
Article:
Hall, R.J., Wei, H.L. orcid.org/0000-0002-4704-7346 and Hanna, E. (2019) Complex systems modelling for statistical forecasting of winter North Atlantic atmospheric variability: a new approach. Quarterly Journal of the Royal Meteorological Society, 145 (723). pp. 2568-2585. ISSN 0035-9009

https://doi.org/10.1002/qj.3579

Accepted for publication by *Quarterly Journal of the Royal Meteorological Society*
Accepted 30 May 2019.
First published: 01 June 2019, https://doi.org/10.1002/qj.3579

Complex Systems Modelling for Statistical Forecasting of Winter North Atlantic Atmospheric Variability: a New Approach

Richard J. Hall[1], Hua-Liang Wei[2], Edward Hanna[1]

[1] School of Geography and Lincoln Centre for Water and Planetary Health,

University of Lincoln,

Think Tank, Ruston Way,

Lincoln LN6 7FL, UK

[2] Department of Automatic Control and Systems Engineering,

University of Sheffield,

Mappin Street, Sheffield S1 3JD, UK

Running Head: North Atlantic Seasonal Forecasting

Abstract

Seasonal forecasts of winter North Atlantic atmospheric variability have until recently shown little skill. Here we present a new technique for developing both linear and non-linear statistical forecasts of the winter North Atlantic Oscillation (NAO) based on complex systems modelling, which has been widely used in a range of fields, but generally not in climate research. Our polynomial NARMAX models demonstrate considerable skill in out-of-sample forecasts and their performance is superior to that of multiple linear regression models, albeit with small sample sizes. Predictors can be readily identified and this has the potential to inform the next generation of dynamical models and models allow for the incorporation of non-linearities in interactions between predictors and atmospheric variability. In general there is more skill in forecasts developed over a shorter training period from 1980 compared with an equivalent forecast using training data from 1956. This latter point may relate to decreased inherent predictability in the period 1955-1980, a wider range of available predictors since 1980 and/or reduced data quality in the earlier period and is consistent with previously identified decadal variability of the NAO. A number of predictors such as sea-level pressure over the Barents Sea, and a clear tropical signal are commonly selected by both linear and polynomial NARMAX models. Both approaches can be extended to developing probabilistic forecasts and to other seasons and indices of atmospheric variability such as the East Atlantic pattern and jet stream metrics.

Key Words: NAO, seasonal forecast, NARMAX, predictability, jet stream, North Atlantic, winter

1. Introduction

Winter North Atlantic (NA) atmospheric variability is dominated by the North Atlantic Oscillation (NAO; Hurrell and Deser, 2009; Hanna and Cropper, 2017). The NAO index gives a measure of the pressure difference between semi-permanent high pressure over the Azores and a semi-permanent low pressure over Iceland. There is a see-saw of atmospheric mass between these two nodes. The greater (smaller) the pressure difference, the more positive (negative) the NAO index. A positive NAO is associated with mild, wet and often stormy winters over northwestern Europe while a negative NAO is linked with cold, dry conditions in this region, but with wetter weather in the Mediterranean (Xoplaki et al., 2004).

The NAO can be regarded as arising from storm-track and jet-stream variability (Vallis and Gerber, 2008; Stendel et al., 2016), and is an indicator of the zonality of the atmospheric flow. A positive winter NAO tends to arise when the tropospheric jet and storm track are shifted further northwards, driving storms towards western Europe, with a more zonal jet stream, whereas a negative NAO indicates a southerly displacement with an increased meridonal jet-stream component (Stendel et al., 2016) which can steer storms towards the Mediterranean and enable cold air outbreaks from the Arctic to lower latitudes. The NAO is more closely associated with shifts in jet latitude than it is with jet speed variability (Woollings et al., 2010a).

The NAO is a mode of internal atmospheric variability in idealised modelling experiments (e.g. James and James, 1989) and until recently it was considered that NAO variability on intraseasonal and interannual timescales was a result of internal atmospheric variability, or

climate noise (e.g. Feldstein 2000) and largely unpredictable (e.g. Johansson, 2007; Kim et al., 2012). However, there is recent evidence that the storm track and jet stream are subject to forcing from slowly varying boundary conditions such as ocean temperatures and sea-ice changes, together with solar variability and influences from the winter stratosphere (for a review see Hall et al., 2015).

A number of recent studies indicates significant potential for winter seasonal forecasting in the NA region based on the influence of slowly-varying boundary conditions (Scaife et al., 2014; Riddle et al., 2013; Kang et al., 2014; Dunstone et al., 2016; Stockdale et al., 2015). There are also a number of older studies which identify significant skill in seasonal forecasting of the winter NAO using climate models (Palmer et al., 2004; Müller et al., 2005; Derome et al., 2005) and empirical approaches (Fletcher and Saunders, 2006). Scaife et al. (2014) report a correlation skill of 0.62 for hindcasts of the winter NAO over the period 1993-2012, based on the UK Met Office GloSea5 seasonal forecasting system (MacLachlan et al., 2015). Statistical forecasts are quick and cheap to implement (e.g. Cohen et al., 2018) and therefore complement the dynamical forecasts. They allow for identification of sources of potential predictability and may help to explain particular instances of poor forecasts in dynamical NWP models and inform their future development. Recent studies (Dunstone et al., 2016; Wang et al. 2017, Hall et al., 2017) have shown promising skill in predicting the winter NAO using a linear statistical framework and Folland et al. (2012) produced skillful forecasts of winter European temperatures based on a number of these factors. However, these studies only use linear combinations of predictors, not considering non-linear, cross-product and interaction terms. Often statistical models can be constructed for a training period with a very good fit, subsequently failing when making out-of-sample forecasting, due to non-stationary relationships, internal variability and overfitting in the training period.

The slowly varying boundary conditions may act to reinforce or oppose one another and numerous studies examine remote causes of NA atmospheric variability, mostly in a linear framework. However, the interaction of the atmosphere with boundary forcing can be non-linear (Petoukhov and Semenov, 2010), so a purely linear approach may only capture a limited portion of the variability. Here we further develop statistical seasonal forecasting by using a novel application of NARMAX (Non-linear Auto-Regressive Moving Average with eXogenous inputs) methodology (e.g. Billings, 2013), comparing linear and polynomial regression-based forecasting models. We aim to investigate whether the inclusion of non-linear interactions help to explain changes in the NAO. NARMAX modelling can reveal and characterize non-linear dynamic relationships among signals from recorded data, and produces transparent models which demonstrate how a response variable (system output signal) is linked to a number of candidate explanatory variables (system input signals) and their combined interactions. The NARMAX approach will construct the simplest model to explain the system: therefore if a linear model provides a good representation of the system, the NARMAX method will go no further (Billings, 2013, p9). NARMAX modelling was first introduced to solve non-linear dynamical system identification and modelling problems in engineering, and it has been successful in revealing linear and non-linear relationships at a wide range of scales within the engineering, biological, ecological, medical, geophysical, and environmental sciences (e.g. Billings, 2013; Bigg et al., 2014; Ayala-Solares et al., 2018).

Here we review some of the drivers of North Atlantic climate variability identified in previous research. A number of studies identify a sea-ice influence, particularly from the Barents-Kara Sea (BK) region (e.g. Koenigk et al., 2016; Garcia-Serrano et al., 2017). A likely pathway of influence is due to constructive interference of the atmospheric warming related to autumn sea-

ice loss with climatological planetary wave patterns (Screen et al., 2018, Wu and Smith, 2016). The location of the BK region is close to the climatological ridge of the zonal wave-1 and wave-2 planetary waves, with localised warming acting to reinforce this pattern (Zhang et al., 2018) which enhances vertical wave propagation. This then can weaken the stratospheric polar vortex with a subsequent downward propagation of this signal over a number of weeks (Baldwin and Dunkerton, 2001). A stronger (weaker) vortex is associated with a positive (negative) NAO, as a strengthened (weakened) circumpolar stratospheric jet induces a poleward (equatorward) shift in the tropospheric jet and storm tracks (Kidston et al., 2015). A cryospheric influence has also been detected from Siberian snow anomalies, which may enhance the Siberian high-pressure region, resulting in vertical wave propagation into the stratosphere and a weakening of the stratospheric vortex (Cohen et al., 2007). This effect has been observed in models, but can be weak, and may require modulation by the Quasi-biennial Oscillation (QBO) to be more effective (e.g. Tyrrell et al. 2018). The strength of the vortex can also be perturbed by factors such as the QBO phase (Boer and Hamilton 2008), the solar cycle (Ineson et al., 2011) and tropical volcanic eruptions (Robock and Mao,1995; Driscoll et al., 2012).

Sea-surface temperatures (SST) are an important element of boundary layer forcing. The interaction between atmosphere and ocean is complex, with the atmospheric variability characterised by the NAO forcing SST variability in the NA, to produce the distinctive tripole pattern of SSTs (Deser et al., 2010). However, there is evidence for feedback of this SST pattern to the atmosphere at time lags of a few months, as the spring tripole anomalies are preserved beneath the summer mixed layer, re-emerging in winter as the mixed layer deepens (Rodwell et al., 1999, Deser et al., 2003; Czaja and Frankignoul, 1999). A complementary SST pattern is the North Atlantic Horseshoe (NAH; Czaja and Frankignoul, 2002), which may evolve from the tripole anomalies and where SST anomalies may lead the NAO by up to six months. These patterns fluctuate on a decadal scale, but Atlantic SSTs also experience multidecadal variability known as the Atlantic Multidecadal Oscillation (AMO, e.g. Enfield et al., 2001) with warm and cool phases and a period of 65-80 years.

Other recently identified associations involve SSTs in the BK and Greenland-Norwegian (GIN) Seas (Kolstad and Årthun, 2018). In addition, sea-level pressure (SLP) in the BK region can precondition autumn sea-ice extent, and so itself may be a potential predictor of European winter weather variability (King and Garcia-Serrano, 2016). Furthermore, the temperature variability of the NA subpolar gyre (SPG) has been associated with changes in jet speed. A weakened gyre circulation can lead to increased poleward transport of warmer subtropical waters, and a decrease in meridional temperature gradient (Häkkinen et al., 2011; Woollings et al., 2018).

The role of solar fluctuations in North Atlantic climate variability is a subject of considerable debate. A solar cycle signal has been detected in European winters (Lockwood et al., 2010; Woollings et al., 2010b), with lower solar activity associated with colder European winters . Recent studies suggest that there is a lag of 3-5 years between the solar signal and its impact on the atmosphere, measured by the NAO (Scaife et al., 2013; Gray et al., 2013; 2016; Andrews et al., 2015), possibly as a result of integration of the solar signal over time by SSTs.

The influence of tropical teleconnections is also evident. The El-Niño-Southern Oscillation (ENSO) signal can propagate via troposphere and stratosphere (Toniazzo and Scaife, 2006; Bell et al., 2009) and there is evidence of non-linearity, with stronger El Niño events not having the NAO-like impact of moderate events (Folland et al., 2012; Rao and Ren, 2016a; 2016b).

This may be a consequence of a more eastward centre of action for stronger events (Takahashi and Dewitte, 2015) with a different pathway of propagation. In addition, links between Indian Ocean SST anomalies and the NAO, (Hoerling et al., 2004; Li et al., 2010) with the NAO lagging SST anomalies by a month or more and between the NAO and the Madden-Julian Oscillation (Garfinkel et al., 2014 Tseng et al., 2018) have also been identified. Yu and Lin (2016) find an NAO response to tropical heating anomalies in the Indian and Atlantic regions, although not necessarily to SSTs.

Influences from these boundary conditions at a range of lead times means there could be a significant component of predictability within the winter NA atmospheric circulation (Smith et al., 2016). As the NAO is such a significant factor in determining the winter weather around the Atlantic basin, skillful seasonal forecasts of the NAO will have considerable economic and societal benefits. The NAO is related to hydrological outlooks and flood risk (e.g. Svensson et al., 2015; Bell et al., 2017) and is significantly related to energy demand (Thornton et al., 2017; Clark et al., 2017).

Section 2 presents the data used, and section 3 explains the methods, including how the NARMAX methodology is applied. Results are presented in section 4 and interpreted in section 5. This is followed by some concluding remarks in section 6.

2. Data

Data used are summarised in Table I with additional information in Table S1. HadISST1 (Rayner et al., 2003) is used for SST-based and sea-ice variables. We select monthly predictor variables taken at lead times of one month up to a maximum of seven months (the preceding May) preceding the winter in question. For this study it is assumed that there is no prediction skill derived from the previous year's NAO, although November values are available for selection. The sea-ice regions are taken from Screen (2017), where nine distinct sectors are identified with limited covariability suggesting a large degree of regional independence. The NA SST tripole index is constructed following Fan and Schneider (2012). Similarly, the NAH and SST gradient indices are derived by calculating mean SSTs over two regions and subtracting them. See Table S1 for details. Blanca Ayarzaguena provided the T100 index which is a measure of the strength of the stratospheric polar vortex (SPV). This is an index of daily temperature anomalies at 100hPa, averaged over 65-90N, derived from JRA-55 reanalysis data (Kobayashi et al., 2015). Monthly mean values are constructed from an average of daily means. Tropical rainfall anomalies provide an indication of convective activity and divergence aloft, which can generate Rossby waves which propagate away from the source and are capable of influencing extratropical atmospheric circulation (Hoskins and Karoly, 1981). Data are obtained from the Global Precipitation Climatology Project (GPCPv2.3, Adler et al., 2003). However, such data are not available prior to 1979 so tropical SSTs for similar regions are used in the 1956 models, although there is not always a causal link between SSTs and tropical rainfall, depending on the region under consideration. The MJO is another index which can provide a tropical signal and is obtained from the Climate Prediction Center (CPC). Here ten phases are used rather than the more common eight, and a negative value indicates the active convective state (Baxter et al., 2014). QBO data (Naujokat, 1986, updated) use the 30hPa level following Hamilton (1984). All data are normalised to the period 1981-2010.

To capture the non-linear relationship between the NAO and the ENSO signal, we use the discontinuous N3.4 index of Folland et al. (2012) alongside the conventional index, where values are set to zero for normalised values between ±1. More negative values are set to -1, values in the range 1-1.75 are set to one, while values greater than 1.75 are set to zero. The

volcanic index used by Folland et al. (2012) is also applied. Here, for two years following a volcanic eruption, values are set to one, and are zero for all other years, as Robock and Mao (1995) identify the effects of eruptions in the first and second winters after the eruption. This simulates the persistence of volcanic aerosols in the stratosphere after a major eruption. Major tropical eruptions are as identified by Stenchikov et al. (2006).

