The Forced Response and Decadal Predictability of the North Atlantic Oscillation: Nonstationary and Fragile Skills

BO CHRISTIANSEN,a SHUTING YANG,a AND DOMINIC MATTEb,c

a Danish Meteorological Institute, Copenhagen, Denmark
b Physics of Ice, Department of Climate and Earth, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
c Ouranos, Montréal, Québec, Canada

(Manuscript received 12 October 2021, in final form 25 April 2022)

ABSTRACT: We investigate the forced response of the North Atlantic Oscillation (NAO)—calculated as the ensemble mean—in different large ensembles of climate models including simulations with historical forcings and initialized decadal hindcasts. The forced NAO in the CMIP6 historical ensemble correlates significantly with observations after 1970. However, the forced NAO shows an apparent nonstationarity with significant correlations to observations only in the period after 1970 and in the period before 1890. We demonstrate that such apparent nonstationarity can be due to chance even when models and observations are independent. For the period after 1970 the correlation to the observed NAO continues to increase while the amplitude of the forced signal continues to decrease—although both with some signs of saturation—when the ensemble size grows. This behavior can be explained by a simple statistical model assuming a very small signal-to-noise ratio in the models. We find only rather weak evidence that initialization improves the skill of the NAO on decadal time scales. The NAO in the historical ensembles including only natural forcings, well-mixed greenhouse gases, or anthropogenic aerosols show skill that is not significantly different from zero. The same holds for a large single-model ensemble. The skills of these ensembles, except for the well-mixed greenhouse gas ensemble, are also significantly different from the skill of the larger full historical ensemble even though their ensemble sizes are smaller. Taken together, our results challenge the possibility of useful NAO predictions on decadal time scales.

KEYWORDS: Time series; Ensembles; Climate models

1. Introduction

The North Atlantic Oscillation (NAO) dominates the North Atlantic winter variability with its positive phase indicating stronger westerlies than usual leading to wet and mild winters in northern Europe and dry winters in southern Europe (Hurrell et al. 2003). Oppositely, the negative phase is associated with dry and cold winters in northern Europe. The NAO has time scales from days to decades and its variability is basically generated by internal processes in the atmosphere–ocean system, although there is evidence for a NAO response to volcanic eruptions (Shindell et al. 2004; Stenchikov et al. 2006; Christiansen 2008), forcing from increased greenhouse gas concentrations (Kuzmina et al. 2005; Gillett and Fyfe 2013), and, more disputed, the 11-yr solar cycle (Ineson et al. 2011; Chiodo et al. 2019).

Ensemble modeling has now become a standard tool to gauge the uncertainties originating from uncertain initial conditions and errors due to deficiencies in model physics. A simple and common way to estimate the model skill is to compare the ensemble mean to observations. For large ensembles the averaging suppresses the noise coming from the chaotic behavior of the (modeled) climate system. For historical experiments the ensemble mean then represents the forced signal and for initialized forecasts it represents the forced signal combined with the possible systematic effect from initialization. Recently, the NAO has been claimed to be predictable on decadal time scales. However, the signal in individual model experiments is very weak and it requires averaging over very large ensembles to obtain significant positive correlations with observations (Smith et al. 2019, 2020; Athanasiadis et al. 2020; Klovans et al. 2021). The weak signal combined with a realistic total variance leads to the signal-to-noise paradox: When measured with correlations the ensemble mean predicts the observations better that it predicts the individual ensemble members. This paradox has been observed on time scales ranging from seasons to decades and in both initialized and historical simulations. The paradox is not restricted to the NAO, but also found in other geographical areas and in different variables such as surface pressure, temperature, wind, and precipitation [see the review by Scaife and Smith (2018)].

The weak response to forcings in the models has been attributed to a number of reasons including a general overestimation of uncertainty in initial conditions (Mayer et al. 2021), a too weak teleconnection between the quasi-biennial oscillation and the NAO (O’Reilly et al. 2019), lack of persistence in surface temperature in particular over oceans (Sevèlec and Drijfhout 2019), underestimation of regime behavior (Strommen and Palmer 2019), lack of eddy feedbacks (Scaife et al. 2019), and nonstationarities in combinations with brief hindcasts periods (Weisheimer et al. 2019). While these explanations mainly address the initialized forecasts, Zhang and Kirtman (2019); Zhang et al. (2021) point to more general model problems, such as errors in

Denotes content that is immediately available upon publication as open access.

Corresponding author: Bo Christiansen, boc@DMI.DK

DOI: 10.1175/JCLI-D-21-0807.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
ocean–atmosphere coupling, as the paradox seems to be common to coupled models.

The weak signal-to-noise ratio implies that very large ensembles are needed to isolate the forced signal. The analysis of such samples can often be simplified using the blessings of dimensionality (Christiansen 2018, 2021). In such situations analytical results can be obtained for the different skill measures in simple statistical models. These models can then help us understand important features of the ensemble such as the behavior with ensemble size.

In this manuscript we will focus on the robustness of the skill of the NAO in the models. Even if the modeled signal is statistically significant when compared to the observed NAO in the recent period, it cannot be ruled out that this is a chance occurrence. We will use a set of large model ensembles, including the historical CMIP6 multimodel ensemble and a single-model initial condition ensemble, to investigate the degree to which they agree on NAO skill. We will use the full historical period back to 1850 to track possible nonstationarities and to see if the NAO skill is only found in the recent period. We will also investigate if the NAO signal is only due to the forcing or if initialization will improve the skill. Even if the NAO skill found in the recent period is physical, nonstationarity of the skill might limit the possibility for decadal predictions in the future, as we will not know how long the present favorable conditions will last.

In section 2 we describe the model ensembles and observations, the definition of the NAO, and a simple analytical model. In section 3 we consider the forced response in the historical CMIP6 ensemble: first in the period after 1970 and then in the period back to 1850. In section 4 we consider the effect of ensemble size. In section 5 we look at other CMIP6 ensembles including ensembles driven only by natural or anthropogenic forcings and an ensemble of initialized historical forecasts.

