Received: 7 May 2025 | Revised: 28 August 2025 | Accepted: 1 September 2025

DOI: 10.1002/qj.70036

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Evaluating seasonal forecast improvements over the past two decades

Christopher H. O'Reilly[1] | **David MacLeod**[2] | **Daniel Befort**[3] | **Theodore G. Shepherd**[1] | **Antje Weisheimer**[4,5]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]School of Earth and Environmental Sciences, Cardiff University, Cardiff, UK

[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany

[4]National Centre for Atmospheric Science, Department of Physics, University of Oxford, Oxford, UK

[5]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

Correspondence
Christopher H. O'Reilly, Department of Meteorology, University of Reading, Reading, UK.
Email: c.h.oreilly@reading.ac.uk

Funding information
The Royal Society, Grant/Award Number: URF\R1\201230

Abstract

Seasonal forecasting systems have been operational for over two decades. Here we present a systematic analysis of the performance of operational seasonal forecasting models since their inception. We analyse seasonal forecasting systems from three major international operational centres that have produced and coordinated continuously on operational seasonal forecasts over the past 20 years. Due to the small sample size of available forecasts, it is difficult to draw meaningful conclusions using historical operational forecasts alone, therefore we focus primarily on available model hindcasts. Our analysis, which accounts for differences in ensemble size and period across the forecasting systems, demonstrates that there have been clear improvements in some regions through the different model eras. For both the boreal winter and summer hindcasts, there have been significant improvements in forecasting the tropical regions, which are concurrent with improvements in the skill of tropical sea-surface temperature (SST) forecasts. These improvements in the Tropics are associated with increased predictability of temperature and precipitation across various continental regions on seasonal timescales. For the extratropics, the picture is more mixed, with strong improvements only evident during the boreal winter season over the North Pacific and North America. The sources of improvement over the winter extratropics are found to be strongly related to improvements in tropical SST skill and related improvements in the strength of the El Niño/Southern Oscillation (ENSO) teleconnection to the Pacific/North America pattern (PNA). Improvements of seasonal forecast skill over the rest of the extratropics, such as over Eurasia, are generally absent or patchy in individual models. The improvements that are found are most pronounced in the newest era models and are broadly associated with improvements in atmospheric model resolution. These improvements in skill are also evident in representative multi-model ensembles that represent more closely how operational forecasts are used in practice.

KEYWORDS

climate modelling, ENSO, predictability, seasonal forecasts, teleconnections

1 | INTRODUCTION

Seasonal forecasting systems have now been used to issue operational forecasts routinely for more than 20 years (e.g., Anderson, 2006; Barnston *et al.*, 2003; Palmer *et al.*, 2004). These systems, similar to numerical weather prediction, are based on initialising coupled ocean–atmosphere General Circulation models (GCMs) using the observed state of the climate system and integrating forwards in time to make predictions of the upcoming season, typically with a lead time of 1–6 months. Over the past 20 years, it has been demonstrated that the skill of operational numerical weather prediction, for example, for seven-day forecasts, has increased steadily at a rate of around one day per decade, linked to improvements in the underlying modelling and observing systems (Bauer *et al.*, 2015). However, since their inception, systematic improvements in the skill of seasonal forecasting systems have been less clear, particularly in the extratropics, where observed skill levels have typically been much lower (e.g., Smith *et al.*, 2012).

Nevertheless, improvements in the performance of seasonal forecasting systems have been demonstrated. Some of the most notable progress has been made in the forecasting capabilities of the Tropics. The importance of the El Niño/Southern Oscillation (ENSO: e.g., McPhaden *et al.*, 2006) in controlling seasonal climate variability prompted the earliest development of seasonal ENSO forecasts, using intermediate-complexity coupled models (e.g., Zebiak & Cane, 1987). Subsequent seasonal forecasting systems have implemented fully coupled ocean–atmosphere GCMs, and these operational forecasts, beginning around the year 2000, have demonstrated substantial ENSO forecast skill (e.g., Barnston *et al.*, 2012). The developments in tropical forecasts have resulted in improved utility of seasonal forecasts for various applications in tropical regions (e.g., Arsenault *et al.*, 2020; Jain *et al.*, 2019; MacLeod *et al.*, 2023).

Seasonal forecast skill in the extratropics has proven to be more elusive. Some forecasting systems have demonstrated significant skill for some large-scale extratropical circulation indices. Substantial skill has been demonstrated in seasonal hindcasts for the large-scale circulation over the Pacific/North American sector (e.g., Johansson, 2007; Kim *et al.*, 2012). Some specific forecasting systems have also demonstrated hindcast skill for the North Atlantic Oscillation and Arctic Oscillation indices (e.g., Dunstone *et al.*, 2016; Scaife *et al.*, 2014; Stockdale *et al.*, 2015), though there appears to be substantial variation across different systems (e.g., Baker *et al.*, 2018, 2024). The aim of the present study is to examine how skill has improved systematically across operational forecasting systems over the past two decades, with a particular focus on the extratropical regions, which have historically proved to be challenging.

In this study we evaluate the performance of operational dynamical seasonal forecasts produced by three major WMO Global Producing Centres over the past 20 years. Our objective is to quantify systematically any improvements in forecast skill over the operational period, where these improvements are most substantial, and which aspects remain challenging for operational systems. We examine individual systems, but, in order to streamline the analysis and provide a picture of the overall trajectory of the forecast performance, we also compare groups of systems from different "eras". We extend this analysis to evaluate the performance of representative multi-model ensembles through this period, as this is the type of data that is commonly used in practical applications. Finally, we use the combined dataset of the forecast systems to examine potential sources of improved seasonal forecast skill.

2 | DATASETS AND METHODS

2.1 | Seasonal forecasting systems from the past two decades

Since the inception of operational seasonal forecasts over 20 years ago, there have been coordinated efforts to produce forecasts across multiple research and forecasting centres, beginning with the DEMETER project (Palmer *et al.*, 2004). Since then, the European Centre for Medium-range Weather Forecasts (ECMWF), along with the UK Met Office (UKMO) and Meteo-France (MF), the respective national weather centres of the UK and France, have all developed systems consistently, which have been included in various coordinated multi-model efforts. Other international centres have also developed seasonal forecasting systems, but ECMWF, Meteo-France, and the UK Met Office are the only centres that have contributed continuously to multi-system products over the past two decades, so we focus here on these centres to analyse the development of seasonal forecast performance. We analyse seasonal forecast datasets from each system from these centres that have either been used operationally or have contributed to coordinated multi-model project/frameworks (all of which are archived in the MARS database at ECMWF). An overview of the systems analysed in this study, including details of the hindcast/operational period and ensemble size of each system, is shown in Figure 1.

Following the initial DEMETER project, the European Union (EU) funded ENSEMBLES project focused on developing coordinated initialised ensemble climate

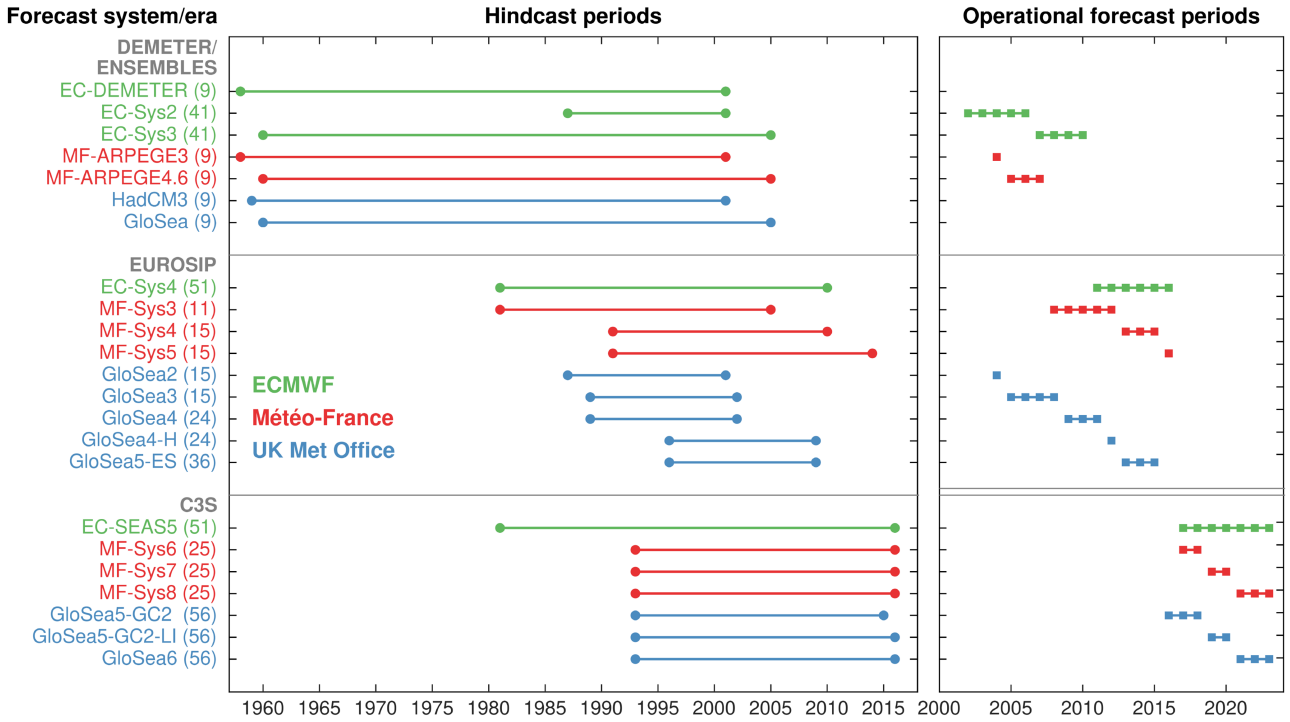

FIGURE 1 Overview of seasonal forecasting systems from ECMWF, Meteo-France, and the UK Met Office. The numbers in parentheses indicate the number of ensemble members in the hindcasts for each of the models. [Colour figure can be viewed at wileyonlinelibrary.com]

predictions further, as well as testing various approaches to initialisation, ensemble generation, and implementing model physics (van der Linden, 2009). From around the mid-2000s, the three operational centres coordinated their operational forecasts through the wider multi-system EUROSIP project, which issued operational seasonal forecasts until around 2017. Since 2018, the three operational centres have all contributed to the multi-system seasonal forecasts issued through the EU's Copernicus Climate Change Service (C3S). To streamline some of the analysis below, the seasonal forecast systems over this period have been grouped into three different model "eras", based on which project they contributed to and the associated hindcast period: referred to as "C3S", "EUROSIP", and "DEM/ENS" hereafter. It should be noted that, for many of the systems, there is not necessarily a clear separation between subsequent versions of the systems (specific system details, including model components and resolutions, are provided in Table S1).

Whilst we will analyse briefly the operational forecasts made over the past 20 years, the bulk of our analysis is of model hindcasts. We will focus our analysis on the seasonal forecasts of the boreal winter (DJF) and boreal summer (JJA) seasons. All of the hindcasts were run for six months, initialised on or before November 1 (for the DJF boreal winter forecast) and May 1 (for the JJA boreal summer forecast). We focus on these initialisation dates, as these are available for all of the forecasting systems.

2.2 | Observational reference datasets

We analyse the forecast skill of geopotential height at 500 hPa ($Z500$) and sea-level pressure (SLP), as these are useful indicators of the large-scale circulation skill; we also analyse the skill for sea-surface temperature (SST), 2-m temperature (T_{2m}), and precipitation. For verification we use the ECMWF Reanalysis v5 (ERA5) dataset (Hersbach *et al.*, 2020) as the observational reference for SLP, $Z500$, and T_{2m}. For SST, we use HadISST as the observational reference dataset (Rayner *et al.*, 2003). For precipitation we use the Global Precipitation Climatology Project (GPCP) gridded dataset, which provide integrated monthly precipitation from a range of satellites over the ocean and gauge-based data over the land (Adler *et al.*, 2003). These datasets all cover the full hindcast periods for all models, with the exception of the GPCP precipitation, which is only available from 1979 onwards due to the reliance on satellite data. In the precipitation analysis below, we truncate the hindcasts to use only periods where the GPCP data are available.

2.3 | Regional indices

In addition to the area-averaged and grid-point skill metrics outlined above, we also consider some common regional circulation indices that have been shown to be

important in modulating seasonal climate in different regions.

