Author Contributions:
Conceptualization: Antje Weisheimer, Magdalena A. Balmaseda, Tim N. Stockdale
Data curation: Michael Mayer
Formal analysis: Antje Weisheimer
Investigation: Antje Weisheimer
Methodology: Antje Weisheimer
Software: Antje Weisheimer
Validation: Antje Weisheimer
Visualization: Antje Weisheimer
Writing – original draft: Antje Weisheimer
Writing – review & editing: Antje Weisheimer, Magdalena A. Balmaseda,

Variability of ENSO Forecast Skill in 2-Year Global Reforecasts Over the 20th Century

Antje Weisheimer[1,2] , Magdalena A. Balmaseda[1] , Tim N. Stockdale[1] , Michael Mayer[1,3] , S. Sharmila[4,5] , Harry Hendon[5,6], and Oscar Alves[5]

[1]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK, [2]University of Oxford, Department of Physics, National Centre for Atmospheric Science (NCAS), Oxford, UK, [3]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria, [4]Centre for Applied Climate Sciences, University of Southern Queensland, Toowoomba, Australia, [5]Bureau of Meteorology, Melbourne, Australia, [6]Monash University, Melbourne, Australia

Abstract In order to explore temporal changes of predictability of El Niño Southern Oscillation (ENSO), a novel set of global biennial climate reforecasts for the historical period 1901–2010 has been generated using a modern initialized coupled forecasting system. We find distinct periods of enhanced long-range skill at the beginning and at the end of the twentieth century, and an extended multi-decadal epoch of reduced skill during the 1930s–1950s. Once the forecast skill extends beyond the first spring barrier, the predictability limit is much enhanced and our results provide support for the feasibility of skillful ENSO forecasts up to 18 months. Changes in the mean state, variability (amplitude), persistence, seasonal cycle and predictability suggest that multi-decadal variations in the dynamical characteristics of ENSO rather than the data coverage and quality of the observations have primarily driven the reported non-monotonic skill modulations.

Plain Language Summary El Niño Southern Oscillation (ENSO) is an irregular periodic variation in ocean temperatures and winds over the tropical Pacific Ocean that occurs every few years with varying intensity of its several-month-long cold and warm phases. Using a modern forecast model as well as observations of the climate system at the start of the forecast, seasonal forecasts draw on the predictability arising from these slows ENSO fluctuations to provide regular predictions of the atmosphere and ocean for the whole globe 1–2 seasons ahead. Such forecasts are important to warn us about potentially dangerous weather and climate conditions in the months to come. How well can we predict ENSO and what confidence do we have in our future abilities to forecast it? Here we present results from an unprecedented extensive model data set of retrospective ENSO forecasts 2 years into the future which were repeated every year from 1901 to 2010. By comparison with observations, we find that our ability to forecast ENSO did not simply degrade the further back in time we went. Rather, the skill had a distinct minimum during the 1930s to 1950s with better forecasts before and after. We explore factors that play a role in explaining this unexpected behavior.

1. Introduction

El Niño Southern Oscillation (ENSO), the large-scale fluctuation of the equatorial atmosphere and ocean over the tropical Pacific region with widespread climatic consequences, is arguably the most predictable climate mode at seasonal timescales and provides the scientific basis for global seasonal climate predictions (Barnston et al., 2010; Shukla et al., 2000; Timmermann et al., 2018; Trenberth et al., 1998). With significant progress made during the last decades in our understanding of the complexity of ENSO, the development of the observing systems, coupled general circulation models and data assimilation techniques for improved forecast initialization schemes, current forecast models can provide effective predictions of ENSO warm and cold events 6–12 months ahead (Barnston et al., 2019; Jin et al., 2008; Johnson et al., 2019; Tang et al., 2018; Tippett et al., 2019; Weisheimer et al., 2009).

However, the challenge of forecasting ENSO is not solved yet. Estimates of ENSO predictability are commonly based on retrospective forecasts (reforecasts) for the last two or three decades. This results in an overall small sample size of ENSO events with substantial uncertainties in skill estimates. In addition, decadal-scale changes in the Pacific background state can make ENSO development in certain decades less predictable than others (Kirtman & Schopf, 1998). These background fluctuations can be seen as spontaneously generated multi-decadal variations in ENSO diversity (Wang et al., 2019; Wittenberg et al., 2014). For example, changes of the tropical Pacific mean state related to the equatorial thermocline have been suggested as a source of interdecadal modulations

Michael Mayer, S. Sharmila, Harry Hendon, Oscar Alves

of the amplitude of interannual variability (Gu & Philander, 1997) with an intensification of the ENSO signal observed in the second half of the twentieth Century.

How can changes in the ENSO amplitude affect its predictability? Kirtman and Schopf (1998) proposed a relationship between the amplitude of interannual variability, forecast skill, and limit of predictability in idealized experiments with a simple coupled model of the tropical Pacific and atmosphere: during decades with a stronger ENSO signal, the forecast skill is higher and the predictability limit longer than during periods with lower amplitudes of interannual variability. This behavior can be explained with the classical theory of the delayed oscillator mechanism which robustly maintains self-sustained oscillations that drive sea surface temperature (SST) anomalies during high-predictability periods. Due to the delayed oscillator's chaotic yet deterministic nature, predictability is largely limited by the growth of initial errors, and the associated potential forecast lead time is likely to be of the order of years. On the other hand, stochastic noise from the atmosphere, for example, westerly wind bursts, can influence the SST anomalies more strongly during epochs when the delayed oscillator is damped by cold SST and easterly wind stress anomalies, leading to reduced predictability.

