npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-025-01198-3

Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data

Check for updates

Chris Kent[1] ✉, Adam A. Scaife[1,2], Nick J. Dunstone[1], Doug Smith[1], Steven C. Hardiman[1], Tom Dunstan[1] & Oliver Watt-Meyer[3]

Machine learning weather models trained on observed atmospheric conditions can outperform conventional physics-based models at short- to medium-range (1–14 day) forecast timescales. Here we take the machine learning model ACE2, trained to predict 6-hourly steps in atmospheric evolution and which can remain stable over long forecast periods, and assess it from a seasonal forecasting perspective (1–3 month lead time). Applying persisted sea surface temperature (SST) and sea-ice anomalies centred on 1st November each year, we initialise a lagged ensemble of seasonal predictions covering 1993/1994 to 2015/2016. Over this 23-year period there is remarkable similarity in the patterns of predictability with a leading physics-based model. The ACE2 model exhibits skilful predictions of the North Atlantic Oscillation (NAO) with a correlation score of 0.47 ($p = 0.02$), as well as a realistic global distribution of skill and ensemble spread. Surprisingly, ACE2 is found to exhibit a signal-to-noise error as seen in physics-based models, in which it is better at predicting the real world than itself. Examining predictions of winter 2009/2010 indicates potential limitations of ACE2 in capturing extreme seasonal conditions that extend outside the training data. This study reveals that machine learning weather models can produce skilful global seasonal predictions and provide new opportunities for increased understanding, development and generation of near-term climate predictions.

In recent years a revolution in weather prediction has occurred in which machine learning-based models can match or outperform physics-based models over a range of metrics[1–5]. Learning the 1–6-hour evolution of the atmospheric state, these models can produce skilful forecasts for several days by feeding the predictions back into themselves, as dynamical models do, known as “autoregressive” forecasting[6]. Recent studies suggest skilful forecasts can be made covering several weeks[5,7–9] and very large ensembles can provide improved estimates of extreme events[10]. Beyond these timescales, instabilities can grow, or the predictions blur and smooth, restricting their application to long-range climate predictions at monthly or seasonal time scales[11]. Some models are stable for long autoregressive rollouts and can capture the climatological state and aspects of interannual variability[7,12–15], however to date, their ability for skilful seasonal predictions has not been established.

Machine learning predictions at seasonal timescales (1–3 month lead times) often utilise more direct approaches in learning relationships between predictors and specific predictands, or resort to using model data for training. For example, skilful predictions have been demonstrated for the El Niño-Southern Oscillation (ENSO) as well as some regional scale climate variability[16–22]. Understanding the mechanisms underpinning such predictions can be difficult and developing methods to provide explainability is a key topic of research[23,24]. With only one event per season, a key limitation at longer forecast periods is the relatively small sample size available for training. This restricts the ability to learn complex relationships while at the same time keeping a suitable number of years separate for testing, as needed for dynamical models[25]. One approach to overcome this is to utilise model data for training[19,26,27], but the errors and biases found in physics-based models are inevitably inherited.

In this study we assess the newly developed machine learning weather model ACE2[13] from a seasonal forecasting perspective. This model predicts the atmospheric evolution at 6-hourly time steps and can remain stable for long autoregressive forecast periods, enabling it to provide seasonal simulations even though it was not explicitly trained to provide such predictions. It is trained only on historical conditions from the ERA5 dataset[28]. We

[1]Met Office Hadley Centre, Exeter, UK. [2]Department of Mathematics and Statistics, University of Exeter, Exeter, UK. [3]Allen Institute for Artificial Intelligence (Ai2), Seattle, WA, USA. ✉e-mail: chris.kent@metoffice.gov.uk

initialise ACE2 during autumn each year from 1993 to 2015 and assess the seasonal skill of December-January-February (DJF) conditions, a lead time of 1–3 months. To provide boundary conditions, the SST and sea-ice anomalies at the time of initialisation are persisted throughout the forecast period each year. The influence from large-scale drivers such as ENSO are therefore preserved, but any coupled ocean-atmosphere processes are missing. We compare the ACE2 seasonal forecasts to those from GloSea, a leading physics-based coupled ocean-atmosphere ensemble prediction system[29,30].

Results

Skilful data-driven seasonal forecasts

Over the 23-year assessment period the pattern of seasonal skill (1-3 month lead) demonstrated by ACE2 closely resembles that of the dynamical model for mean sea level pressure (MSLP, Fig. 1a, b). This is remarkable considering ACE2 was designed for stable climate simulations, with no deliberate attempt to capture seasonal predictability. While much of the tropical skill is due to the persistence of slowly evolving processes such as ENSO from the initialisation of the tropical oceans[31,32], ACE2 also exhibits skill across the tropical land and the extratropics, including the North Atlantic and North Pacific. Interestingly, ACE2 also exhibits reduced skill over Eurasia, as seen in the physics-based model GloSea. In most regions the ACE2 correlation is weaker than that for GloSea. For example, the area-average correlation across the northern hemisphere extratropics (20°N to 90°N) is 0.39 in ACE2 and 0.44 in GloSea, while over the tropics (20°S to 20°N) the scores are 0.79 and 0.82, respectively. In comparison, a persistence forecast using October monthly mean conditions scores 0.17 across the northern hemisphere and 0.52 across the tropics. Subsampling predictions across years indicates no evidence that these results are biased by predictions based on initial conditions seen during the training of ACE2 (Supplementary Figs. 2 and 3).

For temperature (Fig. 1c, d) we continue to see large regions of skill from ACE2, including South America, Africa, Australia and parts of North America. As seen for MSLP, GloSea outperforms ACE2 across many parts of the world with the area-weighted mean correlation across the northern hemisphere extratropics at 0.41 in ACE2 and 0.45 in GloSea, and 0.68 and 0.77 respectively across the tropics. The skill for both systems is lower for precipitation, however the ACE2 model (Fig. 1e) once again closely resembles that of GloSea (Fig. 1f), particularly across the tropics, the Caribbean and east Asia.

These results demonstrate that the ACE2 model can skilfully predict seasonal variability across many parts of the world with a lead time of 1-3 months.