- Niño-3.4: a common index that measures the phase of the ENSO phenomenon (e.g., Trenberth, 1997), defined as the SST anomaly averaged over 170°–120°W, 5°S–5°N in the Tropical Pacific.
- Pacific/North America pattern (PNA): a common index that measures the strength of a dominant mode of large-scale circulation variability over the eastern extratropical North Pacific and North America during winter, defined as

$$\begin{aligned} PNA \;=\; & 0.25 \times [Z'(20°\mathrm{N}, 160°\mathrm{W}) - Z'(45°\mathrm{N}, 165°\mathrm{W}) \\ & + Z'(55°\mathrm{N}, 115°\mathrm{W}) - Z'(30°\mathrm{N}, 85°\mathrm{W})], \end{aligned}$$

  where Z' is the normalised 500-hPa geopotential height anomaly (following Wallace & Gutzler, 1981).
- North Atlantic Oscillation (NAO): a measure of the state of the large-scale circulation over the Euro-Atlantic sector during the boreal winter season that exhibits a strong control over Eurasian seasonal climate variability, defined as the normalised difference between the SLP anomaly over Iceland and the Azores (e.g., Jones *et al.*, 2003).

These indices provide a useful shorthand for the performance of forecasts during some key large-scale phenomena. The results presented below are not sensitive to the specific definitions of these indices, however, and the conclusions drawn from these indices are consistent with mapped distributions of relative skill improvements (as will be shown below).

2.4 | Skill metrics

We analyse and compare the skill of the systems by focusing primarily on the ensemble mean signals. We focus on the ensemble mean, because many of the systems have limited ensemble sizes and analysing the ensemble mean is the most straightforward approach to comparing multiple systems.

In this study we focus largely on temporal correlation metrics, based on the Pearson correlation coefficient, r. The explained variance, or r^2, is defined as follows:

$$\text{Exp. Var. } (\%) = \begin{cases} 100 \times r^2 & \text{for } r \geq 0, \\ 0 & \text{for } r < 0. \end{cases} \tag{1}$$

This definition ensures that negative correlation skill is not identified spuriously as skilful performance.

The correlation is also used to define a large-scale metric of skill that we will analyse as the *total explained variance* (*TEV*: e.g., O'Reilly *et al.*, 2020), which is defined here as

$$TEV(\%) = 100 \times \frac{\int\int_A r^2 Z'^2 \, dA}{\int\int_A Z'^2 \, dA}. \tag{2}$$

Here, r^2 is the ensemble mean correlation skill squared at each grid point (equal to zero where $r < 0$) and Z'^2 is the observed interannual variance of variable Z at each grid point, such that this is the area-averaged skill weighted by the underlying observed variability (to avoid over-weighting regions with relatively little variance). We calculate this for extratropical (30°–90°N) and tropical (30°S–30°N) regions for SST, SLP, and $Z500$. In addition, we analyse grid-point comparisons of the ensemble mean correlation skill between different models.

Alongside the temporal correlation metrics, we also use pattern correlation when examining the historical operational forecasts in Section 3.1. We use a pattern-correlation skill (%) metric that is defined as $100 \times r|r|$, where here r is the Pearson correlation skill between area-weighted grid-point anomalies over a given region/-variable; this formulation allows for the identification of particularly poor forecast years, which can be negative.

The focus on correlation metrics in this study is also in part to account for the weak ensemble mean signals in seasonal forecast models, most notably in the extratropics, which have been documented in several recent studies (e.g., Hardiman *et al.*, 2022; O'Reilly *et al.*, 2019a; Scaife & Smith, 2018; Weisheimer *et al.*, 2024). Whilst these apparent signal-to-noise errors are important and interesting in their own right, here we focus on understanding the development of skilful ensemble mean signals in the seasonal forecasts. However, it is clear that relative improvements in the statistical *reliability* of the forecasts (e.g., Jolliffe & Stephenson, 2012; MacLeod *et al.*, 2018) is also important and is left as a topic for future investigation.

2.5 | Model comparison and uncertainty estimates

Comparison of the skill of the models is not straightforward, due to differences in ensemble size and hindcast period (Figure 1). Specifically, this is because ensemble mean skill tends to increase with increasing ensemble size (e.g., Murphy, 1988; Scaife *et al.*, 2014) and different years and periods demonstrate substantial differences in their inherent predictability (e.g., Baker *et al.*, 2024; O'Reilly *et al.*, 2019b; Weisheimer *et al.*, 2017). To compare fairly across a pair of forecasting systems, we select the longest common period that overlaps across the hindcasts

and discard years that only exist for one of the systems. To account for different ensemble sizes, we randomly subsample (without replacement) a fixed number of members from each model, with that number being equal to two-thirds of the size of the smaller ensemble (where ensemble sizes are different); the minimum for this is six members and the maximum is 37 members (i.e., Figure 1). We tested the sensitivity to this sampling method by repeating the analysis for a bootstrap resampling with replacement and the results and conclusions that follow are not qualitatively sensitive to this choice. We can estimate the relative impact of different hindcast periods or ensemble sizes by relaxing these constraints individually—we include an example of this below—but most of our analysis is based on equal ensemble sizes and hindcast periods. The ensemble subsampling is repeated 10,000 times and the relevant skill measures are calculated and compared for each of these subsamples.

In some of the analysis below, we report the median of these sampled model difference distributions. The uncertainty in the differences between each pair of models—and thereby, evidence that one model is more skilful than another—is estimated by calculating Bayes factors, which provide a measure of the evidence that eschews some of the common pitfalls of null-hypothesis significance testing (e.g., Ambaum, 2010; Shepherd, 2021). The Bayes factor, B, is defined here as

$$B = \frac{p(D|\text{Model A})}{p(D|\text{Model B})}, \tag{3}$$

which is the ratio of the probabilities, p, that each of the two models (i.e., Model A or Model B) is better, based on the hindcast and observational data (i.e., D). These probabilities are estimated from the 10,000 ensemble subsamples used to calculate each model pair. In the results, below we present the Bayes factors as positive-definite and indicate which model the evidence favours in each case.

We use a similar bootstrap approach to calculate Bayes factors in the meta-analysis in which we compare the differences between model eras; in this case, the resampling is performed 10,000 times with replacement over different model pairs within the model eras.

3 | RESULTS

3.1 | Historical operational seasonal forecasts

In previous studies, the improvement of numerical weather prediction (NWP) has been demonstrated by analysing historical operational forecasts (e.g., Bauer *et al.*, 2015). For seasonal forecasts, analysing the historical operational forecasts in a meaningful way is challenging, due to the small sample size and limited common forecast period between the different operational forecasts. For example, for NWP there are typically 365 initialised forecasts per year, whereas for seasonal forecasts seasonality is critical, such that we effectively have a single forecast/observation pair per year. Nonetheless, we will begin by briefly outlining attempts to estimate improvements in the operational skill for seasonal forecasts, similar to the approach more commonly used to analyse operational NWP forecasts.

The extratropical $Z500$ anomaly pattern correlation from the multi-system operational seasonal forecasts for boreal winter (DJF, initialised November 1) is shown in Figure 2a. The year-to-year variability in the pattern correlation is high and there is no clear increase over time, unlike that seen for NWP, although there is a hint of an upward trend. When comparing the first and second decades more formally, there is some indication that the $Z500$ pattern-correlation skill has increased in the most recent decade (Figure 2b), but this is not clearly reflected in the SLP pattern correlation or in measures of the total explained variance (Figure 2b).

There are several major reasons to be cautious in drawing conclusions from the operational forecasts in Figure 2. Firstly, there are very few independent operational forecasts to analyse and the sample size is very small. Secondly, the $Z500$ pattern correlation is not commonly used to analyse seasonal predictability, likely due to several drawbacks. Firstly, there are limited spatial degrees of freedom and the lower latitudes tend to dominate, even though $Z500$ may be less relevant for surface anomalies at these latitudes. Secondly, the anomalies are, necessarily, based on an eight-year climatological hindcast period (1993–2001) and further tests indicate that the pattern correlation results exhibit a clear sensitivity to this short period (not shown). Thirdly, the underlying predictability is known to be non-stationary, varying from year to year as well as exhibiting trends on multi-decadal timescales, even in the same forecasting system (e.g., O'Reilly *et al.*, 2017; Weisheimer *et al.*, 2017). Overall, we observe some apparent improvements in correlation over the period (see Figure 2b), which may reflect incremental improvements in the ability of operational forecasts to represent key processes driving variability. However, we cannot rule out the possibility that this increase in correlation is due simply to underlying changes in the predictability of the climate system. Therefore, on its own, this does not provide robust evidence of improved model skill. To assess this more rigorously, we must analyse the impact of model changes while controlling for potential shifts in underlying predictability.

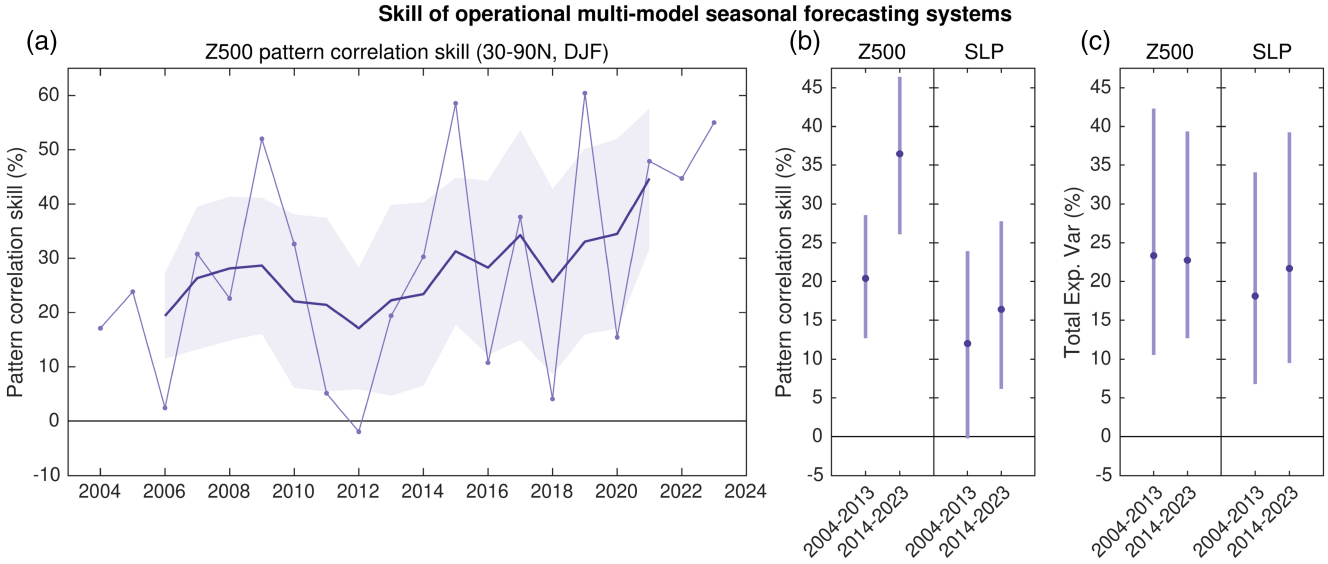

FIGURE 2 Skill of operational multi-model seasonal forecasts for boreal winter (DJF), forecast from November 1. (a) Pattern correlation of the extratropical $Z500$ anomaly between the observations and the multi-model ensemble mean $Z500$ anomaly forecast. (b) Decadal averages of the pattern correlation over the northern extratropics for 2004–2013 and 2014–2023 for both $Z500$ and SLP. (c) As in (b) but for the total explained variance skill. The thick solid line and shading in (a) show the five-year running mean pattern correlation and ± 1 standard error around this mean. The error bars in (b) and (c) show the 90% confidence interval of the relevant estimates based on a bootstrap resampling over years performed 10,000 times. [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Comparison of historical model hindcasts

Given the limitations in analysing the historical operational seasonal forecasts, we instead focus on comparing the performance of the *hindcasts* from the different models, which are available for some of the same years (i.e., Figure 1). A major challenge in comparing the different forecasting systems is that the hindcasts cover a variety of different periods, with only a limited six-year common period, 1996–2001. An example plot of hindcast data for this common period is shown in Figure 3 for the ensemble mean Niño-3.4 index of the hindcasts (these are not bias-corrected due to the very short common period). The diversity in behaviour from model to model is evident even over this short period. Qualitatively, it is clear that there is skill in predicting ENSO at these lead times in all the models; the errors of the raw hindcast output with respect to the observational data seem to be much reduced in the C3S systems compared with the systems in earlier eras. In particular, the systematically cold predictions apparent in these earlier eras seem to be much reduced in the C3S models (perhaps with the exception of some of the GloSea systems), though there is still some drift evident, consistent with the behaviour highlighted in recent studies (Beverley *et al.*, 2023).

Whilst these initial comparisons may be interesting, this short common period is inadequate to compare the hindcasts of the different systems. However, when comparing any two individual models, there are many more hindcast years than the overall six-year overlap (i.e., Figure 1), such that we can make a more meaningful comparison between different pairs of models by utilising as many years of hindcast data as possible. As well as controlling for the hindcast period, we also subsample the ensemble to control for ensemble size (since increased ensemble sizes tend to increase ensemble mean forecast skill).