Two versions of the NAO are used: the PC-based NAO (HPC; Hurrell et al., 2003) and a station-based index derived from SLP data from Reykjavik and Ponta Delgada, Azores (station NAO). This is taken as the difference between SLP at the two stations, which is then normalised. This is the approach followed by Scaife et al. (2014), although other indices are typically constructed by normalising the pressure at each station before subtracting (e.g. Hurrell, 1995; Cropper et al., 2015). The correlation between these two indices for winter is 0.90 (1956-2017); however, this does vary slightly over time (13-year running correlations between the two reach a low of 0.82 for a period in the early 1990s). The PC-based NAO better captures the spatial patterns of the NAO, although is dependent on the area selected for analysis, whereas due to shifts in the NAO centres of action a fixed-point index will not always represent this optimally. A disadvantage of the PC-based approach is that the whole index time series needs recalculating every time a new value is incorporated, and being a mathematical construct, does not necessarily represent climate physics (Dommenget and Latif, 2002)

Data are not detrended as we aim to forecast as closely as possible to the real world. In order to address any trends present in the data, indices of atmospheric carbon dioxide and global temperature are available for incorporation in the models. The winter season is defined as December-January-February (DJF), where the winters refer to the year of the January.

3. Methods

Models are constructed for training periods from 1956-2010 and 1980-2010. The year 2010 is used as the cutoff, as in that year there are extreme values of the NAO and these are included in the training dataset, as simulation experiments show that models trained without extreme values perform poorly when used for prediction. Such extreme values are particularly important in view of the relatively small sample sizes available. This leaves 2011-2018 for use as a validation period (2011-2017 for the station-based NAO due to data availability). We predict values for this period without adapting the model. An alternative, retroactive verification approach (Mason and Baddour, 2008) constructs a model over the training period, then forecasts the next year only based on that model. Forecasts for subsequent years are based on models which incorporate the previous year's observation, and the models are allowed to evolve, both in terms of coefficients and predictors selected. This approach was tested for linear models and produced almost identical forecasts with only slight changes in coefficients, but the same predictors, with no appreciable improvement in forecast quality.

3.1 The NARMAX Method

One of the most attractive features of the NARMAX model, distinguishing it from other non-linear data-driven modelling techniques, is its power to build transparent and interpretable models where the mathematical significance of each model term is meaningful (e.g. Billings, 2013, p10) and in most real applications the selected model terms are physically interpretable (e.g. Billings, 2013, Chapter 14 Case Studies). Essentially, this approach treats each of the candidate predictors as a possible underlying cause of change in the response variable of interest and uses a set of model detection algorithms to automatically identify and pick out the most important predictors, based on which it then establishes a quantitative relationship that best relates the possible forcing variables to the response variable. Note that the dependence

relation of response on potential predictors can be either linear or non-linear. While traditional linear modelling approaches such as ARMA and ARMAX might be able to capture main linear relationships, they fail to reveal or capture any non-linear dynamics that are inherent in weather and climate (e.g. Easterling et al., 2000; Hoerling et al.,2001; Dell et al., 2013; Burke et al., 2015).

Taking the case of a one input (u) and one output (y) problem as an example, the NARMAX model for y is written as (Billings, 2013):

$$\begin{aligned} y(k) = F(&y(k-1), y(k-2), \cdots, y(k-n_y), \\ &u(k-d), u(k-d-1), \cdots, u(k-d-n_u), \\ &e(k-1), e(k-2), \cdots, e(k-n_e)) + e(k) \end{aligned} \tag{2}$$

where $y(k)$ and $u(k)$ are the measured system output (response) and input (explanatory), respectively, at time k; $e(k)$ is a noise sequence which is not measurable but can be estimated once a model is built; n_y, n_u, and n_e are the maximum lags for the system output, input, and noise; $F(\bullet)$ is some non-linear function to be determined; and d is a time delay (typically d=0 or d=1). For an identified model, the noise $e(k)$ can be estimated as the prediction errors: $e(k) = y(k) - \hat{y}(k)$, where $\hat{y}(k)$ is the predicted value at time instant k generated by an estimated model. The noise terms are included to accommodate the effects of measurement noise, modelling errors, and/or unmeasured disturbances. For the purposes of this study, we consider the predictability that can be obtained from the input (predictor) variables of preceding months, and so do not consider past outputs (previous winter NAO values) further.

There are many model subset selection methods such as the traditional forward selection (Faraway, 2002; Wilks, 2011). NARMAX uses an orthogonal forward selection algorithm, called Forward Regression Orthogonal Least Squares (FROLS) algorithm (Billings, 2013), to select the important terms. The efficiency of FROLS can be attributed to its use of mutual information (in addition to the correlation function), to measure not only linear but also potential nonlinear dependent correlation of the target signal and the candidate explanatory signals. Furthermore, unlike traditional stepwise selection which uses hypothesis-tests and p-values to measure the significance of variables (terms), FROLS uses an effective index called the error reduction ratio (ERR) to measure the contribution made by each of the individual terms to explaining the change in the response variable, based on which the most significant term will be selected in each search step. The number of model terms can be determined using either the APRESS (Billings and Wei, 2008) statistics or the penalized error-to-signal ratio (PESR, a variant of APRESS) proposed in Wei et al. (2010). This is similar in principle to the AIC and BIC but is developed specifically for non-linear systems. A leave-K out cross-validation is normally used with NARMAX, where K is around 10% of the training sample. A common model is identified from subsets of the training period, and common model parameters are then estimated using all the training data. In this study, a leave-one-out approach is employed due to the small size sample in the training data. We consider both linear and polynomial NARMAX models, which belong to the family of multiple regression models.

3.1.1. Linear models

NARMAX includes several linear model structures, e.g. autoregressive with exogenous inputs (ARX) and autoregressive moving average with exogenous inputs (ARMAX) as a special case. A general linear model structure of NARMAX is as follows:

$$\begin{aligned} y(k) = &a_1 y(k-1) + \ldots + a_{n_y} y(k-n_y) + b_0 u(k-d) + b_1 u(k-d-1) \ldots + b_{n_u} u(k-n_u), \\ &+ e(k) + c_1 e(k-1) + \ldots + c_{n_e} e(k-n_e) \end{aligned} \tag{2}$$

where $a_1,a_2,...,a_{n_y}$, $b_0,b_1...,b_{n_u}$, and $c_1,c_2,...,c_{n_e}$ are model parameters. The above single input single output (SISO) case model (2) can be extended to multiple input single output (MISO) or multiple input multiple output (MIMO) cases. A special case of the MISO linear model is the commonly used multiple linear regression. For example, for a case where there are *r* inputs, u_1, u_2,…, u_r, by setting d=0, $n_y = 0$, $n_u = 0$, and $n_e = 0$, yields:

$$y(k) = a_1u_1(k) + a_2u_2(k) + ... + a_ru_r(k) + e(k) \quad (3)$$

3.1.2. Polynomial NARMAX models

In practice, many types of model structures are available to approximate the unknown function $F(\bullet)$ in (1), including power-form polynomial models and rational models (Chen and Billings, 1989), radial basis function networks (Wei et al., 2007), and wavelet neural networks (Billings and Wei, 2005; Wei et al., 2010). Power-form polynomial models are the most commonly used representation because such models have a number of unique, attractive properties (Billings 2013, pp35-37): for example for most applications the resulting models are transparent, physically interpretable and simple (parsimonious). This is the model form used in this study.

In practical applications, it is usual to consider many input signals (or explanatory variables) and investigate how the explanatory variables (e.g. u_1, u_2, ..., u_r) influence the response variable of interest. A case of many inputs can be represented by a special form of the NARMAX model as follows:

$$\begin{aligned} y(k) = f(&u_1(k),u_1(k-1),\cdots,u_1(k-n_u),\cdots,u_2(k),u_2(k-1),\cdots,u_2(k-n_u),\cdots, \\ &u_r(k),u_r(k-1),\cdots,u_r(k-n_u)) + e(k) \end{aligned} \quad (4)$$

where $n_u \geq 0$. In this study, $n_u = 0$, for which the model reduces to a polynomial model which belong to the family of multiple linear regression models:

$$y(k) = f(u_1(k), u_2(k), ... u_r(k) + e(k) \quad (5)$$

For example, with two inputs u_1 and u_2, the initial full model of degree 2 is

$$y(k) = a_0 + a_1u_1(k) + a_2u_2(k) + a_3{u_1}^2(k) + a_4u_1(k)u_2(k) + a_5{u_2}^2(k) + e(k) \quad (6)$$

Note that in practice it may not be necessary for all the six model terms in (5) to be included in a final predictive model, and only those that are important for explaining the variation of the response should be included in the final model.

3.2. Small Sample Size Problem and Model Averaging

The NARMAX method has been successfully applied to solve a wide range of real-world problems, and in most cases the method produces a robust reliable model that can be used for system analysis and prediction (simulation). For small sample-size problems where the number of observations is small and much smaller than the number of regressors (explanatory variables and their cross-product interactions), the identified model can be sensitive to adding or removing a variable or cross-product term. In order to reduce the risk of using a single model and mitigate the uncertainty in the prediction of a single model, this study proposes a simple model-averaging approach to deal with the small sample-size problem. Models used in model averaging only differ from each other in the number of terms used, thus all contain a common core of predictors, with only minor differences in the coefficients.

Assume the training dataset *S* contains a total of *N* data points. Rather than using only a single model to calculate predictions, we use a weighted average of multiple models (in this case three) to carry out predictions. The model averaging scheme is described below.

Let $M_1, M_2, ..., M_s$ be the *s* best models identified by the PESR from the training dataset *S*, and assume the values of the mean squared errors (MSE) of the *s* models over the training dataset are $mse_1, mse_2, ..., mse_s$, respectively. Define

$$c_1 = 1/mse_1, c_2 = 1/mse_2, ..., c_s = 1/mse_s \quad (7a)$$

$$c = c_1 + c_2 + ... + c_s \quad (7b)$$

$$w_1 = c_1/c, w_2 = c_2/c, ..., w_s = c_s/c \quad (7c)$$

Let $\hat{y}_1, \hat{y}_2, ..., \hat{y}_s$ be the predicted output values by the *s* models, the model averaging prediction is defined as:

$$\hat{y} = w_1\hat{y}_1 + w_2\hat{y}_2 + ... + w_s\hat{y}_s \quad (8)$$

3.3. Forecast Verification

Correlation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to give an indication of forecast skill, for both training and testing periods. A further measure is used, based on the mean squared error (MSE), the Mean Square Error Skill Score (MSESS, e.g. Wilks, 2011, p328). This compares the skill of the forecast with a reference forecast, in this case a forecast based on climatology. There is generally lower skill in seasonal forecasting compared with shorter term forecasts and an out-of-sample forecast of 8 years is not large enough to allow robust statistical conclusions to be drawn about the quality of the forecast. A generalized discrimination score, D, (Mason, 2012; Mason and Weigel, 2009) indicates whether the forecasts are potentially useful, despite bias and poor calibration in a low-skill situation (Mason, 2012), based on whether forecast values increase (decrease) with an increase (decrease) in observation values, regardless of the magnitude of the change. For a pair of observations, D gives the probability that a forecaster can discriminate the observations based on the corresponding forecasts. D is related to Kendall's correlation coefficient τ by:

$$\tau = 2D - 1 \quad (9)$$

τ is effectively scaled from 0 to 1, a value of 0.5 means there is no skill in the forecast as the probability of correctly discriminating between the size of two observations is 50%. Values greater than 0.5 suggest that the forecast is potentially useful.

Comparing the forecast methods objectively and determining whether there is a significant difference is difficult with a small out-of-sample testing set (N=8; seven for station NAO). To address this, the longer time series is split into even and odd years for the station NAO data. Models are trained on the even years (N=32), the odd years forming the testing period (N=31). The differences between linear and polynomial NARMAX model correlations for the testing period are assessed for significance with the "cocor" package for R (Diedenhofen and Musch, 2015), using tests for overlapping dependent samples (the correlations both concern the NAO and the compared coefficients have a shared variable, the observed NAO). The package runs six different tests and their variations, for assessing the significance of the difference between the correlations, together with a confidence interval test. For details see Diedenhofen and Musch (2015).

A further test is performed to examine whether the results obtained in the testing period for NARMAX models are likely to have occurred by chance. For the HPC80LIN model, 100 surrogate datasets (Kugiumtzis, 2000; Schrieber and Schmitz, 2000) are created for each of the predictors selected, allowing 100 surrogate NAO timeseries to be constructed for each of the models with different numbers of terms (models with five to nine terms were examined). Further details of this approach are given in the online supplementary information.

4. Results

In this section, NAO models are designated by type (station-based NAO: stat; Hurrell principal-component NAO: HPC), by year, and whether linear (LIN) or polynomial (POLY) NARMAX models. Thus stat80POLY is the polynomial NARMAX model for the station-based NAO, using 1980-2010 as the training dataset.

4.1 Linear Models

The predictors used in the linear models, and their coefficients are summarised in Table II and verification statistics for the averaged models are shown in Table III. It is the averaged models that are subsequently discussed, unless otherwise stated. An example of the model fit is shown for the station-based NAO in Figure 1, with other models shown in Figure S2. There is some consistency among the predictors selected for the different NAO indices, both within and between the different training periods (Table II). October Barents Sea SLP and October Barents-Kara Sea ice are selected for all models. Bering Sea and East Siberian-Laptev Sea ice in November are selected for three of the four models; both 1980 linear models, and HPC56LIN (Bering Sea ice only) and stat56LIN(East Siberian-Laptev Sea ice only). There is some discrepancy in the cryospheric terms, with Greenland Sea ice being selected only for the 1956 models. While Greenland Sea ice is chosen in the 1956 NAO models, October Bering Sea ice is selected for HPC80LIN as an additional term. Tropical influences are identified in both 1980 models, and for stat56LIN, however these show considerable variation, as variables selected for the 1980 models such as the MJO and tropical rainfall, are not available for the 1956 models. Consequently the greater availability of tropical predictors post-1980 is reflected in these models: November MJO (phase 8, HPC80LIN and phase9, stat80LIN) and tropical Atlantic rainfall (July; stat80LIN) and August (HPC80LIN) are present in both. Despite being previously identified as an important predictor of winter extratropical atmospheric variability, N3.4 is not particularly prominent, only being selected for both station-based NAO models (July N3.4I). Extratropical oceanic SST-related terms are selected for all except stat80LIN; October GIN SST for HPC80LIN and NAH (September-stat56LIN, October-HPC56LIN). The only model to detect a stratospheric influence is HPC56LIN, where the stratospheric polar vortex indices for October and November are selected, although interestingly these are of opposite signs and are only weakly correlated (r=0.21).