Predictability of the North Atlantic Oscillation

The NAO is the primary mode of seasonal variability across the North Atlantic[33] and is a key focus for extratropical seasonal prediction[34–36]. ACE2 can predict the DJF-mean NAO[37] with a correlation score of r = 0.47 (Fig. 2a), at a lead time of 1–3 months. This is statistically significant at the 95% level (p = 0.023) and is highly competitive with a range of dynamical

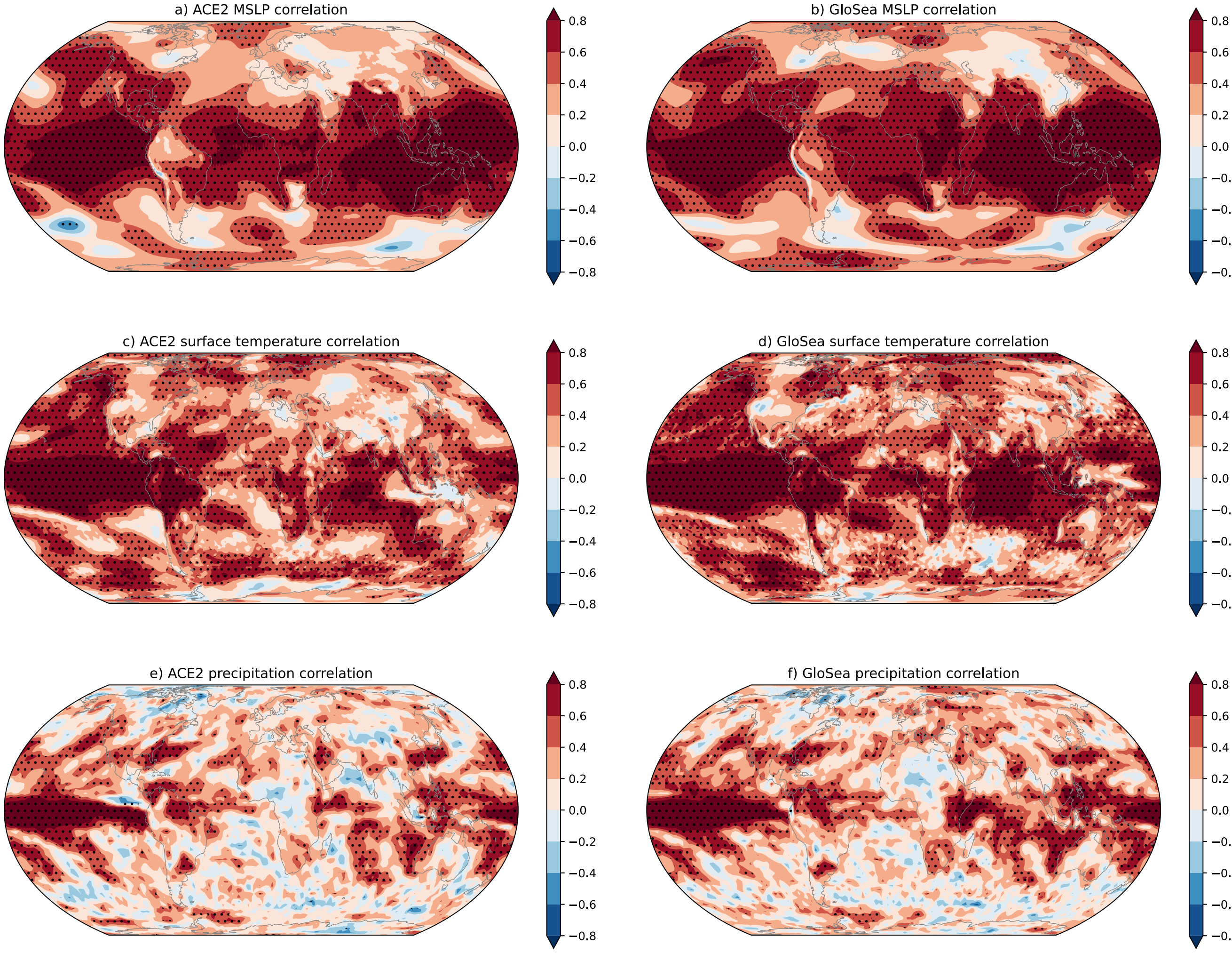
a) ACE2 MSLP correlation
b) GloSea MSLP correlation
c) ACE2 surface temperature correlation
d) GloSea surface temperature correlation
e) ACE2 precipitation correlation
f) GloSea precipitation correlation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

Fig. 1 | Skilful seasonal (DJF) predictions from the ACE2 machine learning and GloSea dynamical models with a lead time of 1-3 months. Correlation score of mean sea level pressure (**a**), surface temperature (**c**) and precipitation (**e**) for ACE2 and GloSea (**b**, **d**, **f**) calculated across 1993/1994 to 2015/2016. Stippling indicates correlations are significantly different to zero (23 years, 95% confidence level).

models. For example, over a shorter 19-year analysis period (1993–2011) ACE2 exhibits higher NAO skill ($r = 0.42$) than 4 operational ensemble prediction systems[36].

It is important to note that only the 9 winters between 2002 and 2010 are fully independent of the ACE2 training period[13]. Over this shorter period the NAO correlation remains high ($r = 0.6$), although with reduced significance due to the smaller sample size ($p = 0.07$). Skill is also high across an extended 1981–2022 period (r = 0.52) and a subsampling analysis suggests that these NAO results are not biased by predictions from years within the ACE2 training period (Supplementary Figs. 1 and 3).

Interestingly, ACE2 gives a poor prediction of the extreme winter in 2009/2010 (see Section "The extreme winter of 2009/2010" below). Nevertheless, given the long autoregressive forecasts, the lack of a well resolved stratosphere, and the use of non-interacting, persisted SSTs, the ACE2 model skilfully predicts the NAO. This is surprising as both stratospheric variability and interactive ocean processes underpin dynamical model skill[38,39].

We also find that the ACE2 and GloSea NAO predictions are not strongly correlated ($r = 0.34$, $p = 0.11$) and so there may be additional value in combining them. Indeed, an ensemble mean constructed from both models results in an NAO correlation score of $r = 0.65$ ($p < 0.01$), matching that estimated by GloSea with an extended ensemble size of 127 members. Furthermore, after removing the climatological mean, the ACE2 and GloSea NAO predictions appear to be drawn from the same underlying distribution (two-sample KS-test, 95% confidence). This indicates that ACE2 could also be utilised to enhance dynamical model ensembles.

In addition to skilful seasonal predictions, the ACE2 ensemble closely matches the dynamical model in terms of NAO variability. Following initialisation, we find that the ACE2 ensemble mean error and ensemble spread increase in line with GloSea (Fig. 2, Equations (1) and (2)). Furthermore, the DJF-mean total standard deviation across all years and members is 4.3 hPa in ERA5, 3.6 hPa in ACE2 and 3.8 hPa in GloSea. For the ensemble mean variability the standard deviation is 1.11 hPa in ACE2 and 1.21 hPa in GloSea. The lagged-ensemble methodology used here therefore enables sufficient ensemble member spread to develop, but other methods for ensemble generation are key topics for future research.

