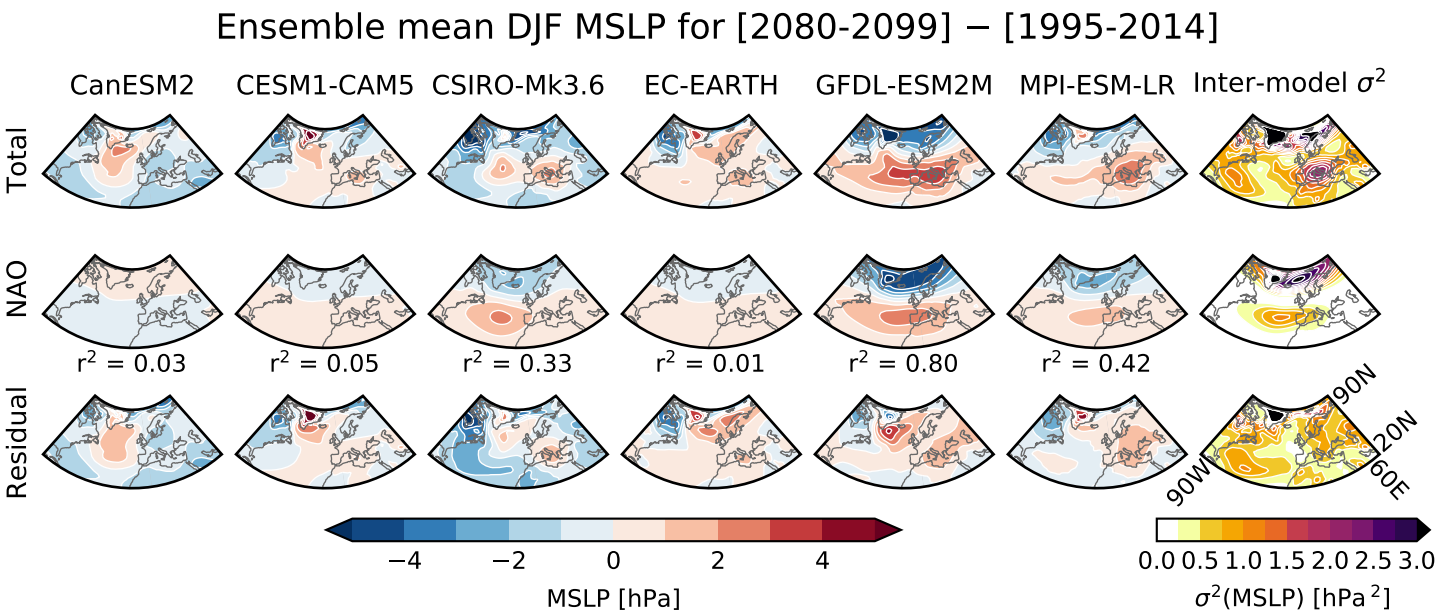

Detecting forced differences in the DJF NAO index
a 2020-2039
b 2040-2059
c 2060-2079
d 2080-2099
NAO index relative to 1995-2014 [hPa]
%U_{IV} = 96
%U_{MD} = 4
%U_{IV} = 74
%U_{MD} = 26
%U_{IV} = 53
%U_{MD} = 47
%U_{IV} = 34
%U_{MD} = 66
GFDL-ESM2M
CSIRO-Mk3.6
MPI-ESM-LR
CESM1-CAM5
EC-EARTH
CanESM2
95% CI
Ensemble mean
e N_{min} for forced response
f N_{min} for a diff in forced response
N_{min}
Difference in NAO index [hPa]
ERA20C
20CRv3
Obs LE
GFDL-ESM2M
EC-EARTH
MPI-ESM-LR
CESM1-CAM5
CanESM2
CSIRO-Mk3.6
CMIP range
CMIP 2.5-97.5% range

Figure 4.


Ensemble mean DJF MSLP for [2080-2099] − [1995-2014]
CanESM2
CESM1-CAM5
CSIRO-Mk3.6
EC-EARTH
GFDL-ESM2M
MPI-ESM-LR
Inter-model σ²
Total
NAO
Residual
r² = 0.03
r² = 0.05
r² = 0.33
r² = 0.01
r² = 0.80
r² = 0.42
90N
120N
90W
60E
−4
−2
0
2
4
MSLP [hPa]
0.0 0.5 1.0 1.5 2.0 2.5 3.0
σ²(MSLP) [hPa²]

Geophysical Research Letters

Supporting Information for

Sources of uncertainty in multi-model large ensemble projections of the winter North Atlantic Oscillation

