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

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:

- Around 66% of spread in North Atlantic Oscillation (NAO) projections is due to model structural differences and 34% to internal variability
- NAO explains a large part of spread in North Atlantic circulation projections due to internal variability, but less due to model differences
- Quantify time horizons and ensemble sizes required to detect a forced NAO response, and model differences in this, from internal variability

Abstract

There is large spread in projections of the winter North Atlantic circulation. Coupled Model Intercomparison Project archives typically provide a few ensemble members per model, rendering it difficult to quantify the contributions of reducible model structural uncertainty and irreducible internal variability (IV) to the spread in projections. We use the Multi-Model Large Ensemble Archive to estimate that model structural differences explain two-thirds of the spread in late 21st century North Atlantic Oscillation (NAO) projections. 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 circulation projections due to IV except in the central North Atlantic. Conversely, spread in North Atlantic circulation projections due to model differences is largely unexplained by the NAO. Therefore, improving understanding of the NAO may not help to constrain the reducible uncertainty in North Atlantic circulation 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 European 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 projections of the NAO. 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 circulation projections in and around the North Atlantic basin, suggesting other dynamical processes must be understood.

1 Introduction

The North Atlantic circulation has a strong influence on European regional climate and is often associated with severe weather events (Buehler et al., 2011; Hurrell et al., 2003). For a given scenario of future greenhouse gas and aerosol forcing, 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 is partly a consequence of 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 due to model structural differences versus internal 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 to sample the effects of IV. This makes it difficult to quantify the contributions of model structural uncertainty and IV to the spread in projections, without making simplifying assumptions such as that IV in a stationary pre-industrial climate can be used to approximate 21st century IV (Hawkins & Sutton, 2009).

This study aims to advance understanding of the roles of model structural error and IV for projections of the North Atlantic circulation. 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 latitudinal position 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, as well as providing guidance for the design of future model experiments.

2 Methods

2.1 Datasets

The MMLEA contains large 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; Sun et al., 2018). This study uses historical and Representative Concentration Pathway (RCP)8.5 simulations from the MMLEA models for the common period 1950-2099. We focus on RCP8.5 because only a small subset of the models is available for other RCPs. GFDL-CM3 is discarded from the MMLEA analysis, since it has a similar formulation to GFDL-ESM2M and gives similar results; GFDL-ESM2M was kept because it has a larger number of ensemble members available. All analyses use monthly mean sea level pressure (MSLP) data averaged over December to February. As in 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 here 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 scenario simulations from 36 CMIP6 models (Table S2; Eyring et al., 2016). The forcing scenarios changed in CMIP6, where SSP5-8.5 is most similar in terms of 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) which should be borne in mind when comparing results.

In general, a small number of ensemble members are available for the CMIP5/6 simulations, so we estimate IV using the pre-industrial control (piControl) runs. Model drift is eliminated in these runs by subtracting the long-term linear trend. Various observation-based datasets are used to evaluate the spread in model projections in the context of observed IV. Since our focus is on multi-decadal timescales, we use two centennial-scale reanalysis datasets: the National Oceanic and Atmospheric Administration Twentieth Century Reanalysis version 3 (20CRv3; Compo et al., 2011; Slivinski et al., 2019) and the European Centre for Medium-Range Weather Forecasts Twentieth 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 consider the full extent of Obs LE (1921-2014) and use 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 trend in MSLP (McKinnon & Deser, 2018).

All model and observation-based data were regridded to a common 2° horizontal grid using bilinear interpolation; this 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). 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. 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). Multiplying the NAO index response by the historical NAO pattern gives the NAO-congruent part of the MSLP response.

2.3 Statistical methods

In each MMLEA model, uncertainty due to IV is estimated as the spread across ensemble members with the same external forcing and different initial conditions (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 t-statistic for p=0.025 and $N-1$ degrees of freedom, σ is the inter-ensemble 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 re-arrange a two-sided Student's t-test for a difference of means (von Storch & Zwiers, 1999):

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

where 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 CMIP5, CMIP6 and MMLEA models. For both CMIP5 and CMIP6 ensembles, the multi-model mean (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. Furthermore, while some models have large positive NAO anomalies that exceed their modelled range of IV, most modelled anomalies are small compared to IV. The range of NAO anomalies is 7 hPa in CMIP5 and 6 hPa in CMIP6, where <90% of models agree on sign (79% in CMIP5 and 86% in CMIP6). This spread is comparable to the range of observed NAO variability (grey box; Figure 1).