Correlations of the 1980 models for the 2011-2018 testing period (Table IIIa) are 0.90 (HPC80LIN) and 0.82 (stat80LIN). However, correlations can mask a systematic negative bias in the forecasts, particularly in the linear model forecasts (with the exception of HPC80LIN), which becomes evident when MAE and RMSE are examined. For example, although stat80LIN correlates very well in the testing period (r=0.82), MAE and RMSE are 1.11 and 1.26 respectively, considerably higher than the training period values. MSESS scores are negative, the forecasts having less skill than a climatological forecast. The correlations indicate that the two NAO models are able to capture aspects of the local maxima and minima of the observed NAO during the testing period, but with a large negative bias (Figure 1), which reduces skill according to MSESS and inflates the MAE and RMSE. However, D-scores suggest that the models for the NAO have considerable potential usefulness (0.80, stat80LIN; 0.75, HPC80LIN, Table IIIa).

For the 1956 linear models, over the testing period the correlation is weaker for both (HPC56LIN, 0.46; stat56LIN, 0.69; Table III) compared with the 1980 models although no significant difference can be determined due to small sample size. Test data MAE and RMSE scores are similar to the 1980 linear model for the station NAO, but greater for the HPC NAO,

but are always larger than those for the training period, reflecting the negative bias of forecasts in the testing period. MSESS are negative, but D-scores indicate potential usefulness (0.80, stat56LIN; 0.68, HPC56LIN).

Many of the terms used in construction of the linear models, for example sea-ice, SST and tropical rainfall terms contain significant trends, which are likely to contribute to the negative bias of the forecasts. It is notable that the station NAO forecasts from 1980 training data appear to have overall negative trends (Figure 1), probably due to the influence of the sea-ice inputs to the models, as found in Hall et al. (2017).

4.2 Polynomial Models

The polynomial models are presented in detail in Table IV and compared with linear models in Figures 1 and S2. The comparisons use the averaged linear and polynomial models and an example of model averaging for the stat80POLY model is shown in Figure S3. Verification statistics are in Table III. For the HPC80 forecast the NARMAX algorithm produces a linear model only, as the inclusion of interaction terms does not significantly improve prediction performance.

There is less consistency amongst the predictor variables selected (Table IV) compared with the linear models, as terms are more complex and often based on interactions between variables, and with short datasets the models are very sensitive to slight differences in predictors. Equally, different predictors selected may be trying to capture the same variability or acting as proxies for some hidden variable. Some predictors are selected for both linear and polynomial models, but may differ by a month or two.

Forecasting verification statistics for the polynomial model averages are shown in Table III. In almost all cases, MAE, RMSE, MSESS and correlation are superior to the linear models in both training and testing periods, for both time periods. The D-score is consistently above 0.7 indicating considerable potential usefulness of the forecasts and MSESS are always positive. Verification statistics are mostly poorer for the 1956 models. Initially this may appear to be surprising as the models are constructed over a longer training set, although data from the earlier part of the time series prior to the satellite era are of poorer quality, and some variables are not available for the earlier period. See also section 5 on variable inherent predictability over the time period. The 1956 polynomial models show considerable improvements in all verification metrics compared with linear models on the longer time-scale, although results from linear models are broadly comparable to results obtained using the shorter training set. Due to the small sample size, the statistical significance of any improvements achieved by polynomial over linear models cannot always be satisfactorily determined. As would be expected for such a small sample size, when correlations in the testing period are compared using the suite of tests in the "cocor" package, differences are not significant.

In order to create a longer testing period and provide greater confidence in the performance of the models in out-of-sample prediction, the 1956-2017 period was split into even years for the training dataset and odd years for the testing dataset, using the station-based NAO index. The predictive skill of NARMAX polynomial and linear models when even years only are used as the training period is shown in Figure S4 and verification statistics are presented in Table S2. Correlations for the testing period (odd years) are 0.45 (linear) and 0.60 (polynomial). While both are significant ($p<0.05$), the polynomial correlation coefficient is appreciably higher, consistent with results from the shorter testing datasets, yet still not deemed significantly so by the range of comparison tests in "cocor". However, this does not tell the whole story. For the

earlier part of the testing period (1957-1979), correlations for the linear and polynomial models are very similar (0.6, linear; 0.66, polynomial), however in the later period (1981-2017), the correlations are significantly different using "cocor" (0.04, linear; 0.44 polynomial, $p<0.1$, one-tailed test). In addition both MAE and RMSE for 1981-2017 are around two thirds of the values for 1957-1979. This is suggestive of some change in the NAO in the later period, which is better captured by the polynomial models.

Verification statistics taken over the 100 surrogate models for HPC80LIN show that model performance is much poorer for the surrogates in the testing period (Table S3), with significant differences between model performance and the surrogate data for all verification metrics. This indicates that the good prediction performance obtained with NARMAX is not a result of chance but is due to the efficacy of the algorithm.

Figure 2 allows closer inspection of model performance in the testing period. Both linear and polynomial 1980 models capture the local maxima and minima of both versions of the NAO for the first three years (Figure 2a,b), although the amplitude of the response is much reduced in the stat80LIN model (Figure 2b), which also display a negative bias, as discussed above, while any bias in stat80POLY is minimal and non-systematic. Similarly1956 linear models for both NAO indices reproduce the interannual change from 2011-2014, after which the response is damped. In general, polynomial models better match the local maxima and minima, with the exception of HPC56POLY which forecasts a negative rather than positive NAO in 2014 (Figure 2c).

Figures 3 and 4 give a visual representation of the selected predictors for the different NAO models, for both training periods, in order to highlight common predictor variables, although some of these are used in interaction terms in the NARMAX models. For 1980 (Figure 3), Barents Sea SLP and Bering Sea ice are used in all three models, Barents-Kara Sea ice, East Siberian-Laptev Sea ice and tropical Atlantic rainfall are selected by both linear models while MJO phase 8 occurs for both linear and polynomial HPC models. For the 1956 models (Figure 4), some predictors (Greenland Sea ice, NAH, Barents Sea SLP) are included in all four models, while October Barents-Kara Sea ice and the N3.4I index are selected in three. A number of predictors occur in only one model, due to them occurring as interaction terms, and the sensitivity of the model to slight changes in variables given the short time series. For the 1956 training data, linear and polynomial models will select predictors that were suboptimal in the 1980 models, as the best predictors are not always available for the longer time series. Thus NARMAX selects N3.4I for the 1956 models, whereas for the 1980 models the optimal tropical influences seems to come from the MJO and tropical rainfall. While models are capturing essentially the same signals, there is high sensitivity to small variations in input data due to small sample size: hence different predictors are selected, which represent slightly different aspects of a common signal.

From the results, it is remarkable that while linear models show more limited skill, particularly over the longer training dataset, polynomial model forecasts, whether linear or non-linear, particularly those based on the 1980 training set, are able to replicate local maxima and minima, the amplitude of the observation and have minimal bias for the testing dataset.

5. Discussion

5.1 Differences when using longer and shorter training periods

Based on the verification statistics (Table III), polynomial models generally outperform the equivalent linear model (lower MAE and RMSE; higher correlation, MSESS and D-score),

although the significance of this is hard to assess on a small sample size. 1956 models generally perform less well than the equivalent from 1980. There is evidence from other studies to suggest that from the 1950s to 1960s there is less skill in dynamical forecasting hindcasts, with the possibility that during this period the atmospheric circulation was inherently less predictable (e.g. Weisheimer et al., 2016), associated with a more prevalent negative NAO. However, dynamical models appear to be skillful in predicting strongly negative NAO events, but less skillful regarding weaker negative NAO events. O'Reilly et al. (2017) concur that in the mid 20th century, forecast skill for the NAO is reduced, and attribute this to weaker forcing from tropical Pacific SSTs during this period. An alternative perspective is presented by Woollings et al. (2015). Decadal variability of the NAO is identified, which is associated with changes in strength of the jet and different dynamical behaviour (eddy-mean flow interaction, Rossby wavebreaking and blocking), the period from 1980 being associated with relatively higher jet speeds. Therefore when statistical models are trained on a longer period, with the assumption of stationarity, these decadal variations are averaged out and predictors that are selected are likely to be sub-optimal. Furthermore, if there is a return to relatively weak forcing from Pacific SSTs, or a period of reduced jet speeds, forecasts trained on the recent period may perform less well (O'Reilly et al., 2017). Weisheimer et al. (2016) report that forecast skill actually increases further back in the 20th century, in the 1930s and 1940s, suggesting the issue may not be attributed to reduced data quality. This corresponds to the decadal fluctuations identified by Woollings et al. (2015), suggesting that periods of lower jet speed may contribute to the reduced predictability. In this study, a longer training dataset does not equate to a better forecasting model, in agreement with the studies above. It is also notable that when tested on all the odd years, a polynomial model significantly outperforms a linear model for the period 1981-2017, better capturing the NAO for both early and late periods.

Greenland Sea ice is a predictor in all 1956 models (Figure 4) but is not selected for models based on the shorter training period. The months selected (May-July) represent the time of maximum decrease in the regional annual sea-ice cycle (October is selected as the 8th and final term in stat56LIN, Table IIc, and makes a minimal contribution). Interannual variability and mean values of Greenland Sea ice are notably larger prior to 1980; consequently contributions of the predictor terms to the models are greatest prior to 1980. Given the relatively small coefficients used in the models, contributions between 1980 and 2010 are negligible, also indicating why this variable is not selected for the 1980 models. Figure S5 shows that the four Greenland Sea ice terms used in the models essentially make similar contributions to each model, and can be regarded as different, somewhat imperfect representations of some unknown predictor variable. It is possible that the association in the 1960s and 1970s is related to the Great Salinity Anomaly (e.g. Dickson et al., 1988), which circulated in the North Atlantic at this time and had its origins in increased ice export and freshwater release through the Greenland Sea region. The association between the Greenland Sea ice and NAO is physically plausible at this time, particularly as the predictors capture the transition from a more negative to a positive NAO over the period.

5.2. Tropical forcing and interaction terms

Evidence for forcing of the winter NAO from the tropics is evident in all models; the 1980 training data favours the selection of the MJO and tropical rainfall, whereas these are not available for the 1956 training data, so the N3.4 Index becomes more prominent, along with West Indian Ocean SST in stat56LIN and HPC56POLY. It is also notable that HPC56LIN includes no tropical forcing, although the use of stratospheric polar vortex terms here could incorporate aspects of tropical forcing, such as the ENSO-stratospheric-mid-latitude teleconnection (e.g. Bell et al., 2009). Tropical variables selected are somewhat inconsistent,

in terms of both month and predictor, which is likely to be a result of model sensitivity to slight variations due to the short time series used. However, the evidence for tropical signals is clear.

There are multiple interaction terms identified in the polynomial models, which can be very difficult to explain effectively without subsequent further analysis. N3.4I terms are selected in both 1956 polynomial models, possibly due to the unavailability of MJO indices as discussed above. However, in each case it occurs as an interaction term with a high latitude predictor (terms 9 and 10, stat56POLY; terms 4 and 8, HPC56POLY, Table IVb,c). A closer inspection of the interaction terms' time series reveals that the main input parameter is N3.4I, with the high latitude term modulating the signal in terms of amplitude and sometimes polarity (not shown). This, along with term 1 in stat56POLY (Table IVb), is suggestive of a tropical signal, propagated by Rossby waves, which is then modulated by slowly-varying boundary conditions at higher latitudes (e.g. Ding et al., 2014). Term 5 in HPC56POLY (June Hudson Bay ice x volcanic index; Table IVc) operates in a similar way, the magnitude of the tropical volcanic signal being the dominant input, modulated by the sea-ice term.

A composite plot of low minus high years for MJO phase 8 in October shows a wave train emanating from the central Pacific, over North America and to the Atlantic (Figure S6). Poleward propagating Rossby waves emanate from only a few source regions, whose activity varies with tropical rainfall variability from year to year (Scaife et al., 2017). Of note in this wavetrain is a node boundary over the Labrador Sea. An interaction term between September Labrador Sea ice and the October MJO8 is identified for stat80POLY (Figure 3, Table IVa). It is possible that in years with high sea-ice in this region, there is an interaction with the MJO signal, and the effect of this interaction is sustained into winter. As the interaction term is multiplicative, then it is the sign of each index which is particularly important. Both high (low) sea-ice and high (low) MJO8 will combine to make the NAO prediction more positive. If however, one input is positive and the other negative, that will combine towards a more negative NAO forecast.

5.3. Higher latitude forcing and interaction terms.

Both cryospheric and extratropical forcings are included in all models, often as interaction terms with tropical forcing as discussed above. Key sea-ice predictors are October Barents-Kara Sea (five out of seven models), November East Siberian -Laptev Sea (four out of seven), Bering Sea (five out of seven). The Greenland Sea has been discussed in Section 5.1 above. Sea-ice in the Barents-Kara Seas has frequently been identified as a key source of potential winter NAO predictability (Scaife et al., 2014; Garcia-Serrano et al., 2015; Wang et al., 2017; Hall et al., 2017). The importance of Bering Sea ice is more surprising, but could well be a proxy for atmospheric variability in the region, such as the Pacific-North American pattern (PNA), with associations with North Atlantic atmospheric variability (e.g. O'Reilly et al., 2017), which in turn is linked to tropical Pacific SST variability.

Another input variable worthy of more detailed discussion is Barents SLP. This is selected by all NAO models (Figures 3, 4). The most commonly selected month is October (four times). In polynomial models Barents Sea SLP occurs as an interaction term with tropical signals (stat80POLY, Table Iva, HPC56POLY, Table IVc), as a single term (stat56POLY, Table IVb), an interaction term with SSTs (stat56POLY, Table IVb) and the QBO (HPC56POLY, Table IVd). A cyclonic (anticyclonic) anomaly in the BK region in October can lead to positive (negative) sea-ice anomalies there in November (King and Garcia-Serrano, 2016). In other words, the preceding pressure/geopotential height anomaly is a precursor of the Barents-Kara Sea ice ice which has been frequently identified as a predictor in other studies (e.g. Garcia-

Serrano et al., 2015; Hall et al., 2017). However, it also has the advantage that it is not subject to the same long-term dramatic trends as autumn sea-ice. The interaction terms support the concept of geopotential height anomalies over the BK Seas in autumn modulating signals from the tropics, stratosphere and cryosphere in a non-linear way (e.g. Vihma et al., in review, International Journal of Climatology). There are a number of input terms from these high latitude regions that are combined in polynomial models: [November east Siberian Sea ice x November GIN seas SST], [October Barents SLP x November GIN seas SST] (stat56POLY); [September GIN seas SST x October Atlantic SST gradient] (HPC56POLY). It seems likely that these terms are capturing different aspects of autumn variability in the northern seas, that would merit further investigation as important predictors of the winter NAO.