A first attempt to reforecast historical ENSO events was made by Chen et al. (2004) who run the ENSO model of intermediate complexity of Zebiak and Cane (1987) from 1857 to 2003 by initializing the simulations with reconstructed SST data. They showed that ENSO predictability depends on the time period over which it is estimated: periods with high skill were dominated by strong ENSO events. Large El Niño events were found to be predictable up to 2 years in advance, raising the prospect of skillful long-lead forecasts of ENSO. These results are consistent with the notion that the predictability of ENSO may reside in the prediction of its mean amplitude and not the uncertainty or spread (Kumar & Hu, 2014).

In this paper we move these early studies with simplified models and initialization techniques forward to explore ENSO predictability during the twentieth century in a state-of-the-art dynamical seasonal forecasting system based on an initialized fully coupled global circulation model. CERA-20C (Laloyaux et al., 2018), the Coupled European Reanalysis of the twentieth century from 1901 to 2010, provides an ideal opportunity to initialize long-range reforecast experiments and to study state-dependent predictability over an unprecedented 110-year period. We introduce a new biennial (24-month long) reforecast experiment with a lower-resolution version of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational seasonal prediction system SEAS5 (Johnson et al., 2019) for the CERA-20C period and label these reforecasts as SEAS5-20C, see also Weisheimer et al. (2021). The purpose of SEAS5-20C is twofold: (a) to examine the predictability of ENSO in the presence of multi-decadal climate change and variability in long continuous reforecasts, and (b) to assess the feasibility of extending skillful seasonal forecasts to longer lead times by exploring the predictability limit of ENSO up to 2 years. As such SEAS5-20C substantially extends and advances previous work of reforecasts run over a shorter historical period (Huang, Shin, et al., 2017) or with shorter forecast lead times (Weisheimer et al., 2020). New insight obtained through these experiments is expected to be valuable for the development and interpretation of future ENSO prediction systems and their associated uncertainties.

The paper is structured as follows: Section 2 describes the reforecast experiments, data and metrics. In Section 3 we analyze ENSO forecast skill and its variations throughout the hindcast period. Section 4 contrasts the ENSO mean state and variability in the model reforecasts with those from the verification, and in Section 5 we discuss and conclude our results.

2. Model Experiments, Data, and Skill Metrics

The CERA-20C reanalysis aimed at reconstructing the weather and climate of the coupled atmosphere, ocean, land, ocean waves and sea-ice system for the past period 1901–2010 (Laloyaux et al., 2018). It is used here to provide initial conditions (ICs) for coupled reforecast experiments, and its SST and atmospheric analysis provides the verification data for our ENSO predictions. CERA-20C assimilated observed subsurface temperature and salinity profiles in the ocean and only conventional surface observations (surface pressure and marine winds) in the atmosphere. No satellite-derived data went into the reanalysis. The SSTs were relaxed toward monthly HadISST2 observed data (Titchner & Rayner, 2014). In addition, verification of the ENSO predictions was also performed against ERSSTv5 data (Huang, Thorne, et al., 2017), see Supporting Information S1.

The reforecast experiments were performed with a low-resolution configuration of ECMWF's operational fully coupled forecasting system SEAS5 (Johnson et al., 2019), labeled SEAS5-20C. The experiments were run at atmospheric resolution $T_{co}199$ (approximately 50 km) horizontally with 91 vertical levels and with a 1-degree horizontal ocean resolution (with equatorial meridional refinement) using 42 vertical levels. Historical reconstructions of greenhouse gases (CMIP5) and volcanic stratospheric sulfate aerosol (GISS) were used as radiative forcings similar to SEAS5. The reforecasts have an ensemble size of 10 members, sampling the 10 CERA-20C realizations of ocean ICs. SEAS5-20C reforecasts were started on each 1 November and 1 May from 1901 to 2010 and have a forecast length of 24 months.

Throughout the paper we use the NINO3.4 SST index (Trenberth, 1997) defined as the average SST over the central equatorial Pacific (5°N–5°S, 170°–120°W) and the atmospheric equatorial Southern Oscillation Index (SOI) defined as the difference in standardized mean sea-level pressure between the equatorial western Pacific area (5°N–5°S, 90°–140°E) and an equatorial eastern Pacific area (5°N–5°S, 130°–80°W) as a proxy of ENSO.