2. Data and methods

In section 2a we present the model ensembles and the observations and in section 2b we discuss the calculation of the NAO index. In section 2c we describe the simple statistical model and the analytical results obtained using the blessings of dimensionality.

a. The data

We will mainly investigate the CMIP6 historical multimodel ensemble (Eyring et al. 2016). Considering the ensemble members r1i1p1f1, r2i1p1f1, ..., r10i1p1f1, we have 213 ensemble members for sea level pressure from 49 different models (considering models with different names as different; e.g., NorESM2-LM and NorESM2-MM are counted as different). We include only the first 10 ensemble members to get a balanced ensemble that is not dominated by a few models. We also look at the smaller CMIP6 hist-nat, hist-GHG, and hist-aer ensembles, which are historical ensembles driven by only natural forcings, well-mixed greenhouse gases, and anthropogenic aerosols forcings, respectively (Gillett et al. 2016). Again using the ensemble members r1i1p1f1, r2i1p1f1, ..., r10i1p1f1, these ensembles have 55 members from 12 models, 51 members from 11 models, and 58 members from 12 models for sea level pressure, respectively. We also have data from the preindustrial control experiments (piControl) from 39 different models (r1i1p1f1). These experiments have constant forcings and are of different lengths. The multimodel ensembles are summarized in Table 1. As a single-model ensemble, we use the Max Planck Institute Grand Ensemble (MPI-GE) (Maher et al. 2019), which is a 100-member initial condition ensemble with historical forcings performed with the Max Planck Institute Earth System Model. The historical experiments cover the period 1850–2015.

Additionally, we will consider the CMIP6 decadal hindcasts (dcpp-A) (Boer et al. 2016). Here we have 101 ensemble members for sea level pressure from 11 different models again using the ensemble members r1i1p1f1, r2i1p1f1, ..., r10i1p1f1. The experiments are initialized in the end of each year in the period 1960–2008 and run for 10 years. We use the forecasts for the period beginning at 1970 to have the same number of forecasts for each lead time.

We use surface pressure from three reanalyses 20CRv2c (Compo et al. 2011), CERA-20C (Laloyaux et al. 2018), and JRA-55 (Onogi et al. 2007) all collected in the Collaborative Reanalysis Technical Environment (CREATE) project (Potter et al. 2018). We also use Hurrell’s station-based NAO index (1865–2020) (Hurrell 1995).

For all ensembles and observations we have used monthly means. Prior to the analysis models and observations are interpolated to a common 2.5° × 2.5° global grid using a simple nearest neighbor procedure.

b. The NAO

The NAO is based on monthly sea level pressure anomalies. We calculate the NAO index as the difference between normalized anomalies of Azores (mean over 20°–28°W, 36°–40°N) and Iceland (mean over 16°–25°W, 63°–70°N). The winter means are calculated over December–February. Here normalizations are done independently for each model experiment and are based on variances over 1960–2012. The centering is based on the same period. For the decadal forecasts the anomalies are calculated for each lead time (Boer et al. 2016) adjusting for possible biases due to model drift, the size of which might depend on the lead time.

We want to use the full historical period in our analysis and to get an estimate of the errors in the observed NAO, we first compare Hurrell’s index and the NAO calculated from the different reanalyses. We find an excellent agreement for the NAO calculated from the different datasets. This agreement holds even for the earliest periods. For example, the correlation between Hurrell’s index and the NAO from the 20CRv2c reanalysis is 0.97 when calculated over the common winters 1866–2012 and 0.95 for the winters 1866–1900. For CERA-20C and JRA-55 the correlation is 0.999 for the common winters 1959–2010.

We will apply the same two-point definition of the NAO to both observations and models. This definition could be expected to be somewhat restrictive as it does not allow for possible displacements of the centers of action. See Baker et al. (2018) for a
<table>
<thead>
<tr>
<th>Model</th>
<th>Historical</th>
<th>Hist-nat</th>
<th>Hist-GHG</th>
<th>Hist-aer</th>
<th>piControl</th>
<th>Decadal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESS-CM2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ACCESS-ESM1-5</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AWI-CM-1-1-MR</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AWI-ESM1-1-1-LR</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCC-CSM2-MR</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>BCC-ESM1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CAMS-CSM1-0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CAS-ESM2-0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CESM1-1-CAM5-CMIP5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>CESM2</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CESM2-FV2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CESM2-WACC-MR</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CESM2-WACC-MF2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CIESM</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CMCC-CM2-HR4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CMCC-CM2-SR5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>CMCC-ESM2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CanESM5</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CanESM5-CanOE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E3SM-1-0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E3SM-1-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E3SM-1-1-1-ECA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EC-Earth3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>EC-Earth3-AerChem</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EC-Earth3-CF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EC-Earth3-ER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EC-Earth3-Veg</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EC-Earth3-Veg-LR</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FGOALS-f3-L</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>FGOALS-g3</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FIO-ESM-2-0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GFDL-CM4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GFDL-ESM4</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GISS-E2-1-G</td>
<td>10</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GISS-E2-1-G-CC</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GISS-E2-1-H</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GISS-E2-2-G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IITM-ESM</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>INM-CM4-8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INM-CM5-0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IPSL-CM6A-LR</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>KACE-1-0-G</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KIOST-ESM</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MCM-UA-1-0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MIROC-ES2H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MIROC-ES2L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MIROC6</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>MPI-ESM-1-2-HAM</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MPI-ESM1-2-HR</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>MPI-ESM1-2-LR</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>MRI-ESM2-0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>NESM3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NorCPM1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>NorESM2-LM</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NorESM2-MM</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SAM0-UNICON</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TaiESM1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
discussion and references. However, we have repeated the analysis below with the alternative definition based on principal component analysis (finding the leading empirical orthogonal function from detrended monthly winter surface pressure anomalies in the area 20°–80°N, 90°W–40°E and then projecting this pattern back on the undetrended surface pressure) and found only small differences.