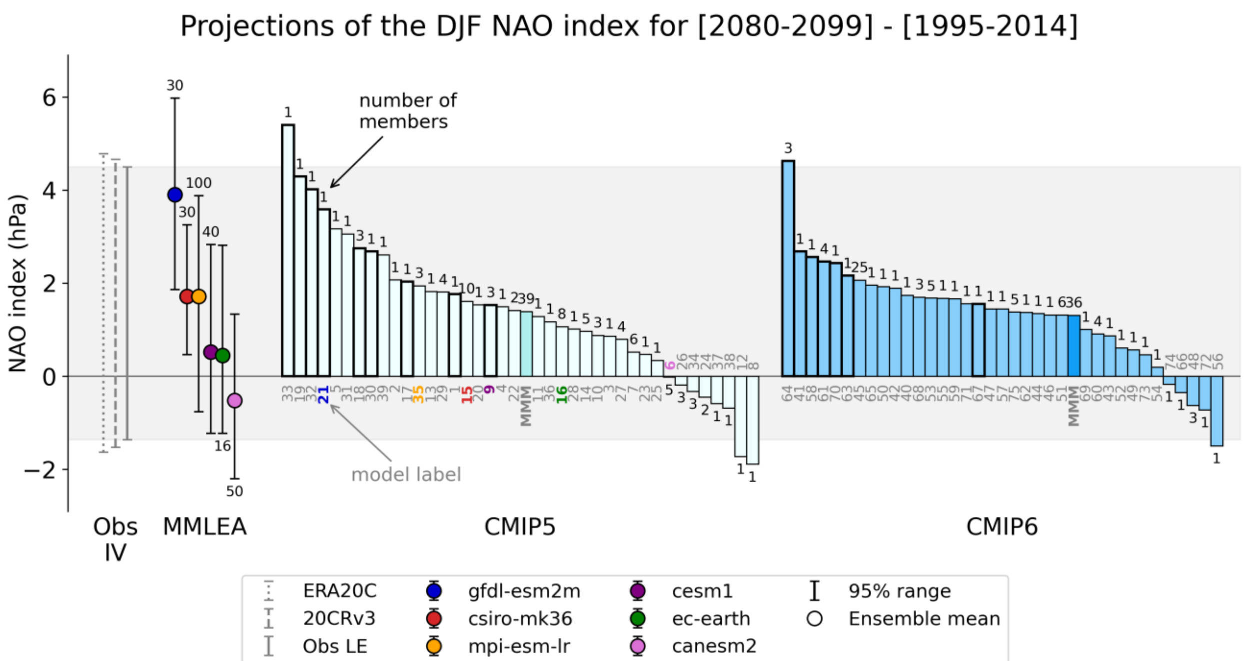

Figure 1. **Projections of the DJF NAO index for [2080-2099]−[1995-2014] in the CMIP5, CMIP6 and MMLEA models.** For CMIP5/6 models, ensemble means are shown if more than one ensemble member is available. Bold bar outlines indicate a CMIP5/6 model response that is larger than 2 standard deviations of IV (Text S2). Whiskers for MMLEA models indicate the 2.5-97.5% range of responses across the ensemble members. Grey whiskers show the 2.5-97.5% range of 10^5 differences in 20-year epoch means of different observation-based records selected by randomly resampling with replacement. Grey shaded box denotes this range for Obs LE. To compare the inter-model spread with the observation-based estimates of IV, the latter are shifted by the CMIP6 MMM anomaly.

Given that 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. The MMLEA models suggest there are indeed substantial structural differences in the forced response between models of up to 5 hPa (coloured circles; Figure 1). 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 due to internal variability

In several MMLEA models, the forced winter NAO response is small compared to 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, in agreement with 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 the observation-based datasets (black and grey whiskers in Figure 1; Tables S1, S2). The model projections may therefore be overconfident: i.e., in the real-world a larger part of the uncertainty in future NAO responses may be due to IV. When model-based estimates of IV are adjusted to an observation-based estimate of IV (Text S1), IV and model structural differences each contribute to around 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 will discuss 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 part (third row). The total uncertainty due to 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 due to IV in north-eastern North America and in most of continental Europe. The NAO contributes to a large proportion (>50%; Figure 2, bottom row) of the uncertainty in MSLP projections at high latitudes. It also contributes a substantial portion (up to 50%) of the uncertainty in Southern Europe, the Mediterranean, and north Africa. However, the remaining uncertainty in projections due to IV is largely not NAO-congruent. In the central Atlantic and western Europe this uncertainty 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).