Labrador Sea ice is only selected as a predictor in polynomial models, as an interaction term with tropical (see section 5.2 above) or extratropical forcings. An interaction term selected in both HPC56POLY and stat80POLY is [October Labrador Sea ice x October NAH SST pattern]. The NAH pattern of SSTs is associated with forcing of the winter NAO by persistent SST anomalies at up to six months lead time (Czaja and Frankignoul, 2002). The separate input terms and the resulting multiplicative term are shown in Figure S7a. An examination of the 13-year running correlation between the October NAH and winter NAO reveals a consistent negative correlation, except for the period from 1995 to 2005 when the correlation coefficient increases sharply, becoming positive, before a rapid return to negative values (Figure S7b). This positive excursion coincides with large positive Labrador Sea ice anomalies in October, these being negative for the rest of the time series. The regions for both these variables have a partial overlap. Therefore when there is a negative ice anomaly in the Labrador Sea, the inverted NAH index provides a predictor of the winter NAO interannual variability; however for the brief period with positive ice anomalies, there is a positive relationship between the NAH and NAO interannual variability (Figure S7a). The sea-ice in the Labrador Sea modulates the NAH/NAO interaction, or there is a hidden North Atlantic variable for which Labrador Sea ice is a proxy.

6. Summary

The NARMAX approach shows appreciable potential skill in out-of-sample forecasting, albeit with small testing datasets, for both linear and polynomial models which both outperform a more conventional ordinary least squares approach to multiple regression (e.g. Hall et al., 2017). There are strong correlations with observations, reproducing local maxima and minima of the observations, and the amplitude of the observed signal. Model fits are improved when based on a shorter training dataset from 1980-2010. This may partly relate to a wider range of potential predictors being available for this period, but is also because of reduced inherent predictability of circulation indices during the 1950s and 1960s. The skill of polynomial models is greater than that of equivalent linear models, and error statistics are reduced, but small sample size means that further work is needed to establish the significance of this result. However, an analysis based on using odd years for the testing data is strongly suggestive of a better performance by a polynomial model over that of a linear model, particularly in that it better represents the transition from the early more negative NAO period to the end of the 1970s to the more positive phase post-1980. NARMAX can identify important predictors of winter North Atlantic atmospheric variability. Discrepancies between predictor selection in models is likely to arise through increased sensitivity to small fluctuations in input, due to the small sample sizes available. This means that the models capture the same signals, but select them in slightly different ways.

An important result of the study is that polynomial NARMAX models are capable of revealing the potential modulation of tropical forcing by higher latitude boundary conditions. Barents

Sea SLP can also play a crucial role in modulating cryospheric and extratropical signals in addition to those from the tropics. This could be significant in developing the next generation of NWP models.

Our NARMAX approach can be extended to other circulation indices such as the East Atlantic and Scandinavian Patterns and jet latitude and speed, and to other seasons. This may be especially beneficial for summer seasonal forecasts, where there is currently relatively little predictability from dynamical models (e.g. Ossó et al., 2018; Dunstone et al., 2018). It is also possible to extend the approach to probabilistic forecasting and - by utilising links between North Atlantic circulation patterns and, for example, UK regional temperature and precipitation patterns (Hall and Hanna, 2018) - provide enhanced seasonal forecasts that should be useful for a wide range of stakeholders. NARMAX can also be used to assess how contributions from different atmospheric circulation predictors vary over time using a moving window approach. Future work will use Coupled Model Intercomparison Project (CMIP) 5/6 output to construct models using a longer timeseries, enabling the use of a longer testing dataset, to confirm whether polynomial forecasts are significantly improved compared to linear versions. The models can be further extended to include previous years' predictor values, and by increasing the lead-time at which forecasts are issued for a given season.

Acknowledgements

We thank Blanca Ayarzaguena for providing the stratospheric polar vortex data, and the providers of the numerous datasets used in the study. We thank the three anonymous reviewers for their insightful comments which greatly improved the manuscript.

References

Adler, R. F., Huffmann, G. J., Chang, A., Ferrar, R., Xie, P., Janowiak, J., … Arkin, P. (2003): The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). Journal of Hydrometeorology, 4, 1147–1167, doi:10.1175/1525-7541(2003)004, 1147:TVGPCP.2.0.CO;2

Andrews, M. B., Knight, J.R., & Gray, L.J. (2015). A simulated lagged response of the North Atlantic Oscillation to the solar cycle over the period 1960–2009. Environmental Research Letters, 10, 054022, doi:10.1088/1748-9326/10/5/054022

Ayala-Solares, J.R., Wei, H.-L., & Bigg, G.R. (2018). The variability of the Atlantic meridional circulation since 1980, as hindcast by a data-driven nonlinear systems model. Acta Geophysica 66, 683-695, doi: 10.1007/s11600-018-0165-7

Baxter, S., Weaver, S., Gottschalck, J., & Xue, Y. (2014). Pentad evolution of wintertime impacts of the Madden-Julian Oscillation over the contiguous United States. Journal of Climate, 27: 7356-7367, doi:10.1175/JCLI-D-14-00105.1

Baldwin, M. P., & Dunkerton, T.J. (2001). Stratospheric harbingers of anomalous weather regimes. Science, 294, 581–584, doi:10.1126/science.1063315

Bell, C. J., Gray, L.J., Charlton-Perez, A.J., & Joshi, M.M. (2009). Stratospheric communication of El Niño teleconnections to European winter. Journal of Climate, 22, 4083–4096, doi:10.1175/ 2009JCLI2717.1

Bell, V.A., Davies, H.N., Kay, A.L., Brookshaw, A., & Scaife, A.A. (2017). A national-scale seasonal hydrological forecast system: development and evaluation over Britain. Hydrological and Earth System Sciences Discussions, doi: 10.5194/hess-2017-154

Bigg, G.R., Wei, H.-L., Wilton, D.J., Zhao, Y., Billings S.A., Hanna, E., & Kadirkamanathan, V. (2014). A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change. Proceedings of the Royal Society Series A, 470. doi:10.1098/rspa.2013.0662.

Billings, S. A. (2013). *Non-Linear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. London: Wiley.

Billings, S.A. & Wei, H.-L. (2005). The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions. International Journal of Systems Science, 36: 137–52.

Billings, S.A. & Wei, H.-L. (2008). An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. International Journal of Control, 81(5), 714-724, doi: 10.1080/00207170701216311.

Boer, G.J., & Hamilton, K. (2008). QBO influence on extratropical predictive skill. Climate Dynamics, 31: 987-1000, doi: 10.1007/s00382-008-0379-5

Burke, M., Hsiang, S.M., & Miguel, E. (2015). Global non-linear effect of temperature on economic production. Nature, 527: 235–23.

Chen, S., & Billings, S.A. (1989). Representation of non-linear systems: The NARMAX model. International Journal of Control, 49: 1013–32.

Clark, R.T., Bett, P.E., Thornton, H.E., & Scaife, A.A. (2017). Skilful seasonal predictions for the European energy industry. Environmental Research Letters 12: 024002. doi: 10.1088/1748-9326/aa57ab

Cohen, J., Barlow, M., Kushner, P.J., & Saito, K. (2007). Stratosphere-troposphere coupling and links with Eurasian land surface variability. Journal of Climate, 20: 5335-5343, doi: 10.1175/2007JCLI1725.1

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., & Tziperman, E. (2018). S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. WIREs Climate Change 2018e00567, doi: 10.1002/wcc.567.

Cropper, T., Hanna, E., Valente, M.A., & Jónsson, T. 2015. A daily Azores-Iceland North Atlantic Oscillation index back to 1850. Geoscience Data Journal, 2(1): 12-24, doi: 10.1002/gdj3.23

Czaja, A., & Frankignoul, C. (1999). Influence of the North Atlantic SST on the atmospheric circulation. Geophysical Research Letters, 26: 2969-2972.

Czaja, A., & Frankignoul, C. (2002). Observed impact of Atlantic SST anomalies on the North Atlantic Oscillation. Journal of Climate, 15: 606-623.

Dell, M., Jones, B.F., & Olken, B.A. (2013). What do we learn from the weather? The new climate-economy literature. Journal of Economic Literature, 52(3): 740:798.

Derome, J., Lin, H., & Brunet, G. (2005). Seasonal Forecasting with a simple general circulation model: predictive skill in the AO and PNA. Journal of Climate 18:597-609.

Deser, C., Alexander M.A., & Timlin, M.S. (2003): Understanding the persistence of sea surface temperature anomalies in midlatitudes. Journal of Climate, 16, 57-72.

Deser C., Alexander, M.A., Xie, S.-P., & Phillips, A.S.(2010). Sea surface temperature variability: patterns and mechanisms. Annual Reviews in Marine Science, 2: 115-143, doi: 10.1146/annurev-marine-120408-151453

Diedenhofen, B., & Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. PLoS ONE 10(4): e0121945, doi: 10.10371/journal.pone.0121945

Dickson, R.R., Meincke, J., Malmberg, S.-A., & Lee, A.J. (1988). The "Great Salinity Anomaly" in the Northern Atlantic, 1968-1982. Progress in Oceanography 20, 103-151.

Ding, Q., Wallace, J.M., Battisti, D.S., Steig, E.J., Gallant, A.J.E., Kim, H-J, &Geng, L. (2014). Tropical forcing of the recent rapid Arctic warming in northeastern Canada and Greenland. Nature 509, 209-212, doi: 10.1038/nature13260

Dommenget, D., & Latif, M. (2002), A cautionary note on the interpretation of EOFs. Journal of Climate 15, 216-225.
Driscoll, S., Bozzo, A., Gray, L.J., Robock, A., & Stenchikov, G. (2012). Coupled Model Intercomparison Project 5 (CMIP5) simulations of climate following volcanic eruptions. Journal of Geophysical Research, 117: D17105, doi: 10.1029/2012D017607
Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., …Knight, J. (2016). Skilful predictions of the winter North Atlantic Oscillation one year ahead. Nature Geoscience, 9: 809-814. doi: 10.1038/NGEO2824
Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Fereday, D., O'Reilly, C., …Belcher, S. 2018. Skilful seasonal predictions of summer European rainfall. Geophysical Research Letters, 45: 3246:3254. doi: 10.1002/2017GLO76337
Easterling, D.R., Meehl, G.A., Parmesan, C., Changnon, S.A., Karl, T.R., & Mearns, L.O. (2000). Climate extremes: observations, modeling, and impacts. Science, 289: 2068–2074.
Enfield, D.B., Mestas-Nuñez, A.M., & Trimble, P.J. (2001). The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S. Geophysical Research Letters, 28(10): 2077-2080.
Fan, M., & Schneider, E.K. (2012). Observed decadal North Atlantic tripole SST variability. Part i: weather noise forcing and coupled response. Journal of the Atmospheric Sciences, 69: 35-50, doi: 10.1175/JAS-D-11-018.1
Faraway, J (2002) Practical Regression and ANOVA Using R. Available at: http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf. Accessed November 2018.
Feldstein, S. (2000). The timescale, power spectra, and climate noise properties of teleconnection patterns. Journal of Climate, 13: 4430-4440
Fletcher, C.G., & Saunders, M.A. (2006). Winter North Atlantic Oscillation Hindcast Skill: 1900-2001. Journal of Climate, 19: 5762-5776.
Folland, C.K., Scaife, A.A., Lindesay, J., & Stephenson, D.B. 2012. How predictable is northern European winter climate a season ahead? International Journal of Climatology, 32: 801-818, doi: 10.1002/joc.2314
García-Serrano, J., Frankignoul, C., Gastineau, G., & De La Cámara, A. (2015). On the predictability of the winter Euro-Atlantic climate: lagged influence of autumn sea ice. Journal of Climate 28, 5195-5216, doi: 10.1175/JCLI-D-14-00472.1.
García-Serrano, J., Frankignoul, C., King, M.P., Arribas, A., Gao, Y., Guemas, V., …Sanchez-Gomez, E. (2017). Multi-model assessment of linkages between eastern Arctic sea-ice variability and the Euro-Atlantic atmospheric circulation in current climate. Climate Dynamics**,** 49: 2407-2429, doi: 10.1007/s00382-016-3454-3.
Garfinkel, C.I., Benedict, J.J., & Maloney, E.D. (2014). Impact of the MJO on the boreal winter extratropical circulation. Geophysical Research Letters, 41: 6055-6062, doi: 10.1002/2014GL061094
Gray, L.J., Woollings, T.J., Andrews, M., & Knight, J. (2016). Eleven-year solar cycle signal in the NAO and Atlantic/European blocking. Quarterly Journal of the Royal Meteorological Society, 142: 1890-1903, doi: 10.1002/qj.2782
Gray, L.J., Scaife, A.A., Mitchell, D.M., Osprey, S., Ineson, S., Hardiman, S., …Kodera K. (2013). A lagged response to the 11 year solar cycle in observed winter Atlantic/European weather patterns. Journal of Geophysical Research: Atmospheres, 118: 13 405-13-420, doi: 10.1002/2013JD020062
Häkkinen, S., Rhines, P.B., & Worthen, D.L. 2011. Atmospheric blocking and Atlatnic multidecadal ocean variability. Science, 334: 655-659, doi: 10.1126/science.1205683
Hall, R.J., & Hanna, E. (2018). North Atlantic circulation indices: links with summer and winter UK temperature and precipitation and implications for seasonal forecasting.