In line with dynamical models[34,40,41], ACE2 NAO skill also increases strongly with ensemble size (solid line, Fig. 2c). This is encouraging as it is much cheaper and quicker, in computational terms, to increase the ensemble size of data-driven models compared to dynamical models. However, it can also be seen that when the ACE2 ensemble mean is used to predict one of its own individual members (so-called 'perfect model' skill), the skill is markedly lower (r = 0.25, dashed lines in Fig. 2c) than the ACE2 skill in predicting the observed NAO (thick solid lines, Fig. 2c). The ratio of predictable components (Equation (3)) provides a measure of observed and modelled predictability and variance. For ACE2 this quantity is found to be 1.6, only slightly less than the 1.8 for GloSea, but still greater than 1 (90% confidence). This indicates that for ACE2, the ensemble mean variance is small compared to the total ensemble variance given its skill in predicting the observed NAO[42].

Therefore, despite having been trained only on reanalysis data, the ACE2 predictions also exhibit a signal-to-noise error which resembles that

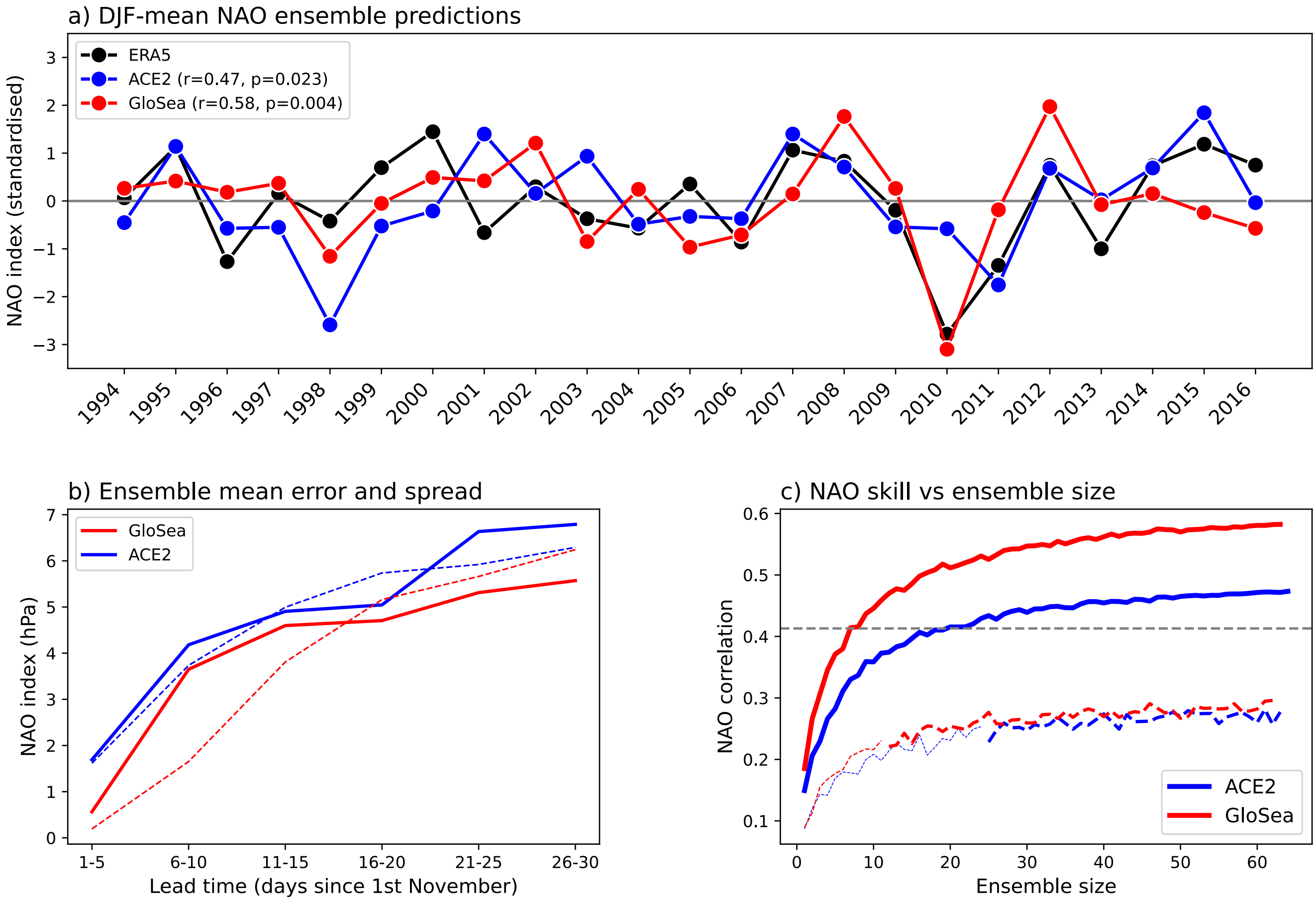

Fig. 2 | Skilful predictions of the DJF-mean North Atlantic Oscillation (NAO). **a** DJF-mean NAO index, standardised to unit variance, from ERA5 (black), GloSea (red) and ACE2 (blue). **b** ensemble mean RMSE (solid) and spread (dashed) for the NAO (hPa) as a function of lead time during November each year, averaged over all years. **c** Relationship between NAO correlation score and ensemble size (solid lines) and skill in predicting individual withheld ensemble members (dashed lines) based on 1000 random samples with no replacement. The dashed lines are thickened when significantly below the corresponding solid line (outside 95% sampling range). The horizontal dashed grey line indicates the 95% significance level for a sample size of 23 years.