C. M. McKenna[1] and A. C. Maycock [1]

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

Corresponding author: Christine McKenna (C.McKenna1@leeds.ac.uk)

Contents of this file

Introduction

This document contains additional text, figures, and tables that provide more technical detail on the methods/datasets used and investigate the sensitivity of the results to our methodological choices. Text S1 provides more detail on the uncertainty decomposition method of Maher et al. (2021). Text S2 explains how N_{min} is calculated in Figure 3. Text S3 explains how internal variability (IV) in the DJF NAO index was quantified for each model and observation-based dataset used. Figure S1 shows the historical NAO patterns used in Figures 2 and 4 to decompose an MSLP anomaly map into an NAO-congruent part and a residual. Figure S2 shows the effect of including the EA pattern in this decomposition for Figure 2. Figure S3 shows the effect of adjusting the model-based estimates of IV used in Figure 3a-d to an observation-based estimate of IV. Figures S4 and S5 are versions of Figures 2 and 4, respectively, but for the zonal wind at 850 hPa. Tables S1 and S2 respectively provide a detailed list of the MMLEA model simulations and CMIP5/6 model simulations used in the study.

Text S1. Separating uncertainty into parts due to IV and model structural differences

The total uncertainty (U) in projections of the DJF NAO index (X) across the MMLEA models is separated into a part due to IV (U_{IV}) and a part due to model structural differences (U_{MD}) using the method of Maher et al. (2021). This method is described in detail below.

The projected change in X in a single ensemble member (i) of a single MMLEA model (m) is given by:

$$\Delta X_{m,i} = \bar{X}_{m,i,\mathrm{fut}} - \bar{X}_{m,i,\mathrm{pres}}$$

where overbars indicate a time mean over a future (fut) or near-present-day (pres) 20-year epoch. The forced response in X in a single model (m) is given by the ensemble mean projected change:

$$\Delta X_{m,\mathrm{F}} = \frac{1}{N_m} \sum_{i=1}^{N_m} \Delta X_{m,i}$$

where N_m is the ensemble size for the model. The spread in ΔX across a model (m) due to IV is calculated as the inter-member standard deviation of the projected change:

$$\sigma(\Delta X_m) = \sqrt{\frac{1}{N_m - 1} \sum_{i=1}^{N_m} (\Delta X_{m,i} - \Delta X_{m,F})^2} \quad .$$

The uncertainty in ΔX due to IV (U_{IV}) is then given by the average of the IV across the models:

$$U_{\mathrm{IV}} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \sigma^2(\Delta X_m)}$$

where M is the number of MMLEA models.

The MMM forced response in ΔX for the MMLEA models is calculated as the mean of the forced responses for each model:

$$\Delta X_{\mathrm{F}} = \frac{1}{M} \sum_{m=1}^{M} \Delta X_{m,\mathrm{F}} \quad .$$

The variance in the forced response across the models is then estimated as:

$$\sigma_{\mathrm{F}}{}^2 = D^2 - E^2$$

where D^2 is the variance in the ensemble means:

$$D^2 = \frac{1}{M-1} \sum_{m=1}^{M} (\Delta X_{m,\mathrm{F}} - \Delta X_{\mathrm{F}})^2$$

and E^2 removes the contribution of IV to the variance in the ensemble means. E^2 is equal to the average mean squared error of the models:

$$E^2 = \frac{1}{M} \sum_{m=1}^{M} \frac{\sigma^2 (\Delta X_m)}{N_m} \quad .$$

The uncertainty in ΔX due to model structural differences (U_{MD}) is then estimated as:

$$U_{\mathrm{MD}} = \sqrt{{\sigma_{\mathrm{F}}}^2} \quad .$$

We quantify the contribution of U_{MD} and U_{IV} to the total uncertainty in projections (U), by calculating the percentage variance contribution of each ($\%U_{\mathrm{MD}}$ and $\%U_{\mathrm{IV}}$) to the sum of U_{MD} and U_{IV}. To estimate the contributions of U_{MD} and U_{IV} to real-world uncertainty in the future NAO response, the model-based estimate of U_{IV} is replaced with an observation-based estimate of IV. Specifically, the IV in each MMLEA model, $\sigma(\Delta X_m)$, is replaced with an estimate of IV from Obs LE calculated as described in Text S3. Note that there are minimal differences to the results when using 20CRv3 or ERA20C.

Text S2. Calculation of N_{min}

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 the method of Screen et al. (2014).