As a metric of the skill of the forecasts the ensemble-mean anomaly correlation coefficient (ACC) will be used. It estimates the interannual correlation over N forecast years of the ensemble mean forecast anomaly and the corresponding observed anomaly. For ensemble-mean forecast $\overline{FC}_i$ and observation OBS_i pairs with $i = 1 \dots N$, the ACC is defined as

$$ ACC = \frac{\frac{1}{N}\sum_{i=1}^{N} \overline{FC}'_i * OBS'_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \overline{FC}'^2_i * \frac{1}{N}\sum_{i=1}^{N} OBS'^2_i}} \qquad (1) $$

with $\overline{FC}'_i = \overline{FC}_i - \frac{1}{N}\sum_{i=1}^{N} \overline{FC}_i$ and $OBS'_i = OBS_i - \frac{1}{N}\sum_{i=1}^{N} OBS_i$ being the ensemble-mean forecast and observed anomalies, respectively. The ensemble-mean forecast $\overline{FC}_i$ is given by $\overline{FC}_i = \frac{1}{M}\sum_{j=1}^{M} FC_{i,j}$ where $FC_{i,j}$ with $j = 1 \dots M$ indicates the individual ensemble members of the forecast at year i.

The "*perfect model*" is defined as the concept where instead of observations, each individual ensemble member is, in turn, used as the verification so that the "*perfect model*" tries to forecast itself. The skill of the "*perfect model*" which is not affected by model biases, is also called "*potential skill*" and can be interpreted as the model's estimate of realizable skill in the hypothetical world where the statistics of the observations are identical to the statistics of the model predictions (Kumar et al., 2014). In analogy to Equation 1, the perfect model anomaly correlation coefficient ACC_{PM} is thus defined as

$$ ACC_{PM} = \frac{1}{M}\sum_{j=1}^{M} \frac{\frac{1}{N}\sum_{i=1}^{N} \overline{FC}'_i * FC'_{i,j}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \overline{FC}'^2_i * \frac{1}{N}\sum_{i=1}^{N} FC'^2_{i,j}}} \qquad (2) $$

3. ENSO Forecast Skill

The historical evolution of reforecast skill (ACC, see Equation 1) in predicting seasonal-mean ENSO anomalies as a function of the hindcast period and forecast lead time is shown in Figure 1 for start dates in November (left) and May (right). Here, the skill has been computed for 30-year windows moved across the period 1901–2010 by 1 year, with the resulting scores plotted at the central year of the corresponding window. Significance of the skill is estimated using a one-sided *t*-test at significance level $\alpha = 0.05$. Figures 1a and 1b display the skill of NINO3.4 SSTs, whereas Figures 1c and 1d show the corresponding skill of the atmospheric manifestation of ENSO through the SOI.

Distinct periods of enhanced and statistically significant skill stand out against extended multi-decadal epochs of reduced skill. On the shortest seasonal forecast lead time scales, skill during DJF in Figure 1a (see also Figure 1g for a cross-section at forecast lead times 2–4 months) remains both robust and steadily high from the 1960s onwards with correlations above 0.95, in agreement with recent findings (Huang, Shin, et al., 2017; Liu et al., 2021; Weisheimer et al., 2020). The period between the 1930s and 1950s saw a decline in seasonal forecast
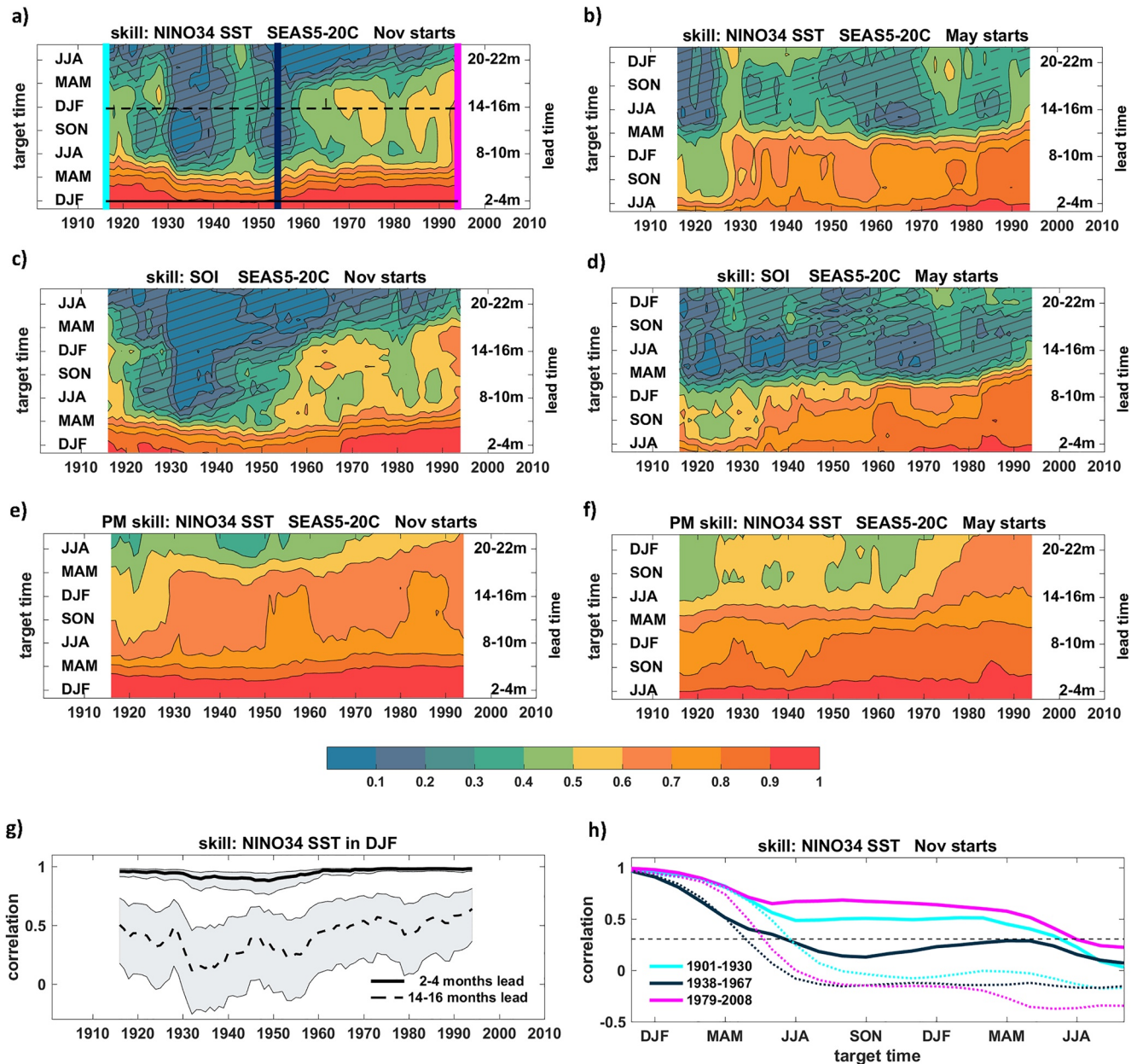
Figure 1. ENSO ensemble-mean correlation skill of SEAS5-20C with CERA-20C as a function of hindcast period on the horizontal axis and forecast lead time on the vertical axis for 1 November (left) and 1 May (right) forecast start dates. (a) and (b) NINO3.4 SST. The solid and dashed horizontal lines indicate cross-sectioned time series of skill shown in (g). The cyan, dark blue and magenta vertical lines indicate cross-sectioned skill as a function of forecast lead time shown in (h). (c) and (d) Southern Oscillation Index (SOI). (e) and (f) NINO3.4 SST perfect model skill. (g) NINO3.4 SST correlation skill in DJF as a function of hindcast period for two distinct lead times. (h) NINO3.4 SST correlation skill as a function of target time for three distinct periods in SEAS5-20C (solid lines) and for simple persistence forecasts (dotted lines). The dashed black line indicates the significance level at $\alpha = 0.05$. For (a)–(g), skill is estimated for 30-year moving windows and plotted at the central year. Hatching in (a)–(f) indicates non-significant skill at $\alpha = 0.05$.

