
This is a repository copy of *Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/209254/

Version: Published Version

---

**Article:**
Wei, H. orcid.org/0000-0002-4704-7346 (2024) Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models. Meteorological Applications, 31 (1). e2178. ISSN 1350-4827

https://doi.org/10.1002/met.2178

---

**Reuse**
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.


eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Check for updates

Received: 24 October 2023 | Revised: 28 December 2023 | Accepted: 23 January 2024

DOI: 10.1002/met.2178

Meteorological Applications
Science and Technology for Weather and Climate
Open Access
RMetS

RESEARCH ARTICLE

# Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models

**Yiming Sun**[1] | **Ian Simpson**[2] | **Hua-Liang Wei**[1] | **Edward Hanna**[2]

[1]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, UK

[2]Department of Geography and Lincoln Climate Research Group, College of Health and Science, University of Lincoln, Lincoln, UK

**Correspondence**
Hua-Liang Wei, Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, UK.
Email: w.hualiang@sheffield.ac.uk

**Funding information**
The Natural Environment Research Council, United Kindom, Grant/Award Number: NE/V001787/1

## Abstract

Dynamical seasonal forecast models are improving with time but tend to underestimate the amplitude of atmospheric circulation variability and to have lower skill in predicting summer variability than in winter. Here, we construct Nonlinear AutoRegressive Moving Average models with eXogenous inputs (NARMAX) to develop the analysis of drivers of North Atlantic atmospheric circulation and jet-stream variability, focusing on the East Atlantic (EA) and Scandinavian (SCA) patterns as well as the North Atlantic Oscillation (NAO) index. New time series of these indices are developed from empirical orthogonal function (EOF) analysis. Geopotential height data from the ERA5 reanalysis are used to generate the EOFs. Sets of predictors with known associations with these drivers are developed and used to formulate a sliding-window NARMAX model. This model demonstrates a high degree of predictive accuracy, as indicated by its average correlation coefficients over the testing period (2006–2021): 0.78 for NAO, 0.83 for EA and 0.68 for SCA. In comparison, the SEAS5 and GloSea5 dynamical forecast models exhibit lower correlations with observed circulation changes: for NAO, the correlation coefficients are 0.51 for SEAS5 and 0.34 for GloSea5, for EA they are 0.15 and 0.09, respectively, and for SCA, they are 0.28 and 0.24, respectively. Comparison of NARMAX predictions with forecasts and hindcasts from the SEAS5 and GloSea5 models highlights areas where NARMAX can be used to help improve seasonal forecast skill and inform the development of dynamical models, especially in the case of summer.

**KEYWORDS**

ensemble forecasts, forecasting, machine learning, miscellaneous, NARMAX, NorthAtlantic atmospheric circulation, probabilistic forecasts, probabilistic seasonal forecast, seasonal, verification, weather prediction

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. *Meteorological Applications* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

## 1 | INTRODUCTION

The North Atlantic jet stream strongly influences the weather in Northwest Europe and has a significant role in determining the strength and sign of North Atlantic atmospheric circulation indices such as the North Atlantic Oscillation (NAO), East Atlantic (EA) pattern, and Scandinavian (SCA) pattern; the anomalous weather patterns of a particular season can be described by the interplay of these modes of variability (Hall & Hanna, 2018). Recent extreme seasons have been characterized by distinctive jet-stream configurations, and jet strength and location are intimately linked with extreme weather conditions (e.g., in temperature and precipitation) experienced across Northwest Europe (Hall & Hanna, 2018). Extreme seasonal weather has important socio-economic implications, in terms of risk avoidance, with costs to the insurance industry (e.g., ~£1.5 billion across the UK in winter 2013/14 (Davies, 2014)), and impacts on agriculture, food security, energy supply, public health/well-being and severe weather planning.

Until relatively recently, North Atlantic atmospheric variability was thought to be largely due to unpredictable fluctuations (Stephenson et al., 2000). However, dynamical seasonal forecasting systems have been used to develop skillful seasonal forecasts for UK winter weather from a few months ahead (Scaife et al., 2014). Many factors (drivers) appear to influence the NAO and jet-stream changes, and these potential drivers can be broadly grouped into cryosphere effects from variations in sea-ice extent and snow cover, oceanic effects from North Atlantic sea-surface temperatures (SST), tropical influences such as the El-Niño Southern Oscillation (ENSO), and stratospheric effects due to stratospheric circulation variability, solar variability, volcanic eruptions and the Quasi-Biennial Oscillation (QBO) (Hall et al., 2015). These drivers of jet-stream variability can oppose or reinforce one another, and there are indications of interactions between them (Hall et al., 2019). Drivers of jet-stream variability show seasonal variation and distinctive drivers of jet-stream variability operate in different seasons. In addition to these identifiable drivers, a significant part of North Atlantic jet changes is driven by internal unforced variability due to chaotic internal dynamical processes (Kushnir et al., 2006; Lorenz, 1963). While a consensus has now been reached that some observed drivers can be reproduced in climate models, improved understanding of more recently identified drivers of the North Atlantic extratropical jet stream is crucial for making progress in UK seasonal climate predictions (Hall et al., 2015).

The focus of government-funded research is on dynamical forecast systems; however, such forecasts are not always accurate, such as in winter 2004–2005 (Hall, Scaife, et al., 2017) and more recently in 2013–2014, when dynamical model forecasts did not well predict the positive winter NAO, and furthermore did not consider the accompanying positive EA pattern and hence the exceptionally heavy rain and flooding in southern England (Maidens et al., 2021). While dynamical seasonal forecast models are sensitive in winter to tropical forcing such as El Niño events, some evidence suggests that they may be relatively insensitive to Arctic variability (Cohen et al., 2019). Compared with winter, dynamical model forecasts show relatively little skill in summer, when there is less forcing from the tropics (Hall et al., 2015). Recent work on seasonal prediction with dynamical models has also revealed an intriguing conundrum called the signal-to-noise paradox: this is where such models reasonably well predict the year-to-year variability of the winter NAO but underpredict its amplitude, due to a systematic underestimation of the mechanisms influencing mid-latitude atmospheric circulation (Eade et al., 2014; Scaife et al., 2014; Siegert et al., 2016; Stockdale et al., 2015). Rare events are usually forecast to have a low probability partly due to the signal-to-noise paradox, meaning that if a model predicts a low, but above-average, chance of a rare event happening, this does not necessarily constitute a missed event (Legg & Mylne, 2004). Supplementing dynamical seasonal forecasting systems, statistical methods identify slowly varying boundary conditions such as sea-ice variability, ocean temperatures, and influences from the stratosphere, which are capable of 'nudging' the jet stream and providing elements of predictability (e.g., Baker et al., 2018; Hall, Jones, et al., 2017; Hall, Scaife, et al., 2017). In the mid-latitudes, statistical forecasting has been relatively neglected compared with the tropics; however, recent developments in statistical techniques, under the umbrella of 'machine learning' (e.g., Billings, 2013; Hall et al., 2019) have taken place mainly outside the climate-science community and are relatively quick and cheap to implement.

The novel application of these advanced statistical techniques and systems science methods has significant potential to improve forecast skills and help inform the development of the next generation of dynamical seasonal forecasting systems. Here, we use a Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) systems identification approach (Billings, 2013; Hall et al., 2019), which is an interpretable machine learning method, to identify and model linear and nonlinear dynamic relationships between a range of meteorological and related variables. In addition to its ability to delineate nonlinear relations, NARMAX is able to identify non-stationary associations that arise from changes in forcings over time, building on studies where dynamical models have suggested changes in NAO forecast skill over periods

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of several decades (Weisheimer et al., 2019), and so NARMAX, therefore, has significant potential to help improve Northwest Europe seasonal weather forecasts. In a pioneering application of NARMAX in this area, Hall et al., 2019 found significant skill in NAO winter forecasting and identified key sources of predictability. Here, we extend skillful seasonal forecasting to the summer season, where dynamical seasonal prediction models currently remain the most problematic, and identify factors that contribute skill to the forecast. In a further innovation, we also consider two other principal North Atlantic atmospheric circulation patterns that complement the NAO. Our results form a firm basis for improving Northwest European regional seasonal weather prediction and should therefore be of interest to potential end-users as well as to model developers and the broader seasonal prediction scientific community.

## 2 | DATA

Updated versions of the three principal EOFs of European and North Atlantic atmospheric circulation variability (NAO, EA and SCA) were generated using 500 hPa geopotential height data from the European Centre for Medium-Range Forecasts (ECMWF) ERA5 reanalysis (Hersbach et al., 2020), combined with the Python eofs package (Dawson, 2016). ERA5 is based on the Integrated Forecasting System (IFS) and replaces ECMWF's earlier ERA-40 and ERA-Interim reanalysis products. We obtained ERA5 data, specifically of sea-surface temperatures (SSTs), sea-ice cover, sea level pressure, and 500 hPa geopotential heights, from the Copernicus Climate Data Store.

The summer EOFs are based only on high summer (July and August), as using the full June/July/August data generates a poorly defined pattern for EOF1. This is consistent with previous analysis, for example, Folland et al. (2009)), which suggest there is a strong summer NAO signal that characterizes July and August, while the summer NAO behaves differently during June. The three winter EOFs are based on seasonal data for the full winter quarter (December–February). Maps for the winter and summer EOFs are shown in Figure 1. These are largely similar to the EOFs obtained by Hall and Hanna

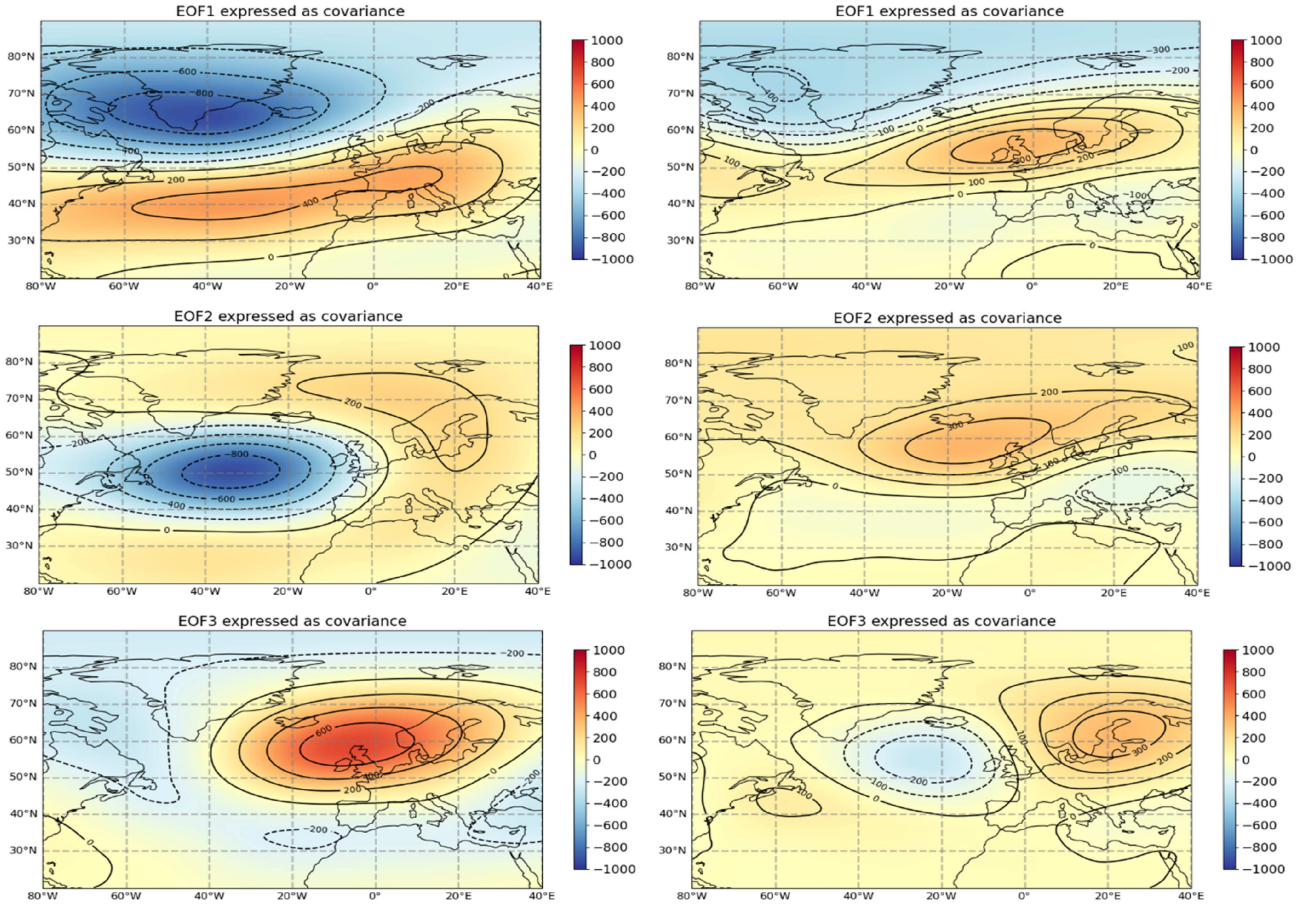



FIGURE 1 The three primary empirical orthogonal functions (EOFs) of atmospheric circulation variability (at the 500 hPa geopotential height level) from ERA5 reanalysis based on 1950 to 2021 for winter (DJF, left) and high summer (JA, right).