First, we calculate the Student's t-statistic, t, for many different ensemble sizes, N, using a Student's t-test for a difference of means (von Storch & Zwiers, 1999):

$$t = \frac{X}{\sigma\sqrt{2/N}}$$

where σ is the standard deviation of 20-year epoch means due to IV (Text S3). Note it is assumed that σ is constant for all N, which Screen et al. (2014) show is a reasonable assumption.

Second, we define the difference, X, as statistically significant when $t \geq t_c$, where t_c is the cutoff value of the Student's t-distribution for a two-sided p-value of 0.025 and $2N$-2 degrees of freedom. N_{min} is the smallest value of N for which this is satisfied. Hence, N_{min} is calculated by re-arranging the Student's t-test, and replacing t with t_c and N with N_{min}:

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

Text S3. Methods for calculating IV

All methods described below are applied to the DJF NAO index.

For each CMIP5/6 model, the IV in 20-year epoch means (Table S2) is calculated as the standard deviation in non-overlapping 20-year epoch means from the piControl simulation. This is multiplied by the square root of 2 when a difference in 20-year epoch means is of interest; this assumes the two 20-year epochs are independent and have the same variance (Collins et al., 2013). As in Collins et al. (2013), the median IV across all models is used for the MMM. Non-overlapping 20-year epochs are used to ensure each sample is independent.

The IV in 20-year epoch means for each MMLEA model (Table S1) is calculated as the inter-member standard deviation of a 20-year epoch mean, where this is pooled (i.e., averaged) for all possible 20-year epochs over 1951-2099. The same method is used for Obs LE, but over the period 1922-2014. For ERA20C and 20CRv3, we use the standard deviation of all possible 20-year epoch means over 1901-2010; given the limited temporal extent of these records, non-overlapping segments could not be used. For consistency between all datasets considered, the IV for a difference in 20-year epoch means is the IV in 20-year epoch means multiplied by the square root of 2. In all cases we assume the IV is constant in time; analysing timeseries of the inter-member standard deviation in 20-year epoch means for the MMLEA models suggests this is a reasonable assumption.