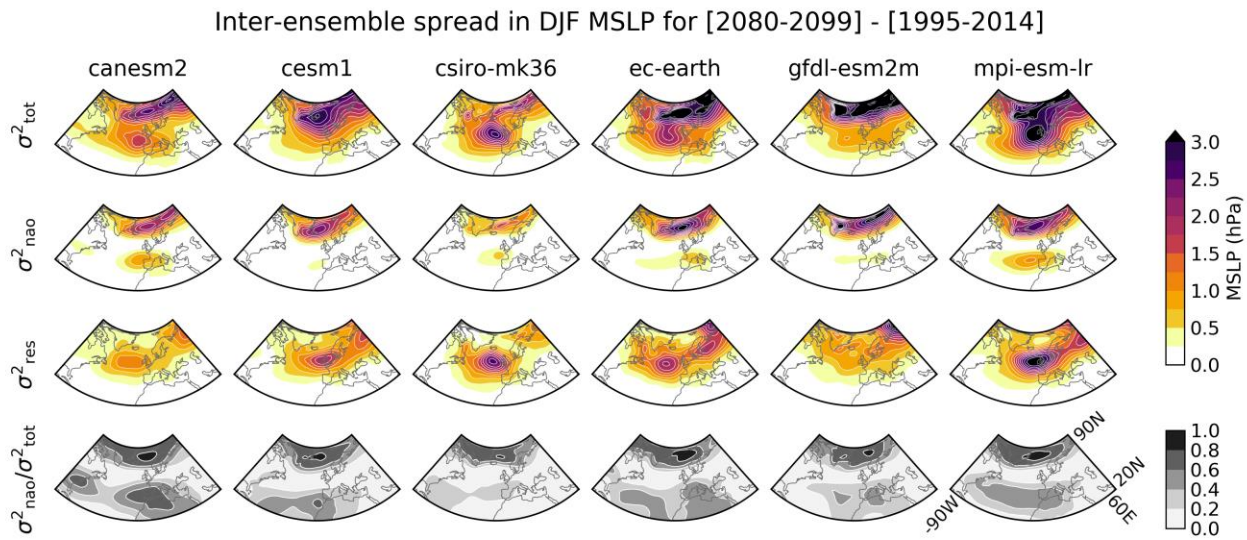

Figure 2. **Inter-ensemble spread in projections of DJF MSLP for [2080-2099]−[1995-2014] for each MMLEA model.** [Top row] Total variance (σ^2_{tot}); [Second row] Variance explained by NAO (σ^2_{nao}); [Third row] Residual variance (σ^2_{res}); [Bottom row] Proportion of total variance explained by NAO. σ^2_{nao} is obtained by regressing the total 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.

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 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 its small response and smaller ensemble size (Figure 3c). Regarding detection of model differences in response, in the early 21st century only the relatively large positive NAO response in GFDL-ESM2M is robustly distinguishable from the other models (Figure 3a). By 2060-2079, the only model with a negative NAO response (CanESM2) 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 model structural differences in the forced response are only detectable by 2080-2099; this is also when $\%U_{MD}$ first dominates over $\%U_{IV}$ (Figure 3a-d). This is largely still the case when the model-based estimates of IV are adjusted to an observation-based estimate of IV (Figure S3).