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(2018)) but also show some notable differences. For example, the winter SCA pattern has the high-pressure anomaly further west than the (Hall & Hanna, 2018) version, centred to the north and north-east of the British Isles, while the winter NAO has the low-pressure anomaly centred to the south of Greenland, rather than over Iceland. The differences are mostly down to the different methodology used to calculate the EOFs, rather than different time periods.

Monthly 500 hPa geopotential height forecasts from SEAS5 (provided by ECMWF) and GloSea5 (provided by the UK Met Office) were analysed and projected onto the EOFs using projectField from the eofs package. Seasonal forecast runs were provided by the Copernicus Climate

**TABLE 1** Potential drivers of winter and summer atmospheric circulation variability that are used as predictors for NARMAX models.

| Dataset | Variable used and their abbreviations as used in this study | Region selected | Dates |
|---|---|---|---|
| Atlantic Multidecadal Oscillation (AMO) | ERA5 SST | 7–75 W, 25–60 N, regional SST anomaly minus global SST anomaly | 1956–2021 |
| Sea-surface temperature | Nino 3.4 | 170–120 W, 5S–5N | 1956–2021 |
| | Tropical Atlantic (TASST) | 50 W–0E,5S–5N | 1956–2021 |
| | W. Indian Ocean (WISST) | 50–85E, 5S–5N | 1956–2021 |
| | E. Indian Ocean (EISST) | 85–120E, 5S–5N | 1956–2021 |
| | W. Pacific (WPSST) | 120–170E, 5S–5N | 1956–2021 |
| | E. Pacific (EPSST) | 140–90 W, 5S–5N | 1956–2021 |
| | North Atlantic Horseshoe (NAH) | 40–15 W, 15–30 N minus 60–40 W, 30–45 N | 1956–2021 |
| | North Atlantic dipole (DIP) | 52–40 W, 42–52 N minus 35–20 W, 35–42 N | 1956–2021 |
| | North Atlantic tripole (TRI) | 60-40 W, 40-55 N minus 80–60 W, 25–35 N | 1956–2021 |
| | Sub-Polar Gyre (GRE) | 60–10 W, 50–65 N | 1956–2021 |
| | Barents Sea SST (Bar_SST) | 25–70E, 75–80 N | 1956–2021 |
| | Greenland/Iceland Norwegian Seas (GIN) | 20 W–20E, 65–80 N | 1956–2021 |
| | North Atlantic SST gradient (SST_grad) | 60–30 W, 20–40 N minus 60–10 W, 50–65 N | |
| | Sub-Polar Gyre (SPG_SST) | 60–10 W, 50–65 N | |
| Sea-ice concentration | Barents–Kara Seas (BK) | 10–100E, 65–85 N | 1956–2021 |
| | E. Siberian/Laptev Seas (ESL) | 100–180E, 68–85 N | 1956–2021 |
| | Beaufort/Chukchi Seas (BC) | 180–120 W, 68–85 N | 1956–2021 |
| | Canadian Archipelago/Baffin Bay (ArB) | 120–45 W, 63–80 N | 1956–2021 |
| | Greenland Sea (GRE) | 45–0 W, 63–85 N | 1956–2021 |
| | Bering Sea (BER) | 195–155 W, 55-68 N | 1956–2021 |
| | Hudson Bay (HUD) | 100–70 W, 50:63 N | 1956–2021 |
| | Labrador Sea (LAB) | 70–45 W, 40–63 N | 1956–2021 |
| Tropical precipitation | Tropical Atlantic Rainfall (TAR) | 50 W–0E, 5S–5N | 1979–2021 |
| | W. Indian Ocean Rainfall (WIR) | 50–85E, 5S–5N | 1979–2021 |
| | E. Indian Ocean Rainfall (EIR) | 85–120E, 5S–5N | 1979–2021 |
| | W. Pacific Rainfall (WPR) | 120–170E,5S–5N | 1979–2021 |
| | E. Pacific Rainfall (EPR) | 140–90 W, 5S–5N | 1979–2021 |
| Stratospheric polar vortex | Temperature 100 hPa | 65–90 N | 1956–2021 |
| Sea level pressure | Barents SLP | 60–120E, 67.5–90 N | 1956–2021 |
| Carbon dioxide | Annual $CO_2$ level | NA | 1959–2021 |
| QBO | Mean zonal wind, 30 hPa | NA | 1956–2021 |
| Sunspots | Sunspot no. | NA | 1956–2021 |
| Snow cover extent | Eurasian snow | 55–150E, 45–80 N | 1979–2021 |
| HadCRUT5 | 2 m Temperature anomaly | 90 W–90E, 20–80 N | 1955–2021 |
| MJO Indices | 200 hPa velocity potential anomalies | | 1979–2021 |

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Change Service (C35) via the Climate Data Store website. For both models, hindcasts are available from 1993 to 2016 inclusive. For SEAS5, a complete set of seasonal forecast runs is also available from 2017 onwards, but at the time of analysis for GloSea5 C35 only provided an incomplete set of seasonal forecast runs covering the winters of 2017/2018 to 2019/2020 inclusive and the summers of 2018 and 2019.

A number of variables that may be used to predict the North Atlantic jet-stream and atmospheric circulation variability, and by extension temperature and precipitation over Northwest Europe, have been collected for both winter (DJF) and summer (JJA), building from the drivers identified by (Hall, 2016; Hall & Hanna, 2018; Hall, Scaife, et al., 2017). A wide range of potential drivers have been assembled, so as to be able to select from a wide range of variables for inclusion in NARMAX. SST anomaly patterns are used, including the ENSO and the Atlantic Multidecadal Oscillation (AMO), plus sea-ice anomalies, snow cover anomalies and tropical precipitation anomalies. The stratospheric polar vortex and QBO are used as predictors for winter atmospheric circulation, but not summer, due to a lack of evidence for them having a strong influence on summer atmospheric circulation.

SST anomalies, sea-ice coverage anomalies, the AMO, tropical precipitation anomalies and the strength of the stratospheric polar vortex were calculated based on ERA5 reanalysis data. This version of the AMO is based on the region from 7 to 75° W, 25 to 60° N, subtracting the global SST anomaly from the regional SST anomaly to remove biases that would result from the upward trend in global SSTs. For summer, an SST-based North Atlantic dipole index is used, based on Ossó et al. (2018)), who provided evidence for a link between this and a high-pressure anomaly to the west of the British Isles, resulting in relatively anticyclonic weather over Britain. The North Atlantic Horseshoe SST pattern (Cassou, Terray, et al., 2004) is linked with the winter NAO. The North Atlantic tripole is based on the methodology of Marshall et al. (2001), who provided evidence for this being linked especially with the NAO in winter. Snow cover data are based on Estilow et al. (2015). Monthly sunspot numbers were obtained from the Solar Influences Data Analysis Center (Center, 1956–2021). The QBO data are obtained from the Free University of Berlin (Naujokat, 1986). A full list of the drivers is provided in Table 1. Predictors are sourced from a range of preceding months, up to 8 months in advance in some cases, to account for possible lagged teleconnections. For example, for the winter season, the predictors from March to October are considered to be the inputs, while for the summer season, the predictors from last September to April are set to be the inputs of the modelling.

# 3 | METHODS

## 3.1 | The NARMAX model

The Nonlinear Autoregressive Moving Average with exogenous input (NARMAX) model for multiple-input and single-output (MISO) systems is generally represented as follows (Wei, 2019; Wei & Billings, 2022):

$$\begin{aligned} y(k) = F\Big[&y(k-1), y(k-2), \ldots, y(k-n_y), \\ &u_1(k-d), u_1(k-d-1), \ldots, u_1(k-d-n_u), \\ &u_2(k-d), u_2(k-d-1), \ldots, \\ &u_2(k-d-n_u), \ldots, u_r(k-d), u_r(k-d-1), \\ &\ldots, u_r(k-d-n_u), \quad e(k-1), e(k-2), \ldots, \\ &e(k-n_e)\Big] + e(k) \end{aligned} \tag{1}$$

where $y(k)$, $u_i(k)$ (i = 1,2, ..., r) and $e(k)$ are the system output, input and noise sequences, respectively; $n_y$, $n_u$, and $n_e$ are the maximum lags for the system output, input and noise, respectively; $F[\cdot]$ is some nonlinear function, and $d$ is a time delay, typically set to $d = 0$ or $d = 1$. The noise sequence $e(k)$ is nearly always unknown for real-world modelling, and it is usually estimated using the prediction error $\xi(k) = y(k) - \hat{y}(k|k-1)$. In practice, there are many model structures that can be used to approximate the unknown mapping $F[\cdot]$, including power-form polynomial models, neural networks, radial basis function networks and wavelet expansions (Billings, 2013; Wei et al., 2010). Power-form polynomials, due to their desirable properties, especially their transparency and interpretability, are commonly used to construct NAMARX models (Billings, 2013).

The Nonlinear Autoregressive with eXogenous input (NARX) model presented below is a special case of the NARMAX model (1), which does include the lagged noise variables $e(k-1), e(k-2), \ldots, e(k-n_e)$,

$$\begin{aligned} y(k) = F\Big[&y(k-1), y(k-2), \ldots, y(k-n_y), \\ &u_1(k-d), u_1(k-d-1), \ldots, u_1(k-d-n_u), \\ &u_2(k-d), u_2(k-d-1), \ldots, u_2(k-d-n_u), \\ &\ldots, u_r(k-d), u_r(k-d-1), \ldots, \\ &u_r(k-d-n_u)\Big] + e(k). \end{aligned} \tag{2}$$

In many real-world applications, the output $y(k)$ in (1) and (2) is assumed to be irrelevant to previous output

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

values $y(k-1), y(k-2), \ldots, y(k-n_y)$. For such cases, model (2) reduces to the nonlinear infinite-impulse response model (NFIR, also known as Voterra series model) (Billings & Wei, 2008; Wei & Billings, 2009) as follows:

$$\begin{aligned} y(k) = F\Big[ & u_1(k-d), u_1(k-d-1), \ldots, u_1(k-d-n_u), \\ & u_2(k-d), u_2(k-d-1), \ldots, u_2(k-d-n_u), \ldots, \\ & u_r(k-d), u_r(k-d-1), \ldots, \\ & u_r(k-d-n_u)\Big] + e(k). \end{aligned} \quad (3)$$

In (4), if the time delay and the maximum lags are all set to be zero, that is, $d = n_y = n_u = 0$, then the NFIR model (4) reduces to a nonlinear multiple regression (NMR) model (Hall et al., 2019),

$$y(k) = F[u_1(k), u_2(k), \ldots u_r(k)] + e(k). \quad (4)$$

For a simple illustration, consider a system with three inputs $u_1$, $u_2$ and $u_3$, for which the full initial NFIR model comprising all linear and quadratic terms is

$$\begin{aligned} y(k) = {} & \theta_0 + \theta_1 u_1(k) + \theta_2 u_2(k) + \theta_3 u_3(k) + \theta_4 u_1(k)^2 \\ & + \theta_5 u_2(k)^2 + \theta_6 u_3(k)^2 + \theta_7 u_1(k) u_2(k) \\ & + \theta_8 u_1(k) u_3(k) + \theta_9 u_2(k) u_3(k) + e(k). \end{aligned} \quad (5)$$

The degree of nonlinearity or nonlinear degree of model (5) is 2, as it contains a number of quadratic model terms. If a polynomial model contains at least one cubic cross-product term, then its nonlinear degree is 3. However, model terms in (5) are usually not equally important for explaining the system and interpreting the change of the output $y(k)$, meaning that some terms that make a tiny or negligible contribution to explaining the variation in the response $y(k)$ may be removed from the model. With the help of a model-selection algorithm, called the Forward Regression with Orthogonal Least Squares (FROLS) (Billings, 2013), the most important model terms in (3) can be determined and used to generate a concise and compact model. FROLS uses an effective but simple measure, called the error reduction ratio (ERR), to evaluate the contribution of each candidate model term to explaining the variation of the system output. The number of model terms in the final model can be determined using several statistics criteria, such as the Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978) and the penalized error-to-signal ratio (PESR) (Wei et al., 2010).

## 3.2 | The sliding-window NARMAX models

Usually, the NARMAX model can generate reliable predictions and reveal convincing transparent relationships between the system input and output over the entire period of the dataset. However, it is more desirable to pay attention to and make use of local information in the dataset, especially for a complex dynamic time-varying system or process like climate change. Therefore, in the following, we introduce the sliding-window NARMAX model (SW-NARMAX) for seasonal forecasts.

The proposed framework of the sliding-window NARMAX model is shown in Figure 2. Take a single-input, single-output system as an example, where the measured input signal of $N$ samples is denoted by $U = [u_1, \ldots, u_N]^T$ and the corresponding output is denoted by $Y = [y_1, \ldots, y_N]^T$. In a matrix format, the dataset of the system can be represented as $D = [U, Y] = \left[(u_1, \ldots, u_N)^T, (y_1, \ldots, y_N)^T\right]$.

Let $W = [1, 1, \ldots, 1]_{w\times 1}$ be a window of length $w$. With the one-step forward sliding window (which is shown in the dotted rectangles in Figure 2), the original dataset $D$ can be resampled into $s$ subsets, where $s = N - w + 1$ as follows:

$$\begin{cases} \widehat{D}_1 = \left[(u_1, \ldots, u_w)^T, (y_1, \ldots, y_w)^T\right] \\ \widehat{D}_2 = \left[(u_2, \ldots, u_{w+1})^T, (y_2, \ldots, y_{w+1})^T\right] \\ \ldots \\ \widehat{D}_s = \left[(u_{N-w+1}, \ldots, u_N)^T, (y_{N-w+1}, \ldots, y_N)^T\right] \end{cases}. \quad (6)$$

For each windowed dataset $\widehat{D}_i$ $(i = 1, 2, \ldots, s)$, a NARMAX model $M_i$ can be generated using the method discussed in Section 3.1. Thus, based on (6), there will be $s$ NARMAX models $M = \{M_1, M_2, \ldots, M_s\}$, which are represented by the green rectangles in Figure 2, and $s$ predictions of the system output over the testing period $\hat{y}^{test} = [\hat{y}_1^{test}, \ldots, \hat{y}_s^{test}]$ will be calculated accordingly.