skill and an increase in uncertainty. Interestingly, skill recovered in the very early decades of the century to reach similarly high values as during the most recent period, suggesting that the reduced skill in the mid-twentieth century period is not simply due to poorer quality of the ICs compared to the later periods.

Significant skill up to lead times of approximately 18 months is found for November initialization from the late 1960s onward and, although to a slightly reduced degree, during the first two decades of the twentieth century. However, during the extended period from the 1920s to the 1960s the skill on all lead times is markedly reduced:

once the forecasts hit the first boreal spring season and its associated predictability barrier (Liu et al., 2019; Webster & Yang, 1992), skill drops sharply and becomes non-significant. The time series of NINO3.4 SSTs during DJF for forecast lead times 14–16 months is shown in Figure 1g and demonstrates the overall skill levels and their variations for forecasts started the previous November. We note that the decadal skill modulations are not sensitive to the choice of the observational data set, and similar results are obtained by verifying against ERSSTv5, see Figure S1 in Supporting Information S1.

Multi-decadal variations in ENSO forecast skill are not exclusively related to SST predictions and can also be found in the ACC characteristics for the large-scale atmospheric sea-level pressure based Southern Oscillation Index (SOI, Figures 1c and 1d), supporting the conclusion that the ENSO skill variations are robust in both the surface temperatures of the equatorial Pacific and the atmosphere above.

The "*perfect model*" approach provides a testbed for analyzing how well the model can predict itself, and, because of its independence from observations, is sometimes interpreted as potential skill (Kumar et al., 2014). Figures 1e and 1f showing the perfect model ACC_{PM} (see Equation 2) for both start dates illustrate several interesting points: (a) The skill levels are, in general, higher than for the real-world cases in Figures 1a and 1b indicating that the real-world prediction skill is hampered by initialization and model formulation problems which indicate some potential for improvements; and (b) the multi-decadal variability of perfect model skill is much smoother and temporally not coherent with the *ACC* real-world skill variations. The skill reduction seen between 1930 and 1950 is not present in the perfect model skill, implying that it is an issue related to performance compared to observations and not to the intrinsic predictability in the model itself. The reasons for the discrepancy in decadal variations between the perfect model and observed skill can reside on deficiencies in both model and observations alike. We return to this point later in Section 4 when discussing possible drivers of decadal variations of skill.

The variations of forecast skill throughout the century are further illustrated in Figure 1h showing the lead-time dependent skill evolution for three distinct 30-year reforecast periods together with a benchmark persistence forecast started in November. Here, persistence means the SST anomaly of the ICs has been persisted throughout the forecast. The dynamical forecasts have higher skill during early and late century periods, but noticeably reduced skill from the start of the forecasts during the period 1931–1960. For all three periods skill remains roughly constant for approximately 12 months after the first spring, at levels between 0.4 and 0.7, extending skillful predictions to at least 18 months for certain periods (see also Dunstone et al., 2016). The skill of persistence forecasts for all three epochs falls to zero already after the first spring season, with persistence during the middle period showing a similar rapid skill loss during the first forecast months as the dynamical model.