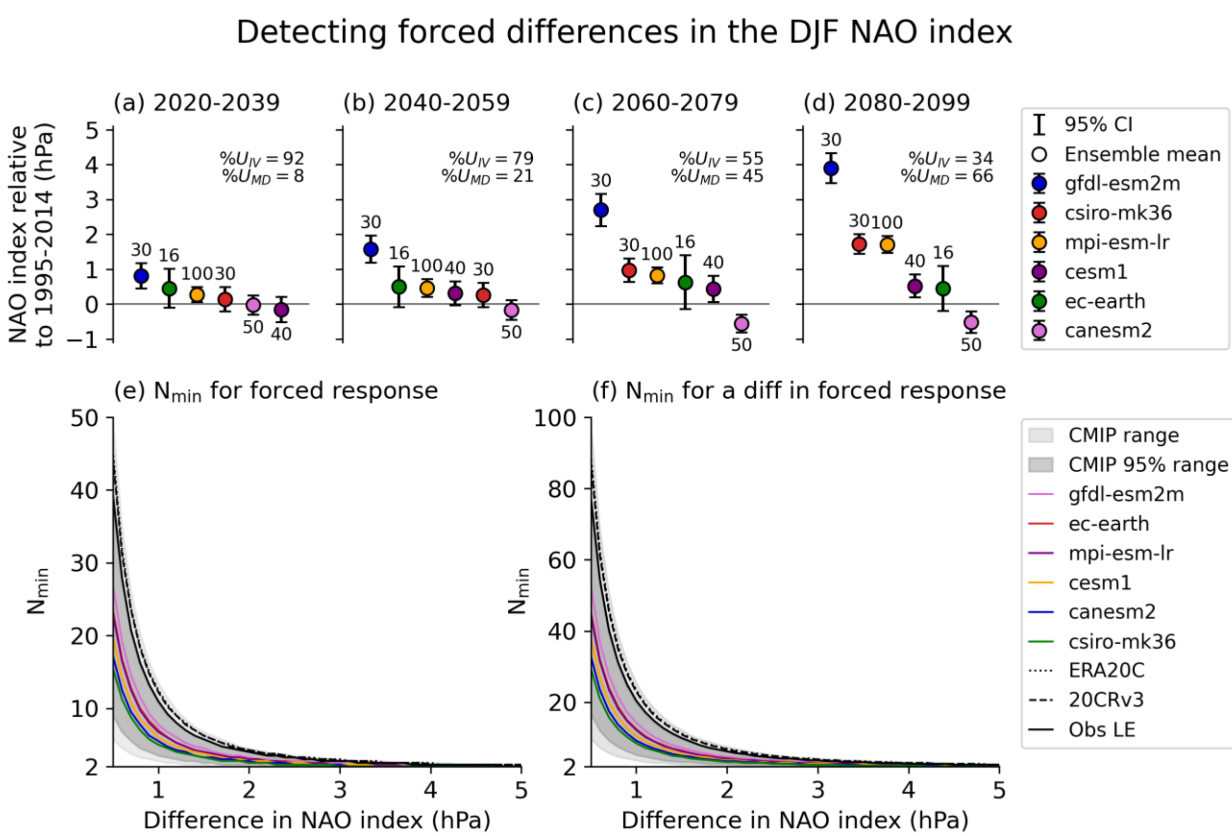

Figure 3. **Detecting a forced response in DJF NAO index and differences in this response between models. a-d**, NAO anomalies in MMLEA models for 20-year epochs relative to 1995-2014. Whiskers are 95% confidence intervals and numbers indicate ensemble size. Section 2.3 defines $\%U_{IV}$ and $\%U_{MD}$. **e,** N_{min} required to detect a forced NAO response of a given magnitude at the 95% confidence level based on estimates of IV from models and observation-based datasets (Text S2). **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.

These findings are generalised by calculating 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 that magnitude in one model (Figure 3e-f). This explains why structural differences in forced responses emerge from IV later. To detect a small NAO index response of 0.5 hPa – the typical limit of model responses in the early to mid 21$^{\mathrm{st}}$ century (Figure 3a-b) – requires N_{min} = 10, 20 or 40 for a model with low (2.5$^{\mathrm{th}}$ percentile), median, or high (97.5$^{\mathrm{th}}$ percentile) IV based on the CMIP5/6 multi-model ensembles. For context, the interannual standard deviation in the DJF NAO index is around 3.7 hPa on average 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 estimates of IV. All subsequent results use the high estimate of IV as this provides an upper bound on N_{min}. To detect larger NAO responses of 1 hPa and 2 hPa – typical of MMLEA responses in the late 21$^{\mathrm{st}}$ century (Figure 3c-d) – no more than 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 around 4 hPa in 2080-2099 (Figure 3d) require 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. This suggests that in the context of more realistic estimates of IV, 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 the NAO explains differences in the forced response of North Atlantic circulation. 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 dipole in pressure anomalies between high and low latitudes, while this is not the case in CanESM2, CESM1 and EC-EARTH. This can be attributed to inter-model spread in the NAO response (Figure 4, middle row and far right column). However, while the NAO explains a substantial portion of the forced North Atlantic MSLP response in some models (e.g., 81% in GFDL-ESM2M), it explains almost none of it in 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 are associated with the majority of the inter-model spread in the forced MSLP response (Figure 4; far right column). This is particularly the case for the large spread over Greenland, eastern North America, and central Europe.