Note that the whole available observations are split into two parts: (1) around 80% of the data are used for model training, and (2) the remaining 20% are used for model testing. Each windowed dataset is a subset of the training dataset. To find the most appropriate model for each window, the associated windowed dataset is further partitioned into training and validation sub-datasets. For each window, the role of the validation data is threefold: (1) to test and validate the model performance using 'unseen' data during the training process; (2) to optimize and adjust model parameters and hyper-parameters, such as the window size and the model structures, where necessary; and (3) to avoid overfitting.

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

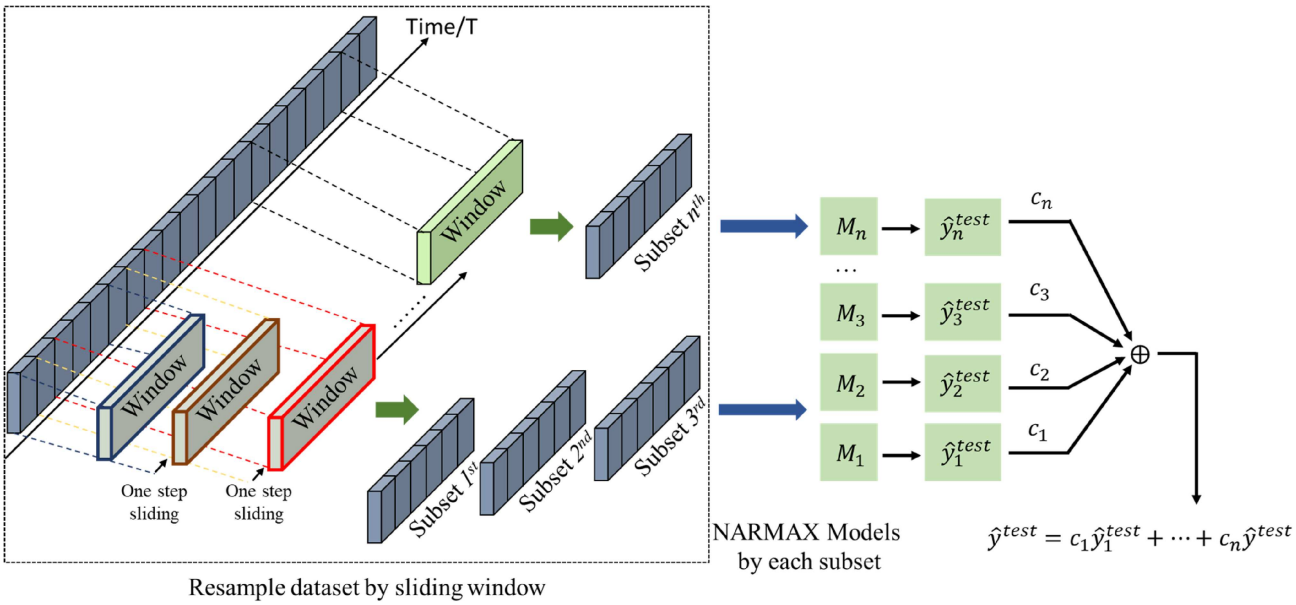



FIGURE 2 The framework of the sliding-window NARMAX models.

## 3.3 | Model selection and averaging

NARMAX modelling encompasses both linear and nonlinear models. In this case, as discussed in Section 3.1, the polynomial form of the model structure, including both linear and nonlinear forms, is considered. Note that different nonlinear degrees can lead to quite different models, which will affect the forecast and the explanation of the system. A model with a higher nonlinear degree may produce a more accurate representation of the system and hence produce better prediction. However, such a model can become very complex, and it needs a large number of samples for model training and estimation. Therefore, the present modelling challenge is a typical small-size problem, where the number of observations is smaller than the number of regressors. Taking a system with $n$ input variables as an example, if the nonlinear degree of the model is deg, then the number of generated model terms would be $(n+\deg)!/n!\deg!$, where the symbol '!' denotes the factorial function. This implies that models with a high nonlinear degree deg (4 or higher) are intractable, especially when the number of inputs is large, as the number of generated model terms would be tremendous.

In order to avoid overfitting and reduce the sensitivity of the model but meanwhile guarantee a reliable model structure, the highest nonlinear degree is set to be three in this study. Then, for each windowed dataset, there are three candidate NARMAX models: the linear model $M_{m,li}$, the quadratic model $M_{m,q}$ and the cubic model $M_{m,c}$, where the number of each candidate NARMAX models is defined as $M$. Therefore, with the window of length $w$, there are a total of $3M$ NARMAX models including linear and nonlinear model forms. To select the best models from the huge number of potential model candidates, the validation set is applied as discussed in Section 3.1. For each model $M_{m,i}$ (identified from the $m$-th windowed dataset), where $i \in \{li, q, c\}$, the prediction for the validation set is denoted by $\hat{y}_{m,i}^{v}$. The values of the mean squared error (MSE) of each of the $3M$ identified models are denoted by, $mse_{m,li}^{v}$, $mse_{m,q}^{v}$ and $mse_{m,c}^{v}$. The best model is selected by comparing the MSE in each data group. Thus, for each window of length $l$, there are $M$ best NARMAX models.

NARMAX methods are generally robust for system analysis and prediction, but using a single 'best' model may be risky in some applications. Therefore, it is reasonable to apply a model-averaging algorithm to reduce the risk associated with depending solely on a single model, especially when dealing with small sample size applications; this can also mitigate the sensitivity of the model to noise or uncertainties (Hall et al., 2019). In this study, the weighted mean scheme is also considered to deliver the predicted value. However, unlike the method in (Hall et al., 2019), the weights are calculated based on the MSE of the M models over the validation period, rather than over the training period.

Assume the values of mean squared errors (MSEs) of $n$ MARMAX models over their respective training periods are known as $mse_1, ..., mse_s$, respectively. Let

$$l_1 = 1/mse_1, \ ..., \ l_n = 1/mse_s, \tag{7}$$

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$l = l_1 + \ldots + l_n, \tag{8}$$

$$c_1 = l_1/l, \ \ldots, \ c_s = l_s/l. \tag{9}$$

Then, the averaged model prediction can be defined as

$$\hat{y}^{test} = c_1\hat{y}_1^{test} + \ldots + c_n\hat{y}_s^{test}. \tag{10}$$

### 3.4 | Prediction verification

Some typical model validation criteria, including correlation coefficients, mean absolute error (MAE) and root mean square error (RMSE), are used to evaluate model performance. The number of forecasts needs to be sufficiently large to make the statistical conclusions about the skill of the forecast robust and convincing, while the sliding-window NARMAX (SW-NARMAX) method can generate several models in the training set and produce valid statistical conclusions. As depicted in Figure 3, the original dataset is meticulously segmented into a training set, a validation set and a testing set. This methodological approach facilitates the initial training of models on the training set. Subsequently, these models are refined and optimized within the validation set. Finally, the efficacy and robustness of the models are comprehensively assessed in the (fully independent) testing set, ensuring a rigorous evaluation of their performance. In addition, the continuous ranked probability score (CRPS) (Leutbecher & Haiden, 2021) is used in this study to evaluate the quality of the seasonal forecast models. The CRPS estimates the difference between the observed and expected outcomes and can be viewed as an integral over the possible Brier scores (Bradley et al., 2008). It is herein defined as

$$CRPS\left(N\left(\overline{x},s^2\right),y\right) = \frac{(s)}{\sqrt{\pi}}\left\{\sqrt{\pi}\frac{y-\overline{(x)}}{s}erf\left(\frac{y-\overline{x}}{\sqrt{2}}\right) + \sqrt{2}\exp\left(-\frac{(y-\overline{x})^2}{2s^2}\right) - 1\right\}. \tag{11}$$

Normally, $x$ takes the value 1 or 0 according to whether or not the event occurred in the predefined class, especially a binary classification forecast, while $p_i$ is the forecast probability for such occasion $i$. To clearly calculate the CRPS of the SW-NARMAX models, we define the two class as follows:

$$class\ 1: \ x = 1, \ p_i \in [-0.5, 0.5], \tag{12}$$

$$class\ 2: x = 0, \ p_i \in [-3, -0.5) \cup (0.5, 3]. \tag{13}$$

To calculate the CRPS of the SW-NARMAX models, the forecast probability $p_i$ should be obtained first. By

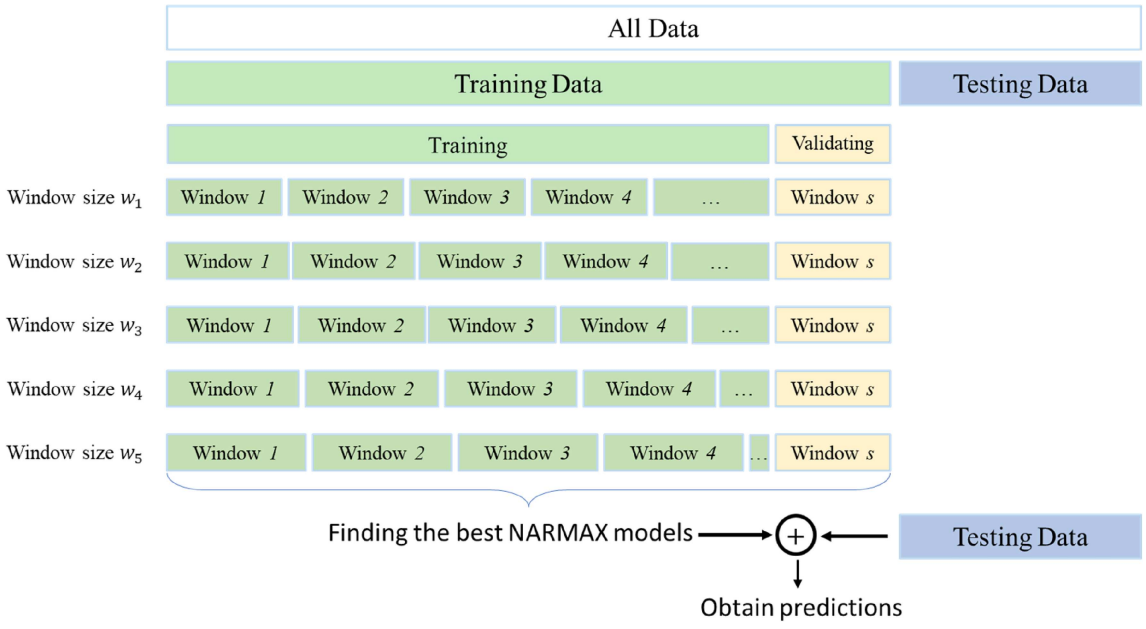



FIGURE 3 The schematic illustration of the NARMAX model resampling using the sliding-window approach.

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

applying the sliding-window methods in the training period, there will be $s$ SW-NARMAX models identified as discussed above. Thus, for each year in the testing set, there are $s$ predictions as defined above. We can calculate the forecast probability by

$$p_i = \frac{count(\widehat{y}_i \in [-0.5, 0.5])}{s}, \tag{14}$$

where *count* means function to calculate the quantity that meets the condition and $\widehat{y}_i$ $i = 1, 2, ..., s$ indicates the predictions of $i$th SW-NARMAX model.

Based on the definition of the classes, the CRPS verifies the accuracy of the predictions of SW-NARMAX models matching the class of the observation. The smaller the CRPS, the more consistent the category of SW-NARMAX predictions is with the category of observed values; otherwise, they are inconsistent.

Correlation coefficients and significance are assessed using Pearson's correlation coefficient and associated p-values to assess the probability of finding the result if the correlation was zero, where $p < 0.05$ and $p < 0.01$ are commonly selected values to assess significance.

To evaluate the statistical significance of the NARMAX models, a Monte Carlo sampling method with the autoregressive (AR) model analytical framework is implemented in the supplementary material. This approach involves generating 100 simulated databases derived from the AR models. By repeatedly sampling from these databases, a distribution of outcomes that reflects the inherent variability is constructed, which in turn allows us to estimate the statistical significance of the empirical results. The insights gained from these simulations provide a robust basis for evaluating the statistical significance of the NARMAX findings, ensuring that conclusions are not only grounded in empirical evidence but also resilient to the stochastic nature of the underlying processes we investigate.

### 3.5 | Dynamical models

To help assess the accuracy and utility of NARMAX-generated seasonal forecasts, comparisons are made with the seasonal forecasts from two commonly used dynamical models. Dynamical model seasonal forecasts are generated based on runs from up to 1 month in advance. Monthly forecast runs are obtained from the C35 Copernicus Climate Change Service. For this analysis, hindcast data for 1993 to 2016 are used from the ECMWF SEAS5 model (Johnson et al., 2019) and the Met Office GloSea5 model (MacLachlan et al., 2015). The GloSea5 outputs are based on seven ensemble members, while SEAS5 has 25 ensemble members over this period. The monthly runs are aggregated to produce a seasonal forecast that corresponds to the seasonal prediction from 1 month out (e.g., the winter forecasts are based on December with 1 month lead time, January with 2 months lead time and February with 3 months lead time, corresponding to a seasonal forecast issued in November).