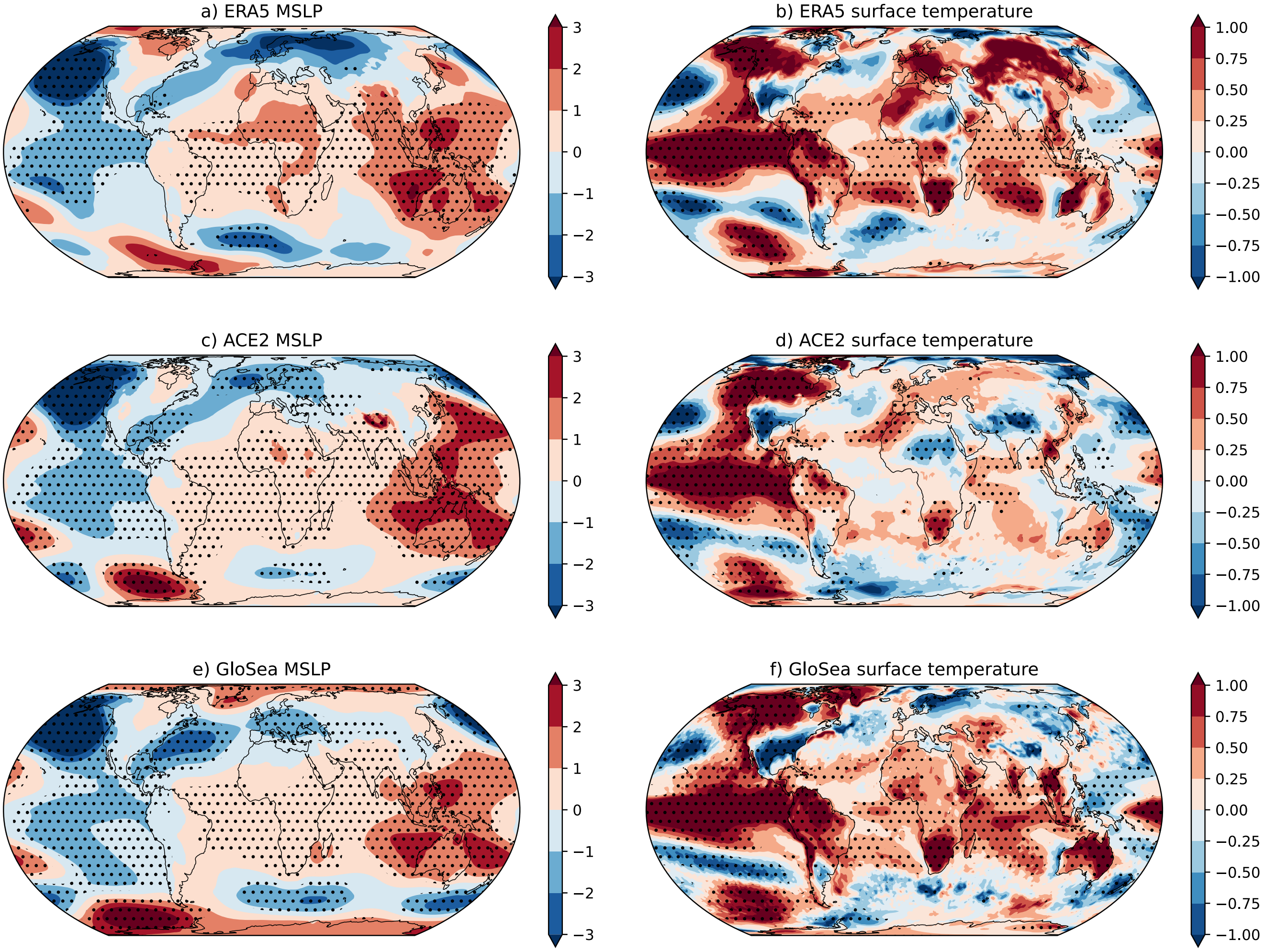

Fig. 3 | Influence of ENSO on DJF surface conditions. Composite maps of El Niño years ($n = 8$) minus La Niña years ($n = 9$) for mean sea level pressure (hPa) and surface temperature (K) anomalies for ERA5 (**a**, **b**) ACE2 (**c**, **d**) and GloSea (**e**, **f**). Shaded contours show the DJF mean anomaly. Stippling indicates significant differences (two-tailed T-test, 95% confidence level).

found in dynamical models[34,40,42–44]. This is somewhat surprising as it may suggest that the signal-to-noise error is not restricted to a physical model error and instead occurs due to some other damping effect on the predictable signal. For example, weak eddy forcing and feedback are one hypothesised cause of the error[45], however these characteristics are not weak within the reanalysis used to train ACE2. Further investigation of ACE2 characteristics is needed, but we note that machine learning predictions can also exhibit damping and smoothing of the kinetic energy spectrum[11,46] potentially leading to similar errors in forecast anomaly amplitude. It is possible that the same qualitative behaviour occurs for different reasons in the ACE2 and GloSea models, but further research is needed to understand if this is the case.

ENSO as a driver of seasonal skill

ENSO is the primary mode of interannual climate variability and is a key driver of seasonal skill across many parts of the world[47,48]. In this section we investigate whether ACE2 is correctly capturing ENSO teleconnections.

Composite differences between El Niño and La Niña years (Fig. 3) reveal that ACE2 exhibits very similar teleconnection patterns to those seen in ERA5 and GloSea for both MSLP and surface temperature. In particular, we find El Niño deepens the Aleutian low and influences the North Atlantic jet, extending eastward from the Caribbean. This suggests that ACE2 is capturing the ENSO relationship on the subtropical jet, an important mechanism underpinning the global influence of ENSO[47,49]. In terms of the surface temperature response, ACE2 also exhibits very similar ENSO teleconnections to ERA5 and GloSea, particularly over North America, South America, southern Africa and Australia. These composites indicate that ACE2 is correctly capturing the regional interannual variability associated with ENSO across many parts of the world despite being trained only on the 6-hourly evolution of the atmosphere.

The extreme winter of 2009/2010

As a final part of our assessment we focus on predictions for the extreme northern hemisphere winter of 2009/2010, which is part of the independent dataset withheld during the training of ACE2. This winter is characterised by a record negative NAO, well beyond the anomalies seen in other years. It was also subject to a minor and a major sudden stratospheric warming (SSW), a strong El Niño and an easterly Quasi Biennial Oscillation (QBO)[50]. The winter mean MSLP anomaly (Fig. 4a) exhibits a very zonal negative NAO which is well captured by GloSea (Fig. 4c). However, the ACE2 ensemble mean prediction does not appear to capture this signal with only slightly above average pressure across the Arctic (Fig. 4b). This is surprising given the strong tropical forcing and potentially indicates a limitation of ACE2 in predicting extreme, out of sample conditions. Exploring this further, we find that both ERA5 and GloSea exhibit a weakened stratospheric polar vortex Fig. 4d, f), while ACE2 exhibits near-normal vortex strength (Fig. 4e).