3.2.1 | Boreal winter seasonal hindcasts

A plot comparing all the different seasonal forecasting systems with one another, in terms of $Z500$ total explained variance over the northern extratropics, is shown for the boreal winter hindcasts in Figure 4. The plot shows the median difference in total explained variance calculated between each pair of models (following the sampling method outlined in Section 2.5). Scanning from left to right across the plot, for the C3S systems (towards the bottom), it is clear that they generally show an improvement over the systems from previous eras, with the largest improvements seen in comparison with the EUROSIP era hindcasts. However, there are also instances where there is less improvement and even decreased skill. It is important to highlight here that there is considerable uncertainty in the differences between many of the pairs—indicated by the size of the circles, which represent the Bayes factor of

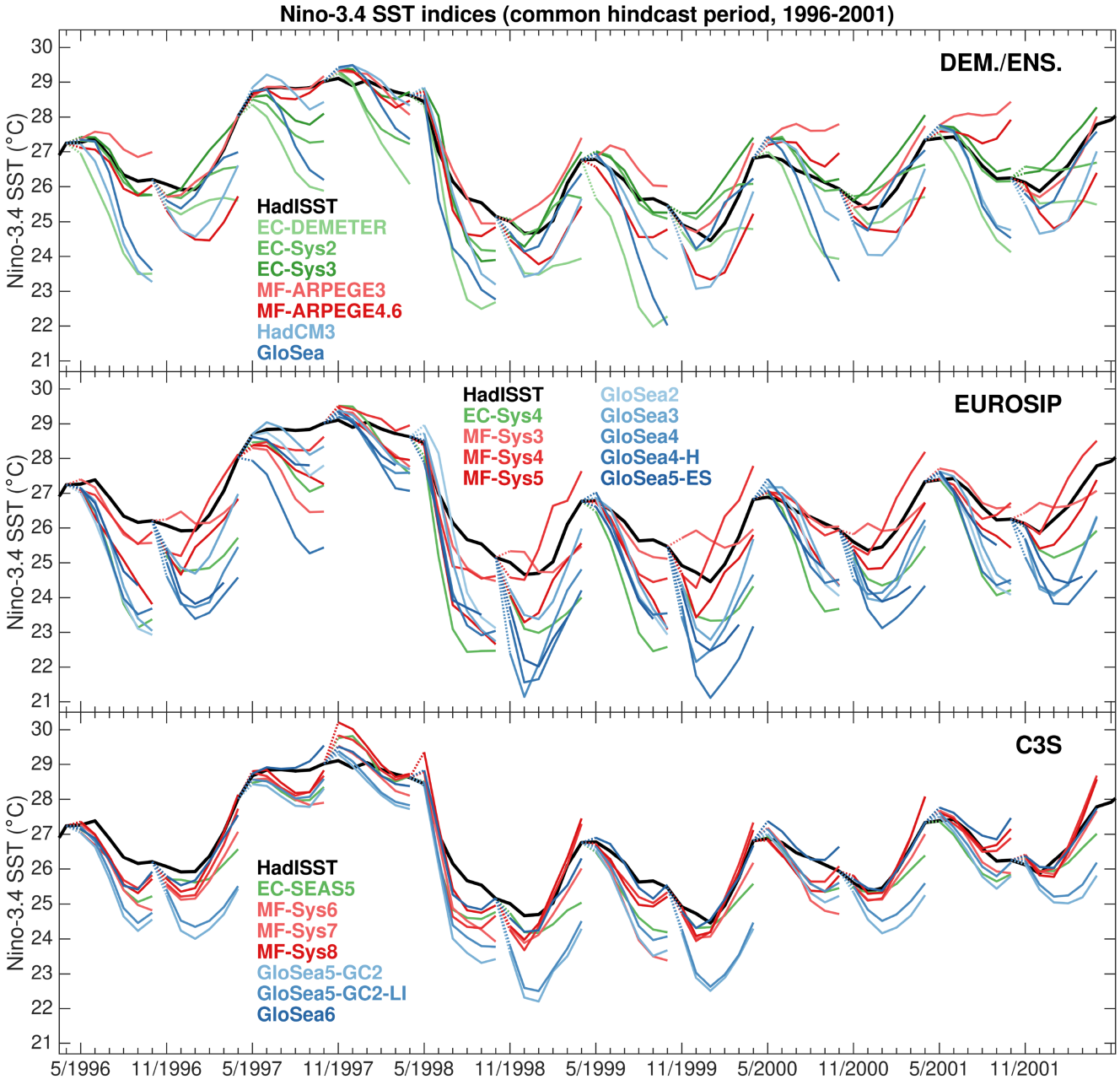

FIGURE 3 Ensemble mean monthly Niño-3.4 indices from all hindcasts over the common period, along with the HadISST observational dataset (in black). [Colour figure can be viewed at wileyonlinelibrary.com]

the pair-wise comparisons. For many of the model pairs, the Bayes factor (estimated from the ensemble subsampling) is not much greater than one, indicating only modest evidence to favour one model over another. However, in other instances–primarily where the skill differences are larger–the Bayes factors are 10 or larger, providing strong evidence that one model is more skilful than another. Overall, this plot indicates that there has been substantial improvement for the wintertime seasonal forecasts of $Z500$ anomalies in the northern extratropics in the C3S era compared with older systems.

We can formalise the comparison between the different model eras by plotting the pairwise difference between the relevant models, which is shown in Figure 5. We focus initially on the comparison for equal ensemble size and hindcast period (highlighted in red text). When comparing across these eras, for the extratropical $Z500$ total explained variance in the Northern Hemisphere winter, we find strong evidence that the C3S era models are more skilful overall than the EUROSIP and DEM/ENS era models. Interestingly, there is relatively little difference between the EUROSIP and DEM/ENS era models. There

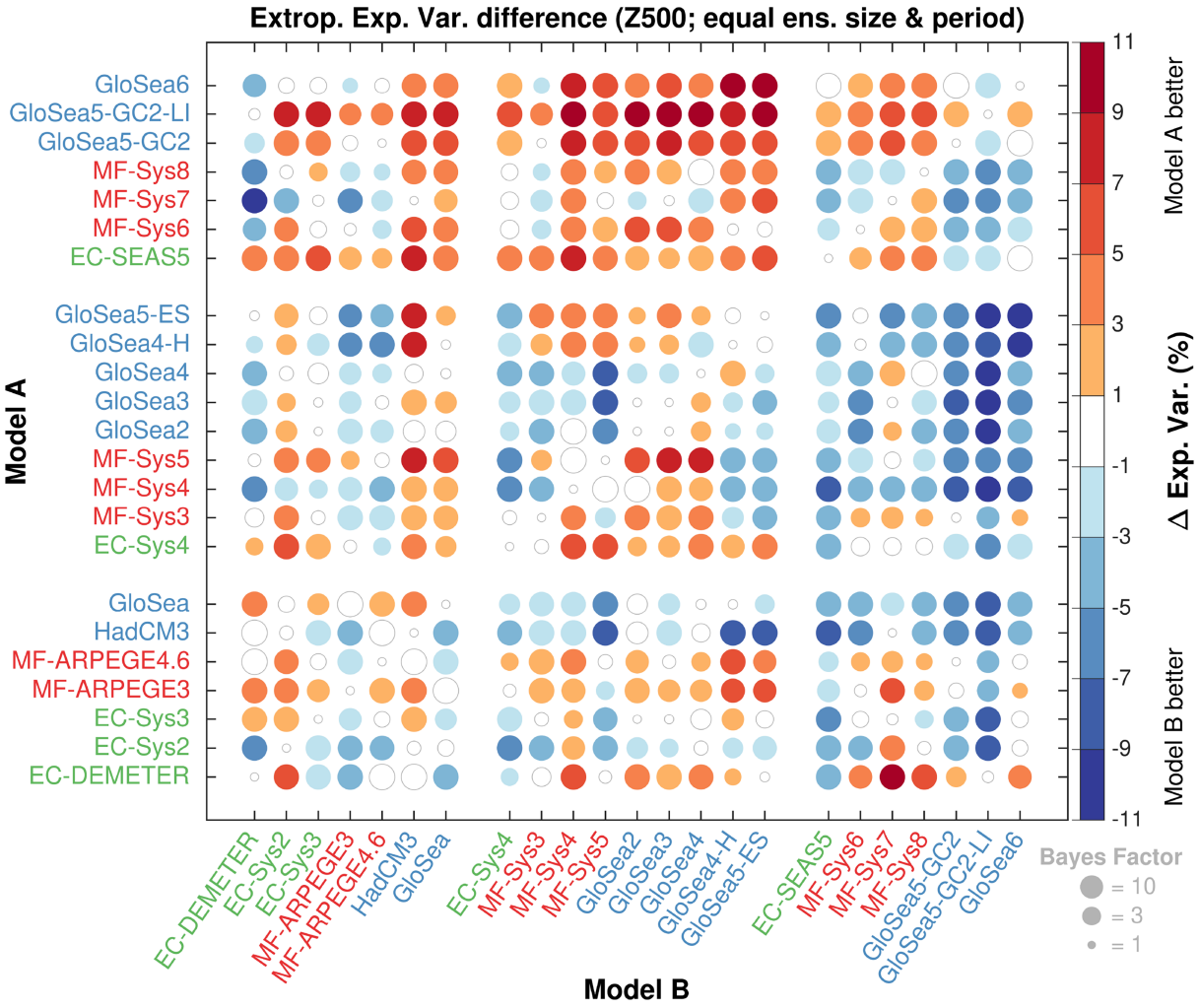

FIGURE 4 Example of inter-model skill comparisons, shown for extratropical Northern Hemisphere *Z*500 for boreal winter hindcasts (DJF). The comparison between each pair of models was made whilst controlling for the same ensemble size and hindcast period (see Methods, Section 2.3). The colour of the dot indicates the difference between the pair of medians, calculated from the subsampled distributions. The size of the dots varies continuously and corresponds to the Bayes factor of the difference between each pair of models (a legend showing some example sizes is shown in the lower right corner of the figure); the Bayes factor saturates at 10 for the largest dots, though many pairs are substantially greater than 10. [Colour figure can be viewed at wileyonlinelibrary.com]

is not a clear step-change in model components/resolution between all the EUROSIP and DEM/ENS era models (see Table S1), which might explain why there is relatively little difference between the models. We will explore some reasons for differences in skill in a little more detail below, in Section 3.4.

In addition to regularising the model comparison by analysing equal ensemble size and periods, we also relaxed these constraints to examine the sensitivity to these constraints, and examples from the extratropical *Z*500 hindcasts are shown in Figure 5. The biggest differences are found between the DEM/ENS era models and the more recent models. The DEM/ENS era models typically have lower ensemble sizes in the hindcasts, but also they were often run starting from the 1960s (i.e., Figure 1), indicating from this analysis that this early period has lower skill. The mid-century period has generally been found to be less skilful for winter forecasts, at least partly related to the lower ENSO amplitude during this period (e.g., O'Reilly *et al.*, 2017; Palmer *et al.*, 2004; Weisheimer *et al.*, 2017, 2020). The analysis here emphasises the need to use equal ensemble sizes and hindcast periods when analysing the relative skill between different models.