Predictions of 500 hPa heights from the dynamical models are assessed against the three EOFs discussed in Section 2, for both winter and summer. As with NARMAX, in the case of summer, June is considered separately from high summer (July and August), while winter is defined simply as December, January and February and is labelled by the year of the January. Correlation, RMSE and the CRPS score are used together to provide an indication of forecast skill.

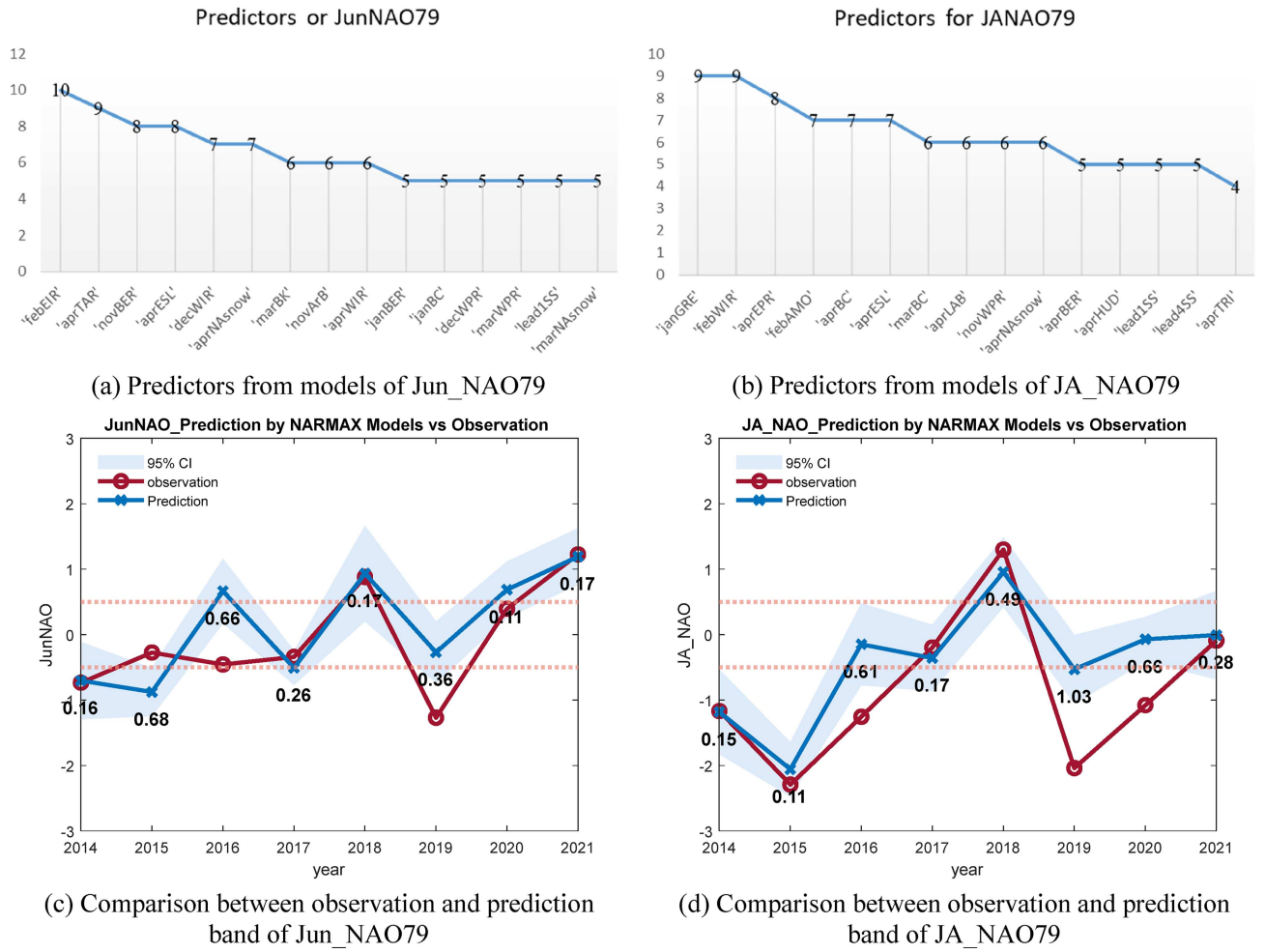
## 4 | RESULTS

In this section, sliding-window NARMAX models are defined by the indices they use (station-based NAO, EA and SCA), by the start year of the predictor dataset (1979) and by season (summer or winter). For example, the Jun_NAO79 (JA_NAO79) summer models are the sliding-window NARMAX models for the summer NAO in June (July and August average), using the 1979–2022 predictor dataset. In the main part of this paper, we focus on the NARMAX prediction results based on the 1979 (start date) datasets, while the rest of the prediction results are presented in the Supplementary Information. For a better evaluation of the performance of different models against observations, model results are based on weighted means of the model ensemble members over validation and test periods.

### 4.1 | Experimental settings

Firstly, we give a brief introduction to the 1979 dataset and the settings for the sliding-window NARMAX models. In this dataset, six indices or system outputs needed to be modelled in summer: that is, Jun_NAO79, JA_NAO79, Jun_EA79, JA_EA79, Jun_SCA79 and JA_SCA79. This dataset contains 43 observations (years) for each index, while the number of input variables or predictors is 130. Therefore, this application is a typical small number modelling and forecasting problem (Section 3.3). In the experiments, the predictors are up to 8 months in advance (Section 2) of the phenomenon being predicted, where the latest month is April for summer and October for winter weather.

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License




(a) Predictors from models of Jun_NAO79

(b) Predictors from models of JA_NAO79

(c) Comparison between observation and prediction band of Jun_NAO79

(d) Comparison between observation and prediction band of JA_NAO79

**FIGURE 4** Results (predictors (a, b) and predictions (c, d) by sliding-window NARMAX) of Jun_NAO79 and JA_NAO79. Predictors are shown according to the month in which that value occurred (e.g., AprTAR = tropical Atlantic rainfall for the month of April). Refer to Table 1 for a full list of predictor names.

In this paper, the original dataset is firstly resampled into training and testing sets with a ratio of 8:2, leading to 35 points in the training set and 9 points in testing set (8 points in testing set for summer season). To avoid overfitting, a validation process is performed over a subset of around eight or nine samples in the training set. The maximum nonlinear degree, deg, of NARMAX models is 3 (Section 3.1). Prediction results from sliding-window NARMAX models are shown in Sections 4.2 and 4.3.

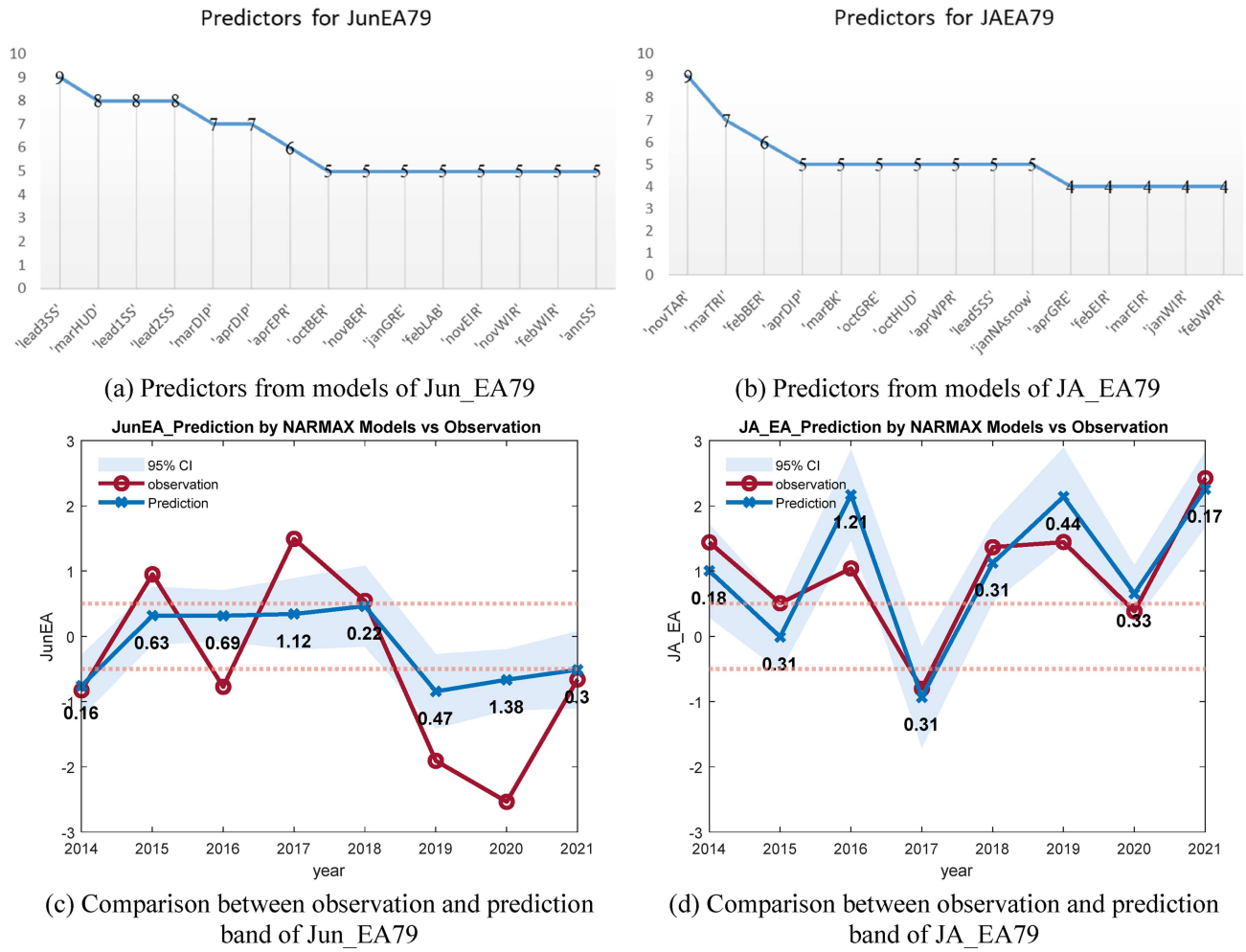
To evaluate how predictors have evolved, a separate NARMAX model for 1956 predictors dataset (i.e., covering the period 1956–2022) was developed. This model undergoes training during the period from 1956 to 2008/2009 is validated in the period 2001– 2008 and tested in the period 2009– 2021, as detailed in Supplementary Information Section 1.1. We refer to these as the 1979 and 1956 NARMAX models, named after the two different time periods they represent.

## 4.2 | Summer seasonal prediction results

### 4.2.1 | NAO summer results

For the June (July and August average) NAO, the 15 most frequent predictors in the sliding-window NARMAX models (16 models for Jun_NAO79 and 19 models for JA_NAO79) are listed in Figure 4a,b. Where the same predictor is shown for different months in the same graph (Figures 4–9), that refers to separate ensemble NARMAX models. To avoid overfitting, different months, such as records in February and April for summer, for a particular predictor are not used in the same NARMAX model. In models of Jun_NAO79, the most selected predictors are 'WIR' and' BER', which considering across all relevant months are both selected 13 times, followed by 'NAsnow', 'EIR' and 'WPR', which are selected 12, 10 and 10 times, respectively. For the JA_NAO79 models, the most selected predictor is 'BC', selected 13 times, followed by 'GRE' and

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License




(a) Predictors from models of Jun_EA79

(b) Predictors from models of JA_EA79

(c) Comparison between observation and prediction band of Jun_EA79

(d) Comparison between observation and prediction band of JA_EA79

**FIGURE 5** Results (predictors [a, b] and predictions [c, d] by sliding-window NARMAX) of Jun_EA79 and JA_EA79. Predictors are shown according to the month in which that value occurred (e.g., aprDIP = North Atlantic dipole for the month of April). Refer to Table 1 for a full list of predictor names.

'WIR', which are each identified nine times. In addition, in models of Jun_NAO79 and JA_NAO79, some predictors, specifically 'WIR', 'BER', 'ESL', 'NAsnow' and 'lead1SS', are relatively frequently selected.

The predictions made by the sliding-window NARMAX models are presented alongside the observed NAO time series in Figure 4c,d. In Figures 4–9, the red lines with 'o' markers represent the observations of the indices (i.e., the EOF time series defined in Section 2), while the blue lines with crosses represent the weighted mean predictions by the sliding-window NARMAX models. The light blue area is the 95% confidence interval (CI) generated by the identified NARMAX models. As shown in Figure 4c, the weighted average predictions by SW-NARMAX models follow the observations closely in 5 of 8 years and fall out of the CI in the remaining years (2015, 2016 and 2019). Similarly, in Figure 4d, most predictions (5/8) by SW-NARMAX models are close to the observations, while the rest of the years' (2016, 2019, and 2020) observations fall outside the CI.

As shown in Figure 4c,d, the two-digit decimal values are the CRPS based on the NARMAX ensemble weighted mean prediction for each year, while the two horizontal dashed lines indicate the values of −0.5 and 0.5. As the definition above, the area between −0.5 and 0.5 is considered to be 'true', while areas outside are set as 'false' for the purpose of CRPS calculations defined in Equations (12) and (13). Smaller CRPS values indicate that the predictions by SW-NAMRAX models are relatively more similar to the observations with the same class, while larger CRPS values indicate a larger departure between the predicted and observed values.