Figure 1 shows the time series of Hurrell’s NAO index and the ensemble mean NAO for all the ensembles.

c. A simple model

As mentioned in the introduction, the analysis of climate models can often be simplified using the blessings of dimensionality (Christiansen 2018, 2021). Using the nonintuitive properties of high-dimensional spaces we can systematically obtain analytical results for many quantities. The properties of such spaces include that vectors drawn independently from the same distribution have the same lengths and that independent vectors are orthogonal. Christiansen (2021) showed that these properties with success can be applied to climatic fields and time series. In the present case the high dimensionality is related the length of the NAO time series or, more precisely, the effective degrees of freedom in the series. The blessings of dimensionality are very general and do not require that the time series are Gaussian distributed or that their variances are constant in time.

The behavior of correlations and amplitudes as a function of ensemble size can be understood within such a simple framework. We write the observations \( \mathbf{o} \), as \( \mathbf{o} = \mathbf{O} + \mathbf{\xi} \). Here, boldface symbols denote vectors (here time series) of length \( N \). In this equation \( \mathbf{\xi} \), describes the weather noise on top of the forced signal \( \mathbf{O} \). Let the ensemble members all share a part of the forced signal in the observations: \( \mathbf{x}_i = \mathbf{X} + \mathbf{\xi}_i \), \( i = 1, 2, \ldots, K \), where \( \mathbf{X} = \lambda \mathbf{O} \) and \( \mathbf{\xi}_i \) is independent noise. Without lack of generality, we assume that all the vectors have zero means. Similar simple models have been used before (e.g., Siegert et al. 2016).

Let \( \sigma^2_\mathbf{O} \) and \( \sigma^2_\mathbf{\xi} \) be the variances of \( \mathbf{O} \) and \( \mathbf{\xi}_i \), respectively. Likewise, \( \sigma^2_\mathbf{X} = \lambda^2 \sigma^2_\mathbf{O} \) is the variance of \( \mathbf{X} \), and \( \sigma^2_\mathbf{\xi} \) is the variance of the noise \( \mathbf{\xi}_i \).

The ensemble mean is \( \bar{\mathbf{x}} = \sum \mathbf{x}_i / K \). It can then be shown (Christiansen 2021) that for large \( N \) the mean-square amplitude of the ensemble mean behaves like

\[
\|\bar{\mathbf{x}}\|^2 / N = \sigma^2_\mathbf{X} + \sigma^2_\mathbf{\xi} / K. \tag{1}
\]

Then the variance of \( \mathbf{\xi} \) is \( \sigma^2_\mathbf{\xi} = \sigma^2_\mathbf{\xi} \) and the variance of \( \mathbf{X} \) is \( \sigma^2_\mathbf{X} \). We therefore get

\[
\text{cor}(\mathbf{o}, \mathbf{x}) = \frac{1}{\sqrt{1 + \sigma^2_\mathbf{\xi} / \sigma^2_\mathbf{O} \sqrt{1 + \sigma^2_\mathbf{\xi} / \sigma^2_\mathbf{X}}} \tag{2}
\]

The correlation increases with ensemble size but with ever decreasing rate. The noise terms will drive the correlation down, and the rate of the convergence will be determined
by \( \sigma_\xi^2 / \sigma_\chi^2 \). An expression similar to Eq. (2) was used by Bellomo et al. (2018) in a study of the Atlantic multidecadal variability.

When calculating the correlation between the ensemble mean and an ensemble member we need to distinguish between the in-sample and out-of-sample situations, where the ensemble member is included or excluded from the calculation of the ensemble mean, respectively (see the appendix for a derivation). We get

\[
\text{cor}(x_i, x_j) = \begin{cases} \\
\frac{\sqrt{1 + \frac{1}{N} \sigma_\xi^2 / \sigma_\chi^2}}{\sqrt{1 + \sigma_\xi^2 / \sigma_\chi^2}} & \text{in-sample} \\
\frac{1}{\sqrt{1 + \sigma_\xi^2 / \sigma_\chi^2}} & \text{out-of-sample.}
\end{cases}
\]

This correlation increases with ensemble size in the out-of-sample situation, as observed by Scaife and Smith (2018) and Athanasiadis et al. (2020), and decreases in the in-sample situation. For large \( K \) they both converge to \( 1/\sqrt{1 + \sigma_\xi^2 / \sigma_\chi^2} \).

For the fraction of the correlations—a measure of the signal-to-noise paradox—we get

\[
\text{cor}(o_i, x_j) / \text{cor}(x_i, x_j) = \begin{cases} \\
\frac{\sqrt{1 + \frac{1}{N} \sigma_\xi^2 / \sigma_\chi^2}}{\sqrt{1 + \sigma_\xi^2 / \sigma_\chi^2}} & \text{in-sample} \\
\frac{1}{\sqrt{1 + \sigma_\xi^2 / \sigma_\chi^2}} & \text{out-of-sample.}
\end{cases}
\]

In the out-of-sample situation this expression is independent of ensemble size and is larger than 1 when \( \sigma_\xi^2 / \sigma_\chi^2 > \sigma_\xi^2 / \sigma_\chi^2 \). In the in-sample case it grows with ensemble size to reach the out-of-sample value for large \( K \) (for fixed \( \sigma_\xi^2 / \sigma_\chi^2 \)), while it has a local maximum as function of \( \sigma_\xi^2 / \sigma_\chi^2 \) (for fixed \( K > 2 \)).

The equations above give the values for \( N \to \infty \). For finite \( N \) there will be a spread (standard deviation) around these values. For the mean-square amplitude we get the spread

\[
\sqrt{\frac{2}{N} \frac{\sigma_\xi^2}{\sigma_\chi^2} \frac{1}{1 + 2K \sigma_\xi^2 / \sigma_\chi^2}}.
\]

The spread on \( \text{cor}(o, x) \) becomes (see the appendix for a derivation)

\[
\frac{1}{\sqrt{N}} \frac{1}{\sqrt{1 + K \sigma_\xi^2 / \sigma_\chi^2}}.
\]

We note that these spreads do not depend on the variances of the observations and that they go to zero for large \( K \). For completeness, we mention that the spread of \( \text{cor}(x, \bar{x}) \) is \( \frac{1}{N} (1 + \frac{1}{N} \sigma_\xi^2 / \sigma_\chi^2)^{-1/2} \), which is independent of \( K \). In the presence of serial correlations \( N \) should be replaced with \( N_{eq} \), the effective degrees of freedom of \( \xi \), in this equation and in Eqs. (5) and (6).