Comparisons of the total explained variance skill between different model eras for boreal winter hindcasts (as in Figure 5) are shown in Figure 6a for SLP and SST, for both the Tropics and extratropics. For SLP, on average, there have been improvements through the different model eras in both the Tropics and extratropics, with the clearest improvements coming in the C3S era. For SSTs,

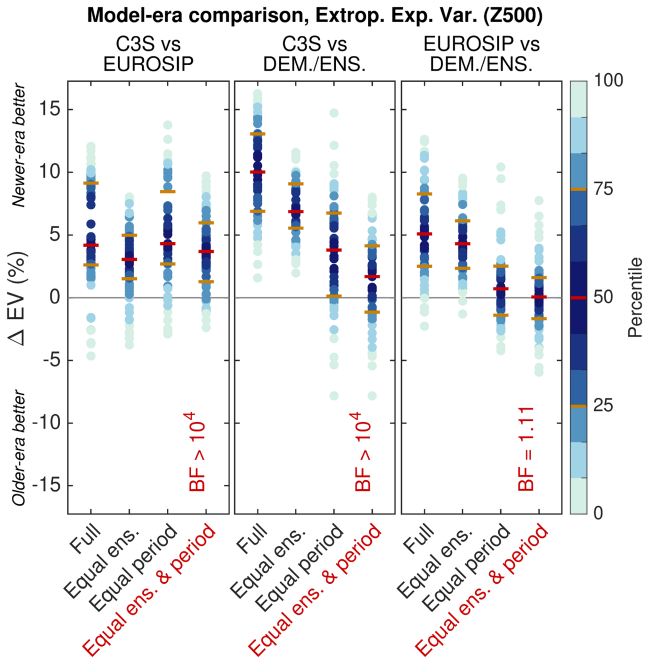

FIGURE 5 Model-era skill comparison for extratropical Northern Hemisphere $Z500$ anomalies for boreal winter hindcasts (DJF). The comparisons are shown for the full ensembles, as well as when controlling for equal ensemble size, equal hindcast period, and both. Each dot shows the difference between one pair of models from the different eras. The colours are shaded to show the percentile of the whole distribution and the horizontal lines show the 25th, 50th, and 75th percentiles. Also shown are the Bayes factors calculated for the median of the equal ensemble size and hindcast period distributions, which were estimated by repeated subsampling (see Section 2). [Colour figure can be viewed at wileyonlinelibrary.com]

the picture is more mixed. Whilst the C3S era models are generally better than older models, there are no clear improvements compared with the oldest DEM/ENS era models. This lack of improvement is perhaps counterintuitive, though there are a few potential reasons. The first is that there are generally a small number of years and ensemble members in the comparisons of the C3S and DEM/ENS era models, with only 13 years and nine members to be subsampled in most instances (i.e., Figure 1), so individual years and members carry substantially more weight in these. With this caveat in mind, another potential reason for this degradation in SST skill is that the initialisation of ocean models has changed substantially over the different eras, with an increase in resolution to $0.25° \times 0.25°$ in the ocean model components of almost all of the C3S models (i.e., Table S1). This increase in resolution has lead to initialisation problems in some regions affecting the ocean forecast skill (Johnson *et al.*, 2019).

In addition to the integrated measures of forecast skill, it is also of interest to analyse the skill of some important ocean and regional circulation indices. Figure 5b shows comparisons of the difference in skill between different model eras for the Niño-3.4, PNA, and NAO indices for the boreal winter seasonal hindcasts. The C3S era models demonstrate a small but robust increase in hindcast correlation skill for the Niño-3.4 SST index compared with the older era models. The small increase is perhaps not surprising, given the Niño-3.4 correlation skill is very high across all models. The improved correlation skill is also associated with a general improvement in the evolution of ENSO events, evident in the lower model drifts and lower absolute errors seen in the Niño-3.4 plumes for all models during the common period, 1996–2001 (Figure 3). For the PNA index, there has been a very strong and consistent increase between the different model eras. The NAO index, however, shows only a very modest, albeit robust, increase in the C3S era compared with EUROSIP, though the skill levels remain low in most models; the exception is GloSea5-GC2-LI, which exhibits higher levels of skill compared with most models (as demonstrated in Scaife *et al.*, 2014). It is remarkable, however, that the model skill for the NAO is not substantially better than in the DEM/ENS era when controlling for ensemble size and hindcast period. It is important to note that the relatively low skill of the NAO overall makes the comparisons in skill particularly uncertain, especially when considering the small number of crossover hindcast years between some of the C3S and DEM/ENS era models.

To explore the geographical distributions of model performance, we now examine the differences in ensemble mean correlation skill at a grid-point level between the different model eras; these are shown for the boreal winter hindcasts of $Z500$ and SST in Figure 7. For $Z500$, there is an increase in skill between C3S and the older eras over most of the Tropics and northern extratropics. The largest increases in skill are seen over the extratropical Pacific and North American continental region. This is consistent with the large increase in PNA skill seen in the C3S era models compared with the older era models seen in Figure 6b and demonstrates that this is not sensitive to the precise definition of the index. There are also modest increases in $Z500$ skill over the extratropical North Atlantic in C3S era models compared with the EUROSIP era. However, the C3S era models on average actually have lower skill over the extratropical North Atlantic compared with the older DEM/ENS era models, though the same caveats regarding the small number of ensemble members and common hindcast years apply. There are also some notable improvements in the Southern Hemisphere for the $Z500$ DJF seasonal forecasts, such as parts of Australasia, South America, and Sub-Saharan Africa. For the hindcasts of DJF SST, the C3S era models show most improvement over the tropical oceans, especially compared with the

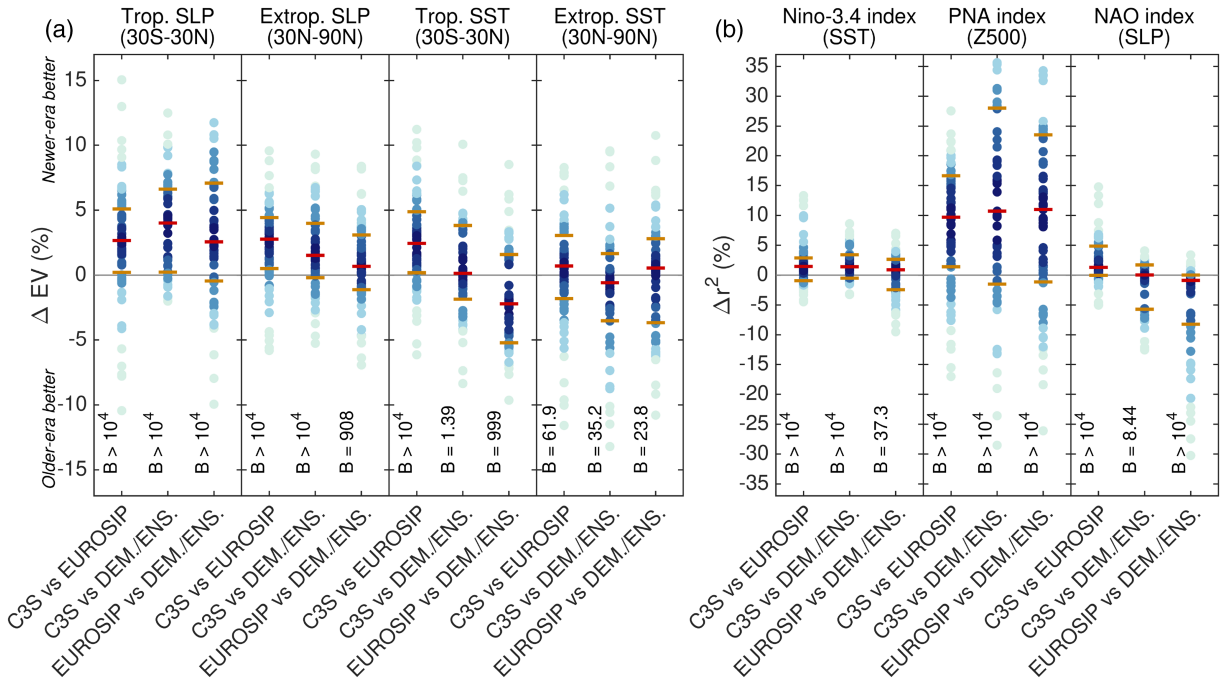

FIGURE 6 As Figure 5, but for (a) different variables/regions and (b) some key regional indices for boreal winter hindcasts (DJF). [Colour figure can be viewed at wileyonlinelibrary.com]

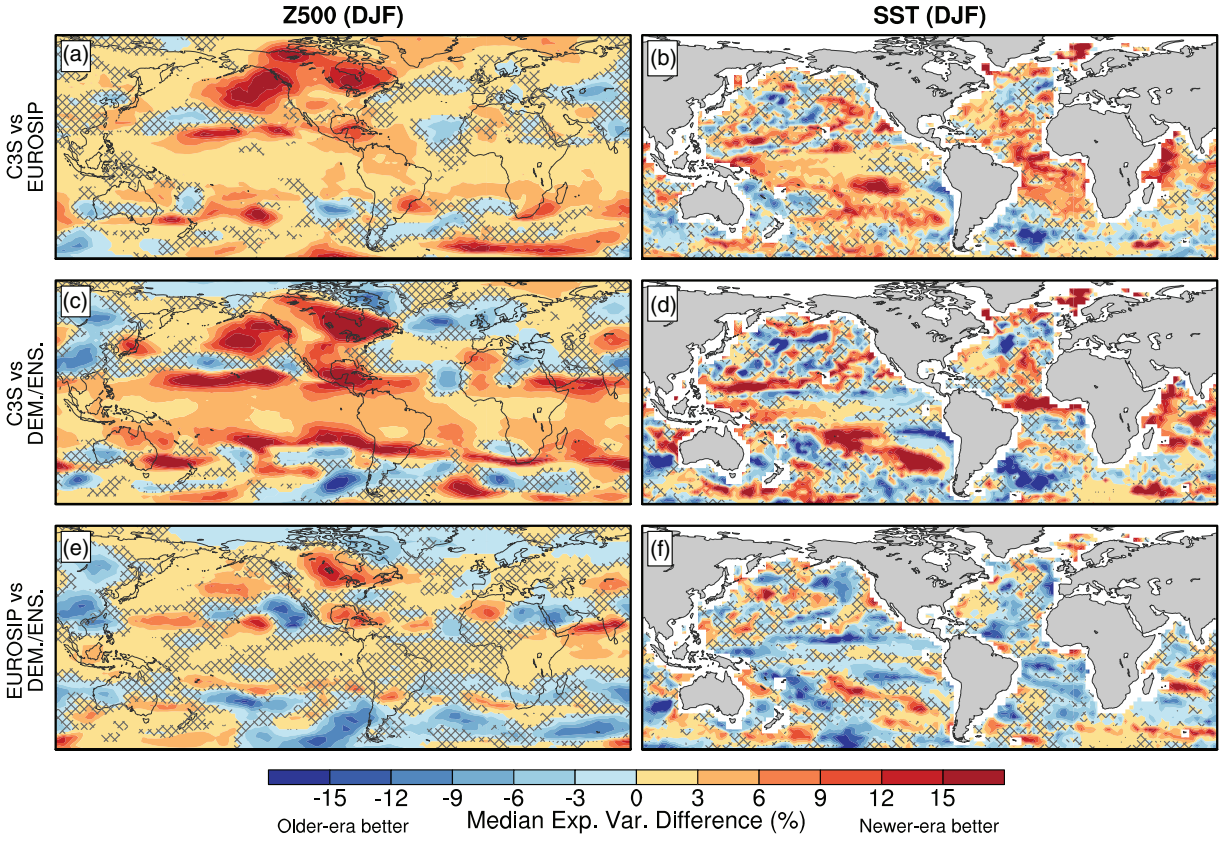

FIGURE 7 Maps of model-era skill comparison for $Z500$ and SST for boreal winter hindcasts (DJF). At each grid point, the colour shows the median difference in explained variance of the ensemble mean hindcast correlation skill (i.e., Δr^2) over the distribution of model era comparisons (similar to the approach in Figures 5 and 6). These were calculated, as before, after controlling for the same forecast period and ensemble size. The hatching indicates grid points where the Bayes factor is less than 100, when comparing between model eras. As before, the Bayes factors were calculated for the median of the distributions by repeated subsampling (see Section 2). [Colour figure can be viewed at wileyonlinelibrary.com]