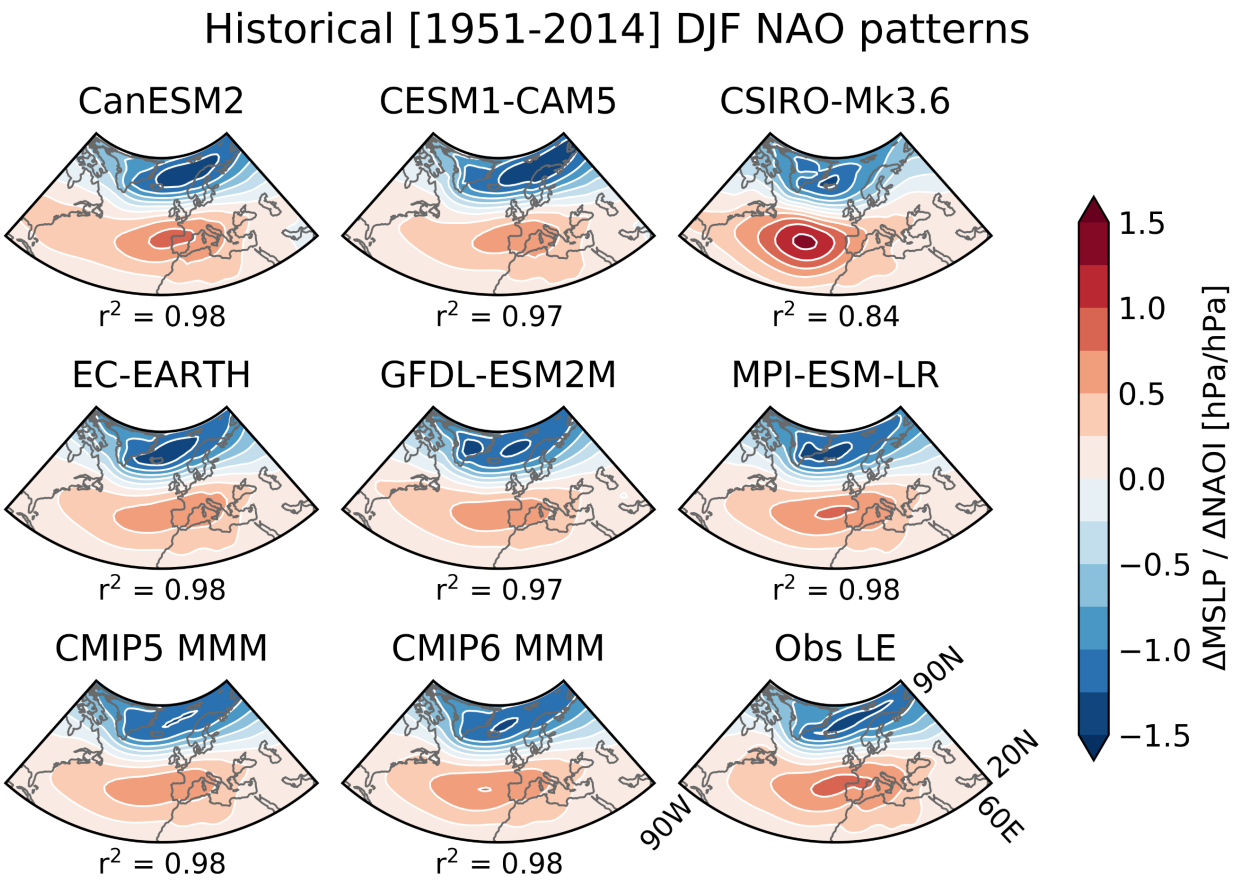


Figure S1. Historical [1951-2014] DJF NAO patterns for the MMLEA models, CMIP5/6 MMM and an observation-based dataset. Shading shows the change in MSLP (hPa) for a 1 hPa positive change in NAO index. Patterns for each CMIP5/6 model and Obs LE are calculated in the same way as for the MMLEA models (Section 2.2). Ensemble means are used to define the patterns to minimise uncertainty in the NAO pattern due to IV (e.g., see Simpson et al., 2020). Obs LE is used for the observation-based NAO pattern because it is designed to be less affected by sampling issues; note that there are minimal differences when using 20CRv3 or ERA20C, or when using a longer historical period. r^2 is the squared area-weighted pattern correlation between the modelled and observation-based NAO patterns. Largely, the modelled and observation-based patterns are highly correlated. The southern centre of action, however, is generally weaker in the models and in CSIRO-Mk3.6 the pattern is westward shifted. There is little improvement in the CMIP6 MMM compared to the CMIP5 MMM.