The reforecasts started in May as shown in Figures 1b, 1d, and 1f also experience substantial multi-decadal modulations of skill, although their temporal variations and lead-time dependence differ from the November reforecasts. The predictability horizon of approx. twelve months is primarily determined by the spring predictability barrier. It is again the most recent period that has the highest skill over the longest lead time, but skill is quickly lost before the 1930s. Forecasts from May are, in general, more difficult because smaller SST anomalies provide weaker constraints on the ICs than for the winter peak ENSO season.

Using moving windows of 20 instead of 30 years for the skill estimation results in overall very consistent patterns, see Figures S2 and S3 in Supporting Information S1. Shorter windows can better detect abrupt skill changes but also introduce more noise into the correlation estimates.

In order to demonstrate the global SST forecast skill of the SEAS5-20C reforecasts, Figure 2 shows the ACC over the full 110-year hindcast period from 1901 to 2010 for different forecast lead times and both start dates. The skill patterns during the first year of the forecast resemble well the skill of ECMWF's operational system SEAS5 (Johnson et al., 2019). While the ENSO skill is substantially reduced on longer lead times beyond 1 year, some areas in the tropical Atlantic and Indian Oceans as well as in the Southern Ocean are skillfully predicted even up to 2 years. We also note that the skill patterns are common for forecasts verifying on a given season, irrespective of the lead time (see for instance forecast verifying in DJF, when ENSO peaks). This suggests that the skill is largely determined by the seasonal phase locking of ENSO (Jin et al., 1996), which is in turn influenced by the properties of the background state.
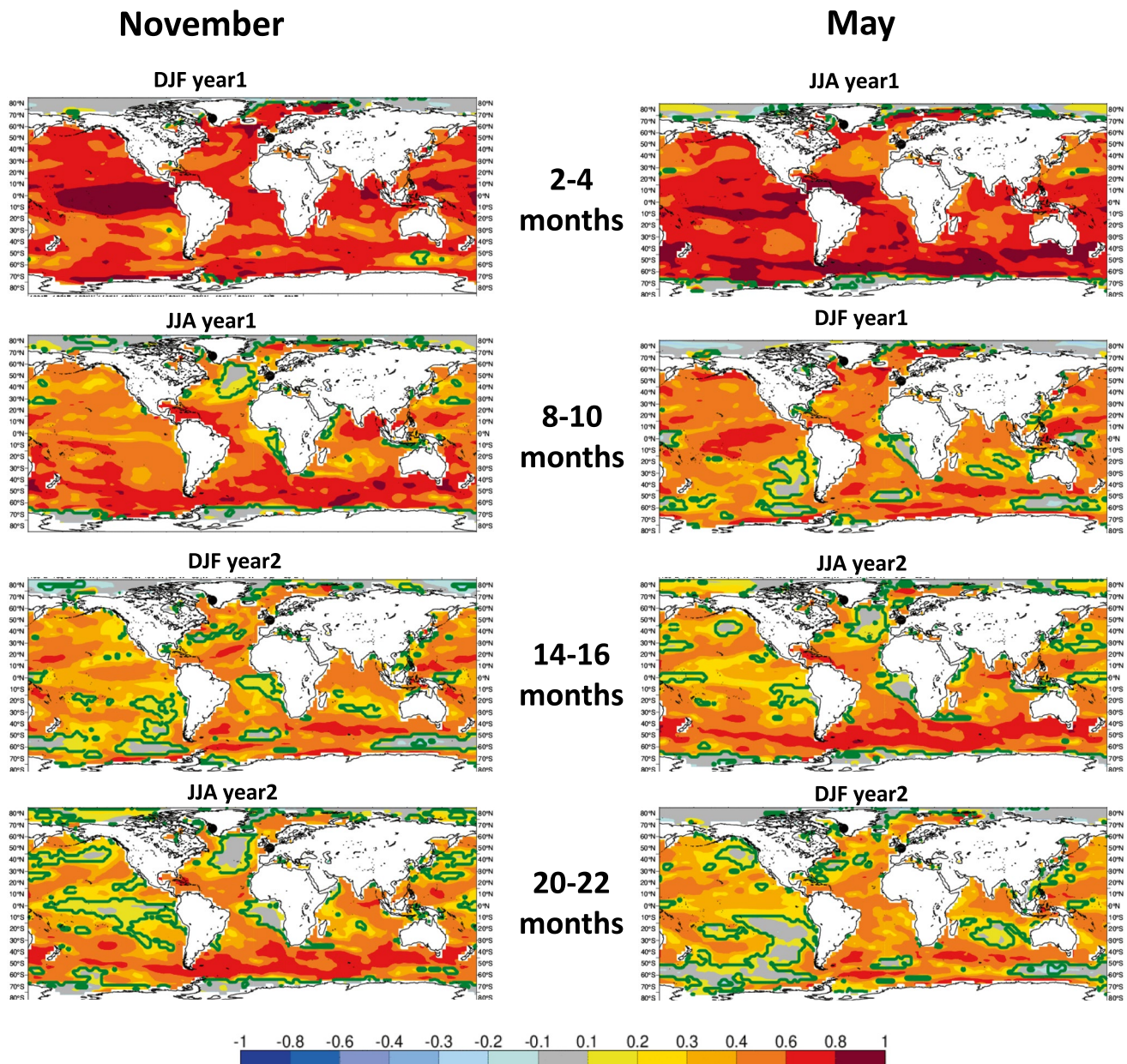
Figure 2. SST correlation skill of SEAS5-20C with CERA-20C over the hindcast period 1901–2010 for different forecast lead times initialized on 1 November (left) and 1 May (right). Green contours indicate where the skill becomes significant at $\alpha = 0.05$.