Expressions resembling Eq. (4) were given in Siegert et al. (2016) and in Zhang and Kirtman (2019) for a simple model with persistency. Expressions as those above can be derived for large \( N \) using the blessings of dimensionality under quite general conditions; we do not, as in previous work, need to assume Gaussianity of the noise terms, \( \xi \) and \( \chi \), or that the variance of \( \xi \) is the same for all \( i = 1, \ldots, N \). These general conditions increase the applicability of the simple model.

### 3. The forced response in NAO

In this section we mainly focus on the CMIP6 historical experiment. In section 3a we consider the period from 1970 and in section 3b we investigate the temporal stationarity by considering the full historical period.

#### a. The recent period

The ensemble mean of the NAO over the more than 200 CMIP6 historical ensemble members—representing the forced response of the NAO—is shown in Fig. 1 together with the observed NAO (top panels) as function of time. Both series are smoothed over seven winters.

Observations and model mean partly agree after 1970. Note that the similarity is not just found for the strong positive phase in the 1990s but also for several smaller fluctuations (e.g., the positive values in the beginning of the 1970s and just after 2010). There is also some weaker correspondence earlier (e.g., the low values around 1915). This is somewhat overshadowed by the trend in the ensemble mean, which is not apparent in the observations. However, the amplitude of the ensemble mean is much lower (by a factor of 10) than in observations, as also found by Smith et al. (2019, 2020).

For the ensemble mean the trend calculated over 1866–2015 is 0.10 per century and highly statistically significant. For the observations the trend in the same period is 0.01 per century and not significant (a similar trend is obtained with the 20CRv2c reanalysis). However, the trend in the individual ensemble members fall in the interval 0.1 ± 0.2 (standard deviation) and therefore the observations do not stand out. We can split the trend in the ensemble mean NAO into contributions from its two centers. We find that 40% come from a positive pressure trend over the Azores and 60% come from a negative trend over Iceland. While there is no general trend in the NH mean pressure in the ensemble mean, we see a general negative trend north of 45°N and a positive trend equatorward of this latitude. Similar long-term trends are found in the hist-GHG and MPI-GE ensembles, but not in the hist-nat and hist-aer ensembles (Fig. 1). This is consistent with the trend being caused by the forcing from increasing greenhouse gases.

The correlations between ensemble members and observations are shown together with the correlations between the ensemble mean and observations in the left panel of Fig. 2 for the period 1970–2015. All time series are 7-yr smoothed and trends are not removed. The correlations for individual
ensemble members are distributed around 0 (mean 0.04) with a standard deviation of 0.27. The distribution is almost symmetrical with only a weak skewness. The correlation of the ensemble mean is 0.57, which is only exceeded by a few ensemble members. This is comparable to what is found in an ensemble of decadal forecasts by Smith et al. (2019, 2020), in noninitialized multimodel ensembles by Klavans et al. (2021) (a 269-member ensemble based on six climate models), and by Zhang and Kirtman (2019) (the CMIP5 ensemble). The fact that the ensemble mean has larger skill than the individual ensemble members is a direct consequence of the properties of high-dimensional spaces. See Christiansen (2018) for a simple explanation and Christiansen (2019) for an analytical result of the fraction of individual ensemble members having better skill than the ensemble mean.

We calculate the significance of the ensemble mean correlations by a Monte Carlo method where surrogates are produced by first phase-scrambling individual detrended ensemble members (Theiler et al. 1992; Christiansen 2001) and then adding the trend back. The phase-scrambling (randomizing the Fourier phases) produces a surrogate time series with the same power spectrum as the original time series, and thus also retains the full autocorrelation spectrum of the original series. The ensemble mean of the surrogates is calculated and the correlation to observations is found. We do this for 5000 realizations to get the distribution of the surrogate correlations. The original correlation is finally compared to this distribution. We find that the correlation between ensemble mean and observations is significant at the $p = 0.01$ level.

**b. The temporal nonstationarity of correlations**

We mentioned above that there is also some match between ensemble mean and observations before 1970, but that this, to some extent, could be hidden by the trend in the ensemble mean. Calculating the correlation for the full period 1866–2015 gives a correlation of 0.11 that is not statistically different from zero ($p = 0.13$) if the long-term trend in ensemble mean is kept included. Removing that trend gives a significant correlation of 0.20 ($p = 0.03$). However, the top panel in Fig. 3 shows the correlations between ensemble mean and observations for different 45-yr periods as function of the start years of the periods. The correlations are very variable and only above 0.5 for start years after 1955 and before 1890. It is also only for these start years that correlations statistically significant from zero are found. Negative correlations (as low as $-0.5$) are found for start years in the interval 1915–45. Removing the low-frequency variability by detrending the time series in each 45-yr period does not change the overall conclusions (not shown).

The preindustrial control experiments with constant forcings can have low-frequency variability due to internal processes such as atmosphere–ocean couplings but are otherwise representative for a stationary climate. Thus, these experiments offer a realistic opportunity to test the stationarity. As mentioned, the experiments are of different lengths, and we organize these data into 108 nonoverlapping 150-yr sections. We now use these sections as our ensemble. For each ensemble member we calculate the winter NAO as above. After a 7-yr smoothing we calculate the correlations between the ensemble mean and “observations” in 45-yr periods. For observations, we use either the real Hurrell’s index for 1970–2015 or a random ensemble member (swapping the ensemble member with Hurrell’s index). The results are shown in the middle panel in Fig. 3. By construction these correlations are expected to be zero in mean (population mean). However, we see an apparent nonstationarity resembling what was found for the historical ensemble in the top panel of the figure.