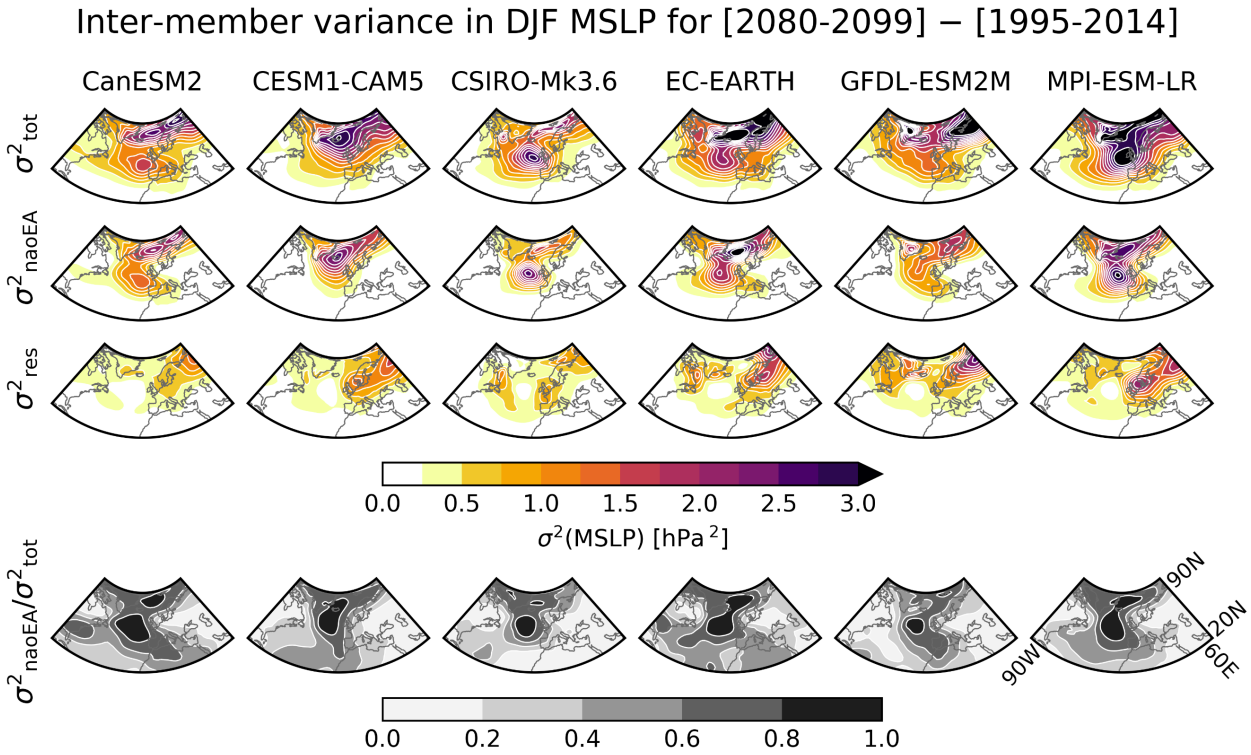


Figure S2. Same as Figure 2, but with the EA pattern included in the regression. [Top row] Total variance (σ^2_{tot}); [Second row] Variance explained by the NAO and EA (σ^2_{naoEA}); [Third row] Residual variance (σ^2_{res}); [Bottom row] Proportion of total variance explained by the NAO and EA. σ^2_{naoEA} is obtained through multivariate regression at each grid-point of the total inter-member spread in MSLP on the spread in NAO-congruent MSLP and spread in EA-congruent MSLP. σ^2_{res} is the variance in the residuals of this regression. The EA pattern is characterised by a monopole in MSLP over the mid-latitude North Atlantic ocean (Barnston & Livezey, 1987; Moore et al., 2011; Wallace & Gutzler, 1981). Following Moore et al. (2011), the EA index is calculated as the anomalous MSLP in the nearest gridbox to (52.5N, 27.5W). The EA-congruent part of the MSLP is obtained using the same procedure as for the NAO-congruent part (Section 2.2).

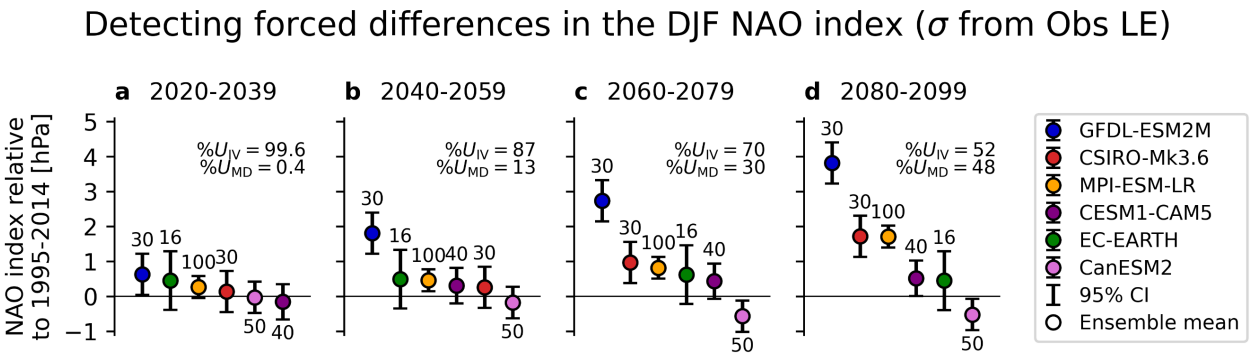


Figure S3. Same as Figure 3a-d, but with confidence intervals calculated by replacing the model-based estimates of IV with observation-based estimates of IV. Obs LE is used for the observation-based IV estimate because it is designed to be less affected by sampling issues; note that there are minimal differences when using 20CRv3 or ERA20C. IV is estimated as described in Text S3. $\%U_{IV}$ and $\%U_{MD}$ are defined as described in Text S1 using Obs LE to estimate IV.