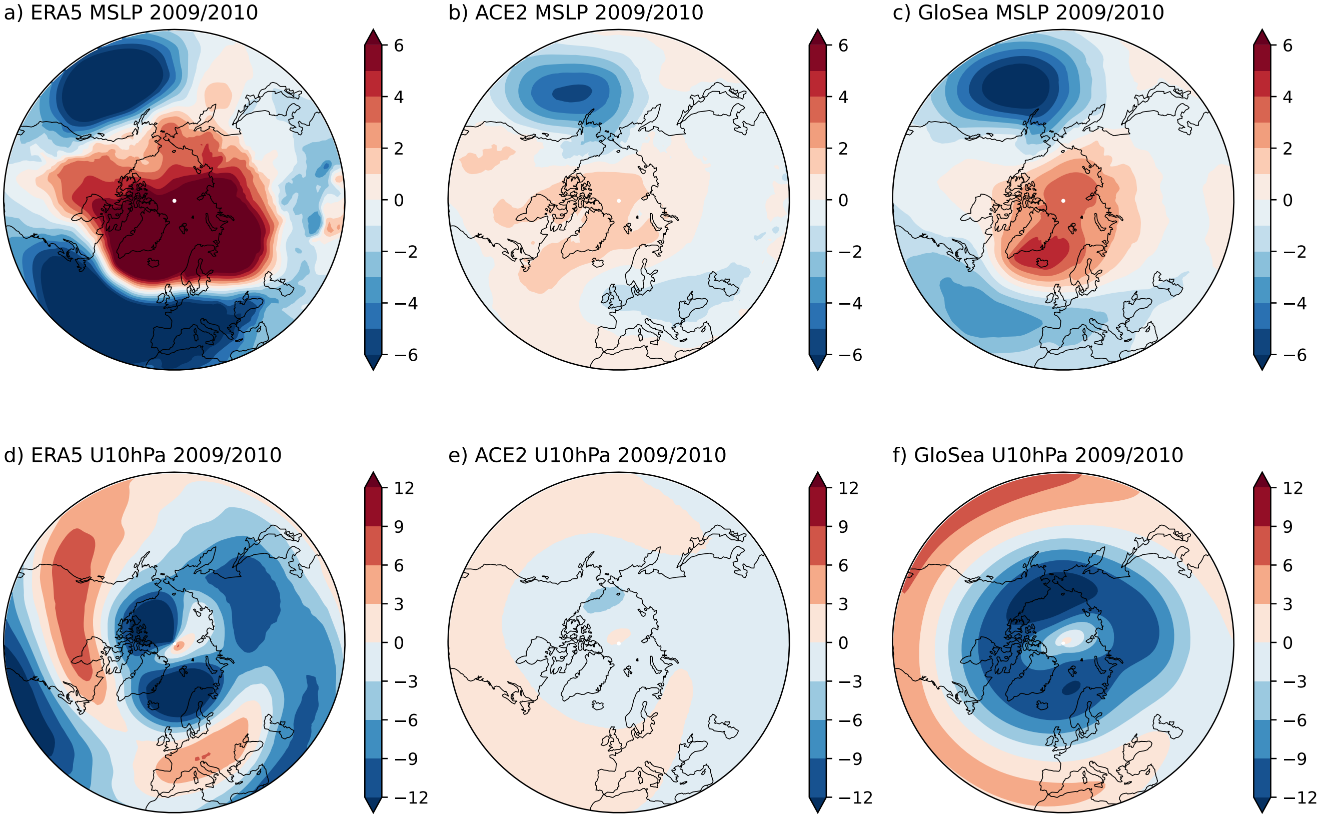

Fig. 4 | Surface and stratospheric anomalies associated with the extreme winter of 2009/2010. Anomalies from the 1994-2016 climatology of MSLP (hPa) and zonal wind at 10hPa (ms-1) for ERA5 (**a**, **d**), ACE2 (**b**, **e**) and GloSea (**c**, **f**). ACE2 stratospheric conditions are model layer 0 (above 50 hPa).

In terms of SSWs, the winter comprised of a minor warming in December 2009 and a major warming in January 2010, reflecting the increased SSW probability due to the El Niño and easterly QBO[50–53]. GloSea appears to capture this increase, with 81% of members (51 out of 63) experiencing easterly zonal winds at 10hPa and 60°N within the winter. This is significantly higher than GloSea's climatological probability of 62% (two proportion Z-test, 95% confidence level). In comparison, only 39% of ACE2 members (25 out of 64) exhibit easterly stratospheric winds in the upper most model layer (above 50mb), which is not significantly different to the climatological rate of 40%. This indicates that the ACE2 model is not correctly capturing the disruption to the stratospheric polar vortex during winter 2009/2010.

Furthermore, the SSW probability within ACE2 is relatively consistent across El Niño (45%) and La Niña (36%) years, neither of which are significantly different from neutral years (41%, one-tailed two proportion Z-test, 95% confidence level). GloSea and ERA5 however exhibit significant differences between active and neutral ENSO years, with a higher chance of an SSW during El Niño[54–57]. This suggests that while the ACE2 can exhibit sub-seasonal stratospheric variability[13] it is not fully capturing the ENSO teleconnection to the stratosphere despite realistic tropospheric teleconnections.

Discussion

This study demonstrates skilful seasonal predictions from a machine learning weather model. Despite being trained only on the 6-hourly observed evolution of the atmosphere, when assessed from a seasonal prediction perspective (i.e. lead time 1-3 months), the ACE2 model exhibits significant skill and is competitive with current dynamical systems. A lagged-ensemble approach is found to generate ensemble spread which closely matches observations and a physics-based ensemble prediction system, a characteristic is it not specifically trained on. The model produces realistic ENSO teleconnections in the troposphere, but the stratospheric pathway is not in line with observations. This may be due to a relatively small sample of observed events (e.g. slower time scales in the stratosphere and limited number of SSWs), the training methodology (e.g. loss weightings applied to different levels or parameters), or model architecture. If the latter, this could potentially be addressed through enhanced vertical resolution in the stratosphere, a characteristic found to be important in dynamical models[54,58–60], providing an opportunity for improved skill in the future.

Dataset independence is an important part of understanding the generalization of machine learning models and our results are based on predictions initialised with conditions both within and independent of the ACE2 training period. However, we find no evidence of bias within our predictions at the global or regional scale. This is potentially due to the use of long (4-month) rollouts and persisted boundary conditions which differ from the 6-hour loss minimalization and time-evolving conditions within the ACE2 training. Understanding the sensitivity of seasonal predictions to different training and test years, particularly over the satellite period, is a key topic for moving towards real-time predictions which occur within a climate outside of the training period.