4. Variations of Mean State and ENSO Amplitude

What could be potential reasons for the pronounced multi-decadal variability of forecast skill throughout the twentieth century? A simple hypothesis could be that the improved data coverage and quality of the observations in more recent decades compared to poorer coverage and perhaps quality at the beginning of the twentieth century, would lead to a reduction in forecast skill the further back in time we go. However, this is clearly not the case, as shown in Figure 1. Instead, the skill variations are highly non-monotonic and the skill levels reached during the early decades of the century are similar in magnitude as for the recent well-observed period. In addition, the skill variability of simple persistence forecasts based on observed anomalies, a measure that is independent of our model and gives highest persistence skill for the early century decades (see Figure 1h), provides further evidence against the data coverage and quality conjecture.
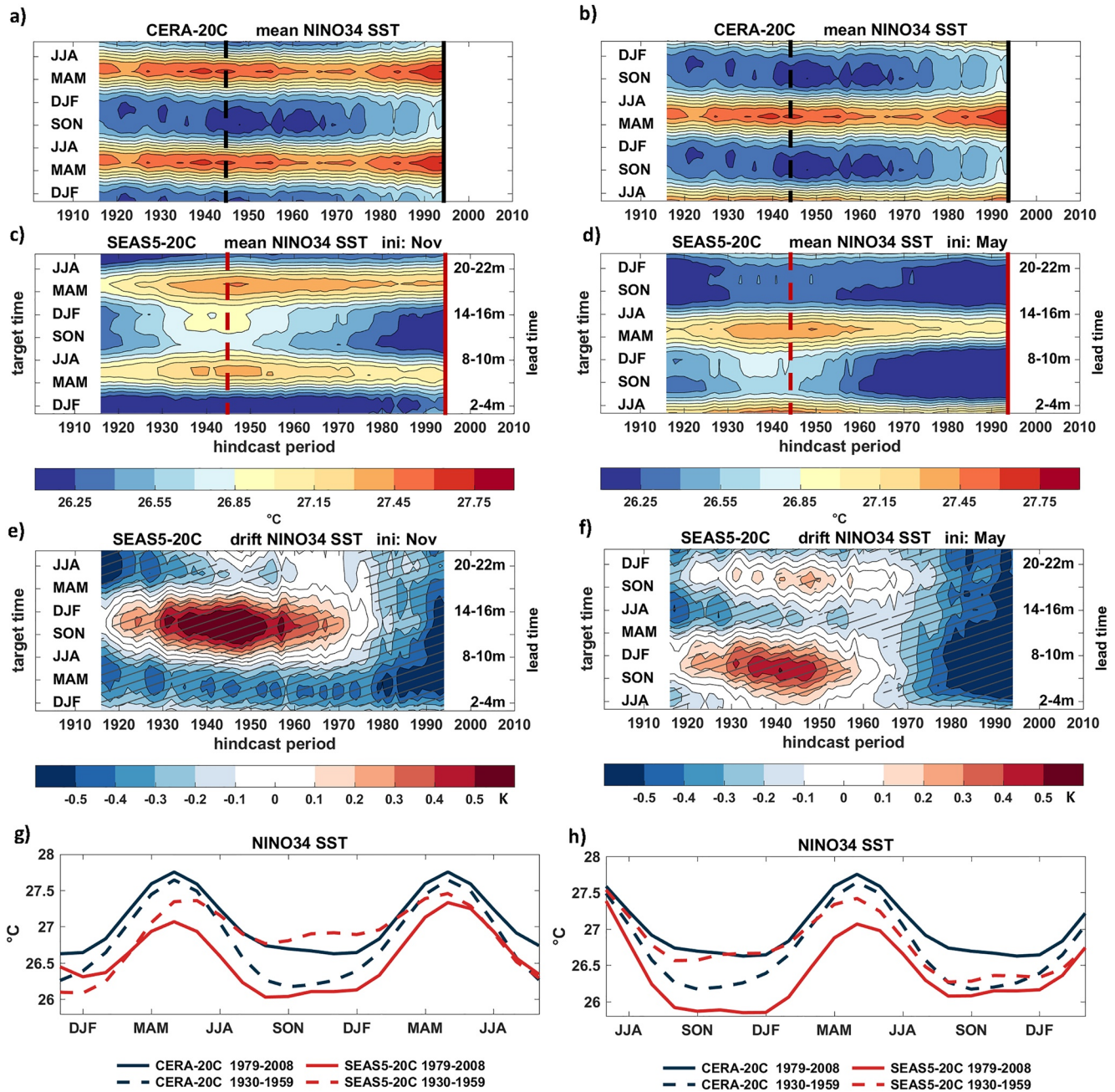
Figure 3. NINO3.4 SST mean-state variations in CERA-20C and SEAS5-20C as a function of hindcast period and season/forecast lead time for 1 November (left) and 1 May (right) forecast start dates. (a) and (b): CERA-20C. (c) and (d): SEAS5-20C. (e) and (f) Bias SEAS5-20C minus CERA-20C. (g) and (h) Mean state for CERA-20C (dark blue) and SEAS5-20C (red) for two distinct 30-year hindcast periods as a function of season (CERA-20C) or forecast target time (SEAS5-20C). Data in (a)–(f) are estimated for 30-year moving windows and plotted at the central year. The solid and dashed vertical lines in (a)–(d) indicate cross-sectioned mean states shown in (g) and (h). Hatching in (e) and (f) indicates significant biases at $\alpha = 0.05$.