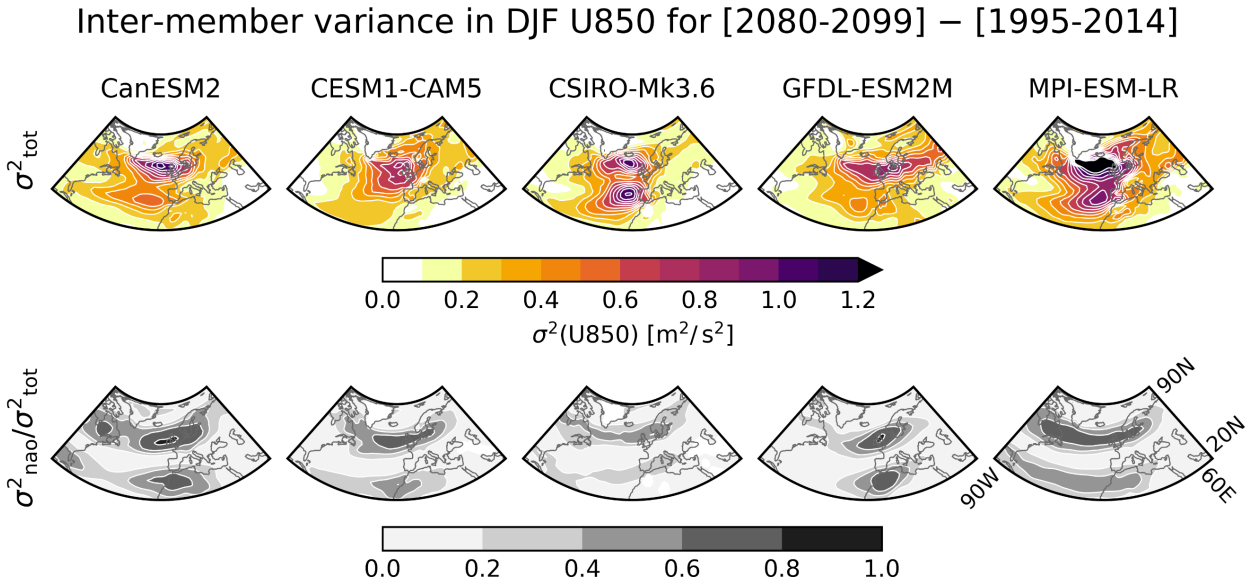


Figure S4. Same as top and bottom rows of Figure 2, but for U850. [Top row] Total variance (σ^2_{tot}); [Bottom row] Proportion of total variance explained by the NAO (σ^2_{nao} / σ^2_{tot}). σ^2_{nao} is obtained by regressing the total inter-member spread in U850 on the spread in NAO-congruent U850 at each grid-point. The NAO-congruent part of U850 is obtained using the same procedure as for MSLP (Section 2.2), but with the historical NAO pattern constructed by regressing historical timeseries of U850 at each grid-point onto the NAO index timeseries. EC-EARTH is omitted from the analysis because there was no three-dimensional zonal wind data available.

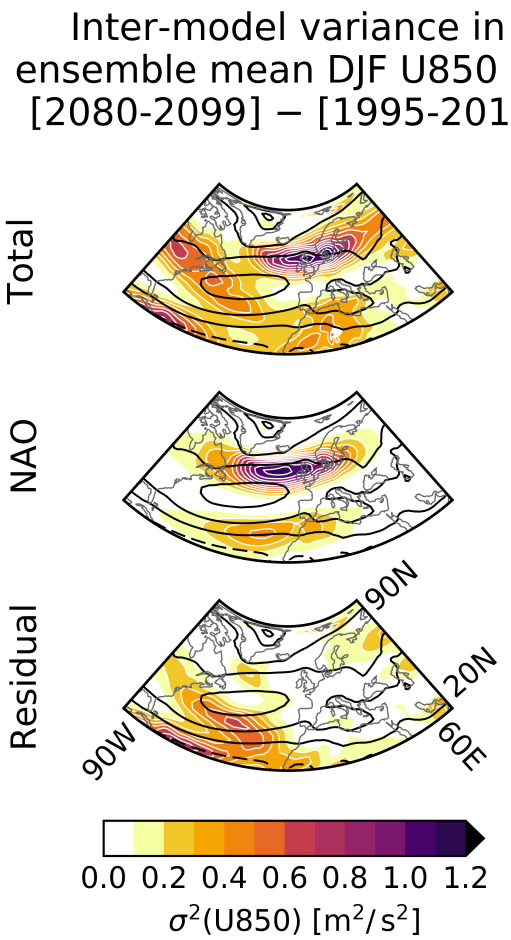


Figure S5. Same as far right column of Figure 4, but for U850. [Top] Total; [Middle] NAO-congruent part; [Bottom] Residual. The NAO-congruent part of U850 is obtained using the same procedure as for MSLP (Section 2.2), but with the historical NAO pattern constructed by regressing historical timeseries of U850 at each grid-point onto the NAO index timeseries. EC-EARTH is omitted from the analysis because there was no three-dimensional zonal wind data available. Note that similar results are obtained for the far right column of Figure 4 when EC-EARTH is removed. Black contours show the MMM near-present-day (1995-2014) U850 climatology with intervals of 10 m/s.

Table S1. List of MMLEA models with historical and RCP8.5 simulations. IV is for 20-year means of the DJF NAO index over 1951-2099 (see Text S3 for details). In all MMLEA models, this IV is underestimated compared to observation-based datasets (1.1 hPa, 1.2 hPa, and 1.2 hPa in Obs LE, 20CRv3, and ERA20C, respectively). Note that while the MMLEA does contain an ensemble for GFDL-ESM2M, there is no three-dimensional zonal wind data available for this model. We therefore use a similar 30 member ensemble from the Princeton Large Ensemble Archive (Schlunegger et al., 2019), which has three-dimensional zonal wind data available. The NAO index and MSLP results are very similar for the two ensembles.