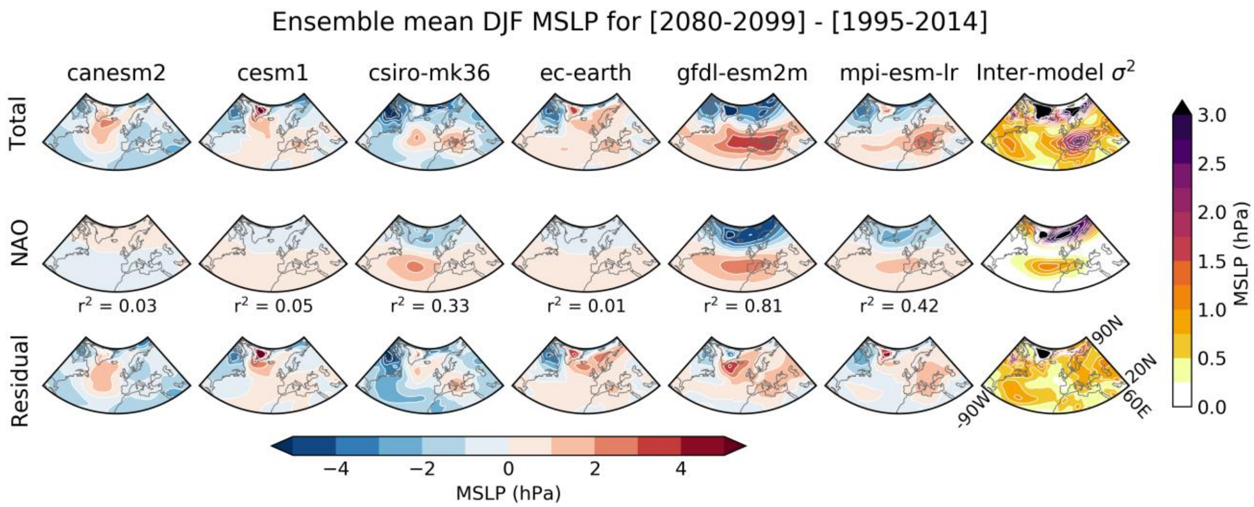

Figure 4. **Projections of the forced response of DJF MSLP for [2080-2099]−[1995-2014], shown for each MMLEA model and in terms of the inter-model variance.** [Top] Total; [Middle] NAO-congruent part; [Bottom] Residual. r^2 is the area-weighted pattern correlation between the total response and the NAO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

4 Discussion and conclusions

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

Firstly, 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, this study is the first to quantify these components of the uncertainty using large initial-condition ensembles performed by a subset of CMIP5 models. The relevance of this separation for the real-world relies on models correctly reproducing the observed magnitude of low frequency IV and forced response in the NAO. We find the former is generally underestimated in models in agreement with previous studies, but note that the latter may also be underestimated; this will be discussed shortly.

Secondly, 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. While individual MMLEA models do have larger NAO responses that are distinct from IV and from other models earlier in the century, this is generally not the case. This highlights a benefit of using the new MMLEA archive in this study, whereas previous studies have been limited to using a Large Initial-Condition Ensemble from a single model to quantify time horizons for the emergence of a forced circulation response (Deser et al., 2012, 2017).

Thirdly, 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. For example, a typical response (or model difference) of 1-2 hPa over the 21st century requires at most 15-5 (30-10) ensemble members for detection. Even for very large responses (model differences) of around 5 hPa (7 hPa), 2 members is only just enough for detection – meaning that the majority of model responses and differences are non-robust in CMIP5/6 models with only 1 ensemble member. These results provide a useful aid for interpreting NAO projections and designing future model intercomparison experiments.

Finally, we have examined the extent to which the NAO explains the spread in North Atlantic MSLP projections. Regarding spread due to IV, this is large in most regions of the North Atlantic and surrounding land areas, where the NAO explains over 50% of the inter-ensemble 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 due to IV is largely explained by dominant modes of atmospheric variability agrees well with Deser et al. (2012). These results build on the results of Deser et al. (2017), who only analysed the NAO contribution to spread in projections due to IV.

Regarding inter-model spread in the forced North Atlantic MSLP response, while this is largely associated with the NAO at high latitudes and in Southern Europe, the majority of the spread is not NAO-congruent. This suggests that improving understanding of the NAO may not help to constrain the reducible uncertainty in North Atlantic circulation projections. This is somewhat 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 suggests that some model differences may be associated with more local effects (e.g., from orography).

There are some caveats to these results. In particular, models have been shown 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). It is possible that this issue also affects multi-decadal NAO projections, though given the limited temporal extent of the observational record this is difficult to assess. If the magnitude of the forced NAO response was underestimated, this would imply an underestimation of model differences in the forced NAO response and therefore the contribution of the NAO to total 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 in order 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 due to IV and/or model differences. The results of Harvey et al. (2020) 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.

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. 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 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

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 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. http://dx.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

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>2.0.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 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 Meteorological and Oceanographic Journal*, *63*(1), 1–13. 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. (2016). 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., & Marotzke, J. (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. https://doi.org/10.1007/s10584-011-0156-z

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

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

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