A significant benefit of machine learning models is the relatively cheap computational cost. For seasonal forecasting timescales, a dynamical model can take hours on a supercomputer for each simulation. In comparison, the ACE2 model can complete a 4-month forecast simulation in under 2 minutes on an Nvidia A100 GPU. Opportunities arising from this include the ability to generate very large ensemble sizes (e.g. over 7000 members[10]), much longer assessment periods, rapid testing of new experimental setups and better exploration of sources of predictability and the signal-to-noise error[44]. Machine learning models are therefore highly applicable for seasonal and climate timescales where large ensembles are needed. Further research is needed on optimal ensemble generation approaches as well as coupling to data-driven ocean models[61] or ocean-atmosphere-coupled dynamical models. However, it is clear from this work that the machine learning models can supplement and support current seasonal forecasting methods.

Overall, these results show that the machine learning revolution is not limited to short-range weather forecasts and can provide several new opportunities for advancing near-term climate predictions.

Methods

Datasets

Historical atmospheric conditions are taken from the ERA5 reanalysis[28]. To persist SST and sea-ice conditions throughout a forecast we create a seasonally varying climatology based on the 6-hourly atmospheric state, for each grid cell, using a rolling-mean gaussian filter with a width (standard deviation) of 10 days. Observed monthly rainfall totals are taken from the Global Precipitation Climatology Project version 2.3 (GPCP)[62].

For comparison with dynamical models, hindcasts (retrospective forecasts) initialised from 1993 to 2015 are taken from the GloSea operational ensemble prediction system with GC3.2 configuration[29,30,63]. A 63-member ensemble is constructed from 21 members initialised on 25th October, 1st November and 9th November each year and the ensemble spread is generated through a stochastic physics scheme[64]. GloSea simulations cover a forecast period of 6 months with an atmospheric resolution of approximately 0.5 degrees and an ocean resolution of 0.25 degrees. It has 85 vertical levels in the atmosphere, covering the entire stratosphere and extending up to 85km (0.01 hPa) as well as 75 levels in the ocean. The GloSea prediction system is one of the top performing dynamical models across sub-seasonal and seasonal timescales for both the tropics and mid-latitudes[32,36,65,66].

For this study we use the machine learning atmospheric model ACE2[13]. The model is trained solely on ERA5 reanalysis atmospheric fields and predicts the evolution of the atmospheric state at 6-hour time steps at a 1° grid resolution. Importantly, ACE2 autoregressive forecasts are stable over multiple years hypothesized to be due to its Spherical Fourier Neural Operator architecture[67], use of user prescribed ocean and sea-ice boundary conditions, and physical constraints on mass conservation, moisture, precipitation rate and radiative fluxes[13].

Of relevance to this study, the 10 years from 2001 to 2010, which lies within our 23-year hindcast period, are withheld during training of ACE2[13] and form an independent test period for the model. The remaining years are used to train the model. However, our experiments (see below) are initialised one month prior to the periods of interest and utilise persisted boundary conditions, while time-evolving boundary data were used for training ACE2. These specific atmospheric and ocean states will therefore be new to the model, although the large-scale patterns will have been seen previously. Combined with this, each forecast involves over 500 autoregressive steps, over which which errors will grow and result in individual trajectories. This is demonstrated through the realistic ensemble spread within ACE2 at seasonal timescales. Quantitative testing of the ensemble (Supplementary Figs. 2 and 3) at global and regional scales found no evidence of bias within the ACE2 predictions between training and independent years.

All ERA5 and GloSea data is bilinearly interpolated to the native 1° x 1° ACE2 grid, except for precipitation, in which ACE2 and GloSea are interpolated to the 2.5° x 2.5° GPCP grid.

Indices and metrics

We define ENSO years based on the DJF Oceanic Niño Index[68] with a threshold of ± 0.5 K. El Niño winters are 1995, 1998, 2003, 2005, 2007, 2010, 2015, and 2016. La Niña winters are 1996, 1999, 2000, 2001, 2006, 2008, 2009, 2011, and 2012.

We define the NAO index[37] as the difference in mean sea level pressure between a southern box (90°W-60°E, 20°N-55°N) and a northern box (90°W-60°E, 55°N-90°N). The results are consistent when applying a smaller regional definition[40] ($r = 0.42$, $p = 0.048$) and a point-based estimate[34] ($r = 0.41$, $p = 0.053$).

To calculate the ensemble mean error and spread as a function of lead time we utilise only ACE2 members initialised between 00:00z on 28th October and 00:00z on 1st November ($n = 20$) each year and GloSea members initialised at 00:00z on 1st November ($n = 21$). Forecasted daily NAO values are aggregated into 5-day means (pentads) and the climatological mean removed. The ACE2 values are therefore partly larger than GloSea's due to the inclusion of longer lead time forecasts. The ensemble mean error for a given 5-day average, $RMSE_p$ is defined as:

$$RMSE_p = \sqrt{\frac{1}{23}\sum_{i=1994}^{2016}\left(model_{i,p} - ERA5_{i,p}\right)^2} \tag{1}$$

The corresponding average ensemble spread is defined as:

$$\sigma_p = \sqrt{\frac{1}{23}\sum_{i=1994}^{2016}\sigma_{ip}^2} \tag{2}$$

Where σ_{ip} is the standard deviation of the model NAO across members for year i and pentad p.

To assess ACE2 and GloSea predictions in terms of signal and noise we compute the ratio of predictable components (RPC,[43]) as

$$RPC = \frac{r}{\sigma_{sig}/\sigma_{tot}} \tag{3}$$

where r is the ensemble mean correlation with ERA5, σ_{sig} is the ensemble mean standard deviation and σ_{tot} is the standard deviation across all members and years. A random resampling procedure is used for significance testing[43].

ACE2 experimental setup

ACE2 seasonal predictions are generated using a lagged ensemble approach. An ensemble member is initialised every 6 hours between 25th October and 9th November each year from 1993 to 2015, creating a total of 64 members per year. The forecast period extends from initialisation through to mid-March the following year, providing a lead time of 1-3 months. For example, a forecast member initialised in November 2001 is rolled out over 500 times until March 2002. Initial conditions for each member are taken from the ERA5 reanalysis dataset[28]. Boundary SST and sea-ice conditions are provided throughout each forecast by calculating the instantaneous anomaly at initialisation for each grid cell and persisting this throughout the forecast using the derived ERA5 6-hourly climatology. This is different to the ACE2 training, in which time-evolving boundary conditions are used.

The 6-hourly climatology is calculated using a gaussian filter with a width (standard deviation) of 10 days, averaged across the 1994-2016 period (23 years). For each initialisation the instantaneous initial condition anomaly is persisted using this climatology, e.g. for a given gridcell at time (t) the SST boundary condition is

$$SST(t) = SST(0) - climatology(0) + climatology(t) \tag{4}$$

Where $t = 0$ indicates the value at initialisation. The same method is used to persist sea-ice concentrations, with all values limited to be between 0 and 1.