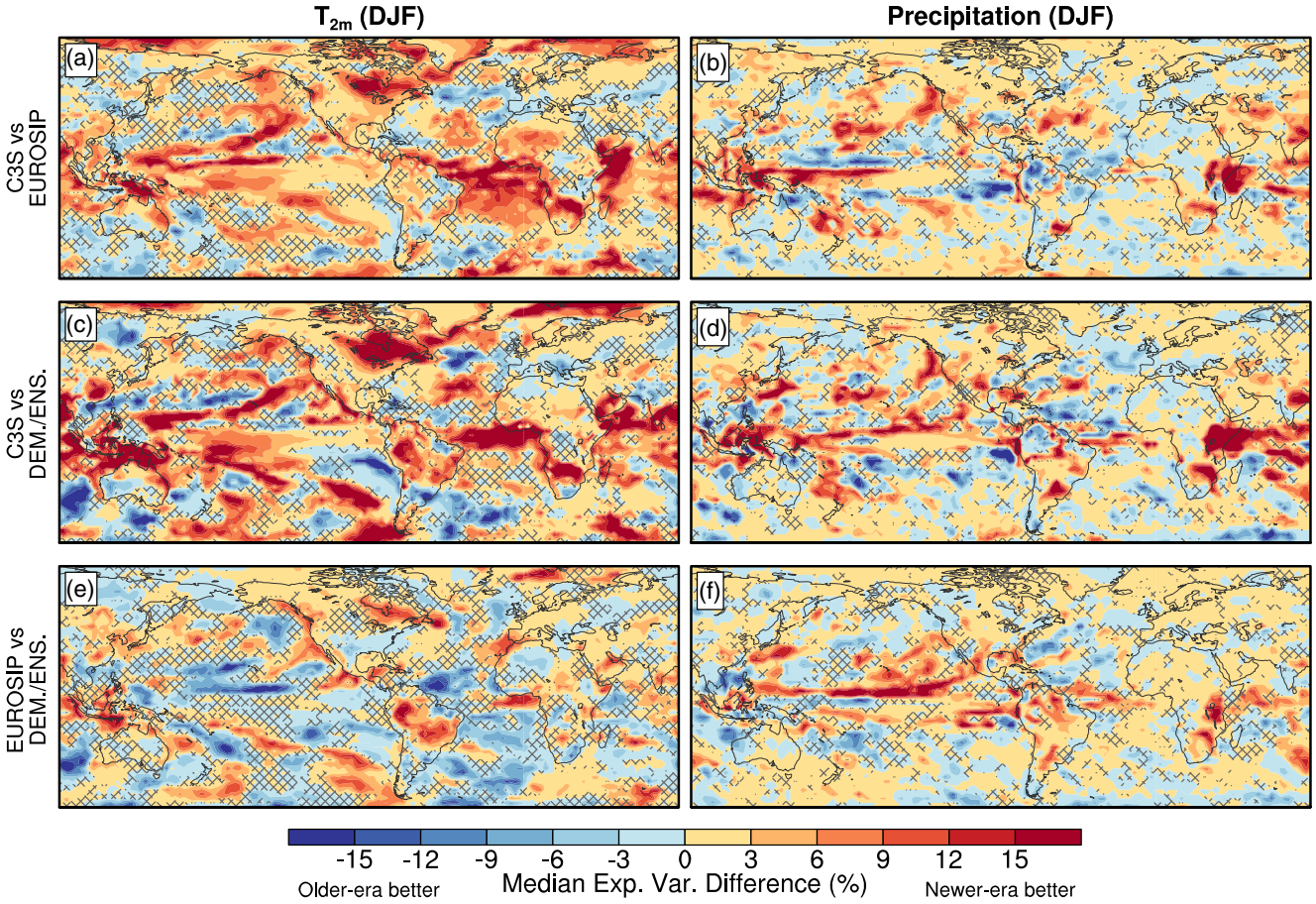

FIGURE 8 As Figure 7, but for T_{2m} and precipitation for boreal winter hindcasts (DJF). [Colour figure can be viewed at wileyonlinelibrary.com]

EUROSIP era models. A particularly notable region of improvement in the C3S models is the Tropical Atlantic, but there are also some modest improvements in correlation skill over the Tropical Pacific and Indian Ocean. From the SST maps, however, it is evident that the improvements in SST hindcast skill have not been very consistent through the model eras, and certainly less consistent than those seen for $Z500$. Some of this may be due to the relatively good underlying skill in the older models, such that there is less scope for SST correlation skill to improve on 2–4 month lead times. Nonetheless, it is still notable that there is lower skill over much of the extratropical North Atlantic in the C3S era models, which is likely related to the aforementioned initialisation issues in higher resolution ocean models (e.g., Johnson *et al.*, 2019).

In addition to $Z500$ and SST, we have also examined the grid-point level correlation skill of T_{2m} and precipitation, shown in Figure 8. Over the ocean, the T_{2m} maps are quite similar to the SST maps (i.e., Figure 7), as one might expect. Over land, there are substantial improvements in T_{2m} hindcast skill in the C3S era models compared with the older eras, particularly over parts of North America, South America, and Africa. Over North America, the areas of improvement are likely related to the strong improvements in ensemble mean skill seen in the large-scale circulation, evident in terms of both the PNA index (Figure 6) and $Z500$ anomalies (Figure 7). Some of the improvements have been reasonably consistent through the model eras, with the EUROSIP era models also demonstrating improvements over the earlier DEM/ENS era models, though not as large as the improvements seen for the C3S era.

There are also some notable improvements in the precipitation skill through the model eras (Figure 8). In the Tropics, there have been consistent improvements in the Maritime continent region, the Tropical Pacific, and over East Africa. These may be related to the relatively skilful ENSO teleconnection to these regions that has been demonstrated in C3S era models (e.g., Macleod, 2019; Macleod *et al.*, 2021). In the northern extratropics, however, there are very small differences between the model eras, with no clear improvements in most regions. The exception is perhaps North America, where there have been some very modest improvements in precipitation skill through the different model eras, likely due to the

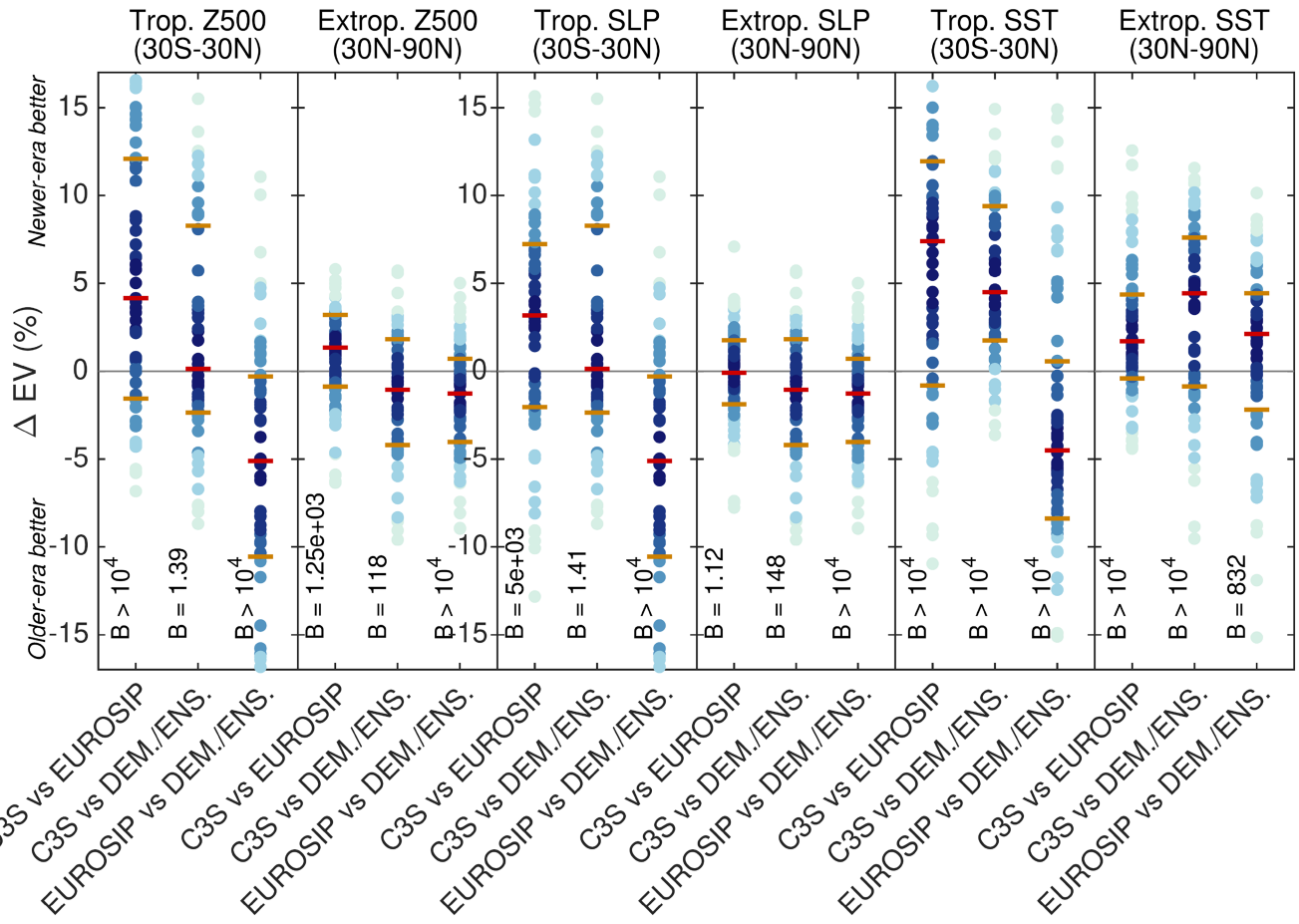

FIGURE 9 As Figure 6a, but for regions/variables in the boreal summer hindcasts (JJA). [Colour figure can be viewed at wileyonlinelibrary.com]

improvements in the large-scale circulation skill over this region. The absence of stronger improvement in precipitation skill over North America, despite the large-scale circulation skill, may be linked to the signal-to-noise errors seen over the North Pacific for the more skilful C3S models (Williams *et al.*, 2023). These signal-to-noise issues imply that greater skill improvements are possible for climate impact variables, such as T_{2m} and precipitation, if this problem can be understood better and eradicated; this is an area of active ongoing research (e.g., O'Reilly, 2025; Weisheimer *et al.*, 2024).

3.2.2 | Boreal summer seasonal hindcasts

We now move to the equivalent analysis of improvements in the boreal summer seasonal hindcasts. Comparisons of the total explained variance skill between different model eras for boreal summer hindcasts (similar to Figures 5 and 6) are shown in Figure 9 for SLP, $Z500$, and SST, for both the Tropics and extratropics. In the Tropics, there have been improvements in SLP and $Z500$ in the C3S era compared with the EUROSIP era models, but compared with the older ENS/DEM era models there are no evident improvements. Some of the modest improvement seen in the Tropics may be related to the strong improvement in tropical SST skill seen in the C3S era models over the EUROSIP era models, and there is also a clear improvement over the DEM/ENS era models. There are no consistent improvements in $Z500$ and SLP over the extratropics in the boreal summer hindcasts; the absence of appreciable improvement in the extratropical large-scale circulation is also evident in the maps of grid-point level correlation skill of $Z500$ (Figure S1). One reason for the lack of obvious improvement in the extratropics in the boreal summer hindcasts is that the overall levels of seasonal hindcast skill are typically much lower. An example of this is the systematically lower levels of total explained variance in $Z500$ in the boreal summer hindcasts, shown for the systems in this study in Figure S1. These lower levels of skill are due, at least in part, to the lower amount of tropically forced circulation variability during boreal summer (e.g., ENSO amplitude peaks during DJF), and the fact that models generally struggle to capture these extratropical

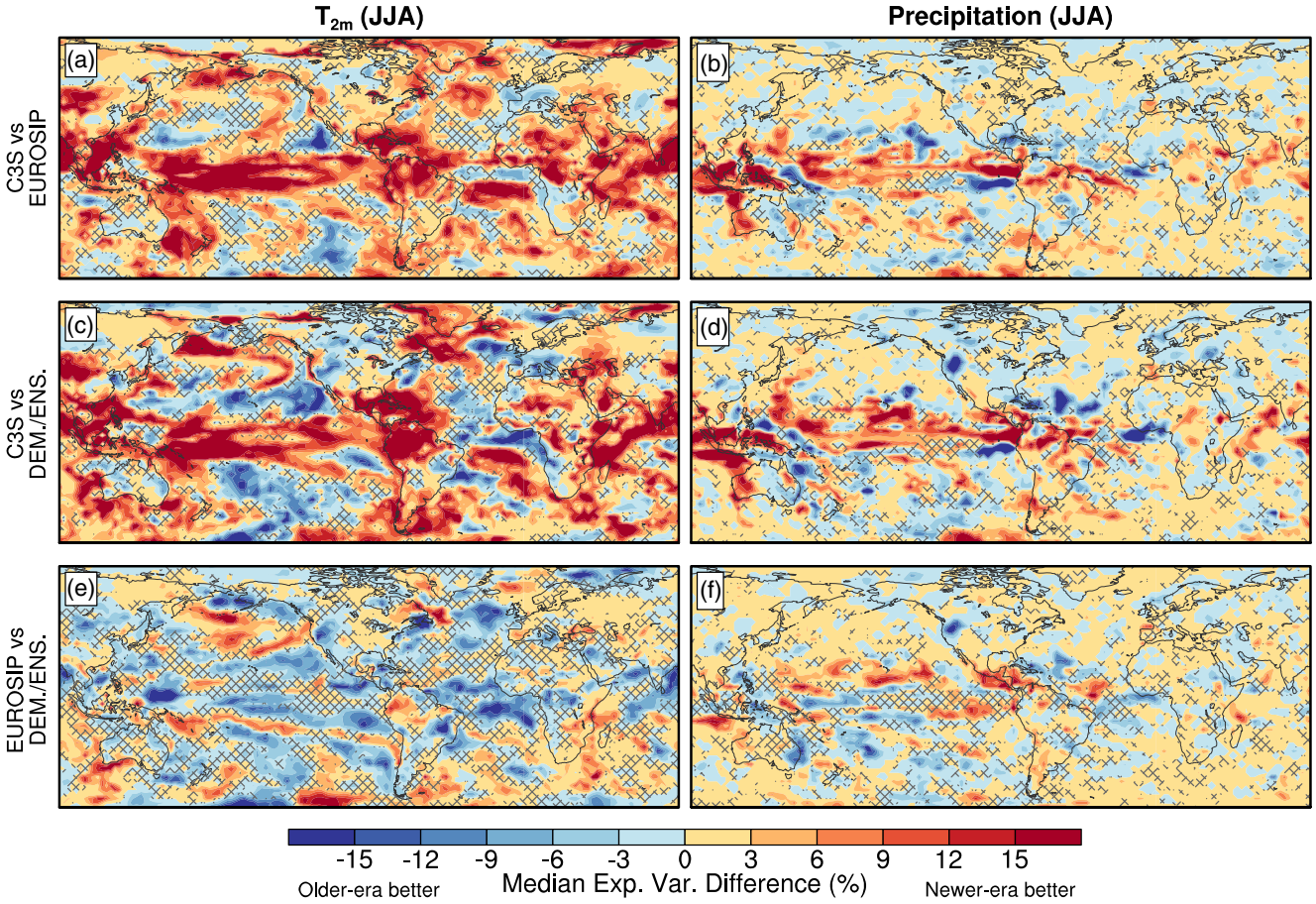

FIGURE 10 As Figure 8, but for the boreal summer hindcasts (JJA). [Colour figure can be viewed at wileyonlinelibrary.com]

teleconnections from tropical forcing (e.g., Beverley *et al.*, 2021; Knight & Scaife, 2024; O'Reilly *et al.*, 2018).