An alternative hypothesis is the following: The background state of the coupled ocean-atmosphere system that defines ENSO underwent changes that led to the varying skill levels. Since the ACC skill is based on forecast and observed anomalies with respect to the mean state, any such background changes would need to have affected the forecasts in a non-linear way, and would thus not be fully removed by using anomalies.

In order to investigate how the background ENSO state evolved throughout the hindcast period, Figure 3 shows variations of moving 30-year window mean NINO3.4 SST conditions in CERA-20C (Figures 3a and 3b) and in

the SEAS5-20C hindcasts for November (Figure 3c) and May (Figure 3d) start dates. The vertical axis describes the seasonal cycle of CERA-20C (note that Figures 1a and 1b show exactly the same data but plotted differently for convenience to compare with the model forecasts) and the lead time for the forecasts (Figures 3c and 3d). The observed absolute NINO3.4 SSTs experienced substantial changes throughout the twentieth century with temperatures in boreal winter being coldest in a period from the 1940s to 1970s and a warming during both summer and winter from the 1970s onward.

In contrast, SEAS5-20C SSTs were warmest in the period mid-1930s to mid-1950s for a range of lead times with a cooling trend in the warm and cold season thereafter (Figures 3c and 3d). Displayed as biases of SEAS5-20C with respect to CERA-20C in Figures 3e and 3f, it becomes very apparent that the model has a significant cold bias for all lead time and both start dates during the most recent well-observed decades, but suffers from a substantial and significant warm bias in boreal autumn and winter during the mid-Century periods.

An illustration of the changes in the observed seasonal cycle is shown with the dark blue lines in Figures 3g and 3h which are vertical cross-sections of Figures 3a and 3b for two distinct climate periods. The colder winter SSTs during 1930–1959 (dashed) compared to 1979–2008 (solid) are not reproduced in the hindcasts (red curves) which show much warmer SSTs in the winter season during the earlier period compared to the later period.

In addition to these variations of the SST mean state, the ENSO amplitude as a measure of interannual ENSO variability, or signal strength, also exhibits strong variations during the hindcast period (Figure 4). Here, the ENSO amplitude is defined as the standard deviation of the SSTs during 30-year moving windows (for the model hindcasts the standard deviation is computed from all individual ensemble members). In the CERA-20C reanalyses, the ENSO amplitude during the mature ENSO season DJF was at a minimum during the 1930s to 1960s, followed by a strong positive trend until the early 2000s (see also dark line in Figure 4g).

Encouragingly, SEAS5-20C reproduced this overall behavior well for the short seasonal forecast lead times when initialized in November (Figure 4c and purple line in Figure 4g). It fails, however, to produce any noticeable long-term variations of the ENSO signal for longer lead times (Figure 4d and lighter-colored lines in Figure 4g) and thus substantially underestimates ENSO variability during the second half of the century. The amplitude ratio of SEAS5-20C (defined as the ratio of the standard deviation in SEAS5-20C to the standard deviation in CERA-20C) is displayed in Figures 4e and 4f. The good agreement of the November initialized forecast amplitudes from the 1950s onwards before the spring barrier can clearly be seen (Figure 4e), whereas for longer lead times and the first half of the century the forecast amplitude is significantly smaller than in the reanalysis. In the May initializations, the amplitudes rapidly drop during the first month of the forecast before the 1950s, while in the second half of the hindcast period they are consistent with CERA-20C for approx. three to four months but become significantly damped for the boreal autumn and winter seasons. For both start dates, however, ENSO forecast amplitudes are too strong at the longest lead times during the 1930s to 1950s.

5. Discussion and Conclusions

New multiannual reforecasts of the twentieth century to study the predictability of ENSO have been presented. The novelty of these experiments lies in the combination of the extensive length of the reforecast period (110 years) with the long forecast lead time of 2 years in a fully coupled and initialized state-of-the-art global circulation model. Two reforecast start dates in May and November each year from 1901 to 2010 with an ensemble of 10 members are available.