Historical downward shortwave radiative flux at the top of the atmosphere and global mean atmospheric carbon dioxide inputs are prescribed throughout the hindcast period[13] as performed for the GloSea simulations. However, understanding the sensitivity of ACE2 predictions to these boundary conditions is a key topic for further research. We find that repeating the hindcast experiment using a climatology derived from 1988–2022 (excluding 1994–2016) produces consistent results (NAO $r = 0.54$) as does utilising the previous year's TOA shortwave flux (NAO $r = 0.43$) and using the previous year's CO_2 (NAO $r = 0.38$). These additional results are in line with a natural variability test (NAO $r = 0.42$) where the initial condition times were manually altered by 6 hours, suggesting a limited sensitivity of these boundary conditions for this application.

Data availability

Initial conditions for ensemble members are taken from the ERA5 reanalysis dataset[28]. CO_2 and solar irradiance forcing data are available at https://huggingface.co/allenai/ACE-2-ERA5. The data used for the figures is available at https://doi.org/10.5281/zenodo.15025230.

Code availability

The trained ACE2-ERA5 model checkpoint used in this study is available at https://huggingface.co/allenai/ACE-2-ERA5. The ACE2 code is available at https://github.com/ai2cm/ace.

Received: 22 April 2025; Accepted: 11 August 2025;
Published online: 25 August 2025

References

1. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
2. Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).
3. Kurth, T. et al. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, 1–11 (2023).
4. Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, 6425 (2024).
5. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
6. de Burgh-Day, C. O. & Leeuwenburg, T. Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Dev.* **16**, 6433–6477 (2023).
7. Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024).
8. Ling, F. et al. Fengwu-w2s: A deep learning model for seamless weather-to-subseasonal forecast of global atmosphere. *arXiv preprint arXiv:2411.10191* (2024).
9. Guo, Y. et al. Maximizing the impact of deep learning on subseasonal-to-seasonal climate forecasting: The essential role of optimization. *arXiv preprint arXiv:2411.16728* (2024).
10. Mahesh, A. et al. Huge ensembles part ii: Properties of a huge ensemble of hindcasts generated with spherical fourier neural operators. *arXiv preprint arXiv:2408.01581* (2024).
11. Karlbauer, M. et al. Advancing parsimonious deep learning weather prediction using the healpix mesh. *J. Adv. Modeling Earth Syst.* **16**, e2023MS004021 (2024).
12. Cresswell-Clay, N. et al. A deep learning earth system model for stable and efficient simulation of the current climate. *arXiv preprint arXiv:2409.16247* (2024).
13. Watt-Meyer, O. et al. Ace2: Accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Clim. Atmos. Sci.* **8**, 1–15 (2025).
14. Bodnar, C. et al. A foundation model for the earth system. *Nature* **641**, 1180–1187 (2025).
15. Yang, S. et al. Generative assimilation and prediction for weather and climate. *arXiv preprint arXiv:2503.03038* (2025).
16. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year enso forecasts. *Nature* **573**, 568–572 (2019).
17. Qian, Q. F., Jia, X. J. & Lin, H. Machine learning models for the seasonal forecast of winter surface air temperature in north america. *Earth Space Sci.* **7**, e2020EA001140 (2020).
18. Kim, J. et al. Spatiotemporal neural network with attention mechanism for el niño forecasts. *Sci. Rep.* **12**, 7204 (2022).
19. Taylor, J. & Feng, M. A deep learning model for forecasting global monthly mean sea surface temperature anomalies. *Front. Clim.* **4**, 932932 (2022).
20. Mu, B., Jiang, X., Yuan, S., Cui, Y. & Qin, B. Nao seasonal forecast using a multivariate air–sea coupled deep learning model combined with causal discovery. *Atmosphere* **14**, 792 (2023).
21. Qian, Q. & Jia, X. Seasonal forecast of winter precipitation over china using machine learning models. *Atmos. Res.* **294**, 106961 (2023).
22. Sun, Y., Simpson, I., Wei, H.-L. & Hanna, E. Probabilistic seasonal forecasts of north atlantic atmospheric circulation using complex systems modelling and comparison with dynamical models. *Meteorological Appl.* **31**, e2178 (2024).
23. Labe, Z. M. & Barnes, E. A. Detecting climate signals using explainable ai with single-forcing large ensembles. *J. Adv. Modeling Earth Syst.* **13**, e2021MS002464 (2021).
24. Eyring, V. et al. Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Clim. Change* **14**, 916–928 (2024).
25. Manzanas, R., Torralba, V., Lledó, L. & Bretonnière, P.-A. On the reliability of global seasonal forecasts: Sensitivity to ensemble size, hindcast length and region definition. *Geophys. Res. Lett.* **49**, e2021GL094662 (2022).
26. Toms, B. A., Barnes, E. A. & Hurrell, J. W. Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophys. Res. Lett.* **48**, e2021GL093842 (2021).
27. Pan, B. et al. Improving seasonal forecast using probabilistic deep learning. *J. Adv. Modeling Earth Syst.* **14**, e2021MS002766 (2022).
28. Hersbach, H. et al. The era5 global reanalysis. *Q. J. R. meteorological Soc.* **146**, 1999–2049 (2020).
29. MacLachlan, C. et al. Global seasonal forecast system version 5 (glosea5): A high-resolution seasonal forecast system. *Q. J. R. Meteorological Soc.* **141**, 1072–1084 (2015).
30. Kettleborough, J. et al. Glosea6: A large ensemble seasonal forecasting system. *Submitted to Monthly Weather Review* (2025).
31. Ehsan, M. A., L'Heureux, M. L., Tippett, M. K., Robertson, A. W. & Turmelle, J. Real-time enso forecast skill evaluated over the last two decades, with focus on the onset of enso events. *npj Clim. Atmos. Sci.* **7**, 301 (2024).
32. Scaife, A. A. et al. Tropical rainfall predictions from multiple seasonal forecast systems. *Int. J. Climatol.* **39**, 974–988 (2019).
33. Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. An overview of the north atlantic oscillation. *Geophys. Monogr.-Am. Geophys. Union* **134**, 1–36 (2003).
34. Scaife, A. et al. Skillful long-range prediction of european and north american winters. *Geophys. Res. Lett.* **41**, 2514–2519 (2014).
35. Smith, D. M., Scaife, A. A., Eade, R. & Knight, J. R. Seasonal to decadal prediction of the winter north atlantic oscillation: Emerging capability and future prospects. *Q. J. R. Meteorological Soc.* **142**, 611–617 (2016).
36. Baker, L. H., Shaffrey, L. C., Johnson, S. J. & Weisheimer, A. Understanding the intermittency of the wintertime north atlantic oscillation and east atlantic pattern seasonal forecast skill in the copernicus c3s multi-model ensemble. *Geophys. Res. Lett.* **51**, e2024GL108472 (2024).
37. Stephenson, D. et al. North atlantic oscillation response to transient greenhouse gas forcing and the impact on european winter climate: a cmip2 multi-model assessment. *Clim. Dyn.* **27**, 401–420 (2006).
38. Scaife, A. A. et al. Long range prediction and the stratosphere. *Atmos. Chem. Phys. Discuss.* **2021**, 1–30 (2021).
39. Meehl, G. A. et al. Initialized earth system prediction from subseasonal to decadal timescales. *Nat. Rev. Earth Environ.* **2**, 340–357 (2021).
40. Dunstone, N. et al. Skilful predictions of the winter north atlantic oscillation one year ahead. *Nat. Geosci.* **9**, 809–814 (2016).
41. Baker, L., Shaffrey, L., Sutton, R., Weisheimer, A. & Scaife, A. An intercomparison of skill and overconfidence/underconfidence of the wintertime north atlantic oscillation in multimodel seasonal forecasts. *Geophys. Res. Lett.* **45**, 7808–7817 (2018).
42. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Clim. Atmos. Sci.* **1**, 28 (2018).