International Journal of Climatology, 38(S1): e660-e677, doi: 10.1002/joc.5398
Hall, R., Erdélyi, R., Hanna, E., Jones, J.M., & Scaife, A.A. (2015). Drivers of North Atlantic polar front jet stream variability. International Journal of Climatology, 35: 1697-1720. doi: 10.1002/joc.4121
Hall, R.J., Scaife, A.A., Hanna, E., Jones, J.M., & Erdélyi, R. (2017). Simple statistical probabilistic forecasts of the winter NAO. Weather and Forecasting 32: 1585-1601. doi: 10.1175/WAF-D-16-0124.S1
Hamilton, K. (1984). Mean wind evolution through the quasi-biennial cycle in the tropical lower stratosphere. Journal of the Atmospheric Sciences 41: 2113-2125.
Hanna, E, & Cropper, T.E. (2017). *North Atlantic Oscillation.* Oxford Research Encyclopedia of Climate Science. doi: 10.1093/acrefore/9780190228620.013.22
Hoerling, M.P., Kumar, A., & Xu, T. (2001). Robustness of the non-linear climate response to ENSO's extreme phases. Journal of Climate 14:1277-1293.
Hoerling, M.P., Hurrell, J.W., Xu, T., Bates, G.T., & Phillips, A.S. (2004). Twentieth century North Atlantic climate change. Part II. Understanding the effect of Indian Ocean warming. Climate Dynamics, 23: 391-405.
Hoskins, B.J., & Karoly, D.J. (1981). The steady linear response of a spherical atmosphere to thermal and orographic forcing. Journal of the Atmospheric Sciences, 38:1179-1196.
Hoskins, B.J., & Valdes, P.J. (1990). On the existence of storm-tracks. Journal of the Atmospheric Sciences, 47: 1854-1864.
Hurrell, J.W. (195) Decadal trends in the North atlantic Oscillation: regional temperature and precipitation. Science, 269: 676-679.
Hurrell, J.W., Kushnir, Y., Visbeck, M., & Ottersen, G. (2003). An overview of the North Atlantic Oscillation. In Hurrell, J.W., Kushnir, Y., Ottersen, G., & Visbeck, M., (eds). *The North Atlantic Oscillation, Climatic Significance and Environmental Impact.* AGU Geophysical Monograph 134, pp1-35.
Hurrell, J.W., & Deser, C. (2009). North Atlantic climate variability: the role of the North Atlantic Oscillation. Journal of Marine Systems, 78: 228-41. doi: 10.1016/j.jmarsys.2008.11.026
Ineson, S., Scaife, A.A., Knight, J.R. Manners, J.C., Dunstone, N.J., Gray, L.J., & Haigh, J.D. (2011): Solar forcing of winter climate variability in the Northern Hemisphere. Nature Geoscience, 4, 753-757, doi: 10.1038/ngeo1282
James, I.N., & James, P.M. (1989). Ultra-low-frequency variability in a simple atmospheric circulation mode. Nature, 342: 53-55.
Johansson, Å. (2007). Prediction skill of the NAO and PNA from daily to seasonal time scales. Journal of Climate, 20: 1957-1975, doi: 10.1175/JCLI4072.1
Kalnay, E., Kanamitsu. M., Kistler, R., Collins, W., Deavon, D., Gandin, L., … Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society, 77: 437-471.
Kang, D., Lee, M.-I., Im, J., Kim, D., Kim, H.-M., Kang, H.-S., …MacLachlan, C. (2014). Prediction of the Arctic Oscillation in boreal winter by dynamical seasonal forecasting systems. Geophysical Research Letters, 41: 3577-3585, doi: 10.1002/2014GL060011
Kidston, J., Scaife, A.A., Hardiman, S.C., Mitchell, D.M., Butchart, N., Baldwin, M.P., & Gray, L.J. (2015). Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. Nature Geoscience, 8: 433-440, doi: 10.1038/NGEO2424
Kim, H.-M., Webster, P., & Curry, J. (2012). Seasonal prediction skill of ECMWF system 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere winter. Climate Dynamics, 39: 2957-2973, doi: 10.1007/s00382-012-1364-6

King, M.P., & García-Serrano, J. (2016). Potential ocean-atmosphere preconditioning of late autumn Barents-Kara sea ice concentration anomaly. Tellus A, 68: 28580, doi:10.3402/tellusa.v68.28580

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., …Takahashi, K. (2015). The JRA-55 reanalysis: general specifications and basic characteristics. Journal of the Meteorological Society of Japan, 93: 5-48, doi: 10.2151/jmsj.2015-001

Koenigk, T., Caian, M., Nikulin, G., & Schimanke, S. (2016). Regional Arctic sea ice variations as predictor for winter climate conditions. Climate Dynamics, 46: 317-337, doi:10.1007/s00382-015-2586-1

Kolstad, E.W., & Årthun, M. (2018). Seasonal prediction from Arctic sea surface temperatures: opportunities and pitfalls. Journal of Climate, 31:8197-8210, doi: 10.1175/JCLI-D-18-0016.1

Kugiumtzis, D. (2000). Surrogate data test for nonlinearity including nonmonotonic transforms. Physical Review E 62, R25-R28, doi: 10.1103/PhysRevE.62.R25

Li, S., Perlwitz, J., Hoerling, M.P., & Chen, X. (2010). Opposite annular responses of the northern and southern hemispheres to Indian Ocean warming. Journal of Climate, 23: 3720-3738, doi: 10.1175/2010JCLI3410.1

Lockwood, M., Harrison, R.G., Woollings, T., & Solanki, S.K. (2010). Are cold winters in Europe associated with low solar activity? Environmental Research Letters, 5: 024001, doi: 10.1088/1748-9326/5/2/024001

MacLachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A., …Madec, G. (2015) Global seasonal forecast system version 5 (GloSea5): a high resolution seasonal forecast system. Quarterly Journal of the Royal Meteorological Society, 141:1072-1084, doi: 10. 1002/qj.2396

Mason, S.J. (2012). Seasonal and longer-range forecasts. In *Forecast Verification: a Practitioner's Guide in Atmospheric Science*. eds: IT Jolliffe, I.T., & Stephenson, D.B. Chichester, Wiley Blackwell, pp203-220.

Mason, S.J., & Baddour, O. (2008). Statistical Modelling. In *Seasonal Climate Variability: Forecasting and Managing Risk*. eds: Troccoli, A. Harrison, M.S.J., Anderson, D.L.T., & Mason, S.J. Dordrecht, Springer pp163-201.

Mason, S.J., & Weigel, A.P. (2009). A generic forecast verification framework for administrative purposes. Monthly Weather Review*,* 137: 331-349, doi: 10.1175/2008MWR2553.1

Morice, C.P., Kennedy, J.J., Rayner, N.A., & Jones, P.D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset. Journal of Geophysical Research*,* 117: D08101, doi: 10.1029/2011JD017187

Müller, W.A., Appenzeller, C., & Schär, C. (2005). Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. Climate Dynamics 24: 213-226, doi: 10.1007/s00382-004-0492-z

Naujokat, B. (1986). An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. Journal of the Atmospheric Sciences, 43: 1873-1877.

van Oldenborgh, G.J., te Raa, L.A., Dijkstra, H.A., & Philip, S.Y. (2009). Frequency- or amplitude-dependent effects of the Atlantic meridional overturning on the tropical Pacific Ocean. Ocean Science*,* 5: 293-301, doi: 10.5194/0s-5-293-2009

O'Reilly, C.H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T.N., Schaller, N., & Woollings, T. (2017). Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. Geophysical Research Letters, 44: 5729-5738, doi: 10.1002/2017GL073736

Ossó, A., Sutton, R., Shaffrey, L., & Dong, B. 2018. Observational evidence of European

summer weather patterns predictable from spring. Proceedings of the National Academy of Sciences, 115: 59-63, doi: 10.1073/pnas.1713146114

Palmer, T.N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., …, Thomson, M.C. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society 85: 853-872, doi: 10.1175/BAMS-85-6-853.

Petoukhov, V., & Semenov, V.A. (2010). A link between reduced Barents-Kara sea ice and cold winter extremes over northern continents. Journal of Geophysical Research, 115: D21111, doi: 10.1029/2009JD013568

Rao, J., & Ren, R. (2016a). Asymmetry and non-linearity of the influence of ENSO on the northern winter stratosphere: 1 observations. Journal of Geophysical Research: Atmospheres, 121: 9000-9016, doi: 10.1002/2015JD024520

Rao, J., & Ren, R. (2016b). Asymmetry and non-linearity of the influence of ENSO on the northern winter stratosphere: 2. Model study with WACCM. Journal of Geophysical Research, 12**1**: 9017-9032, doi: 10.1002/2015JD024521

Rayner, N.A., Parker, D.E., Horton, E.B., Folland, C.K., Alexander, L.V., Rowell, D.P., …Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. Journal of Geophysical Research: Atmospheres, 108(D14): 4407. doi: 10.1029/2002JD002670

Riddle, E.E., Butler, A.H., Furtado, J.C., Cohen, J.L., & Kumar, A. (2013). CFSv2 ensemble prediction of the wintertime Arctic Oscillation. Climate Dynamics, 41: 1099-1116, doi: 10.1007/s00382-013-1850-5

Robinson, D.A., & Estilow, T.W. NOAA CDR program 2012. Accessed 5th March 2017. NOAA Climate Data Record (CDR) of Northern Hemisphere (NH) Snowcover extent (SCE), v01r01. NOAA National Climatic Data Center. doi: 10.7289/V5N014G9

Robock, A., & Mao, J. (1995): The volcanic signal in surface temperature observations. Journal of Climate, 8: 1086-1103.

Rodwell, M.J., Rowell, D.P., & Folland, C.K. 1999. Oceanic forcing of the wintertime North Atlantic Oscillation and European Climate. Nature, 398: 320:323.

Scaife, A.A., Ineson, S., Knight. J.R., Gray, L., Kodera, K., & Smith, D.M. (2013). A mechanism for lagged North Atlantic climate response to solar variability. Geophysical Research Letters 40(2): 434-439. doi: 10.1002/grl.50099

Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N., … Williams, A. (2014). Skillful long-range prediction of European and North American winters. Geophysical Research Letters 41: 2514-2519. doi: 10.1002/2014GL059637

Scaife, A.A., Comer, R.E., Dunstone, N.J., Knight, J.R., Smith, D.M., MacLachlan, C., … Slingo, J. (2017). Tropical rainfall, Rossby waves and regional winter climate predictions. Quarterly Journal of the Royal Meteorological Society, 143: 1-11, doi: 10.1002/qj.2910

Schrieber,T. & Schmitz, A. (2000). Surrogate time series. Physica D 142: 346-382

Screen, J.A. (2017). Simulated atmospheric response to regional and pan-Arctic sea ice loss. 30: 3945-3962, doi: 10.1175/JCLI-D-16-0197.1

Screen, J.A., Deser ,C., Smith, D.M., Zhang, X., Blackport, R., Kushner, P.J., …Sun, L. (2018). Consistency and discrepancy in the atmospheric response to Arctic sea ice loss across climate models. Nature Geoscience. 11: 155-163, doi: 10.1038/s41561-018-0059-y

Smith, D.M., Scaife, A.A., Eade, R., & Knight, J.R. (2016). Seasonal to decadal prediction of the winter North Atlantic Oscillation: emerging capability and future prospects. Quarterly Journal of the Royal Meteorological Society 142: 611-617. doi: 10.1002/qj.247

Stenchikov, G., Hamilton, L., Stouffer, R.J., Robock, R., Ramaswamy, V., Santer, B., & Graf, H.-F. (2006). Arctic Oscillation response to volcanic eruptions in the IPCC AR4 climate models. Journal of Geophysical Research, 111: D07107, doi: 10.1029/2005JD006286

Stendel, M., van den Besselaar, E., Hannachi, A., Kent, E.C., Lefebvre, C., Schenk, F., …Woollings, T. (2016). Recent change-atmosphere in *North Sea Region Climate Change Assessment* (eds: Quante, M., Coliijn, F.). Regional Climate studies, doi: 10.1007/978-3-319-39745-0_2

Stockdale, T.N., Molteni, F., & Ferranti, L. (2015). Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophysical Research Letters, 42: 1173-1179. doi: 10.1002/2014GL062681

Svensson, C., Brookshaw, A., Scaife, A.A., Bell, V.A., Mackay, .JD., Jackson, C.R., …Stanley, S. (2015). Long-range forecasts of UK winter hydrology. Environmental Research Letters 10: 064006. doi: 10.1088/1748-9326/10/6/064006

Takahashi, K., & Dewitte, B. (2016). Strong and moderate non-linear El Niño regimes. Climate Dynamics, 46: 1627-1645, doi: 10.1007/s00382-015-2665-3

Tseng, K.-C., Barnes, E.A., & Maloney, E.D. (2018). Prediction of the midlatitude response to strong Madden-Julian Oscillation events on S2S time scales. Geophysical Research Letters, 45: 463-470, doi: 10.1002/2017GL075734

Tyrell, N.L., Karpechko, A.Y., Räisänen, P. (2018). The influence of Eurasian snow extent on the northern extratropical stratosphere in a QBO resolving model.Journal of Geophysical Research : Atmospheres, 123: 315-328, doi: 10.1002/2017JD27378

Thornton, H.E., Scaife, A.A., Hoskins, B.J., & Brayshaw, D.J. (2017). The relationship between wind power, electricity demand and winter weather patterns in Great Britain. Environmental Research Letters 12: 064017. doi: 10.1088/1748-9326/aa69c6.

Toniazzo, T., & Scaife, A.A. (2006). The influence of ENSO on winter North Atlantic climate. Geophysical Research Letters, 33: L24704, doi: 10.1029/2006GL027881

Vallis, G.K., & Gerber, E.P. (2008). Local and hemispheric dynamics of the North Atlantic Oscillation, annular patterns and the zonal index. Dynamics of Oceans and Atmospheres, 44: 184-212. doi: 10.1016/j.dynatmoce.2007.04.003

Vihma, T., Graversen, R., Chen, L., Handorf, D., Skific, N., Francis, J.A., … Overland, J.E. Effects of the tropospheric large-scale circulation on European winter temperature during the period of amplified Arctic warming. International Journal of Climatology. submitted

Wang, L., Ting, M., & Kushner, P.J. (2017). A robust empirical prediction of winter NAO and surface climate. Science Reports, 7: 279, doi: 10.1038/s41598-017-00353-y

Wei, H.-L., Zhu, D., Billings, S.A., & Balikhin, M.A. (2007). Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks. Advances in Space Research 40: 1863–70.

Wei, H.-L., Billings, S. A., Zhao,Y., & Guo, L. (2010). An adaptive wavelet neural network for spatio-temporal system identification. Neural Networks, 23: 1286–99.

Weisheimer, A., Schaller, N., O'Reilly C., MacLeod, D.A., & Palmer, T. (2016). Atmospheric seasonal forecasts of the twentieth century: multidecadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential for extreme event attribution. Quarterly Journal of the Royal Meteorological Society, 143: 917-926, doi: 10.1002/qj.2976

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences.* Third edition. Academic Press, Oxford.