Consider Figure 4c as an example. When the predicted value approximates the observed value and the 95% confidence interval (CI) coincides with the category of observed values—as was the case in 2014, 2017, 2018 and 2021—the CRPS values are minimal, nearing zero. However, for specific years such as 2015 and 2016, while the observed value falls out of the predicted 95% CI range, the CRPS values approximate 0.6 because most

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

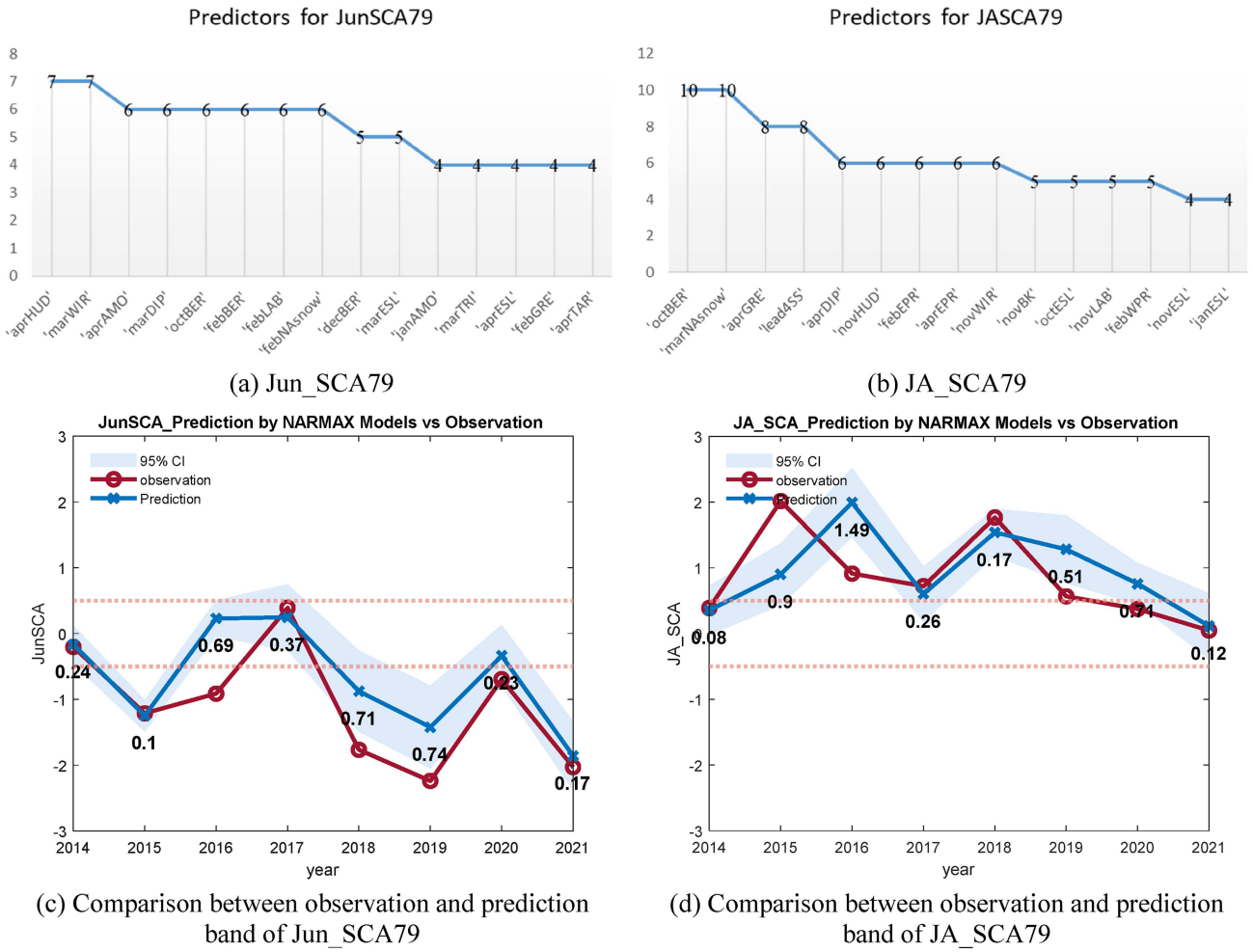



**FIGURE 6** Results (predictors [a, b] and predictions [c, d] by sliding-window NARMAX) of Jun_SCA79 and JA_SCA79. Predictors are shown according to the month in which that value occurred (e.g., aprESL = E. Siberian/Laptev Seas for the month of April). Refer to Table 1 for a full list of predictor names.

predictions do not correspond to the same category as the observed values. This result aligns with our conclusion that the SW-NARMAX weighted mean prediction and observation for that year fall into different categories.

NARMAX verification statistics against observed data for the model testing period are summarized in Table 2. Model-observation correlation coefficients of the Jun_NAO79 models and JA_NAO79 models for the entire testing period (2014–2021) are 0.89 and 0.87, respectively, and are highly significant ($p < =0.01$). Furthermore, the Jun_NAO79 model has more skillful performance due to its smaller RMSE and MAE compared with the JA_NAO79 model.

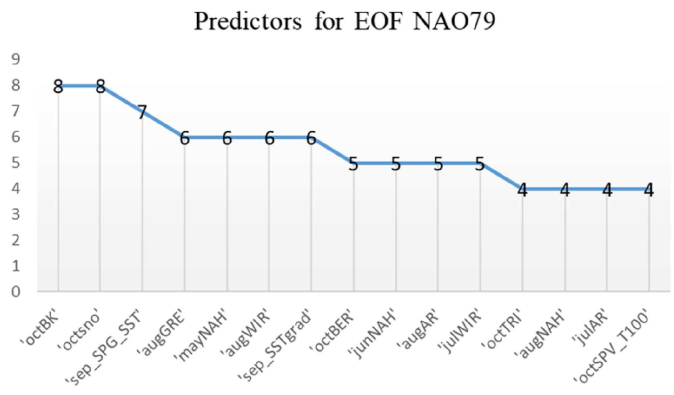
### 4.2.2 | EA summer results

The 15 most frequent predictors in the sliding-window NARMAX models (19 Jun_EA79 models and 14 JA_EA79 models) are shown in Figure 5a,b. There is some consistency in the analysis of the predictors: 'DIP', 'HUD', 'GRE', 'WIR', 'EIR' and 'BER' are selected in models for two indices with high frequencies. Among them, 'aprDIP' is the most frequently selected predictor for the two indices: 7 for Jun_EA79 and 5 times for JA_EA79. Meanwhile, there are several unique predictors in these two NARMAX models, such as 'lead3SS', 'lead2SS', 'lead1SS', 'EPR', 'LAB', and' annSS' for Jun_EA79 and 'TAR', 'BK', 'lead5SS' and 'NAsnow' for JA_EA79.

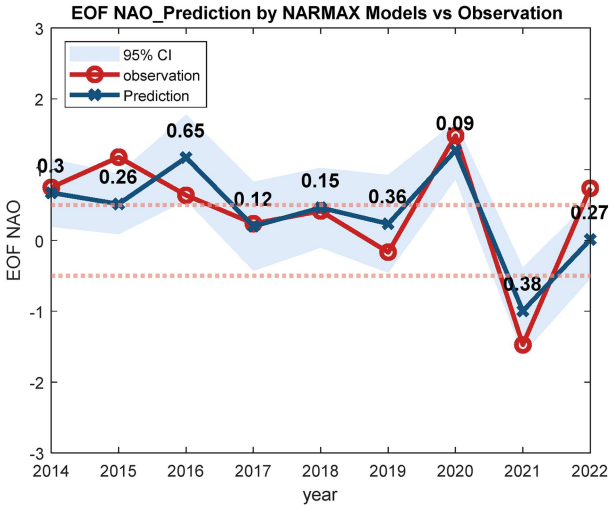
The comparison between the prediction of sliding-window NARMAX models for EA and the observations is shown in Figure 5c,d, while the verification statistics for the validation and testing periods for the mean predictions by the model ensembles are shown in Table 2.

Similarly, Figure 5c,d show that the weighted mean predictions by the NARMAX models usually perform well in following the observed yearly values from 2014 to 2021 (testing period) for Jun_EA79 and JA_EA79. For

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

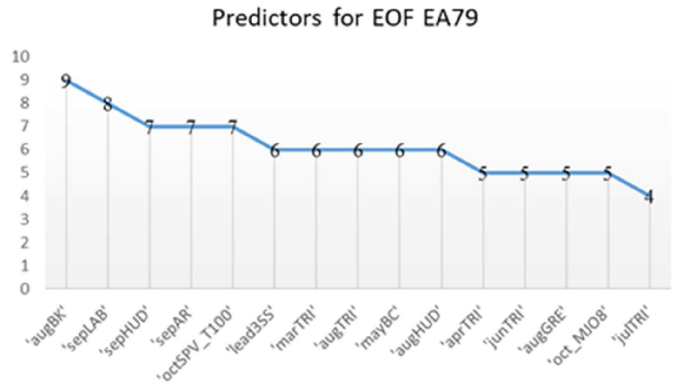



(a) Predictors analysis of models for winter EOF NAO79

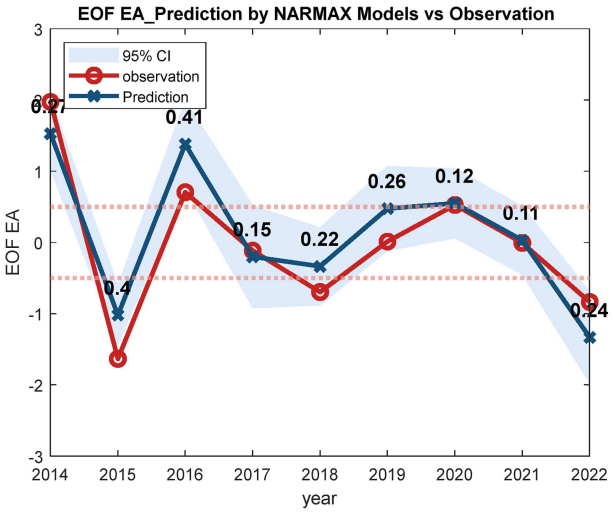



(b) Comparison between observation and averaged prediction for EOF NAO79

**FIGURE 7** Results (predictors [a] and predictions [b] by sliding-window NARMAX) of EOF NAO79 in winter. Predictors are shown according to the month in which that value occurred (e.g., augNAH = North Atlantic Horseshoe for the month of August). Refer to Table 1 for a full list of predictor names.




(a) Predictors analysis of models for winter EOF EA79



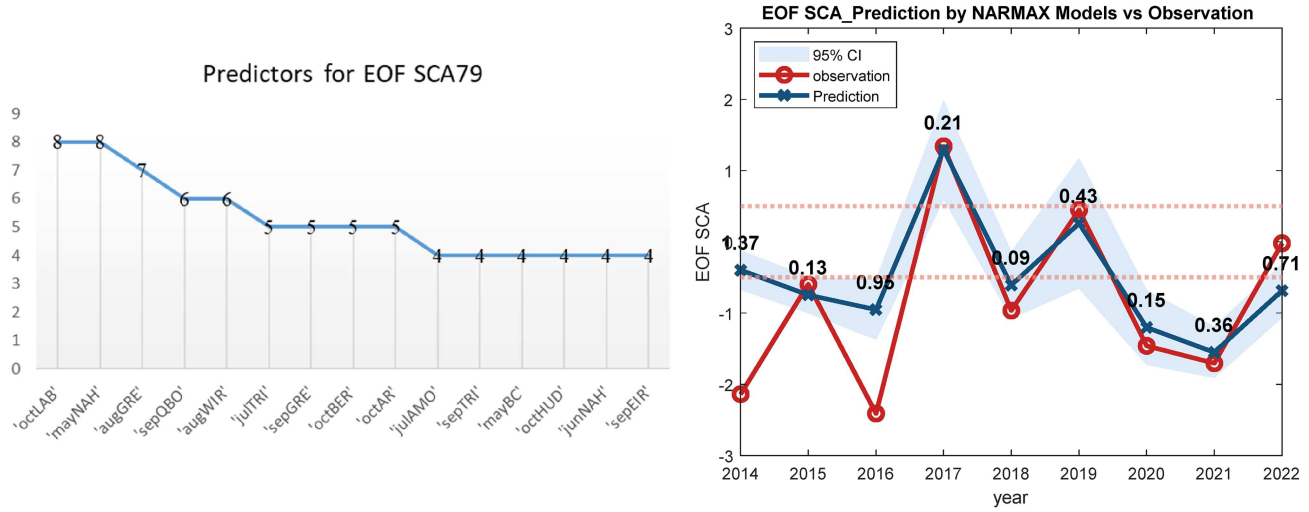

(b) Comparison between observation and averaged prediction for winter EOF EA79

**FIGURE 8** Results (predictors [a] and predictions [b] by sliding-window NARMAX) of EOF EA79 in winter. Predictors are shown according to the month in which that value occurred (e.g., augBK = Barents–Kara Seas for the month of August). Refer to Table 1 for a full list of predictor names.

the Jun_EA79, observations in 2014, 2018 and 2021 are in the 95% CI, while for the JA_EA79, most observations (7/8) fall in the prediction band. Thus, the SW-NARMAX models perform better for JA_EA79 than for Jun_EA79.