In the CMIP6 ensemble we have 26 significant 45-yr periods when the full period 1866–2015 is considered and four significant periods when only periods after 1900 is considered (full black curve in top panel of Fig. 3). To estimate the significance we can do a total of 108 different swaps (five of them are shown in the middle panel of Fig. 3). For each swap we can calculate the number of significant 45-yr periods. Of these swaps 18% have 26 or more significant 45-yr periods when 1866–2015 is considered and 48% have more than four when the period after 1900 is considered.

A direct comparison between CMIP6 historical ensemble and the smaller constructed preindustrial ensemble may be unfair as the ensemble size has an influence on the correlations as seen in Eq. (2). However, similar nonstationarity can be obtained by purely randomness. In the bottom panel in Fig. 3 each of the values in the 150 winters of both the observations and the 213 ensemble members are drawn...
independently from a Gaussian distribution. After a 7-yr smoothing we again calculate the correlations between the ensemble mean and the observations. We see the same picture as in the upper panel.

To estimate the significance we repeat this process many times (500) and calculate the number of significant periods in each repetition. We now find that 17% of these give more than 26 significant 45-yr periods when 1866–2015 is considered. Restricting this to only the period after 1900 then 40% of the random experiments give four or more significant periods.

Although there is no a priori reason to leave out the period before 1900, the results demonstrate that this period, where observations and forcings could be considered less trustworthy, plays a large role for the \( p \) values.

We therefore conclude that the nonstationary and the number of significant 45-yr periods found in the correlations between the CMIP6 historical ensemble mean and the observations easily can be a chance occurrence. This is also in general agreement with the simple analytical arguments. The sample spread (standard deviation) of the correlation of two uncorrelated series is

\[
\frac{1}{\sqrt{N_{\text{ef}}}} \]  

[cf. Eq. (6) with \( \sigma_x^2 = 0 \)],

which for \( N_{\text{ef}} = 20 \) \(^1\) is 0.22.

This indicates that the seeming nonstationarity in the correlations can be obtained for uncorrelated time series without any low-frequency variability. However, the weak variability of the natural forcings in the period 1920–60 (Takemura et al. 2006), where the weakest correlations are found in the historical ensemble, could also play a role. Also, it has been suggested that such nonstationarity can be due to a weakened

\[ N_{\text{ef}} \approx N(1 - r^2)(1 + r^2) \]

we get \( N_{\text{ef}} = 18 \) for \( N = 45 \).

\(^1\) For the historical CMIP6 experiments the autoregressive coefficient at lag one, \( r \), of the smoothed NAO index is around 0.65 on average. According to Bartlett’s (1935) estimate, we get \( N_{\text{ef}} = 18 \) for \( N = 45 \).
influence of El Niño–Southern Oscillation during the middle of the twentieth century (O’Reilly et al. 2020; Rieke et al. 2021).

In Fig. 4 we track the correlations between the models and observed NAO for the 10 individual ensemble members with the strongest correlations in the two 45-yr periods beginning in 1866 and 1970. These are periods with significant correlations between ensemble mean NAO and observed NAO. We do not find any connection between the individual ensemble member’s skill in the different periods. For example, the 10 members with the largest skill in the period beginning in 1970 are well spread out in the previous fully nonoverlapping period beginning in 1925. So we cannot select the skillful ensemble members in one period and expect they will also be skillful in other periods. This is also in agreement with the formulation of the simple statistical model, where a weak signal can be present in all models.

4. Effect of ensemble size

We now want to study the effect of the ensemble size. For an ensemble of size $K_0$ we randomly choose a subset of $K$ ($1 \leq K \leq K_0$) ensemble members and calculate the ensemble mean of these. Now the quantity of interest, the correlation with observations or the mean-square amplitude, is calculated from this ensemble mean. We repeat this process many times ($5000$) for each $K$ to get the distribution of the quantity. The selection of the subsets is done with replacement (bootstrap).

The results are shown in the top panels of Fig. 5 for the historical CMIP6 ensemble for the period 1970–2015. The correlations grow from zero for small ensemble sizes to around 0.5 for the full ensemble (the difference from the correlation of 0.57 in Fig. 1 is because the sampling is done with replacement). The 95% confidence band is wide and an ensemble size of more than 100 is needed for the confidence band not to include zero. On the other hand the amplitude seems to converge toward zero although some saturation is seen. At least some of the saturation is a consequence of the bootstrap procedure (sampling with replacement).

Such behavior where the correlation continues to increase and the amplitude continues to decrease may seem strange but is a consequence of the simple model described in section 2c. We can calculate $\sigma_X^2$ and $\sigma_E^2$ from the model ensemble—more precisely, $\sigma_X^2$ is estimated as the variance of the multimodel mean and $\sigma_E^2$ as the variance of the residuals—and find $\sigma_X^2/\sigma_E^2 = 150$. This estimate is in correspondence with visual inspection of Fig. 1 and in agreement with Smith et al. (2019, 2020) and Klavans et al. (2021). However, this separation cannot be done directly for the observations. Good fits to observed correlations are obtained assuming $\sigma_X^2/\sigma_E^2$ is a factor of 100 smaller. We have also assumed that $\text{cor}(\mathbf{O}, \mathbf{X}) = 1$ by setting $\mathbf{X} = \hat{\mathbf{X}}$ in section 2c; lower values of this correlation require lower values of $\sigma_X^2/\sigma_E^2$. The red curves in Fig. 5 show the analytical values from Eqs. (2) and (6) (correlations) and Eqs. (1) and (5) (amplitude). For the error bars we use $N_{\text{eff}} = 20$ as before.

We note the almost perfect match for both mean-square amplitudes and correlations. The theory explains for the same parameters both the decreasing amplitude and increasing correlation as function ensemble size. The spread of the amplitudes and correlations are also well explained by the theory. The simple model predicts that the correlation saturates at 0.63 for large $K$ [Eq. (2)] in agreement with Fig. 5. For the amplitude the simple model predicts a saturation at 0.001 [Eq. (1)]. Taking the confidence intervals into it is difficult to separate this saturation from the continuous decay of white noise—a straight line with slope $-1$ in the double logarithmic plot.