Maps of the grid-point level correlation skill of some important climate impact variables, T_{2m} and precipitation, are shown for the boreal summer hindcasts in Figure 10. There are some clear improvements through the model eras over the tropical ocean regions, reflecting improvements in SST hindcast skill (i.e., Figures 9 and S1). Over land, relatively strong improvements are seen in some tropical regions, notably over the Maritime Continent, Central/South America, and Central/East Africa. These improvements are isolated to the C3S era of models, potentially linked to the improvements in tropical SST and, specifically, ENSO prediction. In the extratropics, there are only modest improvements in most regions and even deterioration in others. The area of largest improvement in the C3S era models is in the northwestern region of China, where there are also clear improvements in $Z500$ (Figure S1), suggesting that this may be dynamically driven. For precipitation, the largest improvements in model skill are found for C3S in the Tropical Pacific and the Maritime continent. There are also some modest but potentially important improvements in precipitation skill over parts of the Indian subcontinent, consistent with other studies demonstrating improved forecasts of the Asian summer monsoon in the C3S seasonal hindcast models (Jain *et al.*, 2019; Takaya *et al.*, 2023).

3.3 | Examining multi-model ensemble improvements across model eras

To this point, we have focused on comparing different pairs of the models and using the accumulated statistics to examine model improvements. However, in many practical applications, users of operational seasonal forecasting projects make use of multi-model ensembles, so it is of interest to analyse the performance of representative ensembles across model eras. We focus our analysis of multi-model ensembles on hindcasts of the boreal winter season, as this is where we found the strongest and most robust improvements in both the Tropics and extratropics in the previous section (Figures 5–7). To examine the performance of multi-model ensembles, we use hindcasts from a short 13-year period, 1993–2005, where most models are available (Figure 1). We randomly subsampled

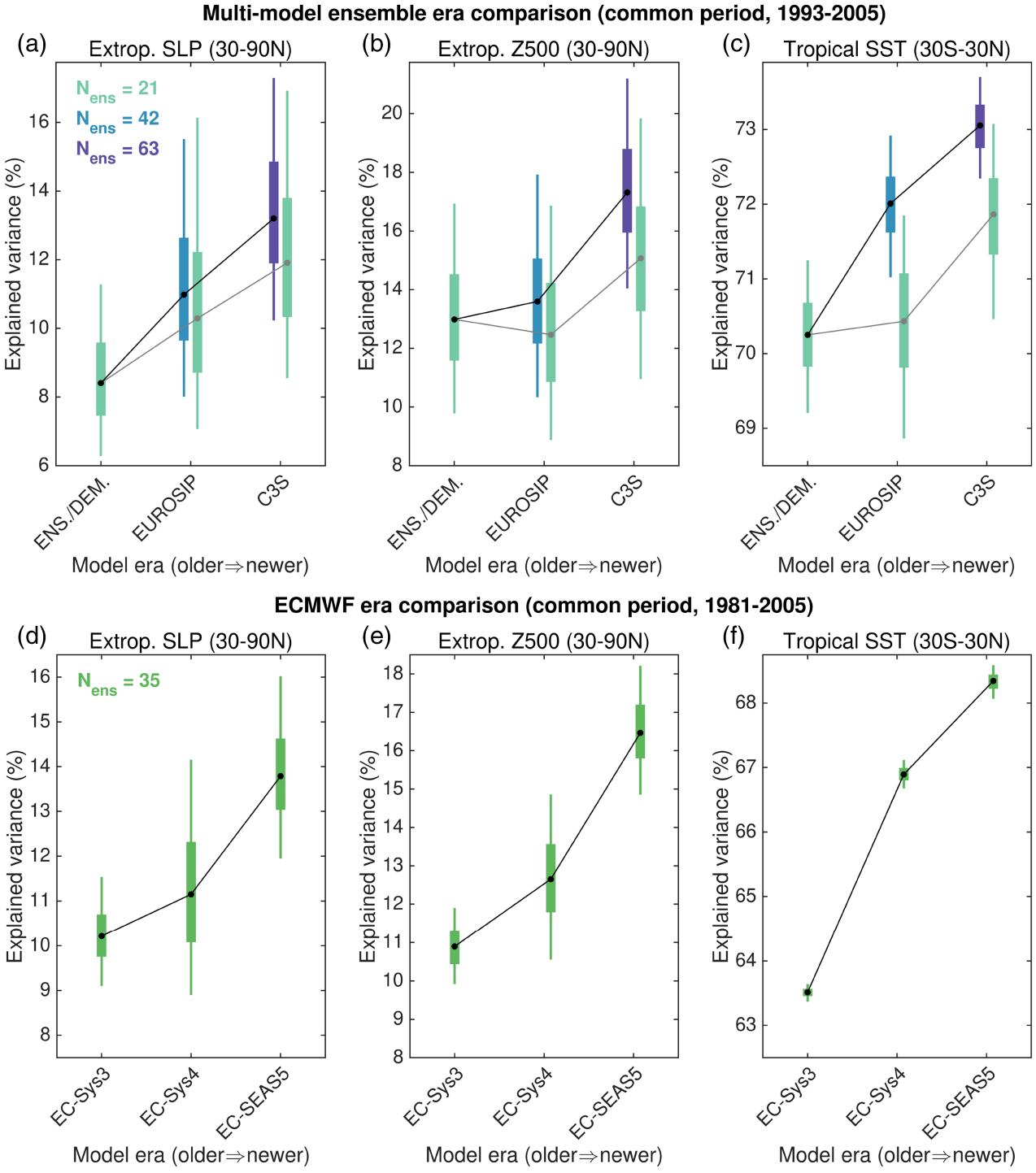

FIGURE 11 (a–c) Multi-model ensemble era comparison and (d–f) ECMWF-only era comparison. The box and whiskers show the 5th, 25th, 50th, 75th, and 95th percentiles of the distribution of subsampled ensembles. Bayes factors measuring the strength of the differences between the eras are shown for each of these metrics in Tables S2 and S3. [Colour figure can be viewed at wileyonlinelibrary.com]

seven members from each forecasting centre and each year, from all the models available over this period, to produce 21-member ensembles. These were then combined into a multi-model ensemble mean (after removing the respective model climatologies) and the process was repeated to produce 10,000 randomly subsampled 21-member ensembles. Measures of the total explained variance skill in the resulting ensembles are shown in Figure 11a–c.

The C3S era multi-model ensembles demonstrate an improvement in extratropical skill in the boreal winter hindcasts over the earlier eras, despite the relatively high

uncertainty. The most robust differences are seen between the C3S era models and the earliest DEM/ENS era models for SLP and $Z500$ in the extratropics and for tropical SSTs. Given the very short common hindcast period and small ensemble sizes considered here, there is substantial uncertainty across the different ensemble subsamples, and some of the Bayes factors are fairly modest (these can be inferred from Figure 11 and are provided in Table S2). Nonetheless, it is useful to be able to illustrate that the overall model improvement seen when analysing the different pairs of models is also found when analysing the types of multi-model ensemble that are most frequently used to communicate operational seasonal forecasts (e.g., by C3S).

In Figure 11a–c, we have also included a measure of the equivalent skill measures calculated for successive increases in ensemble size, in an attempt to approximate the typical increases in *hindcast* ensemble size over the model eras (Figure 1). There is an evident increase in skill associated with the increase in ensemble size compared with the fixed ensemble size calculations. Whilst this is expected, it serves as a useful demonstration that some of the increases in skill that have been reported in the literature, based on hindcasts of more modern seasonal forecasting systems, are due at least in part to the larger ensemble sizes, in addition to underlying model and/or initialisation improvements.

For one of the operational centres, ECMWF, there are hindcasts with larger ensemble sizes and longer periods from all three eras (Figure 1). This provides the opportunity to test whether the improvements seen for the representative multi-model ensemble (Figure 11a–c) are replicated for the ECMWF systems over these different eras. The equivalent results for the ECMWF systems over the common 1981–2005 hindcast period, with 35 subsampled ensemble members, are shown in Figure 11e–f. There is a clear improvement in total explained variance in the extratropical SLP and $Z500$, as well as for tropical SSTs. The longer hindcast period and increased ensemble size reduces the uncertainty (Table S3) compared with the representative multi-model ensemble in Figure 11a–c.

The overall characteristics of the improvements seen across the ECWMF systems resemble those broadly seen in the multi-model ensemble, with the C3S era hindcasts showing more substantial improvements in the extratropics with respect to the earlier model eras, whereas the differences between EUROSIP and DEM/ENS era hindcasts are slightly weaker and associated with lower Bayes factors (Table S3). The equivalent changes in skill of the NAO and PNA indices across the eras are consistent with the findings in the previous section, with notable increases for the PNA through to the C3S era, whereas the NAO shows no clear improvements in the multi-model ensembles (Figure S2). In summary, the analysis in this section demonstrates that an overall improvement in hindcast skill between model eras is evident in the type of multi-model ensembles often used to communicate operational seasonal forecasts, and indicates that the multi-model ensembles of the current operational seasonal forecasts can be considered more skilful than those of previous model eras.

3.4 | Exploring sources of improved seasonal forecast skill

Finally, we analyse briefly some possible sources of the improved seasonal forecast skill that we have seen through the different model eras. Given the overall improvements in skill in both tropical SSTs and the extratropical large-scale circulation (i.e., $Z500$) in the boreal winter hindcasts, it is of interest to compare these across the full set of models we have analysed. Figure 12a shows a scatter plot comparing the difference in total explained variance skill in extratropical $Z500$ and tropical SST across all the different model pairs. There is a clear and robust relationship between the increase in extratropical $Z500$ skill and the increased tropical SST skill (the linear correlation across these points is $r = 0.56$). This alone does not prove a causal link; however, tropical SSTs are widely considered as the dominant source of skill in extratropical seasonal forecasts, and this result provides evidence that improvements in tropical SST predictions have been a source of the improvement in skill in the extratropics.