Similarly, the CRPS values reflect the accuracy of the probabilistic predictions of the SW-NARMAX ensemble models. For Jun_EA79, predictions in 2014, 2018 and 2021 are relatively accurate compared to other years. For the years 2017 and 2020, the CRPS values are greater than 1, as the weighted mean predictions of SW-NARMAX models for 2 years are significantly different. For the year 2015, 2016 and 2019, the CRPS values are less than 1 while greater than 0.4. In these years, the weighted mean predictions of SW-NARMAX models have obvious differences with observations, while the observations have the same category compared to the CI

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License




(a) Predictors analysis of models for winter EOF SCA79

(b) Comparison between observation and averaged prediction for winter EOF SCA79

**FIGURE 9** Results (predictors [a] and predictions [b] by sliding-window NARMAX) of EOF SCA79 in winter. Predictors are shown according to the month in which that value occurred (e.g., octLAB = Labrador Sea for the month of October). Refer to Table 1 for a full list of predictor names.

areas of SW-NARMAX models. For JA_EA79, the SW-NARMAX ensemble models show better accuracy among most years in the testing set, while in 2016 and 2019, the predictions by SW-NARMAX models are not accurate compared to the observations.

Model-observed correlation coefficients of the Jun_EA79 and JA_EA79 models for the testing period (2014–2021) are 0.78 and 0.88, respectively (Table 2), while the correlation coefficient of the Jun_EA79 is significant at the $p < =0.05$ level.

### 4.2.3 | SCA summer results

The statistical analysis of predictors in 15 Jun_SCA79 NARMAX models and 18 JA_SCA79 models are shown in Figure 6a,b. Ranked by frequency of predictor in the models for Jun_SCA79, 'BER' appears most often, 17 times in the analysis, followed by 'AMO' (10 times) and 'HUD' (7 times). For JA_SCA79, 'ESL' is the most frequently selected (appearing 13 times), followed by 'EPR' and 'BER', which appear 12 and 10 times, respectively.

Figure 6c,d show the comparison between the weighted mean predictions and observations over the out-of-sample period for SCA. A majority of the observations (10/16) are, once again, in the prediction band generated by the sliding-window models. However, as before, in a few years, that is, 2016, 2018 and 2019 for Jun_SCA79 and 2015, 2016 and 2019 for JA_SCA79, the probabilistic prediction bands do not encompass the observations. The statistical verification metrics for these models are once again summarized in Table 2. Model-observation correlation coefficients of the Jun_SCA79 models and JA_SCA79 models for the testing period (2014–2021) are, respectively, 0.69 and 0.65 (Table 2). The CRPS values in Figure 6c,d show that the SW-NARMAX models for the summer SCA yield accurate predictions for many years in the testing set, like 2014, 2015, 2017, 2020 and 2021 for Jun_SCA79 and 2014, 2017, 2018, 2020 and 2021 for JA_SCA, while for the rest years in the testing set for the Jun_SCA79 and JA_SCA, the weighted mean predictions and the CI areas of the SW-NAMRAX models have obvious differences with the observations.

**TABLE 2** Verification statistics for averaged sliding-window NARMAX models for summer seasonal prediction (test period = 2014–2021). Correlations that are significant at $p < = 0.05$ are in bold, and correlations that are significant at $p < = 0.01$ are underlined.

| | RMSE | MAE | Correlation coefficient |
|---|---|---|---|
| Jun_NAO79 | 0.38 | 0.27 | **0.89** |
| JA_NAO79 | 0.67 | 0.50 | **0.87** |
| Jun_EA79 | 0.87 | 0.67 | **0.78** |
| JA_EA79 | 0.80 | 0.62 | **0.88** |
| Jun_SCA79 | 0.84 | 0.64 | 0.69 |
| JA_SCA79 | 0.51 | 0.38 | 0.65 |

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

## 4.3 | Winter seasonal prediction results

### 4.3.1 | EOF NAO winter results

The frequency analysis of predictors in the winter EOF NAO79 models are shown in Figure 7a, showing the 15 most chosen predictors among 21 EOF NAO79 models. As listed in Figure 7a, the joint most selected predictors are 'OctBK' and 'Octsno', with 'sep_SPG_SST' as the next most frequently identified predictor. The performance comparison is shown in Figure 7b, while the verification statistics of the weighted mean model is shown in Table 3. For the EOF NAO model, the model-observation correlation coefficient over the testing set (2014–2022) period is 0.78.

The CRPS values in Figure 7b indicate that the SW-NARMAX models show good accuracy over the testing period for winter NAO as all CRPS values are close to zero, implying that observations and predictions are in the same class, while in the year 2016, the CRPS value is relatively high. Moreover, the observations are in the 95% CI area of the predictions proving the accuracy of the models.

### 4.3.2 | EOF EA winter results

Figure 8a shows the relative frequency of predictors that are included in 23 models of the winter EOF EA. 'augBK' is the most commonly identified predictor in winter EOF EA models. Based on the frequency analysis of predictors in EOF EA predictor models, the most commonly selected predictors are 'HUD' and 'BK', indicating they have more influence on the winter EA.

Performance comparison is shown in Figure 8b. The weighted mean prediction by EOF EA models (red line) closely follows the observations in both the validation and training sets, and the model prediction band consistently encompasses the observations. This comparison once again highlights the skillful performance of the NARMAX models, which is particularly evidenced by the significant correlation of 0.87 over the 2014–2022 testing period (Table 3).

As shown by the CRPS values in Figure 8b, the SW-NARMAX models of the winter EA demonstrate high skill over the testing period, as the proximity of all CRPS values to zero suggests that the observations and predictions belong to the same class. Model accuracy is corroborated by nearly all the observations (except 2015) falling within the 95% confidence interval (CI) range of predictions.

**TABLE 3** Verification statistics for averaged sliding-window NARMAX models for winter seasonal weather (testing period = 2014–2022). Correlations that are significant at $p <= 0.05$ are in bold, and correlations that are significant at $p <= 0.01$ are underlined.

| | RMSE | MAE | Correlation (2014–2021) |
|---|---|---|---|
| winter_NAO79 | 0.53 | 0.48 | **0.78** |
| winter_EA79 | 0.49 | 0.42 | **0.87** |
| winter_SCA79 | 0.71 | 0.54 | **0.84** |

### 4.3.3 | EOF SCA winter results

The 15 most frequent predictors in 16 winter EOF SCA models are shown in Figure 9a, where the most selected predictors are 'NAH' and 'TRI'. Figure 9b shows the comparison between observations and weighted mean prediction of identified models in the testing set. Although the RMSE (0.71) and MAE (0.54) of SCA79 in the testing set are larger than those of other indices, most (7/9) observations are in the prediction band, with the predictions in 2014 and 2016 outside the prediction band. However, the correlation coefficient (0.84 for SCA; $p < 0.01$) indicates overall high predictive skill (Table 3).

Figure 9b shows that in most years, the SW-NARMAX models have accurate predictions over the testing period. However, in 2014 and 2016, the performance of the SW-NARMAX models is considerably lower compared with other years, although the models still correctly predict a negative winter SCA.

## 4.4 | Linear and nonlinear NARMAX models

To better demonstrate the performance of the nonlinear relationship between the predictors and the atmospheric circulation indices, we compare experiments carried out using the following two types of NARMAX model: pure linear and pure nonlinear. This allows us to assess the influence of the nonlinear combination of predictors compared with assuming linearity in the seasonal weather system.

As before, the original dataset is divided into two parts: training and testing subsets with a ratio of 8:2. To overcome overfitting, a validation subset is created within the training set. For consistency, we put the statistical results (RMSE, MAE and correlation) from the 1979 dataset in this section with the testing period results for summer and winter, respectively, shown in Tables 4 and 5.

As shown in Tables 4 and 5, the pure nonlinear NARMAX models produce more accurate predictions than

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

| | RMSE | | MAE | | Correlation | |
|---|---|---|---|---|---|---|
| | Linear | Nonlinear | Linear | Nonlinear | Linear | Nonlinear |
| Jun_NAO79 | 0.77 | 0.46 | 0.67 | 0.32 | 0.35 | **0.84** |
| JA_NAO79 | 1.33 | 0.67 | 1.12 | 0.50 | 0.49 | **0.87** |
| Jun_EA79 | 1.34 | 0.87 | 1.12 | 0.67 | 0.46 | **0.78** |
| JA_EA79 | 1.18 | 0.93 | 1.02 | 0.86 | 0.29 | 0.70 |
| Jun_SCA79 | 1.34 | 0.94 | 1.16 | 0.89 | 0.49 | 0.60 |
| JA_SCA79 | 1.05 | 0.51 | 0.84 | 0.46 | 0.59 | 0.65 |

**TABLE 4** Comparison of the numerical performance of linear and nonlinear NARMAX models in summer (testing period: 2014–2021), where significant correlation coefficients are highlighted in bold ($p <= 0.05$).

**TABLE 5** Comparison of the numerical performance of linear and nonlinear NARMAX models in winter (testing period: 2014–2022), where significant correlation coefficients are highlighted in bold ($p <= 0.05$).

| | RMSE | | MAE | | Correlation coefficient | |
|---|---|---|---|---|---|---|
| | Linear | Nonlinear | Linear | Nonlinear | Linear | Nonlinear |
| winter_NAO79 | 0.85 | 0.59 | 0.74 | 0.53 | 0.46 | **0.73** |
| winter_EA79 | 0.81 | 0.68 | 0.54 | 0.40 | 0.59 | **0.87** |
| winter_SCA79 | 1.30 | 0.84 | 1.08 | 0.60 | 0.45 | **0.82** |

pure linear NARMAX models. Compared to the linear NARMAX model, the pure nonlinear NARMAX model has reduced the average RMSE of the predicted values for six indices in summer by 120% and increased the correlation coefficient by 80.65%. Meanwhile, for three indices in winter, it has decreased their average RMSE prediction value by 98.45 and increased their correlation coefficient by 65.3%.

Mixed NARMAX models (including both linear and nonlinear model terms) are usually more robust and show better performance than either pure linear or pure nonlinear NARMAX models, with a reduced average RMSE in summer by 24.85% and in winter by 29.35%, and an improved correlation coefficient in summer by 50.2% and in winter by 35.5%. While models for most indices are dominated by nonlinear elements, for some indices, for example, JA_NAO79, jun_EA79, JA_SCA79 and winter_EA79, linear model terms play a significant role in representing and explaining the variation in the target signals.

## 4.5 | Dynamical models

Figure S12 illustrates how SEAS5 fared when predicting the winter and high summer NAO, EA and SCA, using hindcast data for 1993–2015 and forecast data for 2016–2022, using that were initialized on 1 December (for winter) and 1 June (for summer). Figure S13 shows the corresponding data for GloSea5/6 (GloSea5 was superseded by GloSea6 in 2021), which are based on forecasts where all of the ensembles were initialized during the month leading up to and including 1 December (for winter) and 1 June (for summer). GloSea5 forecast data were not available for 2016 and 2017. There is a consistent tendency to underpredict the amplitude of the extreme seasons, to a greater extent than is observed for the NARMAX predictions, possibly reflecting a greater 'signal-to-noise' problem than we see with NARMAX. Both models showed significant (at $p < 0.05$) skill with the winter NAO, producing correlations of just over 0.4, but showed little skill at predicting the winter East Atlantic pattern. The GloSea5/6 correlation of 0.42 with the winter NAO is lower than 0.62 reported by Scaife et al. (2014), but if the analysis is restricted to the period 1993–2012, the correlation is much closer at 0.56, suggesting that the difference is primarily due to GloSea5/6 having a lower success rate in recent winters, for example, failing to predict the negative NAO of winter 2020/2021. While winters 2009/2010 and 2010/2011 were both predicted to have a negative NAO, none of the ensemble members of SEAS5 or GloSea5 captured the intensity of the anomaly. While SEAS5 correctly predicted a negative East Atlantic pattern for the winter of 2004/05, the observed outcome was more extreme than predicted by any of the ensemble members. SEAS5 particularly tended to underpredict the variability in the East Atlantic pattern in winter. For the SCA pattern, again SEAS5 underpredicted the variability but performed better overall than GloSea5.