Woollings, T., Hannachi, A., & Hoskins, B. (2010a). Variability of the North Atlantic eddy-driven jet stream. Quarterly Journal of the Royal Meteorological Society, 136: 856-868, doi: 10.1002/qj.625

Woollings, T., Lockwood, M., Masato, G., Bell, C., & Gray, L. (2010b). Enhanced signature of solar variability in Eurasian winter climate. Geophysical Research Letters, 37: L20805, doi: 10.1029/2010GL044601

Woollings, T., Barnes, E., Hoskins, B., Kwon, Y.-O., Lee, R.W., Li, C., …Williams, K. (2018). Daily to decadal modulation of jet variability. Journal of Climate, 31: 1297-1314, doi: 10.1175/JCLI-D-17-0286.1

Woollings, T., Franzke, C., Hodson, D.L.R., Dong, B., Barnes, EA, Raible, C.C, & Pinto, J.G. (2015). Contrasting interannual and multidecadal NAO variability. Climate Dynamics 45: 539-556, doi: 10.1007/s00382-104-2237-y

Wu, Y., & Smith, K. (2016). Response of northern hemisphere midlatitude circulation to Arctic Amplification in a simple Atmospheric General Circulation model. Journal of Climate, 29: 2041-2058, doi: 10.1175/JCLI-D-15-0602.1

Xoplaki, E., González-Rouco, J.F., Luterbacher, J., & Wanner, H. (2004). Wet season Mediterranean precipitation variability: influence of large-scle dynamics and trends. Climate Dynamics, 23: 63-78

Xue, Y., Higgins, W., & Kousky, V. (2002). Influences of the Madden-Julian Oscillations on temperature and precipitation in North America during ENSO-neutral and weak ENSO winters. *A Workshop on Prospects for Improved Forecasts of Weather and Short-term Climate Variability on Subseasonal (2 week to 2 month) Time Scales, NASA/Goddard Space Flight Center, April 16-18, 2002.*

Yu, B., Lin, H. (2016). Tropical atmospheric forcing of the wintertime North Atlantic Oscillation. Journal of Climate, 29: 1755-1772, doi: 10.1175/JCLI-D-15-0583.1

Zhang, P., Wu, Y., & Smith, K.L. (2018). Prolonged effect of the stratospheric pathway in linking Barents-Kara Sea sea ice variability to the midlatitude circulation in a simplified model. Climate Dynamics, 50: 527-539, doi: 10.1007/s00382-017-3624-y

Tables

dataset	Obtained from	Variable used
Hurrell PC_NAO	https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-pc-based	DJF NAO index
Station-based NAO	https://rda.ucar.edu/datasets/ds570.0	MSLP, Azores and Iceland
AMO	www.climexp.knmi.nl	HadSST3.1.1 SST
HadISST1	www.climexp.knmi.nl	SST, SIC
GPCPv2.3	www.climex.knmi.nl	Tropical precipitation
Carbon dioxide	www.esrl.noaa.gov/gmd/ccgg/trends/data.html	Annual CO_2 level
QBO	www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html	Mean zonal wind, 30hPa
sunspots	http://sidc.oma.be	Sunspot number
JRA-55 SPV_T100	Blanca Ayarzaguena, University of Exeter	Temperature 100hPa
NCEP/NCAR SLP	www.climexp.knmi.nl	Sea level pressure
Rutgers Global Snow Lab snow cover extent	https://climate.rutgers.edu/snowcover/docs.php?target=datareq	Snow cover extent
HadCRUT4.6	https://www.metoffice.gov.uk/hadobs/hadcrut4/	2m temperature anomaly
MJO Indices	www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_mjo_index/pentad.html	200hPa velocity potential anomalies

Table I. Summary of datasets used. For more detailed information, see Table S1.

a) stat80LIN		Model Parameter		
Model Term		7-term	8-term	9-term
1 July N3.4I		-0.77	-0.39	-0.23
2 July tropical Atlantic rainfall		-0.10	-0.16	-0.22
3 November Bering Sea ice		0.40	0.46	0.55
4 October Barents-Kara Sea ice		0.53	0.67	0.73
5 October Barents Sea SLP		-0.30	-0.38	-0.37
6 November East Siberian-Laptev Sea ice		-0.32	-0.29	-0.34
7 November MJO phase 9		-0.38	-0.53	-0.63
8 July tropical East Indian Ocean rainfall			0.29	0.30
9 lead 4 year sunspot cycle				-0.18
Individual Model Performance				
Training	MAE	0.29	0.27	0.22
data	RMSE	0.37	0.31	0.27
	correlation	0.92	0.94	0.95
Testing data	MAE	1.02	1.25	1.07
	RMSE	1.14	1.38	1.26
	correlation	0.77	0.83	0.81

b)HPC80LIN		Model Parameter		
Model Term		7-term	8-term	9-term
1 October Barents Kara Sea ice		0.91	0.94	0.85
2 October Barents Sea SLP		-0.42	-0.47	-0.48
3 November Bering Sea ice		0.57	0.35	0.32
4 November East Siberian-Laptev Sea ice		-0.44	-0.47	-0.44
5 constant		0.28	0.27	0.28
6 October GIN SST		0.33	0.42	0.39
7 November MJO phase 8		-0.46	-0.62	-0.57
8 October Bering Sea ice			0.35	0.37
9 August tropical Atlantic rainfall				0.23
Individual Model Performance				
Training	MAE	0.33	0.29	0.23
data	RMSE	0.41	0.35	0.28
	correlation	0.93	0.95	0.97
Testing data	MAE	0.33	0.30	0.36
	RMSE	0.49	0.47	0.48
	correlation	0.92	0.88	0.88

c) stat56LIN		Model Parameter		
Model Term		6-term	7-term	8-term
1 May Greenland Sea ice		-0.15	-0.16	-0.14
2 September NAH		-0.27	-0.27	-0.27
3 July N3.4I		-0.61	-0.65	-0.61
4 October Barents-Kara Sea ice		0.37	0.41	0.44
5 October Barents Sea SLP		-0.23	-0.22	-0.25
6 November tropical West Indian Ocean SST		0.14	0.16	0.14
7 November East Siberian-Laptev Sea ice			-0.16	-0.15
8 October Greenland Sea ice				-0.08
Individual Model Performance				
Training	MAE	0.52	0.51	0.51
data	RMSE	0.66	0.65	0.64
	correlation	0.65	0.67	0.69
Testing data	MAE	1.27	0.95	1.03
	RMSE	1.38	1.07	1.14
	correlation	0.56	0.74	0.73

d) HPC56LIN		Model Parameter		
Model Term		5-term	6-term	7-term
1 October NAH		-0.36	-0.33	-0.28
2 November SPV		-0.38	-0.29	-0.28
3 October SPV		0.41	0.34	0.36
4 July Greenland Sea ice		-0.12	-0.22	-0.27
5 October Barents Sea SLP		-0.31	-0.41	-0.41
6 October Barents-Kara Sea ice			0.44	0.44
7 November Bering Sea ice				0.18
Individual Model Performance				
Training	MAE	0.68	0.61	0.58
data	RMSE	0.79	0.72	0.69
	correlation	0.75	0.80	0.81
Testing data	MAE	0.83	1.22	1.29
	RMSE	0.99	1.44	1.56
	correlation	0.41	0.49	0.42

Table II(a)-(d) Selected predictors, model coefficients and verification statistics for linear NARMAX models.

a) models	MAE		RMSE		correlation		MSESS	D
1980 linear	**training**	**testing**	**training**	**testing**	**training**	**testing**		
stat80LIN	0.25	1.11	0.29	1.26	0.95	0.82	-0.96	0.80
HPC80LIN	0.27	0.29	0.31	0.48	0.96	0.90	0.82	0.75
1980 polynomial								
stat80POLY	0.35	**0.43**	0.41	**0.52**	0.89	**0.92**	**0.68**	**0.89**

b) models	MAE		RMSE		correlation		MSESS	D
1956 linear	**training**	**testing**	**training**	**testing**	**training**	**testing**		
stat56LIN	0.51	1.08	0.64	1.19	0.76	0.69	-0.44	0.80
HPC56LIN	0.61	1.10	0.71	1.31	0.81	0.46	-0.28	0.68
1956 NARMAX								
stat56POLY	**0.40**	**0.68**	**0.50**	**0.81**	**0.87**	**0.77**	**0.34**	0.79
HPC56POLY	**0.40**	**0.59**	**0.52**	**0.70**	**0.90**	**0.73**	**0.68**	**0.79**

Table III. Verification statistics for averaged linear and polynomial NARMAX models, for a) periods 1980-2018 and b) 1956-2018. Bold figures show if polynomial outperforms the linear model.

a)		Model Parameter		
stat80POLY	**Model Term**	**4-term**	**5-term**	**6-term**
1	Jul tropical West Pacific Rainfall* Nov Barents Sea SLP	-0.34	-0.31	-0.33
2	Oct Labrador Sea ice LAB * Oct NAH	0.63	0.64	0.58
3	Sep snow * Nov_SST gradient	0.48	0.55	0.56
4	Oct Bering Sea ice * Nov NAH	0.22	0.22	0.27
5	Sep Labrador Sea ice * Oct MJO phase 8		0.51	0.51
6	Jul tropical West Pacific Rainfall * Sep snow			0.33
Individual Model Performance				
Training	MAE	0.40	0.36	0.33
data	RMSE	0.48	0.43	0.38
	correlation	0.84	0.88	0.91
Testing data	MAE	0.43	0.53	0.48
	RMSE	0.50	0.58	0.62
	correlation	0.90	0.88	0.95

b)		Parameter		
stat56POLY	**Model Term**	**8-term**	**9-term**	**10-term**
1	Jul tropical West Indian Ocean SST * Sep GIN SST	-0.06	-0.05	-0.07
2	Nov Bering Sea ice	-0.33	-0.67	-0.49
3	Sep NAH	-0.28	-0.28	-0.27
4	Nov East Siberian-Laptev Sea ice * Nov GIN SST	-0.25	-0.19	-0.18
5	Jul Greenland Sea ice * Jul Greenland Sea ice	-0.03	-0.03	-0.03
6	Oct Barents-Kara Sea ice	0.37	0.33	0.32
7	Oct Barents Sea SLP * Nov GIN SST	-0.21	-0.22	-0.24
8	Oct Barents Sea SLP	-0.22	-0.20	-0.21
9	Jul N3.4I * Sep Canadian Archipelago-Baffin sea ice		0.19	0.18
10	Jul N3.4I * Sep NAH			0.46

Individual Model Performance				
Training data	RMSE	0.53	0.51	0.49
	MAE	0.43	0.40	0.40
	Correlation	0.85	0.86	0.87
Testing data	RMSE	0.74	0.85	0.85
	MAE	0.67	0.68	0.69
	Correlation	0.70	0.63	0.82

c)		**Parameter**		
HPC56POLY	**Model Term**	**6-term**	**7-term**	**8-term**
1	Oct Labrador Sea ice * Oct NAH	0.45	0.41	0.43
2	Aug N3.4 * lead 3 year sunspot cycle	0.63	0.69	0.73
3	atmospheric CO_2 * Jun Greenland Sea ice	0.04	0.04	0.03
4	Nov N3.4I * Sep Barents Sea SLP	0.82	0.97	1.19
5	Jun Hudson Bay sea ice * volcanic index	1.00	1.00	0.99
6	Sep GIN SST * Oct SST gradient	0.25	0.29	0.32
7	Sep QBO * Nov Barents Sea SLP		0.26	0.28
8	Sep N3.4I * Nov GIN SST			-0.56
Individual Model Performance				
Training data	RMSE	0.59	0.55	0.49
	MAE	0.45	0.42	0.37
	Correlation	0.86	0.89	0.91
Testing data	RMSE	0.60	0.68	0.75
	MAE	0.52	0.53	0.57
	Correlation	0.79	0.73	0.68

Table IV(a)-(c). As for Table II, but for polynomial models.

Figure Captions.

Figure 1. Example of hindcasts (solid) and forecasts (dashed) using linear (LIN) and polynomial (POLY) NARMAX models, derived from the 1980 training period for the station-based NAO.

Figure 2. Out-of-sample forecasts compared with observations for linear (LIN) and polynomial (POLY) NARMAX models. Note there is no polynomial model for HPC80 and the observed station NAO is missing in (b) and (d).

Figure 3. Predictors selected for the NAO models based on the 1980 training period. N3.4I =El Niño 3.4 discontinuous index; EIR=tropical East Indian Ocean rainfall, NAH=North Atlantic Horseshoe SST pattern; snow=Eurasian snowcover; BK ice=Barents-Kara Sea ice; GIN SST= Greenland-Iceland Norwegian SST; ES/L= East Siberian/Laptev Sea ice; TAR= tropical Atlantic rainfall; Bering ice=Bering Sea ice; MJO8/9=Phase 8/9 Madden-Julian Oscillation; Barents SLP=Barents Sea regional SLP; WPR=tropical West Pacific Ocean rainfall; SST gradient=North Atlantic SST gradient; LAB ice= Labrador Sea ice; lead 4 SS = sunspot cycle leading by 4 years. Where predictor month is not specified, it is indicated by a white hexagon linked to the variable. This is used where different models select different months of a common predictor.

Figure 4. As for Figure 3 but using the 1956 training period. Additional variables: volc= volcanic index; SPV=Stratospheric polar vortex index; HUD ice=Hudson Bay sea-ice; lead 3 SS=sunspot cycle leading by 3 years; N3.4=standard N3.4 index; CO2=atmospheric carbon dioxide; QBO=Quasi-biennial oscillation; GRE ice =Greenland Sea ice; WISST=tropical West Indian Ocean SST; ARB=Canadian Archipelago/Baffin Bay sea-ice.