The analysis in Section 3.2 showed that the clearest improvements in extratropical large-scale circulation in boreal winter hindcasts were found over the extratropical North Pacific and North America (Figures 6b and 7). The circulation in this area is strongly influenced by ENSO on interannual timescales, so it is of interest to examine how the improvements in the PNA hindcast skill are linked to the strength of the ENSO teleconnection to the PNA region in the different model ensembles—this is shown in Figure 12b. Across the model pairs, the models with better PNA hindcast skill tend to be those that have a stronger ENSO–PNA teleconnection, indicating that the improvement in the teleconnection is a robust source of this improvement (the linear correlation across these points is $r = 0.44$). There is still room for improvement in the modelled teleconnection pathway, however. In general, most of the models exhibit a weak wintertime ENSO–PNA teleconnection compared with reanalysis (Figure 12c) and, whilst this seems to have improved in the C3S era of models, the teleconnection remains broadly too weak. The weak ENSO teleconnection has been demonstrated in various C3S era models from other modelling centres and is a

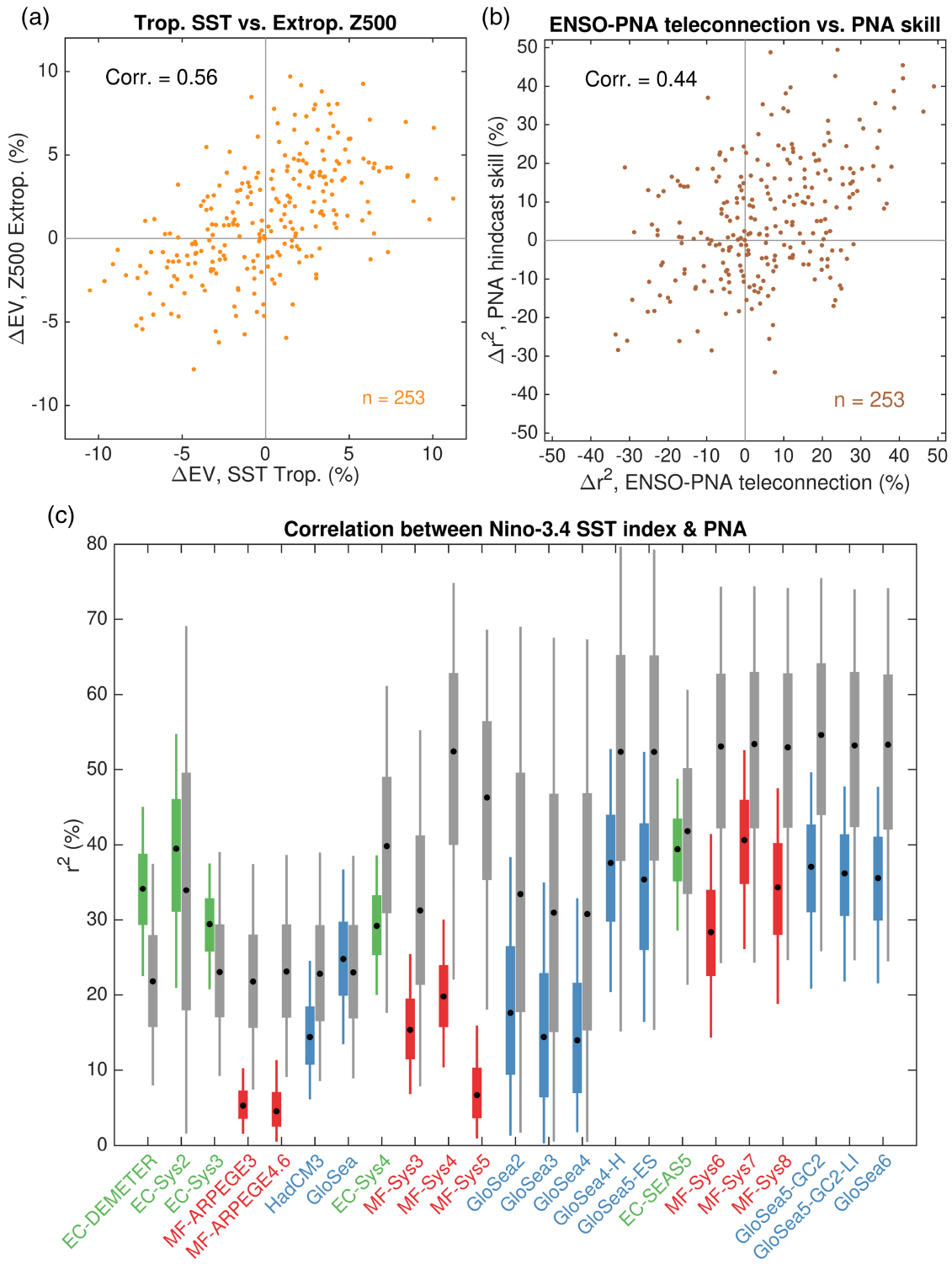

FIGURE 12 (a) Scatter plot of the median difference in extratropical $Z500$ total explained variance against the median difference in tropical SST total explained variance for each of the different model pairs. (b) Scatter plot of the median difference in PNA hindcast correlation skill (measured as explained variance difference, Δr^2) against the median difference in the strength of the ENSO–PNA teleconnections for each of the different model pairs. The ENSO–PNA teleconnection strength is defined as the correlation between Niño-3.4 and PNA indices (see Section 2.3) measured across all individual ensemble members, and expressed here as explained variance difference, Δr^2 (controlling for equal ensemble size and hindcast period in each model pair). (c) ENSO–PNA teleconnection in each model, defined as the r^2 between the Niño-3.4 SST index and the PNA index, across all ensemble members. The equivalent calculation from observational datasets (ERA-5 for the PNA, HadISST for Nino-3.4) is shown in grey. The box and whiskers show the 5th, 25th, 50th, 75th, and 95th percentiles, based on a bootstrap resampling across years. Note that the ERA5 estimate is plotted for each model separately and is calculated using the corresponding hindcast period, which differs across models. [Colour figure can be viewed at wileyonlinelibrary.com]

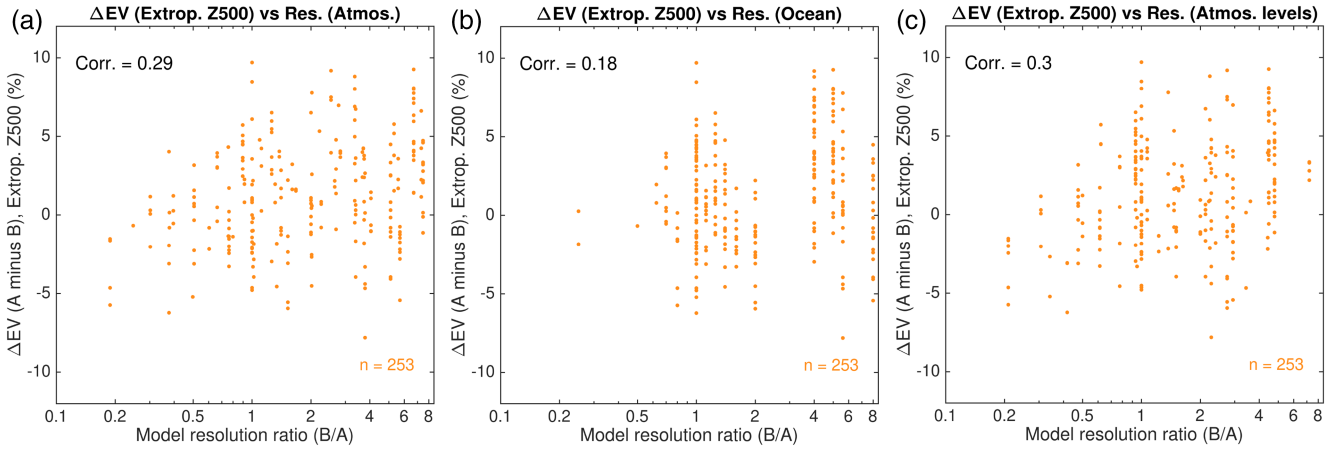

FIGURE 13 Link between model resolution and extratropical $Z500$ hindcast skill (DJF). (a) Horizontal atmospheric resolution, (b) horizontal ocean resolution, and (c) vertical atmospheric levels. The resolutions are given in Table S1. The analysis has been performed whilst controlling for equal ensemble size and hindcast period in each model pair. [Colour figure can be viewed at wileyonlinelibrary.com]

cause of signal-to-noise errors in seasonal hindcasts over the extratropical North Pacific (Williams *et al.*, 2023).

An area of focus over the past two decades of operational seasonal forecasting has been to improve the resolution of models; this has resulted in increases in horizontal and vertical resolution of the atmosphere and ocean components of the models (Table S1). To explore the impact of changes in model resolution on the improved extratropical hindcast skill seen in the boreal winter hindcasts, we compare the change in extratropical $Z500$ total explained variance skill with the ratio of model resolutions between the different model pairs. This is shown in Figure 13. Increased horizontal resolution in the atmosphere is broadly associated with increases in extratropical $Z500$ skill ($r = 0.29$; Figure 13a); a similar relationship is found between improved skill in the extratropics and horizontal resolution in the ocean, though the relationship is substantially weaker and less robust ($r = 0.19$; Figure 13b). Interestingly, there also seems to be a relatively robust relationship between increases in vertical resolution, in terms of the number of atmospheric levels, and increases in extratropical $Z500$ skill ($r = 0.30$; Figure 13c). Comparing the different model resolutions (shown in Table S1), it is obvious that the increases in resolution are not wholly independent, such that models with higher resolution in one component tend also to have higher resolutions in other model components. Nonetheless, it is interesting that increases in the model resolution in the atmosphere seem to stand out as more impactful for seasonal forecasts of extratropical $Z500$ compared with the ocean resolution, which, as noted earlier, may be related to the difficulties in initialising higher resolution ocean models.

In addition to resolution, it is important to highlight that there are numerous other model developments that have been made during this period, such as parameterisation schemes, land-surface models, and sea-ice models, all of which likely contribute to the model performance. In addition, there have been significant developments in the methods—and analysis/reanalysis datasets—used to initialise both the operational forecasts and the hindcasts. These are each expected to have a significant influence on the model performance. However, attributing the specific improvements that have been gained through these model changes is not straightforward using the skill-based metrics we have used here. A process-based analysis that focuses on more straightforward model biases/errors and associated model dynamics may be a fruitful approach in this respect (e.g., Hermanson *et al.*, 2018; O'Reilly, 2025).

4 | SUMMARY AND CONCLUDING REMARKS

Seasonal forecasts have been operational for over two decades. However, the impact of continuous advances in modelling on seasonal forecast skill is not easily demonstrated, as comparisons are challenged by small operational forecast samples, large variations in ensemble sizes between models, and inconsistent hindcast periods. In this study, through a careful and systematic comparison of models from different eras, we present clear evidence of improved seasonal forecast performance.

Our analysis, which accounts for differences in ensemble size and period across the forecasting systems, demonstrates that there have been clear improvements in some regions through the different model eras. For both the boreal winter and summer hindcasts, there have been significant improvements in forecasting in tropical regions, which are concurrent with improvements in the skill of

tropical SSTs. These improvements in the Tropics are associated with increased predictability of temperature and precipitation across various continental regions on seasonal timescales. For the extratropics, the picture is more mixed, with strong improvements only evident during the boreal winter season over the North Pacific and North America. The sources of improvement over the winter extratropics are found to be strongly related to improvements in tropical SST skill and related improvements in the strength of the ENSO teleconnection to the PNA. Improvements of seasonal forecast skill over the rest of the extratropics, such as over Eurasia, are generally absent or patchy in individual models. The improvements that are found are most pronounced in the newest C3S era models, and these improvements in skill are also evident in representative multi-model ensembles (Figure 11) that represent more closely how operational forecasts are used in practice.

One likely source of improvement in seasonal forecasts over the past 20 years is increases in model resolution (Table S2). Here, we find that models with higher atmospheric resolution in both the horizontal and the vertical are associated with improvements in hindcast skill in the extratropics during boreal winter. Increases in horizontal atmospheric resolution have been linked to improved fidelity of simulation of synoptic atmospheric variability and the associated feedback onto the large-scale circulation; however, it is clear that models at present operational resolutions (i.e., $\approx$ 30 km; Table S1) still exhibit clear deficiencies (e.g., O'Reilly, 2025; Scaife *et al.*, 2019). Increases in vertical resolution have been linked to improved simulations of stratospheric processes (Butler *et al.*, 2016), which have been linked to improvements in predictability in the extratropics on seasonal timescales (e.g., Scaife *et al.*, 2022), broadly consistent with the improved skill found here.

Whilst the analysis here has highlighted some clear improvements through the different model eras, we found large uncertainty and variability across the individual model pairs (Figure 4). Much of this uncertainty is caused by the smaller ensemble sizes in the earlier era of models. Another source of this uncertainty is the short hindcast periods that have been used by most modelling centres from the EUROSIP era onwards (e.g., Figure 1). Whilst we have been able to measure some improvements here (though careful subsampling based on hindcast period and ensemble size), the levels of uncertainty associated with the short hindcast sizes make more detailed studies of the sources of model improvement very challenging. With modern reanalyses now routinely being extended to the mid-20th century, the ability to perform hindcasts using coupled prediction systems is evident—for example, the CMIP6 models contributing to the Decadal Climate Prediction Project produce hindcasts from 1960 onwards. However, the hindcast periods for the operational C3S multi-model forecasts cover only 24 years (1993–2016), some of which only have 10 ensemble members. These put strong contraints on the levels of robust, process-based understanding that can be achieved through analysis of the performance of forecast systems. A stronger focus on performing longer hindcasts with large ensemble sizes is crucial to understanding the performance and any potential improvements in future operational seasonal forecasts.

ACKNOWLEDGEMENTS

C. O'Reilly was supported by a Royal Society (URF\R1\201230). To analyse the data and produce the plots, we used MATLAB and NCAR Command Language.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data are publicly available for research purposes (e.g., C3S available online at the Climate Data Store and other data available through the ECMWF MARS archive). The post-processed, regridded data used in this article are available on the Zenodo online repository (https://zenodo.org/records/16949349) for ease of access for related research.

ORCID

Christopher H. O'Reilly https://orcid.org/0000-0002-8630-1650
Antje Weisheimer https://orcid.org/0000-0002-7231-6974

REFERENCES

Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P.P., Janowiak, J. et al. (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, 4(6), 1147–1167.