We can also apply the simple statistical model to describe the distribution of the correlations between the individual ensemble members and the observation in Fig. 2. The red curves in this figure show the distribution assuming Gaussianity and using the mean and standard deviation from Eqs. (2) and (6) for $K = 1$. The figure also shows the approximation for the ensemble mean from Eq. (2) with $K = 213$. Again, we note the very precise predictions from the simple model. In particular the ensemble mean correlation is 0.47 compared to the 0.57 from the raw calculation for the period 1970–2015. The agreement is even better for the negative correlations for the period 1925–69 now with the estimate $\sigma_X^2/\sigma_E^2 = 127$ and $\sigma_X^2/\sigma_E^2$ still a factor of 100 smaller.

Our simple model is different from the “signal-plus-noise” model, where models are considered as observations polluted by noise. The signal-plus-noise model is obtained from our
model when $\sigma_{x^2} = 0$, but this would give different and wrong predictions.

Note that the success of the simple model does not indicate if the weak skill is due to chance or not. The good performance of the simple model is due to the simplifying properties of high-dimensional spaces and indicates that the high correlation for the NAO ensemble mean is not a consequence of a few unique ensemble members (see also Fig. 4) but a consequence of the averaging of an ensemble with low signal-to-noise ratio. In fact, removing the 10 best ensemble members from the CMIP6 historical ensemble reduces the ensemble mean correlation from 0.57 to 0.44. This is somewhat larger than the reduction of 0.07 predicted by the simple model when the 10 best ensemble members are removed. This difference might be related to model dependence and the corresponding breakdown of orthogonality as demonstrated in Christiansen (2021).

5. Other ensembles

In the previous section we focused mainly on the historical CMIP6 ensemble. In section 5a we will consider the other historical experiments, and in section 5b we will compare the initialized decadal forecast ensemble with the uninitialized historical ensemble. The time series of the ensemble mean NAO is shown in Fig. 1 for all the ensembles.

a. Other historical experiments

Now we take a look at some other ensembles. Figure 3—which shows correlations between the ensemble mean NAO and observations for previous 45-yr periods—also includes results for the MPI-GE and smaller hist-nat, hist-GHG, and hist-aer ensembles. Here are also included results for the historical ensemble where NAO is calculated by the definition based on the leading principal component. The bottom panel in Fig. 5 shows how the NAO amplitudes and correlations with observations behave as a function of ensemble size for the hist-nat, hist-GHG, hist-aer, MPI-GE, and initialized forecast (lead time 1) ensembles as well as for a few other ensembles discussed below.

We first note from Fig. 5 that all ensembles show decaying amplitudes with only weak signs of saturation. Considering the error bars (shown in the top panel) the amplitudes of the historical, hist-nat, hist-GHG, hist-aer, and MPI-GE ensembles decay with the same speed.

In contrast to the historical ensemble the hist-nat and the hist-aer ensembles show vanishing correlations for all ensemble sizes. The hist-GHG ensemble shows positive correlations.
that increase with ensemble size, but the correlations are much weaker than for the CMIP6 historical ensemble. Also, the initial condition ensemble MPI-GE fails to produce any positive correlations with the observed NAO signal. This supports the conclusion from the last section that the correlations found in the historical ensemble are due to chance. There are, of course, other explanations. The correlations for the historical ensemble could be due to a nonlinear combination of natural and anthropogenic forcings and not show up in the single forcing experiments.

The previous results were obtained for the period 1970–2015. The top panel in Fig. 3 shows results for other 45-yr periods. We find that none of the hist-nat, hist-GHG, hist-aer, or MPI-GE ensembles show any significant positive correlations in any of these periods. Some of them even show statistically significant negative correlations.

The decay with ensemble size for the CMIP6 historical experiment for the 45-yr period 1925–69 is also shown in Fig. 5. As expected from the right panel in Fig. 2, the correlations are negative, with an absolute value that increases with ensemble size, while the amplitude continues to decrease. Note from Fig. 2 that again the ensemble mean has a stronger (negative) correlation to observations than most of the individual ensemble members.

More encouraging is the CMIP6 decadal forecast of the NAO. Here we show the results when forecasting the mean over the first seven winters for lead time 1. The observations have been aligned and treated similarly. Note that no additional smoothing is applied, so we do not pollute the forecasts with information not available prior to the initializations. The correlations are comparable to those of the historical ensemble: they continue to increase with ensemble size and show no sign of saturation. The amplitude is somewhat larger than for the historical experiment, although it also continues to decay. In the next section we will compare the historical and decadal predictions in more detail.

As described in section 2b we have used the two-point definition of the NAO throughout the paper. As this definition fixes the NAO’s centers of action, it might be expected to be suboptimal. In Fig. 3 we have included the correlations for the principal component based definition (dashed-dotted curve in the upper panel), and we see that there are only small differences compared to the two-point definition.

We end this section with a note on the significance. When estimated with the method we used for the CMIP6 historical experiment in section 3a, we find that neither the single forcing ensembles nor the MPI-GE ensemble have correlations significantly different from zero. This missing significance could be a consequence of the moderate ensemble size, and we therefore also ask if these ensembles could be drawn from the CMIP6 historical ensemble. The confidence intervals shown in the top panels of Fig. 5 are large and may suggest that the behavior of, for example, the single-forcing ensembles is not significantly different from the CMIP6 historical ensemble. But these confidence intervals indicate the uncertainty in the ensemble mean of a single ensemble of size \( K \). However, the full solid curves in the figure show the mean of bootstraps generated from a larger ensemble and are very smooth. We want to estimate the confidence intervals of the mean of the bootstraps from an ensemble of, say, 60 members drawn from the CMIP6 historical ensemble. The larger this ensemble is, the more narrow the resulting confidence levels become; 60 is chosen here to be comparable with the single-forcing experiments. We now proceed as follows: 1) We draw 60 random members from the 213-member CMIP6 historical ensemble. 2) We now choose random subsets for each ensemble size less than 60 and calculate the correlation between the ensemble mean of the subsets and the observations (as described in the beginning of section 4). This will give one curve as in Fig. 5. We repeat these two steps many times to get an ensemble of surrogate curves. From this ensemble we can calculate the 95% confidence interval (blue curves in top left panel of Fig. 5). This confidence interval is considerably narrower than that shown with dashed black curves and has its lower edge near zero. Counting the fraction of the surrogate curves falling below the curves from the single-forcing CMIP6 and MPI-GE ensembles we find the \( p \) values.