43. Eade, R. et al. Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.* **41**, 5620–5628 (2014).
44. Weisheimer, A. et al. The signal-to-noise paradox in climate forecasts: revisiting our understanding and identifying future priorities. *Bull. Am. Meteorological Soc.* **105**, E651–E659 (2024).
45. Hardiman, S. C. et al. Missing eddy feedback may explain weak signal-to-noise ratios in climate predictions. *Npj Clim. Atmos. Sci.* **5**, 57 (2022).
46. Bonavita, M. On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.* **51**, e2023GL107377 (2024).
47. Horel, J. & Wallace, J. Planetary-scale phenomena associated with the southern oscillation. vol. 109. *Mon Weather Rev* 1520–0493 (1981).
48. Taschetto, A. et al. Enso atmospheric teleconnections, el niño southern oscillation in a changing climate, 309–335 (2020).
49. Jiménez-Esteve, B. & Domeisen, D. I. The tropospheric pathway of the enso–north atlantic teleconnection. *J. Clim.* **31**, 4563–4584 (2018).
50. Fereday, D., Maidens, A., Arribas, A., Scaife, A. & Knight, J. Seasonal forecasts of northern hemisphere winter 2009/10. *Environ. Res. Lett.* **7**, 034031 (2012).
51. Garfinkel, C., Butler, A., Waugh, D., Hurwitz, M. & Polvani, L. Why might stratospheric sudden warmings occur with similar frequency in el niño and la niña winters? *Journal of Geophysical Research: Atmospheres***117** (2012).
52. Domeisen, D. I., Garfinkel, C. I. & Butler, A. H. The teleconnection of el niño southern oscillation to the stratosphere. *Rev. Geophysics* **57**, 5–47 (2019).
53. Anstey, J. A. et al. Teleconnections of the quasi-biennial oscillation in a multi-model ensemble of qbo-resolving models. *Q. J. R. Meteorological Soc.* **148**, 1568–1592 (2022).
54. Bell, C. J., Gray, L. J., Charlton-Perez, A. J., Joshi, M. M. & Scaife, A. A. Stratospheric communication of el niño teleconnections to european winter. *J. Clim.* **22**, 4083–4096 (2009).
55. Butler, A. H. & Polvani, L. M. El niño, la niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. *Geophysical Research Letters* **38** (2011).
56. Bett, P. E. et al. Using large ensembles to quantify the impact of sudden stratospheric warmings and their precursors on the north atlantic oscillation. *Weather Clim. Dyn.* **4**, 213–228 (2023).
57. Ineson, S. et al. Statistics of sudden stratospheric warmings using a large model ensemble. *Atmos. Sci. Lett.* **25**, e1202 (2024).
58. Ineson, S. & Scaife, A. The role of the stratosphere in the european climate response to el niño. *Nat. Geosci.* **2**, 32–36 (2009).
59. Cagnazzo, C. & Manzini, E. Impact of the stratosphere on the winter tropospheric teleconnections between enso and the north atlantic and european region. *J. Clim.* **22**, 1223–1238 (2009).
60. Butler, A. H. et al. The climate-system historical forecast project: Do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Q. J. R. Meteorological Soc.* **142**, 1413–1427 (2016).
61. Clark, S. K. et al. Ace2-som: Coupling to a slab ocean and learning the sensitivity of climate to changes in co _2. *arXiv preprint arXiv:2412.04418* (2024).
62. Adler, R. F. et al. The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979–present). *J. Hydrometeorol.* **4**, 1147–1167 (2003).
63. Williams, K. et al. The met office global coupled model 3.0 and 3.1 (gc3. 0 and gc3. 1) configurations. *J. Adv. Modeling Earth Syst.* **10**, 357–380 (2018).
64. Tennant, W. J., Shutts, G. J., Arribas, A. & Thompson, S. A. Using a stochastic kinetic energy backscatter scheme to improve mogreps probabilistic forecast skill. *Monthly Weather Rev.* **139**, 1190–1206 (2011).
65. Vitart, F. Madden-julian oscillation prediction and teleconnections in the s2s database. *Q. J. R. Meteorological Soc.* **143**, 2210–2220 (2017).
66. Feng, P.-N., Lin, H., Derome, J. & Merlis, T. M. Forecast skill of the nao in the subseasonal-to-seasonal prediction models. *J. Clim.* **34**, 4757–4769 (2021).
67. Bonev, B. et al. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, 2806–2823 (PMLR, 2023).
68. NOAA-CPC. Cold and warm episodes by season, accessed on 2025-02-27 https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php (2025).

Acknowledgements

This work was supported by the Met Office Hadley Centre Climate Programme (HCCP), funded by the UK Department for Science, Innovation and Technology (DSIT). We also thank Rowan Sutton, David Walters, Christopher Bretherton and three expert reviewers for their useful comments and feedback.

Author contributions

C.K. performed the ACE2 seasonal predictions and carried out the analysis against GloSea. C.K., A.A.S., N.D., D.S. and S.H. designed the experimental set up and interpreted the results. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-01198-3.

Correspondence and requests for materials should be addressed to Chris Kent.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.