Ambaum, M.H.P. (2010) Significance tests in climate science. *Journal of Climate*, 23(22), 5927–5932.

Anderson, D.L.T. (2006) Operational seasonal prediction. In: Hagedorn, R. & Palmer, T. (Eds.) *Predictability of weather and climate*. Cambridge: Cambridge University Press, pp. 514–531. Available from: https://www.cambridge.org/core/books/predictability-of-weather-and-climate/operational-seasonal-prediction/7B9B8177518A2A50F3BEAD664AE6EE3B

Arsenault, K.R., Shukla, S., Hazra, A., Getirana, A., McNally, A., Kumar, S.V. et al. (2020) The NASA hydrological forecast system for food and water security applications. *Bulletin of the American Meteorological Society*, 101, E1007–E1025. Available from: https://doi.org/10.1175/BAMS-D-18-0264.1

Baker, L.H., Shaffrey, L.C., Johnson, S.J. & Weisheimer, A. (2024) Understanding the intermittency of the wintertime North Atlantic oscillation and East Atlantic pattern seasonal forecast

skill in the copernicus C3S multi-model ensemble. *Geophysical Research Letters*, 51(15), e2024GL108472. Available from: https://doi.org/10.1029/2024GL108472

Baker, L.H., Shaffrey, L.C., Sutton, R.T., Weisheimer, A. & Scaife, A.A. (2018) An Intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, 45(15), 7808–7817. Available from: https://doi.org/10.1029/2018GL078838

Barnston, A.G., Mason, S.J., Goddard, L., DeWitt, D.G. & Zebiak, S.E. (2003) Multimodel ensembling in seasonal climate forecasting at IRI. *Bulletin of the American Meteorological Society*, 84, 1783–1796. Available from: https://doi.org/10.1175/BAMS-84-12-1783

Barnston, A.G., Tippett, M.K., L'Heureux, M.L., Li, S. & DeWitt, D.G. (2012) Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? *Bulletin of the American Meteorological Society*, 93, 631–651. Available from: https://doi.org/10.1175/BAMS-D-11-00111.1

Bauer, P., Thorpe, A. & Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525(7567), 47–55. Available from: https://doi.org/10.1038/nature14956

Beverley, J.D., Newman, M. & Hoell, A. (2023) Rapid development of systematic ENSO-related seasonal forecast errors. *Geophysical Research Letters*, 50(10), e2022GL102. Available from: https://doi.org/10.1029/2022GL102249

Beverley, J.D., Woolnough, S.J., Baker, L.H., Johnson, S.J., Weisheimer, A. & O'Reilly, C.H. (2021) Dynamical mechanisms linking Indian monsoon precipitation and the circumglobal teleconnection. *Climate Dynamics*, 57(9), 2615–2636. Available from: https://doi.org/10.1007/s00382-021-05825-6

Butler, A.H., Arribas, A., Athanassiadou, M., Baehr, J., Calvo, N., Charlton-Perez, A. et al. (2016) The climate-system historical forecast project: do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Quarterly Journal of the Royal Meteorological Society*, 142(696), 1413–1427. Available from: https://doi.org/10.1002/qj.2743

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N. et al. (2016) Skilful predictions of the winter North Atlantic oscillation one year ahead. *Nature Geoscience*, 9(11), 809–814. Available from: https://doi.org/10.1038/ngeo2824

Hardiman, S.C., Dunstone, N.J., Scaife, A.A., Smith, D.M., Comer, R., Nie, Y. et al. (2022) Missing eddy feedback may explain weak signal-to-noise ratios in climate predictions. *NPJ Climate and Atmospheric Science*, 5(1), 1–8. Available from: https://doi.org/10.1038/s41612-022-00280-4

Hermanson, L., Ren, H.-L., Vellinga, M., Dunstone, N.D., Hyder, P., Ineson, S. et al. (2018) Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Climate Dynamics*, 51(4), 1411–1426. Available from: https://doi.org/10.1007/s00382-017-3962-9

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(730), 1999–2049. Available from: https://doi.org/10.1002/qj.3803

Jain, S., Scaife, A.A. & Mitra, A.K. (2019) Skill of Indian summer monsoon rainfall prediction in multiple seasonal prediction systems. *Climate Dynamics*, 52(9), 5291–5301. Available from: https://doi.org/10.1007/s00382-018-4449-z

Johansson, Ã. (2007) Prediction skill of the NAO and PNA from daily to seasonal time scales. *Journal of Climate*, 20, 1957–1975. Available from: https://doi.org/10.1175/JCLI4072.1

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(3), 1087–1117. Available from: https://doi.org/10.5194/gmd-12-1087-2019

Jolliffe, I.T. & Stephenson, D.B. (2012) *Forecast verification: a practitioner's guide in atmospheric science*. New Jersey: John Wiley & Sons.

Jones, P.D., Osborn, T.J. & Briffa, K.R. (2003) Pressure-based measures of the North Atlantic oscillation (NAO): a comparison and an assessment of changes in the strength of the NAO and in its influence on surface climate parameters. *Geophysical Monograph Series*, 134, 51–62. Available from: https://doi.org/10.1029/134GM03

Kim, H.-M., Webster, P.J. & Curry, J.A. (2012) Seasonal prediction skill of ECMWF system 4 and NCEP CFSv2 retrospective forecast for the northern hemisphere winter. *Climate Dynamics*, 39(12), 2957–2973. Available from: https://doi.org/10.1007/s00382-012-1364-6

Knight, J.R. & Scaife, A.A. (2024) Influences on North-Atlantic summer climate from the El Niño-southern oscillation. *Quarterly Journal of the Royal Meteorological Society*, 150(764), 4498–4510. Available from: https://doi.org/10.1002/qj.4826

MacLeod, D. (2019) Seasonal forecasts of the East African long rains: insight from atmospheric relaxation experiments. *Climate Dynamics*, 53, 4505–4520. Available from: https://doi.org/10.1007/s00382-019-04800-6

MacLeod, D., O'Reilly, C., Palmer, T. & Weisheimer, A. (2018) Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmospheric Science Letters*, 19(5), e815. Available from: https://doi.org/10.1002/asl.815

MacLeod, D., Graham, R., O'Reilly, C., Otieno, G. & Todd, M. (2021) Causal pathways linking different flavours of ENSO with the Greater Horn of Africa short rains. *Atmospheric Science Letters*, 22, e1015. Available from: https://doi.org/10.1002/asl.1015

MacLeod, D., Quichimbo, E.A., Michaelides, K., Asfaw, D.T., Rosolem, R., Cuthbert, M.O. et al. (2023) Translating seasonal climate forecasts into water balance forecasts for decision making. *PLOS Climate*, 2(3), e0000138. Available from: https://doi.org/10.1371/journal.pclm.0000138

McPhaden, M.J., Zebiak, S.E. & Glantz, M.H. (2006) ENSO as an integrating concept in earth science. *Science*, 314(5806), 1740–1745. Available from: https://doi.org/10.1126/science.1132588

Murphy, J.M. (1988) The impact of ensemble forecasts on predictability. *Quarterly Journal of the Royal Meteorological Society*, 114(480), 463–493. Available from: https://doi.org/10.1002/qj.49711448010

O'Reilly, C.H. (2025) Signal-to-noise errors in early winter euro-Atlantic predictions linked to weak ENSO teleconnections and pervasive jet biases. *Quarterly Journal of the Royal Meteorological Society*, 151, e4952. Available from: https://doi.org/10.1002/qj.4952

O'Reilly, C.H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T.N., Schaller, N. et al. (2017) Variability in seasonal forecast skill of northern hemisphere winters over the twentieth century. *Geophysical Research Letters*, 44(11), 5729–5738. Available from: https://doi.org/10.1002/2017GL073736

O'Reilly, C.H., Weisheimer, A., MacLeod, D., Befort, D.J. & Palmer, T. (2020) Assessing the robustness of multidecadal variability in northern hemisphere wintertime seasonal forecast skill. *Quarterly Journal of the Royal Meteorological Society*, 146(733), 4055–4066. Available from: https://doi.org/10.1002/qj.3890

O'Reilly, C.H., Weisheimer, A., Woollings, T., Gray, L.J. & MacLeod, D. (2019a) The importance of stratospheric initial conditions for winter North Atlantic oscillation predictability and implications for the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*, 145(718), 131–146. Available from: https://doi.org/10.1002/qj.3413

O'Reilly, C.H., Woollings, T., Zanna, L. & Weisheimer, A. (2018) The impact of tropical precipitation on summertime euro-Atlantic circulation via a circumglobal wave train. *Journal of Climate*, 31, 6481–6504. Available from: https://doi.org/10.1175/JCLI-D-17-0451.1

O'Reilly, C.H., Woollings, T., Zanna, L. & Weisheimer, A. (2019b) An interdecadal shift of the extratropical teleconnection from the tropical Pacific during boreal summer. *Geophysical Research Letters*, 46(22), 13379–13388. Available from: https://doi.org/10.1029/2019GL084079

Palmer, T.N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P. et al. (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, 85, 853–872. Available from: https://doi.org/10.1175/BAMS-85-6-853

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

Scaife, A.A. & Smith, D. (2018) A signal-to-noise paradox in climate science. *NPJ Climate and Atmospheric Science*, 1(1), 1–8. Available from: https://doi.org/10.1038/s41612-018-0038-4

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(7), 2514–2519. Available from: https://doi.org/10.1002/2014GL059637

Scaife, A.A., Baldwin, M.P., Butler, A.H., Charlton-Perez, A.J., Domeisen, D.I.V., Garfinkel, C.I. et al. (2022) Long-range prediction and the stratosphere. *Atmospheric Chemistry and Physics*, 22(4), 2601–2623. Available from: https://doi.org/10.5194/acp-22-2601-2022

Scaife, A.A., Camp, J., Comer, R., Davis, P., Dunstone, N., Gordon, M. et al. (2019) Does increased atmospheric resolution improve seasonal climate predictions? *Atmospheric Science Letters*, 20(8), e922. Available from: https://doi.org/10.1002/asl.922

Shepherd, T.G. (2021) Bringing physical reasoning into statistical practice in climate-change science. *Climatic Change*, 169(1), 2. Available from: https://doi.org/10.1007/s10584-021-03226-6

Smith, D.M., Scaife, A.A. & Kirtman, B.P. (2012) What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? *Environmental Research Letters*, 7(1), 015602. Available from: https://doi.org/10.1088/1748-9326/7/1/015602

Stockdale, T.N., Molteni, F. & Ferranti, L. (2015) Atmospheric initial conditions and the predictability of the Arctic oscillation. *Geophysical Research Letters*, 42(4), 1173–1179. Available from: https://doi.org/10.1002/2014GL062681

Takaya, Y., Ren, H., Vitart, F. & Robertson, A. (2023) Current status and progress in the seasonal prediction of the Asian summer monsoon. *Mausam*, 74(2), 455–466. Available from: https://doi.org/10.54302/mausam.v74i2.5925

Trenberth, K.E. (1997) The definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12), 2771–2778.

van der Linden, P. (2009) *ENSEMBLES: climate change and its impacts at seasonal, decadal and centennial timescales*. Exeter: Met Office Hadley Centre.

Wallace, J.M. & Gutzler, D.S. (1981) Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 109(4), 784–812.

Weisheimer, A., Baker, L.H., Bröcker, J., Garfinkel, C.I., Hardiman, S.C., Hodson, D.L. et al. (2024) The signal-to-noise paradox in climate forecasts: revisiting our understanding and identifying future priorities. *Bulletin of the American Meteorological Society*, 105, E651–E659. Available from: https://doi.org/10.1175/BAMS-D-24-0019.1

Weisheimer, A., Befort, D.J., MacLeod, D., Palmer, T., O'Reilly, C. & Strømmen, K. (2020) Seasonal forecasts of the twentieth century. *Bulletin of the American Meteorological Society*, 101, E1413–E1426. Available from: https://doi.org/10.1175/BAMS-D-19-0019.1

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D.A. & Palmer, T. (2017) Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quarterly Journal of the Royal Meteorological Society*, 143(703), 917–926. Available from: https://doi.org/10.1002/qj.2976

Williams, N.C., Scaife, A.A. & Screen, J.A. (2023) Underpredicted ENSO teleconnections in seasonal forecasts. *Geophysical Research Letters*, 50(5), e2022GL101.

Zebiak, S.E. & Cane, M.A. (1987) A model El Niño-Southern oscillation. *Monthly Weather Review*, 115(10), 2262–2278.

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: O'Reilly, C.H., MacLeod, D., Befort, D., Shepherd, T.G. & Weisheimer, A. (2025) Evaluating seasonal forecast improvements over the past two decades. *Quarterly Journal of the Royal Meteorological Society*, e70036. Available from: https://doi.org/10.1002/qj.70036