Supplementary Material

dataset	Obtained from	Variable used	Region selected	dates	reference
Hurrell PC_NAO	https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-pc-based	DJF NAO index	90W-40E, 20-80N	1955-2018	Hurrell, 1995; Hurrell et al., 2003
Station-based NAO	https://rda.ucar.edu/datasets/ds570.0	MSLP	Reykjavik, Ponta Delgada	1955-2017	
AMO	www.climexp.knmi.nl	HadSST3.1.1 SST	7-75W, 25-60N, minus regression on global mean temperature	1955-2017	van Oldenborgh et al., 2009.
HadISST1	www.climexp.knmi.nl	SST	N3.4: 170-120W, 5S-5N Tropical Atlantic (TASST): 50-0E,5S-5N W.Indian Ocean (WISST): 50-85W, 5S-5N E. Indian Ocean (EISST): 85-120E,5S-5N W. Pacific (WPSST):120-170E,5S-5N E. Pacific (EPSST): 140-90W, 5S-5N N. Atlantic Horseshoe (NAH): 40-15W, 15-30N minus 60-40W,30-45N N. Atlantic tripole: 60-40W, 40-55N minus 80-60W, 25-35N sub-polar gyre (SPG_SST): 60-10W, 50-65N Barents Sea (Bar_SST): 25-70W, 75-80N Greenland/Iceland Norwegian Seas (GIN_SST): 20W-20E, 65-80N N. Atlantic SST gradient (SST_grad): 60-30W, 20-40N minus 60-10W, 50-65N	1955-2017	Rayner et al., 20003
		SIC	Barents-Kara Seas(BK): 10-100E,65-85N E. Siberian/Laptev Seas (ESL): 100-180E, 68-85N Beaufort/Chukchi Seas (BC): 180-120W, 68-85N	1955-2017	

			Canadian Archipelago/Baffin Bay (ArB): 120-45W, 63-80N Greenland Sea (GRE): 45-0W, 63-85N Bering Sea (BER): 195-155W, 55-68N Hudson Bay (HUD): 100-70W, 50:63N Labrador Sea (LAB): 70-45W, 40-63N		
GPCPv2.3	www.climex.knmi.nl	Tropical precipitation	Tropical Atlantic Rainfall (TAR): 50-0E,5S-5N W. Indian Ocean Rainfall (WIR): 50-85W, 5S-5N E. Indian Ocean Rainfall (EIR): 85-120E,5S-5N W.Pacific Rainfall (WPR):120-170E,5S-5N E. Pacific Rainfall (EPR): 140-90W, 5S-5N	1979-2017	Adler et al., 2003
Carbon dioxide	www.esrl.noaa.gov/gmd/ccgg/trends/data.html	Annual CO_2 level	NA	1959-2017	Tans, P. NOAA ESRL, Keeling, R Scripps IOO.
QBO	www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html	Mean zonal wind, 30hPa	NA	1955-2017	Naujokat et al. 1986*
sunspots	http://sidc.oma.be	Sunspot no.	NA	1955-2017	WDC-SILSO, Royal Observatory of Belgium, Brussels
JRA-55 SPV_T100	Blanca Ayarzaguena, University of Exeter	Temperature 100hPa	65-90N	1958-2015	
NCEP/NCAR SLP	www.climexp.knmi.nl		Barents SLP: 60-120E, 67.5-90N	1955-2017	Kalnay et al. 1996
Rutgers Global Snow Lab	https://climate.rutgers.edu/snowcover/docs.php?target=datareq	Snow cover extent	Eurasian snow: 55-150E, 45-80N	1979-2017	Robinson et al., 2012

snow cover extent					
HadCRUT 4.6	https://www.metoffice.gov.uk/hadobs/hadcrut4/	2m Temperature anomaly	90W-90E, 20-80N	1955-2017	Morice et al., 2012
MJO Indices	www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_mjo_index/pentad.html	200hPa velocity potential anomalies		1979-2017	Xue et al., 2012

Table S1. Datasets used in the study.

Model	MAE		RMSE		correlation	
	training	testing	training	testing	training	testing
linear	0.64	1.15	0.79	1.31	0.70	0.24
polynomial	0.55	0.85	0.67	1.00	0.79	0.66

Table S2. Verification statistics for station NAO models using an even-odd years split.

	RMSE		MAE		correlation coefficient	
	training	testing	training	testing	training	testing
7-term model	0.976±0.065 Q=0.41, Z=-8.71	1.235±0.279 Q=0.40, Z=-2.99	0.781±0.062 Q=0.33, Z=-7.27	1.068±0.268 Q=0.33, Z=-2.75	0.484±0.102 Q=0.93, Z=4.37	-0.044±0.441 Q=0.92, Z=2.19
8-term model	0.960±0.077 Q=0.35, Z=-7.92	1.228±0.243 Q=0.47, Z=-3.12	0.777±0.074 Q=0.29, Z=-6.58	1.036±0.223 Q=0.30, Z=-3.3	0.501±0.110 Q=0.95, Z=4.08	-0.024±0.368 Q=0.88, Z=2.46
9-term model	0.947±0094 Q=0.28, Z=-7.10	1.229±0.284 Q=0.48, Z=-2.64	0.760±0.090 Q=0.23, Z=-5.89	1.045±0.254 Q=0.36, Z=-2.70	0.497±0.127 Q=0.97, Z=3.72	-0.016±0.452 Q=0.88, Z=1.98

Note: 1) A constant term is included in each of these models; 2) The statistic Q is calculated from the model for the original data; 3) The statistic Z is for the test value calculated from equation (S1). $\pm$ indicates the standard deviation.

Table S3. The averaged RMSE, MAE, and Correlation Coefficient of 100 models estimated using the 100 surrogate datasets for HPC80LIN using the 8 variables: octBK, oct_barSLP, novBER, novESL, oct_GIN_SST, nov_MJO8, octBER, and augAR.

A surrogate method is used to test whether the models obtained by the proposed model identification algorithm (FROLS) are not achieved by chance, but are due to the efficacy of the algorithm. The null-hypothesis is that the models are achieved by chance through the FROLS algorithm. A set of surrogate data is generated from the original data. If the statistical metrics (e.g. mean squared error and mean absolute error) calculated for models estimated from the original data significantly differ from those for surrogate data, the null-hypothesis should be rejected. We use the following z-test to measure the significance of the difference between the model performances (MSE, MAE, Correlation Coefficient) with respect to the original data and the surrogate data, respectively,

$$Z = \frac{Q^{(O)} - mean(Q^{(S)})}{std(Q^{(S)})} \tag{S1}$$

where $Q^{(O)}$ represents one of the three metrics: RMSE, MAE, Correlation Coefficient, calculated from the model that is estimated from the original data, and $Q^{(S)}$ represents values of the corresponding metric calculated from the models associated with the surrogate data.

Under the null hypothesis, if $|Z| > 1.96$ for a two-tailed test, the null hypothesis should be rejected.

We used an amplitude-adjusted Fourier-transformed surrogate algorithm (Schrieber and Schmitz, 2000; Kugiumtzis, 2000), which generates data from the original time series using a Fourier transform. The method retains some of the most important time and frequency domain properties of the time series (e.g. autocorrelation function and power spectrum).

We generated 100 surrogate datasets from the original raw data (1980-2017) for the 8 variables: octBK, oct_barSLP, novBER, novESL, oct_GIN_SST, nov_MJO8, octBER, and augAR. For each of the surrogate datasets, a linear model consisting of 7, 8, and 9 model terms was estimated for training periods 1980-2010 using the FROLS algorithm, leaving 2011-2018 for use as the testing period. For an illustration, the distribution of the 100 values of RMSE, MAE, and Correlation Coefficient, calculated from the 100 9-term models estimated from surrogate data are shown in Figures S1a, S1b, S1c. Due to space limitations, graphs for the other two cases are not presented.

The average performance of the 100 models for each of three cases (i.e. with 7-9 model terms) is shown in Table S3. All the models estimated from the surrogate data show far worse performance than the models estimated from the original data of the 8 variables; in all cases $Z > 1.96$ and the null hypothesis is rejected. These models show no predictive skill for HPC NAO. In addition, adding further model terms does not help to increase the model performance on either the training or test data.

Therefore, it can be concluded that the three models presented in Table II were not obtained by chance. The 9-term model reported in Table II works well for predicting the HPC NAO on the test data.

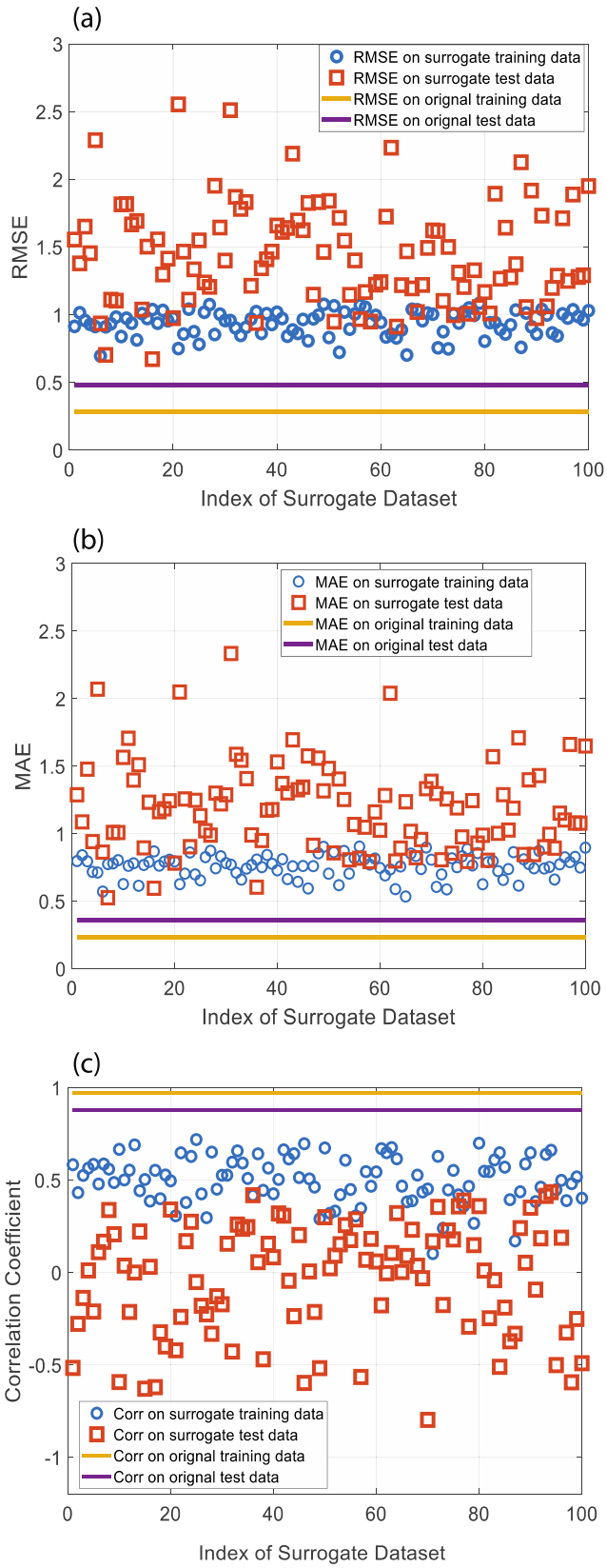

Figure S1. Distribution of the a) RMSE b) MAE and c) correlation coefficient for the 100 9-term models estimated from the surrogate data.

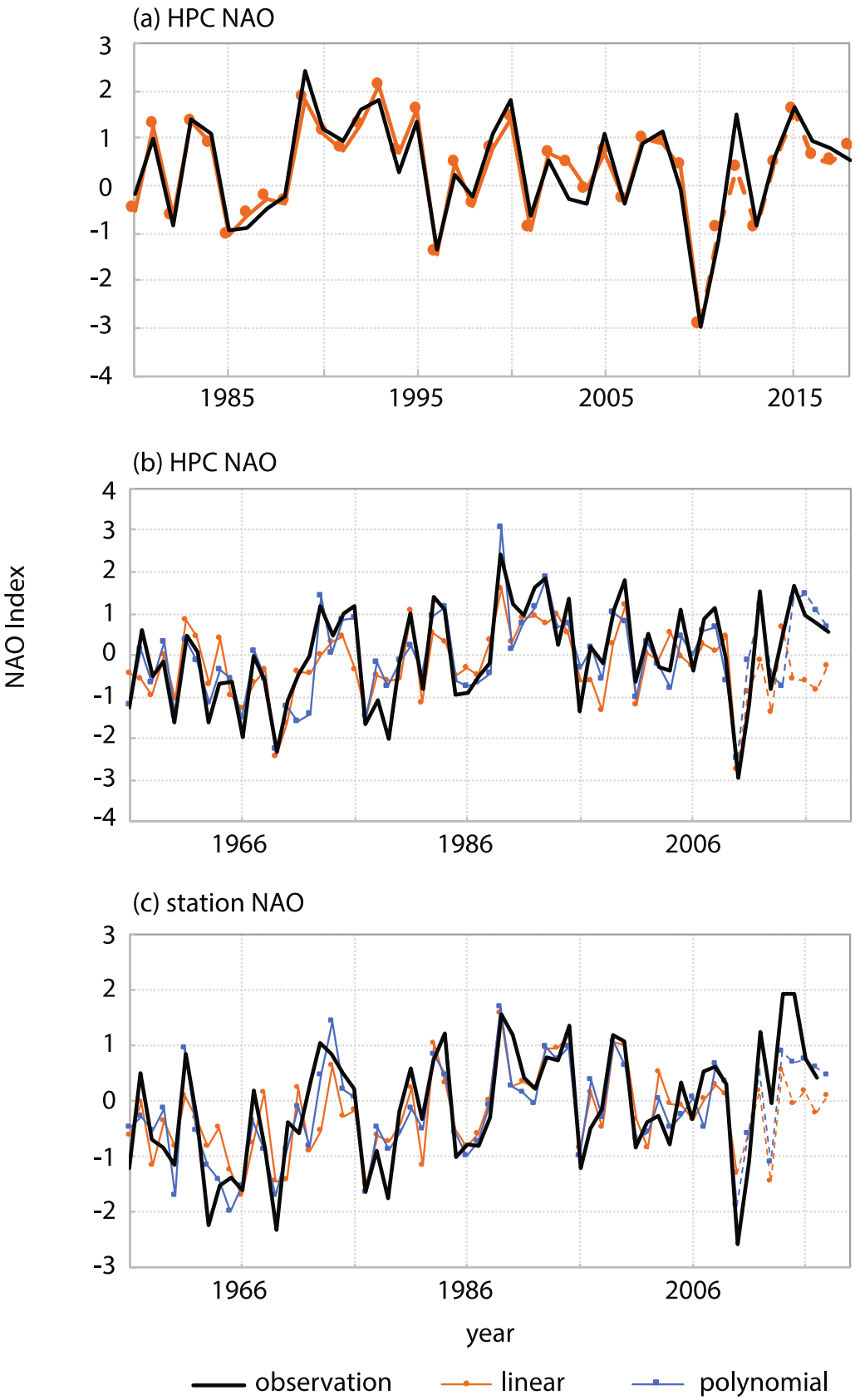

Figure S2. Hindcasts (solid) and forecasts (dashed) of station and HPC NAO using linear and polynomial NARMAX models, derived from the 1980 and 1956 training periods. Note there is no polynomial model selected for HPC80.