We want to consider different values of the lead time and the averaging lengths (\( \Delta \)). Above we considered forecasts of the mean over the first seven winters after the initializations corresponding to \( \Delta = 7 \) and lead time 1. The correlations between the full ensemble means and the observations are shown in the upper panels of Fig. 6 for different \( \Delta \) and lead times. The correlations based on the historical ensemble are statistically significant for some values of \( \Delta \) in the interval between 3 and 7. Likewise, for the forecast ensemble some correlations are significant for small lead times. However, the differences between the correlations in the forecasts ensemble and the historical ensemble are small and mostly negative. Furthermore, the only significant differences are found where the differences are negative (i.e., where the historical ensemble has larger skill than the forecast).

It follows from the discussion in section 4 that the skill of the ensemble mean in general increases with the ensemble size, thus favoring the historical ensemble. When comparing skills of two ensembles it is therefore important that they have the same size.

To make a direct comparison between the historical experiment and the decadal forecast we have identified a common set of 70 ensemble members with the same names and
The behavior of these reduced ensembles as function of ensemble size is shown in Fig. 5 (orange curves for the decadal forecast, pink for historical experiment). For the decadal forecast we find that although the correlations are somewhat smaller than for the full decadal forecast ensemble (blue curve) they do increase with ensemble size. The decay of the amplitude follows that of the full decadal ensemble. The reduced historical ensemble behaves differently compared to the full historical ensemble regarding the correlations, which are in general negative and become more negative with increasing ensemble size.

The correlations for the reduced ensembles are shown in the lower panel of Fig. 6. For the decadal forecasts the correlations are positive but only statistically significant from zero for few combinations of lead time and $\Delta$. The historical ensemble shows zero or weak insignificant, negative correlations. The difference between the correlations of the decadal forecasts and the historical experiments is in general positive and statistically significant for small values of $\Delta$ when the lead time is smaller than 4. The statistical significance of the differences is estimated with a Monte Carlo method where the observed difference between the correlations is compared to correlation differences obtained with surrogate versions of the forecast. These surrogates are generated to have the same population correlation to the observations as well as the same serial correlations as the historical experiment.

Thus, for the ensembles consisting of the same models the decadal forecasts add skill when compared to the historical experiments. However, the improved skill is only because the reduced historical ensemble is atypical as it shows negative correlations with observations. This emphasizes the fragility of

---

3 The 101 member decadal ensemble includes 11 models BCC-CSM2-MR, CESM1-1-CAM5, CMCC-CM2, EC Earth3, FGOALS-G3-L, IPSL-CM6a-LR, MIROC6, MPI-ESM1-2-HR, MPI-ESM1-2-LR, MRI-ESM2-0, and NorCPM1. The common ensemble (70 members) includes these except CESM1-1-CAM5-CMIP5. Only MPI-ESM1-2-HR and IPSL-CM6a-LR are among the 10 models with the strongest NAO correlations in the period after 1970 shown in Fig. 4.
of the NAO response and the need for large ensembles to estimate the significance. We note that Athanasiadis et al. (2020) did find improved skill in the NAO when comparing historical and decadal hindcasts in 40-member ensembles with the Community Earth System Model and that results from Borchert et al. (2021) indicate the possibility that improvements from initialization may be larger for models that poorly simulate the forced response.

6. Conclusions

We have investigated the NAO in several large ensembles including a CMIP6 historical ensemble consisting of 213 individual experiments with the focus on the forced response as estimated by the ensemble mean. Our main findings are the following:

- For the forced NAO signal in CMIP6 we find significant correlation (0.57) with observations for the period after 1970. This is in agreement with the results reported by Smith et al. (2019, 2020) with ensembles of decadal forecasts and by Zhang and Kirtman (2019) and Klavans et al. (2021) with uninitialized ensembles. However, in earlier periods of the same length (45 years) insignificant and even negative correlations are found. Only in periods beginning before 1890 does the situation resemble that for the period after 1970.

- This kind of apparent nonstationarity in the correlations can easily be due to chance as we demonstrated with both the preindustrial control ensemble and pure random numbers.

- In the period after 1970 correlations continue to increase as function of ensemble size while the amplitude continues to decay, although some signs of saturation are found. The behavior with ensemble size can be described by a simple statistical model including only the signal-to-noise ratios. Assuming a very small signal-to-noise ratio, \( \sigma_s/\sigma_n \), of 0.08 (1/\sqrt{150})—in agreement with Smith et al. (2019, 2020) and Klavans et al. (2021)—for the model ensemble we get a close match for both correlations and amplitude. This match includes the increase of the correlation (toward a value less than one) and the weak saturation of the amplitude. Also, the behavior in the period when the ensemble mean has negative skill can be well described by the simple statistical model with comparable parameters. The success of the simple model indicates that the skill is not a consequence of a few unique ensemble members but does not inform us whether the skill is real or due to chance.

- The initialized and noninitialized ensembles show the same skill when the full ensembles are considered. However, using the common subsample of models in CMIP6 historical and decadal forecasts we find that the ensemble-mean forecast improves the skill for most lead times and averaging. Unfortunately, the subsample is not typical for the full CMIP6 ensemble (it has weak negative correlations with observations), and these results might be difficult to generalize. Note also that the skill from the CMIP6 historical ensemble is comparable to the skill of the initialized ensemble in Smith et al. (2019). However, improvements from initialization may be larger for models that poorly simulate the forced response (Borchert et al. 2021).