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

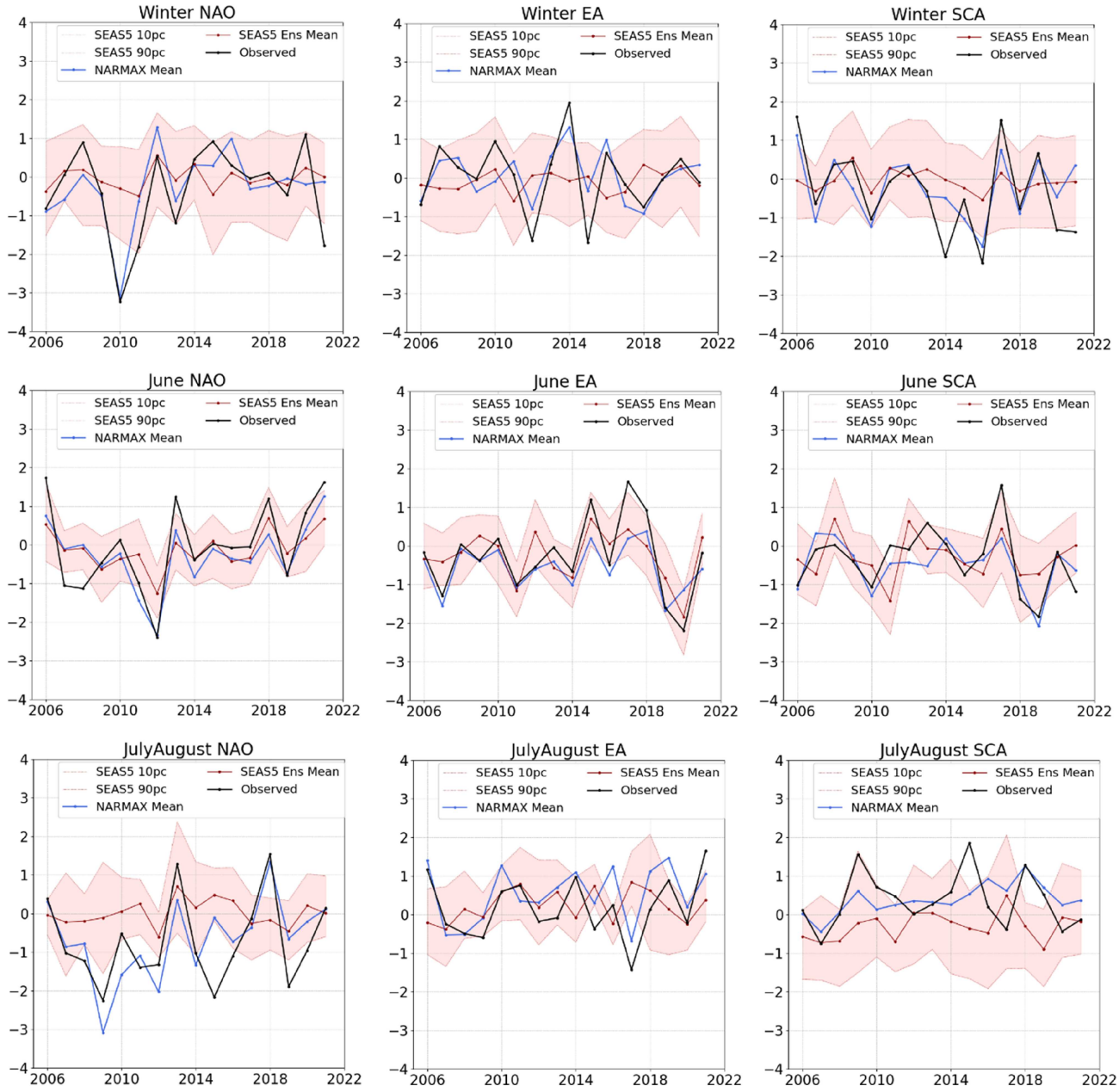



**FIGURE 10** Comparisons between NARMAX predictions and SEAS5 hindcasts and forecasts for the period 2006–2021.

In the case of high summer, covering July and August, no skill is found for the summer NAO or SCA, with both SEAS5 and GloSea5/6 giving small negative correlations with the observed values. However, GloSea5/6 showed evidence of some skill at predicting the summer EA, with a correlation of 0.36 with the observed values, albeit still lower than the values obtained from NARMAX. It appears that high summer is the area where NARMAX may prove especially useful as a means of improving the reliability of our seasonal forecasts.

When June is considered, the dynamical models perform much better, reflecting their greater skill at 1 month out. SEAS5 outperforms GloSea5/6, with a correlation of 0.71 with the June NAO and a correlation of 0.82 with the observed June East Atlantic pattern, while GloSea5/6 showed correlations of 0.51, 0.44 and 0.42 with the NAO, EA and SCA, respectively. SEAS5 correlations with June SCA were also lower, at 0.51.

## 4.6 | Comparison between dynamic models and NARMAX models

In this section, we compare the performance of the sliding-window NARMAX models discussed earlier against the SEAS5 and GloSea5 dynamical models, based

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

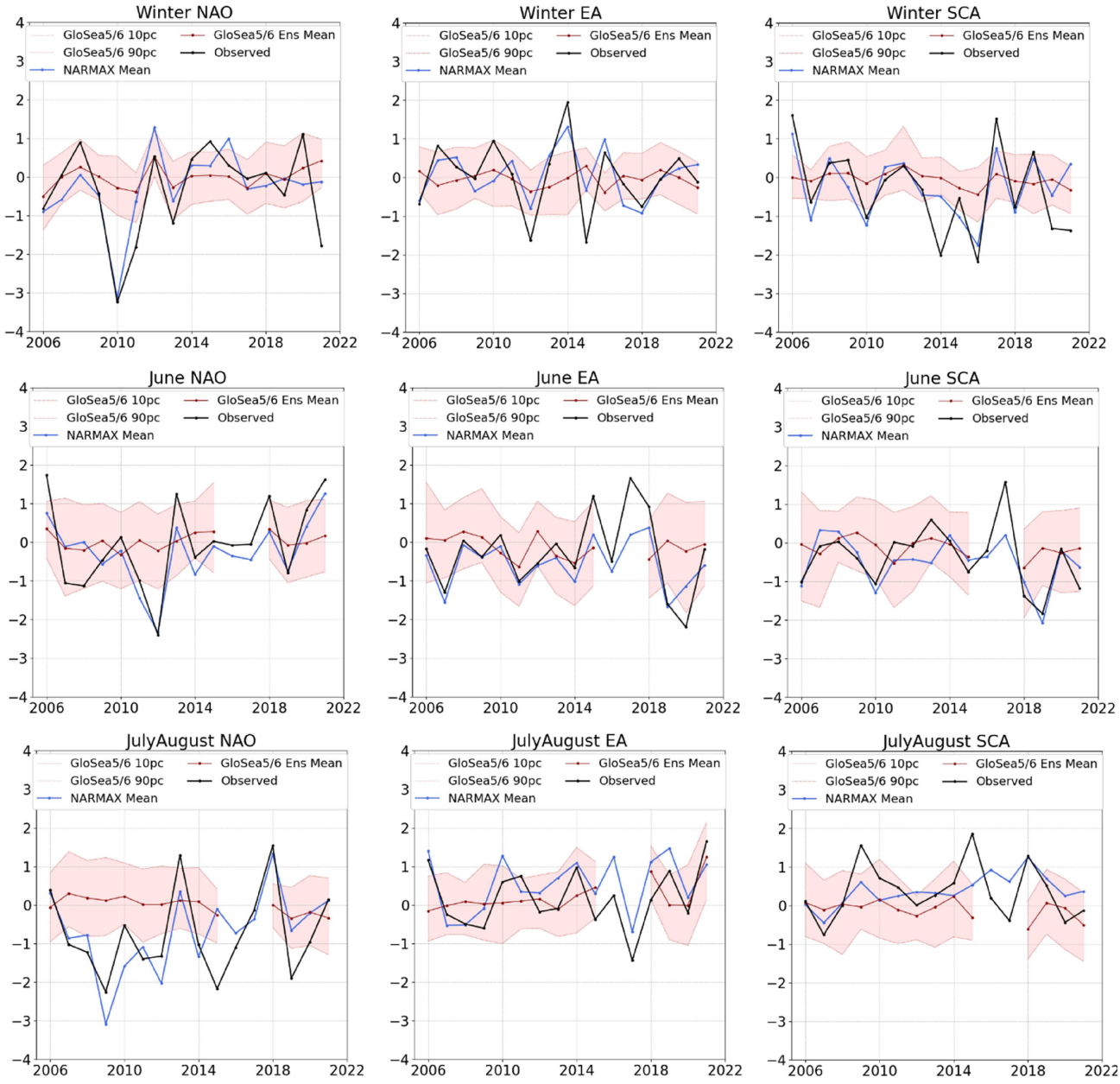



**FIGURE 11** Comparisons between NARMAX predictions and GloSea5/6 hindcasts and forecasts for the period 2006–2021. GloSea5 forecasts issued in the month ending 1 June were missing for 2016 and 2017.

on a training period of 1979–2005 (27 years) and an out-of-sample/testing period of 2006–2021 (16 years). Figures 10 and 11 show the NARMAX mean forecast for each year, compared with the observed values and the SEAS5 forecasts (Figure 10) and GloSea5/6 forecasts (Figure 11). It is apparent that NARMAX suffers less from the 'signal-to-noise' issue than the dynamical models, although in some cases (especially winter EA), there is still evidence of NARMAX underpredicting extreme values. Another comparison between the SEAS5 and GloSea5/6 dynamical models and NARMAX models is shown in Table 6, based on the data shown in Figures 10 and 11. The statistical results for these NARMAX models show higher correlation coefficients and smaller RMSE for all indexes, indicating a more skillful performance than the SEAS5 and GloSea5/6 models.

In Table 6, the predictions by NARMAX are compared with the SEAS5 and GloSea5 models for winter and separately for June and high summer (July and August). NARMAX consistently outperforms the dynamical models, and NARMAX forecasts are significantly correlated with the observed outcome in all cases. With the exception of JA_SCA, all correlations are significant at $p < 0.01$. The dynamical models only showed skill in

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

**TABLE 6** Verification statistics comparing the SEAS5 and GloSea5/6 model hindcasts and forecasts with NARMAX, for the period 2006–2021 (with NARMAX using the training period 1979–2005). Correlations that are significant at $p <= 0.05$ are in bold, and correlations that are significant at $p <= 0.01$ are underlined.

| | Correlation with observed | | | RMSE | | |
|---|---|---|---|---|---|---|
| **Index** | **SEAS5** | **GloSea5/6** | **NARMAX** | **SEAS5** | **GloSea5/6** | **NARMAX** |
| DJF_NAO | **0.52** | 0.47 | **0.76** | 1.07 | 1.10 | 0.75 |
| DJF_EA | −0.16 | −0.11 | **0.78** | 0.98 | 0.94 | 0.57 |
| DJF_SCA | **0.51** | **0.57** | **0.76** | 1.02 | 1.03 | 0.72 |
| Jun_NAO | **0.83** | **0.54** | **0.87** | 0.75 | 1.04 | 0.61 |
| Jun_EA | **0.77** | −0.05 | **0.90** | 0.63 | 0.96 | 0.58 |
| Jun_SCA | 0.41 | 0.27 | **0.77** | 0.77 | 0.71 | 0.53 |
| JA_NAO | 0.18 | 0.01 | **0.71** | 1.30 | 1.34 | 0.80 |
| JA_EA | −0.15 | 0.42 | **0.80** | 0.91 | 0.64 | 0.60 |
| JA_SCA | −0.08 | −0.13 | **0.52** | 1.05 | 0.96 | 0.60 |

some cases, with only SEAS5 predictions of June NAO and June EA proving comparable to the NARMAX predictions. It is worth noting that the GloSea5/6 forecasts of the winter NAO are statistically significant at $p < 0.05$ when a longer time period is considered, as is seen with 1993–2022 (Figure S13), but due to the smaller sample size, the correlation of 0.47 over 2006–2021 falls short of being statistically significant.

## 5 | DISCUSSION

The results presented here highlight the potential for NARMAX to add considerable value to current dynamical model predictions of NAO, EA and SCA, especially in the case of summer, where dynamical models tend to struggle to a greater extent than for winter. The NAO alone only accounts for some of the variability in temperature and precipitation over Northwest Europe, making it useful for predicting other important modes of atmospheric circulation variability. For example, (Hall & Hanna, 2018) attributed the exceptionally high rainfall of winter 2013/2014 over much of the UK primarily to a strongly positive East Atlantic pattern. It is therefore encouraging that the NARMAX results are strongly correlated with the observations in the case of EA and SCA as well as NAO.

Splitting the summer into June and July/August shows that the dynamical models perform better at forecasting June NAO, EA and SCA from 1 month ahead than at forecasting July/August, as would be expected. The highest correlation with the observed data is 0.82, in the case of SEAS5 predictions of June EA (compared with correlations of 0.84 and 0.78 for the 1956 and 1979 NARMAX models for June EA, respectively). Surprisingly, the NARMAX verification rate is similar for June and for July/August, indicating that there is less of an advantage over the dynamical models at a relatively short time range but that the accuracy of NARMAX does not decline as much for 2 and 3 months ahead, at least in the case of summer.

Compared with the dynamical models, NARMAX predictions show a reduced 'signal-to-noise' problem, that is, the year-to-year variability of the NAO, EA and SCA is captured accurately, and while the amplitude of extreme events is at times underpredicted, it is generally underpredicted to a much smaller extent than with the dynamical models. When analysing summer predictions for June and for July/August, it is clear that the 'signal-to-noise' problem with the dynamical models increases markedly when predicting 2 and 3 months ahead as opposed to just 1 month ahead, at least in the case of summer. This is far less apparent with the NARMAX predictions, suggesting that NARMAX-assisted forecasts may be especially useful at reducing the 'signal-to-noise' problem when forecasting further ahead. It is particularly encouraging that this appears to be true for summer, as the dynamical models tend to struggle more with seasonal predictions of summer than of winter.

The lists of predictors that the sliding-window NARMAX chooses for summer are mixed, but some consistent results stand out. The 1956 model (see the supporting information) has March Beaufort/Chukchi Sea SSTs as one of the three most selected predictors for both June and July/August NAO. Sea-ice and SSTs dominate among the most often selected predictors, perhaps suggesting feedback between sea-ice concentrations and SST anomalies. The 1979 model is less likely to select sea-ice concentrations, and tropical rainfall dominates among the most selected predictors for June and July/August NAO. The same is mostly true for the EA, but for June EA, both the 1956 and 1979 models often select solar activity with varying lag times. For July/August SCA, the 1979 model has March North Atlantic snow cover as the second most often selected predictor.

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The recurrence of sea-ice concentrations in the top 10 predictors must be taken with some caution, for as discussed (Hall, Scaife, et al., 2017), the recent sharp decline in sea-ice concentrations could contribute to models overestimating the influence of sea ice and potentially issuing poorer forecasts for recent years. However, despite this issue, the NARMAX predictions consistently outperformed the dynamical models, especially in the case of high summer (July and August).