Model	Modelling Centre	CMIP generation	Years	No. of members	IV (hPa)	Reference
CanESM2	CCCma	CMIP5	1950-2100	50	0.72	Kirchmeier-Young et al. (2017)
CESM1-CAM5	NCAR	CMIP5	1920-2100	40	0.77	Kay et al. (2015)
CSIRO-Mk3.6	CSIRO	CMIP5	1850-2100	30	0.68	Jeffrey et al. (2013)
EC-EARTH	EC-Earth Consortium	CMIP5	1860-2100	16	0.85	Hazeleger et al. (2010)
GFDL-CM3	GFDL	CMIP5	1920-2100	20	0.77	Sun et al. (2018)
GFDL-ESM2M	GFDL	CMIP5	1950-2100	30	0.93	Rodgers et al. (2015); Schlunegger et al. (2019)
MPI-ESM-LR	MPI	CMIP5	1850-2099	100	0.84	Maher et al. (2019)

Table S2. List of CMIP5/CMIP6 models with piControl, historical and RCP8.5/SSP5-8.5 simulations. Numerical labels are for bars in Figure 1. Models are ranked in order of magnitude of IV in 20-year means of the DJF NAO index from the piControl simulations (see Text S3 for details), where rank 1 has the highest IV and rank 75 has the lowest. This enables each model to be located in the grey plumes of Figure 3e-f. In most models the IV is underestimated compared to observation-based datasets (respectively 1.11 hPa, 1.18 hPa, and 1.20 hPa in Obs LE, 20CRv3, and ERA20C). Note that for CMIP5 models that are also MMLEA models, the IV magnitudes listed here do not necessarily match those in Table S1. For example, based on the piControl simulations CESM1-CAM5 has a very low IV, but based on the MMLEA simulations it has an average IV. This likely reflects that the piControl IV is calculated from a relatively short simulation (319 years) with only 15 independent samples of 20-year means, while there are 40 independent ensemble members for the MMLEA simulations. It could also be that there are differences in the magnitude of IV between the pre-industrial state and historical/RCP8.5 state, but this cannot be determined with the limited piControl simulation length.

Label	Model	Modelling Centre	CMIP generation	piControl length (years)	Number of historical/ RCP/SSP members	IV (hPa)	IV rank
1	ACCESS1.0	CSIRO-BOM	CMIP5	500	1	0.62	66
2	ACCESS1.3		CMIP5	500	1	0.84	23
3	BCC-CSM1.1	BCC	CMIP5	500	1	0.77	40
4	BCC-CSM1.1-M		CMIP5	400	1	0.86	18
5	BNU-ESM	BNU	CMIP5	559	1	1.14	2
6	CanESM2	CCCma	CMIP5	996	5	0.68	56
7	CCSM4	NCAR	CMIP5	1051	6	0.83	27
8	CESM1-BGC	NSF-DOE-NCAR	CMIP5	500	1	0.89	14
9	CESM1-CAM5		CMIP5	319	3	0.52	72
10	CESM1-WACCM		CMIP5	200	3	0.45	74
11	CMCC-CESM	CMCC	CMIP5	277	1	1.03	5

Label	Model	Modelling Centre	CMIP generation	piControl length (years)	Number of historical/ RCP/SSP members	IV (hPa)	IV rank
12	CMCC-CM	CMCC	CMIP5	330	1	0.67	59
13	CMCC-CMS		CMIP5	500	1	0.80	32
14	CNRM-CM5	CNRM-CERFACS	CMIP5	850	5	0.78	39
15	CSIRO-Mk3.6.0	CSIRO-QCCCE	CMIP5	500	10	0.67	60
16	EC-EARTH	ICHEC	CMIP5	451	8	0.80	34
17	FGOALS-g2	LASG-CESS	CMIP5	700	1	0.64	63
18	FIO-ESM	FIO	CMIP5	800	3	0.81	30
19	GFDL-CM3	NOAA-GFDL	CMIP5	500	1	0.66	61
20	GFDL-ESM2G		CMIP5	500	1	0.93	11
21	GFDL-ESM2M		CMIP5	500	1	0.74	47
22	GISS-E2-H	NASA-GISS	CMIP5	780	2	0.63	65
23	GISS-E2-H-CC		CMIP5	251	1	0.37	75
24	GISS-E2-R		CMIP5	850	2	0.80	31
25	GISS-E2-R-CC		CMIP5	251	1	0.75	44
26	HadGEM2-CC	MOHC	CMIP5	240	3	0.84	26
27	HadGEM2-ES		CMIP5	576	4	0.86	20
28	INM-CM4	INM	CMIP5	500	1	0.74	46
29	IPSL-CM5A-LR	IPSL	CMIP5	1000	4	0.82	28