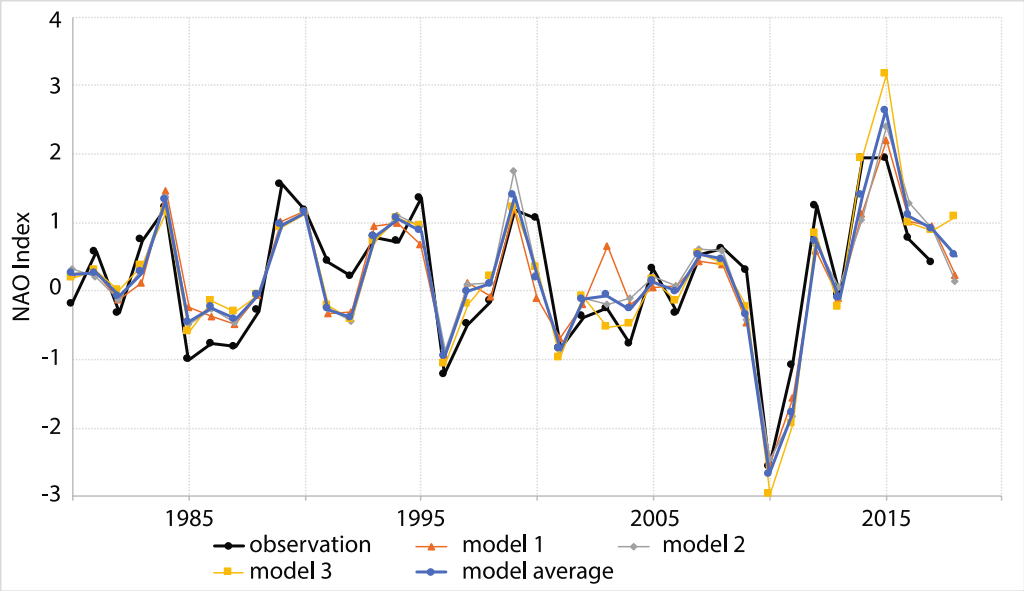

Figure S3. An example of the model-averaging procedure for the station NAO from 1980.

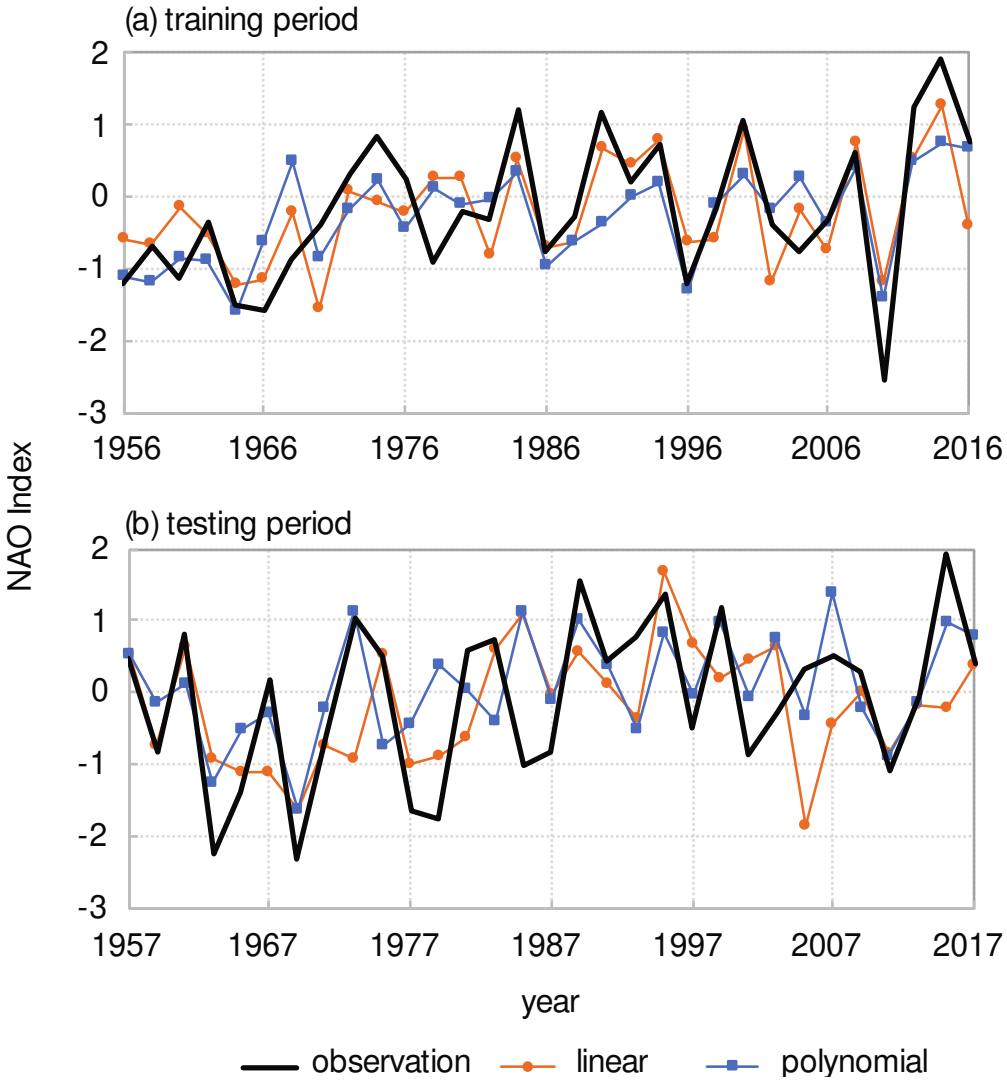

Figure S4. Plots for a) the training period, and b) the testing period, when the training period consists of even years and the testing period of odd years.

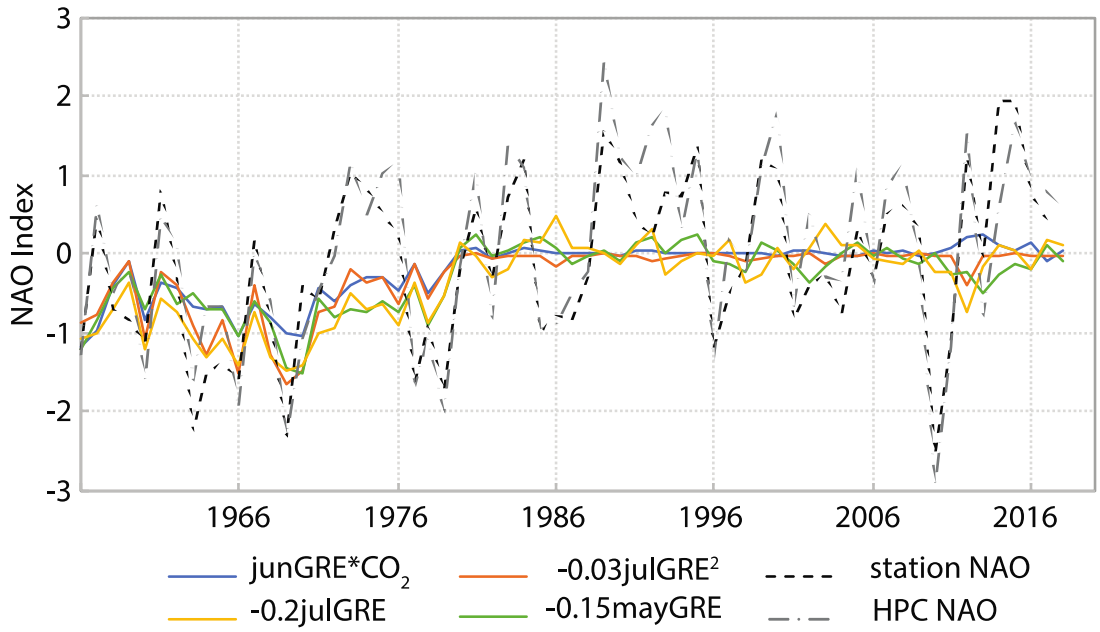

Figure S5. The two observed NAO time series from 1956, with the contributions made to the models by terms including Greenland Sea ice.

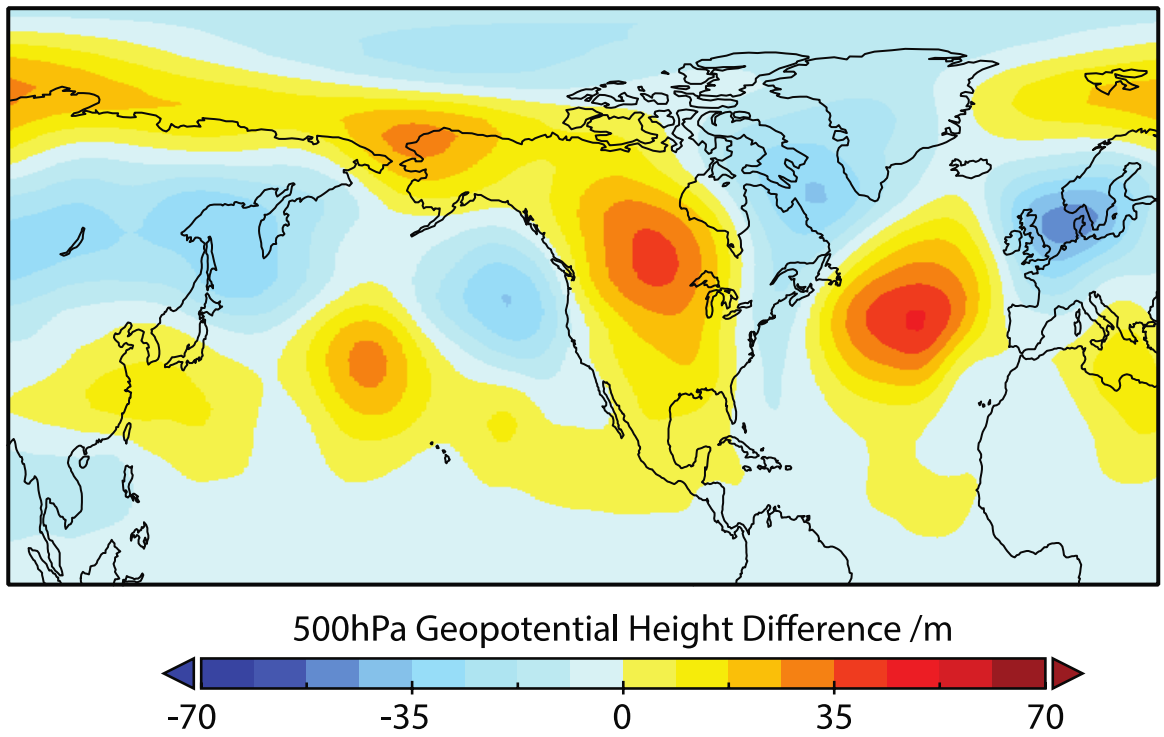

Figure S6 October 500GPH composite plots for October MJO_8 low minus high years.

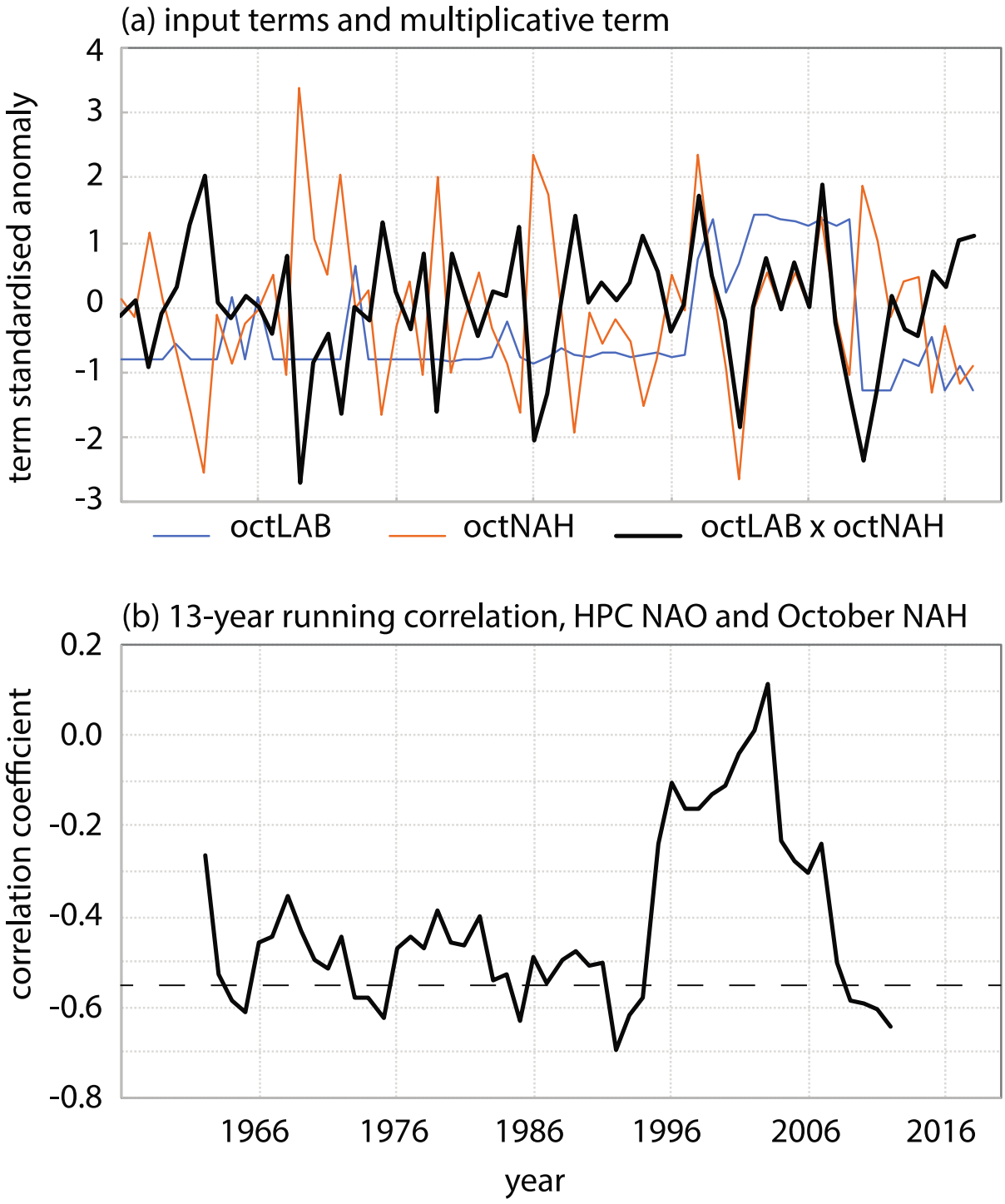

Figure S7. A) October Labrador Sea ice (octLAB) and October NAH SST pattern (octNAH), and the product of the two input terms. B) 13-year running correlation between October NAH and HPC NAO, year is the centre of the moving window. Horizontal dashed line id the 95% significance level