For winter, both sets of models commonly select the North Atlantic Horseshoe (NAH) to predict the winter NAO, especially the 1956 model, which has the May and September NAH as the two most often selected predictors. This is a reassuring result, as it ties in well with the findings of Cassou, Deser, et al. (2004), which identified plausible physical explanations for observed links between the NAH and the subsequent winter NAO, particularly in relation to the NAH during the preceding summer and autumn. The 1979 model most often selects October Barents–Kara Sea-ice concentrations as a predictor for the winter NAO, again an encouraging result, as, for example, Warner et al. (2020) found evidence of interannual winter NAO variability being strongly related to Barents-Kara Sea ice. The 1979 model's second most selected predictor for winter NAO is October European snow cover, which is again a plausible result, as observations and dynamical model simulations also point to the existence of links between the two (Wegmann et al., 2020). September Hudson Bay sea-ice concentrations recur as a commonly selected predictor for the winter EA, and to a lesser extent so does the October stratospheric polar vortex. Hall, Scaife, et al. (2017)) discussed links between solar activity (with a lead time of 6 months to 2 years) and the June tripole and the winter NAO. Neither of those were in the top 10 predictors of the 1956 NARMAX model, although in the model from 1979, the October tripole both featured in the 10 most frequently selected predictors.

Lagged teleconnection links between sea-ice concentrations, SST anomalies, tropical precipitation and subsequent atmospheric circulation patterns have already been found. For example, there may be links between Barents–Kara Sea ice concentrations and extratropical atmospheric circulation via complex teleconnections with the Aleutian low and tropical SST and rainfall variations (Warner et al., 2020). This also ties in well with the 1979 NARMAX model frequently choosing October Barents–Kara sea-ice concentrations as a predictor of the winter NAO. There is also evidence for a link between tropical precipitation anomalies and wintertime European precipitation events (Li et al., 2020) and, correspondingly, the East Atlantic Pattern (Maidens et al., 2021). Some further discussion of the physical interpretability of the NARMAX results is available in the supporting material (see support material section 4).

Ongoing research is downscaling the three principal EOFs used here, in order to determine the links between the EOF time series (both observed and predicted by NARMAX) and Northwest European temperatures and precipitation, including links with persistence and variability indices as well as maximum, minimum and mean values, that are relevant for end-users such as the agri-food, energy and tourism industries.

## 6 | SUMMARY

These results demonstrate that NARMAX models have considerable potential to improve upon purely dynamical model-based seasonal weather predictions, especially in the case of high summer (July and August) and therefore significantly extends the pilot study of Hall et al. (2019), which focused on winter. NARMAX models that are designed based on a sufficiently long training period of at least around 25 years, consistently show skillful performance across the range of atmospheric circulation indices and seasons used here. Links between the individual circulation indices and their potential predictors that are frequently chosen by NARMAX are a basis for future work, both with the aim of evaluating the physical plausibility of such links and using NARMAX to assist the identification of new teleconnection links that have not previously been identified and explored.

## AUTHOR CONTRIBUTIONS

**Yiming Sun:** Formal analysis (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Ian Simpson:** Data curation (equal); formal analysis (equal); resources (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Hua-Liang Wei:** Conceptualization (equal); funding acquisition (equal); methodology (equal); project administration (equal); software (equal); supervision (equal); validation (equal); writing – review and editing (equal). **Edward Hanna:** Conceptualization (equal); data curation (equal); formal analysis (equal); funding acquisition (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); writing – review and editing (equal).

## ACKNOWLEDGEMENTS

We acknowledge the Natural Environment Research Council, United Kindom for funding this research (Grant No. NE/V001787/1), the ECMWF for providing SEAS5

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

forecast and hindcast runs and the ERA5 reanalysis (via Copernicus), and the Met Office for providing GloSea5 forecast and hindcast runs. We also thank Adam Scaife and Jamie Kettleborough for providing useful feedback and assistance with using GloSea5, and Richard Hall and Thomas Cropper for useful feedback and advice on generating the three principal EOFs.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in CORDEX regional climate model data on single levels at http://doi.org/10.24381/cds.bc91edc3, and the climate data store at https://cds.climate.copernicus.eu/#!/home.

## ORCID

*Yiming Sun* https://orcid.org/0000-0003-2685-1615
*Hua-Liang Wei* https://orcid.org/0000-0002-4704-7346
*Edward Hanna* https://orcid.org/0000-0002-8683-182X

## REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. Available from: https://doi.org/10.1109/TAC.1974.1100705

Baker, M., Bergstresser, D., Serafeim, G. & Wurgler, J. (2018) Financing the response to climate change: the pricing and ownership of US green bonds. *National Bureau of Economic Research*. Available from: https://doi.org/10.3386/w25194

Billings, S.A. (2013) *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. London: John Wiley & Sons.

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, 714–724. Available from: https://doi.org/10.1080/00207170701216311

Bradley, A.A., Schwartz, S.S. & Hashino, T. (2008) Sampling uncertainty and confidence intervals for the brier score and brier skill score. *Weather and Forecasting*, 23, 992–1006. Available from: https://doi.org/10.1175/2007WAF2007049.1

Cassou, C., Deser, C., Terray, L., Hurrell, J.W. & Drévillon, M. (2004) Summer sea surface temperature conditions in the North Atlantic and their impact upon the atmospheric circulation in early winter. *Journal of Climate*, 17, 3349–3363. Available from: https://doi.org/10.1175/1520-0442(2004)017<3349:SSSTCI>2.0.CO;2

Cassou, C., Terray, L., Hurrell, J.W. & Deser, C. (2004) North Atlantic winter climate regimes: spatial asymmetry, stationarity with time, and oceanic forcing. *Journal of Climate*, 17, 1055–1068. Available from: https://doi.org/10.1175/1520-0442(2004)017<1055:NAWCRS>2.0.CO;2

Center, S.W.D. (1956–2021) The international sunspot number. International Sunspot Number Monthly Bulletin and Online Catalogue.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S. et al. (2019) S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdisciplinary Reviews: Climate Change*, 10, e00567. Available from: https://doi.org/10.1002/wcc.567

Davies, R. (2014) *The cost of the UK floods* [Online]. UK: The floodlist. Available. Available from: https://floodlist.com/insurance/uk/cost-of-2013-2014-floods [Accessed 2nd March 2016].

Dawson, A. (2016) Eofs: a library for EOF analysis of meteorological, oceanographic, and climate data. *Journal of Open Research Software*, 4, e14. Available from: https://doi.org/10.5334/jors.122

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L. et al. (2014) Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, 41, 5620–5628. Available from: https://doi.org/10.1002/2014GL061146

Estilow, T.W., Young, A.H. & Robinson, D.A. (2015) A long-term northern hemisphere snow cover extent data record for climate studies and monitoring. *Earth System Science Data*, 7, 137–142. Available from: https://doi.org/10.5194/essd-7-137-2015

Folland, C.K., Knight, J., Linderholm, H.W., Fereday, D., Ineson, S. & Hurrell, J.W. (2009) The summer North Atlantic oscillation: past, present, and future. *Journal of Climate*, 22, 1082–1103. Available from: https://doi.org/10.1175/2008JCLI2459.1

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. Available from: https://doi.org/10.1002/joc.4121

Hall, R.J. (2016) *The North Atlantic polar front jet stream: variability and predictability, 1871–1914*. PhD Thesis, University of Sheffield, Sheffield.

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, e660–e677. Available from: https://doi.org/10.1002/joc.5398

Hall, R.J., Jones, J.M., Hanna, E., Scaife, A.A. & Erdélyi, R. (2017) Drivers and potential predictability of summer time North Atlantic polar front jet variability. *Climate Dynamics*, 48, 3869–3887. Available from: https://doi.org/10.1007/s00382-016-3307-0

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. Available from: https://doi.org/10.1175/WAF-D-16-0124.1

Hall, R.J., Wei, H.-L. & Hanna, E. (2019) Complex systems modelling for statistical forecasting of winter North Atlantic atmospheric variability: a new approach to North Atlantic seasonal forecasting. *Quarterly Journal of the Royal Meteorological Society*, 145, 2568–2585. Available from: https://doi.org/10.1002/qj.3579

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049. Available from: https://doi.org/10.1002/qj.3803

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L. et al. (2019) SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12, 1087–1117. Available from: https://doi.org/10.5194/gmd-12-1087-2019

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kushnir, Y., Robinson, W.A., Chang, P. & Robertson, A.W. (2006) The physical basis for predicting Atlantic sector seasonal-to-interannual climate variability. *Journal of Climate*, 19, 5949–5970. Available from: https://doi.org/10.1175/JCLI3943.1

Legg, T.P. & Mylne, K.R. (2004) Early warnings of severe weather from ensemble forecast information. *Weather and Forecasting*, 19, 891–906. Available from: https://doi.org/10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2

Leutbecher, M. & Haiden, T. (2021) Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Quarterly Journal of the Royal Meteorological Society*, 147, 425–442. Available from: https://doi.org/10.1002/qj.3926

Li, R.K.K., Woollings, T., O'reilly, C. & Scaife, A.A. (2020) Tropical atmospheric drivers of wintertime European precipitation events. *Quarterly Journal of the Royal Meteorological Society*, 146, 780–794. Available from: https://doi.org/10.1002/qj.3708

Lorenz, E.N. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20, 130–141. Available from: https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Maclachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A. et al. (2015) Global seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141, 1072–1084. Available from: https://doi.org/10.1002/qj.2396

Maidens, A., Knight, J.R. & Scaife, A.A. (2021) Tropical and stratospheric influences on winter atmospheric circulation patterns in the North Atlantic sector. *Environmental Research Letters*, 16, 024035. Available from: https://doi.org/10.1088/1748-9326/abd8aa

Marshall, J., Kushnir, Y., Battisti, D., Chang, P., Czaja, A., Dickson, R. et al. (2001) North Atlantic climate variability: phenomena, impacts and mechanisms. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21, 1863–1898.

Naujokat, B. (1986) An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. *Journal of Atmospheric Sciences*, 43, 1873–1877. Available from: https://doi.org/10.1175/1520-0469(1986)043<1873:AUOTOQ>2.0.CO;2

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. Available from: https://doi.org/10.1073/pnas.1713146114

Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N. et al. (2014) Skillful long-range prediction of European and north American winters. *Geophysical Research Letters*, 41, 2514–2519. Available from: https://doi.org/10.1002/2014GL059637

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Siegert, S., Stephenson, D.B., Sansom, P.G., Scaife, A.A., Eade, R. & Arribas, A. (2016) A Bayesian framework for verification and recalibration of ensemble forecasts: how uncertain is NAO predictability? *Journal of Climate*, 29, 995–1012. Available from: https://doi.org/10.1175/JCLI-D-15-0196.1

Stephenson, D.B., Pavan, V. & Bojariu, R. (2000) Is the North Atlantic oscillation a random walk? *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 20, 1–18. Available from: https://doi.org/10.1002/(SICI)1097-0088(200001)20:1<1::AID-JOC456>3.0.CO;2-P

Stockdale, T.N., Molteni, F. & Ferranti, L. (2015) Atmospheric initial conditions and the predictability of the Arctic oscillation. *Geophysical Research Letters*, 42, 1173–1179. Available from: https://doi.org/10.1002/2014GL062681

Warner, J.L., Screen, J.A. & Scaife, A.A. (2020) Links between Barents-Kara sea ice and the extratropical atmospheric circulation explained by internal variability and tropical forcing. *Geophysical Research Letters*, 47, e2019GL085679. Available from: https://doi.org/10.1029/2019GL085679

Wegmann, M., Rohrer, M., Santolaria-Otín, M., & Lohmann, G. (2020) Eurasian autumn snow link to winter North Atlantic Oscillation is strongest for Arctic warming periods. *Earth System Dynamics*, 11(2), 509–524. https://doi.org/10.5194/esd-11-509-2020

Wei, H.-L., & Billings, S. A. (2022). Modelling COVID-19 pandemic dynamics using transparent, interpretable, parsimonious and simulatable (TIPS) machine learning models: A case study from systems thinking and system identification perspectives. In Jiang, R., Crookes, D., Wei, H. L., Zhang, L., Chazot, P. (Eds.), *Recent Advances in AI-enabled Automated Medical Diagnosis* (pp. 13–27). CRC Press.

Wei, H.-L. (2019) Sparse, interpretable and transparent predictive model identification for healthcare data analysis. In: Rojas, I., Joya, G. & Catala, A. (Eds.) *Advances in computational intelligence*. Cham: Springer International Publishing, pp. 103–114.

Wei, H.-L. & Billings, S.A. (2009) Improved model identification for non-linear systems using a random subsampling and multifold modelling (RSMM) approach. *International Journal of Control*, 82, 27–42. Available from: https://doi.org/10.1080/00207170801955420

Wei, H.L., Billings, S.A., Zhao, Y.F. & Guo, L.Z. (2010) An adaptive wavelet neural network for spatio-temporal system identification. *Neural Networks*, 23, 1286–1299. Available from: https://doi.org/10.1016/j.neunet.2010.07.006

Weisheimer, A., Decremer, D., Macleod, D., O'reilly, C., Stockdale, T.N., Johnson, S. et al. (2019) How confident are predictability estimates of the winter North Atlantic oscillation? *Quarterly Journal of the Royal Meteorological Society*, 145, 140–159. Available from: https://doi.org/10.1002/qj.3446

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Sun, Y., Simpson, I., Wei, H.-L., & Hanna, E. (2024). Probabilistic seasonal forecasts of North Atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models. *Meteorological Applications*, *31*(1), e2178. https://doi.org/10.1002/met.2178

14698080, 2024, 1, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.2178 by Test, Wiley Online Library on [16/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License