- In none of the hist-nat, hist-GHG, or hist-aer ensembles does the forced NAO signal show positive significant correlations with observations after 1970 or in previous 45-yr periods. The same holds for the 100-member initial-condition, single-model ensemble MPI-GE. The skill of the hist-nat, hist-aer, and MPI-GE ensembles in the period after 1970 can be assumed to be significantly different from the CMIP6 historical ensemble whereas the hist-GHG ensemble cannot.

- There is an overall trend in the forced NAO over the historical period in the historical CMIP6 and MPI-GE ensembles. This trend seems to be related to the greenhouse gas forcing, as it is absent in the hist-nat and hist-aer ensembles but present in the hist-GHG ensemble. See also Kuzmina et al. (2005) and Gillett and Fyfe (2013).

Regarding the predictability of the NAO our results do not paint a consistent picture but rather point in different ways. We do find a significant correlation in the period after 1970 and compared to the original work of Smith et al. (2019, 2020) the similarity of our ensemble mean NAO and observations are not dominated so much by the positive phase in the 1990s but includes more common fluctuations (e.g., the positive phase in 2010–15). On the other hand we find strong apparent nonstationarity in the correlations between observed and model mean NAO. It could be argued that the missing positive correlations in the period 1920–60 are due to the weak variability—or a large uncertainty—of the natural forcings in this period. However, in the period 1925–69 the behavior of the NAO is almost as in the period after 1970 but with negative correlations. Such nonstationarity, where strong positive and negative significant correlations are found in different 45-yr periods, can easily be a chance occurrence. Furthermore, we find no or weak connection between observed and model mean NAO in the hist-nat, hist-GHG, and hist-aer ensembles after 1970. We can therefore not rule out that the significant correlation in the period after 1970 is due to chance or perhaps an indirect result of model tuning as the targets for the tuning are related variables such as the global mean surface temperature, the global mean outgoing radiation, El Niño–Southern Oscillation, the twentieth-century warming, or the meridional overturning circulation (Hourdin et al. 2017).

Finally, we note that even if there is a genuine skillful forced NAO signal in the period since 1970, the apparent nonstationary of the NAO skill, the huge ensemble needed to isolate it, and the seemingly small effect of initialization requiring the future forcings to be known all bring into question the possibility of useful NAO predictions on decadal time scales.

Acknowledgments: This work is supported by the NordForsk-funded Nordic Centre of Excellence project (Award 76654) Arctic Climate Predictions: Pathways to Resilient, Sustainable Societies (ARCPATH) and by the project European Climate Prediction System (EUCP) funded by the European Union under Horizon 2020 (Grant Agreement 776613).
We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy’s Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Data availability statement. The MPI Grand Ensemble Project (https://www.mpimet.mpg.de/en/grand-ensemble/) was downloaded via EFGS from https://esgf-data.dkrz.de/projects/esgf-dkrz/. Hurrell’s NAO index was retrieved from https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-station-based. The reanalyses, compiled by the CREATE project, were downloaded from https://esgf-node.llnl.gov/search/create-ip/.

APPENDIX

Derivations of Eqs. (3) and (6)

As an example, we show how Eq. (3) is derived in the in-sample situation. We assume the series are centered to zero. The simple model assumes that $x_i = X + \xi_i$, $i = 1, 2, \ldots, K$. We now use the features of high dimensions that dot products of independent vectors are zero and that vectors drawn independently from the same distribution have the same lengths. We get

$$x_i \cdot x_i = X \cdot X + \xi_i \cdot \xi_i + 2\xi_i \cdot X = N(\sigma_X^2 + \sigma_\xi^2),$$

and

$$x_i \cdot x = (X + \xi_i) \cdot \left(X + \frac{1}{K} \sum_j \xi_j\right) = X \cdot X + \frac{1}{K} \sum_j \xi_j \cdot \xi_i + \frac{1}{K} \sum_j X \cdot \xi_j + X \cdot \xi_i = N(\sigma_X^2 + \sigma_\xi^2 / K).$$

Likewise, $x \cdot x = N(\sigma_X^2 + \sigma_\xi^2 / K)$. We have

$$\text{cor}(x, x_i) = \frac{x \cdot x_i}{\sqrt{x \cdot x} \sqrt{x_i \cdot x_i}},$$

which leads to Eq. (3) after insertion.

Equation (6) is obtained as follows. We have

$$\text{cor}(\alpha, x) = \frac{1}{N} \sum \alpha \cdot O + \sum \xi_i / K + \xi_0 \cdot \sum \xi_i / K / \sqrt{\sigma_\xi^2 + \sigma_\xi^2 / \sum \xi_i / K}. $$

Only the last two terms contribute to the spread when observations are fixed. The standard deviation of the dot product of two independent vectors of length $N$ and variances $\sigma^2_\xi$ and $\sigma^2_X$ is $\sigma_\alpha \sigma_\beta \sqrt{N}$. Thus the spread of $\text{cor}(\alpha, x)$ becomes

$$\frac{1}{\sqrt{N}} \frac{\sigma_\alpha^2 \sigma_\xi^2 / K + \sigma_\alpha^2 \sigma_\xi^2 / K}{\sqrt{\sigma_\xi^2 + \sigma_\xi^2 / \sum \xi_i / K}} = \frac{1}{\sqrt{N}} \frac{\sqrt{1 + \sigma_\alpha^2 / \sigma_\xi^2 / K} + \sqrt{1 + \sigma_\alpha^2 / \sigma_\xi^2 / K}}{\sqrt{1 + \sigma_\xi^2 / \sum \xi_i / K}}.$$

The spread in the mean-square amplitude, Eq. (5), is obtained similarly. Note that the calculations of the spread assume that the contribution from the denominator is small, which is a very good approximation for $\sigma_\xi^2 / \sigma_X^2 \approx 1$ and $\sigma_\xi^2 / \sigma_\xi^2 \approx 1$.

REFERENCES