Label	Model	Modelling Centre	CMIP generation	piControl length (years)	Number of historical/ RCP/SSP members	IV (hPa)	IV rank
30	IPSL-CM5A-MR	IPSL	CMIP5	300	1	0.61	67
31	IPSL-CM5B-LR		CMIP5	300	1	1.25	1
32	MIROC-ESM	MIROC	CMIP5	630	1	0.79	35
33	MIROC-ESM-CHEM		CMIP5	255	1	0.71	51
34	MIROC5		CMIP5	670	3	0.55	71
35	MPI-ESM-LR	MPI-M	CMIP5	1000	3	0.93	10
36	MPI-ESM-MR		CMIP5	1000	1	0.80	33
37	MRI-CGCM3	MRI	CMIP5	500	1	0.94	9
38	NorESM1-M	NCC	CMIP5	501	1	0.78	38
39	NorESM1-ME		CMIP5	252	1	1.03	7
40	ACCESS-CM2	CSIRO-ARCCSS	CMIP6	500	1	0.84	24
41	ACCESS-ESM1.5	CSIRO	CMIP6	900	1	0.60	69
42	AWI-CM1.1-MR	AWI	CMIP6	500	1	0.78	37
43	BCC-CSM2-MR	BCC	CMIP6	600	1	1.09	3
44	CAMS-CSM1.0	CAMS	CMIP6	500	1	0.92	12
45	CanESM5	CCCma	CMIP6	1000	25	0.86	21
46	CanESM5-CanOE			501	1	0.75	45

Label	Model	Modelling Centre	CMIP generation	piControl length (years)	Number of historical/ RCP/SSP members	IV (hPa)	IV rank
47	CESM2	NCAR	CMIP6	1200	1	0.99	8
48	CESM2-WACCM		CMIP6	499	3	0.81	29
49	CIESM	THU	CMIP6	500	1	0.69	54
50	CMCC-CM2-SR5	CMCC	CMIP6	500	1	0.76	43
51	CNRM-CM6.1	CNRM-CERFACS	CMIP6	500	6	0.85	22
52	CNRM-CM6.1-HR		CMIP6	300	1	0.63	64
53	CNRM-ESM2.1		CMIP6	500	5	1.06	4
54	EC-Earth3-Veg	EC-Earth-Consortium	CMIP6	500	1	0.76	42
55	FGOALS-f3-L	CAS	CMIP6	561	1	0.79	36
56	FGOALS-g3		CMIP6	700	1	0.68	57
57	FIO-ESM2.0	FIO-QLNM	CMIP6	575	1	0.61	68
58	GFDL-CM4	NOAA-GFDL	CMIP6	500	1	0.70	53
59	GFDL-ESM4		CMIP6	500	1	0.90	13
60	HadGEM3-GC3.1-LL	MOHC	CMIP6	500	4	0.67	58
61	HadGEM3-GC3.1-MM		CMIP6	500	4	0.86	19
62	INM-CM4.8	INM	CMIP6	531	1	0.56	70
63	INM-CM5.0		CMIP6	1201	1	0.64	62

Label	Model	Modelling Centre	CMIP generation	piControl length (years)	Number of historical/ RCP/SSP members	IV (hPa)	IV rank
64	IPSL-CM6A-LR	IPSL	CMIP6	2000	3	0.88	15
65	KACE1.0-G	NIMS-KMA	CMIP6	450	1	0.87	16
66	KIOST-ESM	KIOST	CMIP6	500	1	0.72	49
67	MIROC-ES2L	MIROC	CMIP6	500	1	0.51	73
68	MIROC6		CMIP6	800	3	0.70	52
69	MPI-ESM1.2-HR	MPI-M	CMIP6	500	1	0.86	17
70	MPI-ESM1.2-LR		CMIP6	1000	1	0.69	55
71	MRI-ESM2.0	MRI	CMIP6	701	1	0.84	25
72	NESM3	NUIST	CMIP6	500	1	0.77	41
73	NorESM2-LM	NCC	CMIP6	501	1	1.03	6
74	NorESM2-MM		CMIP6	500	1	0.71	50
75	UKESM1.0-LL	MOHC	CMIP6	1880	5	0.73	48