A moving-window analysis of correlation forecast skill shows that distinct periods of enhanced skill stand out against extended multi-decadal epochs of reduced skill (Figure 1). ENSO prediction skill for DJF from initialization in November is steadily high from the 1950s onwards, in agreement with findings of Huang, Shin, et al. (2017) and Weisheimer et al. (2020). The skill temporarily drops during the periods centered between 1930–1950 and regains high levels in the first two decades of the century (see also Chen et al., 2004; Liu et al., 2021), a behavior that is also reflected on longer forecast lead times.

The reduced level of forecast skill seen before the well-observed recent period raises the question whether fewer ocean observations in earlier periods might be the fundamental cause. The fact that skill is similarly high during the early decades of the twentieth century contradicts this suggestion. The interim skill reduction cannot simply be attributed to the inferior quality of the ICs because the skill reached equally high levels at the beginning of
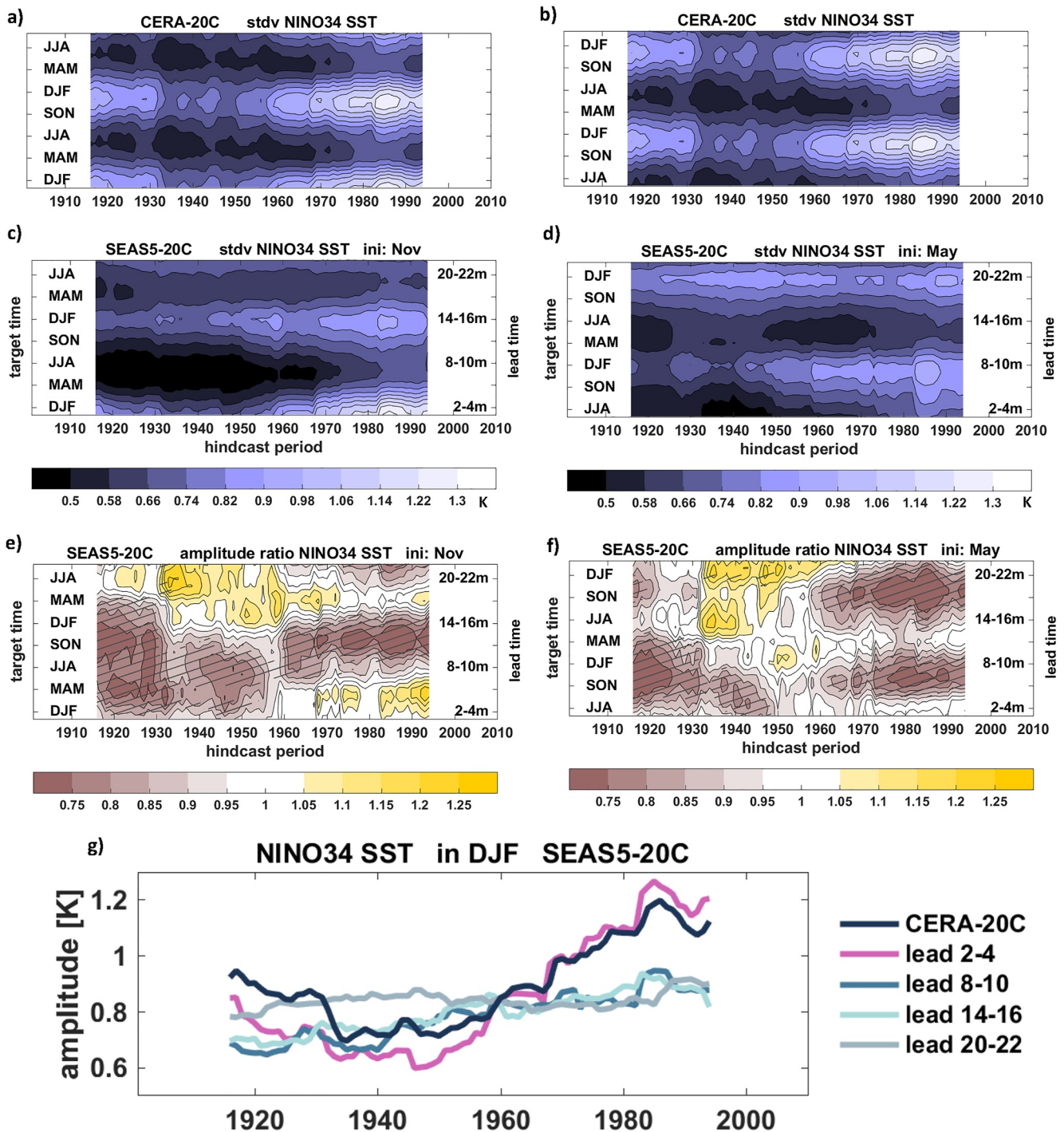
Figure 4. NINO3.4 SST standard deviation (amplitude) variations in CERA-20C and SEAS5-20C as a function of hindcast period and season/forecast lead time for 1 November (left) and 1 May (right) forecast start dates. (a) and (b) CERA-20C. (c) and (d) SEAS5-20C. (e) and (f) Amplitude ratio SEAS5-20C to CERA-20C. (g) Time series of amplitude in DJF for CERA-20C (black) and the SEAS5-20C forecasts with different lead times in months (colored lines). Data in (a)–(d) are estimated for 30-year moving windows and plotted at the central year. Hatching in (e) and (f) indicates amplitude ratios significantly different from 1 at $\alpha = 0.05$.

the twentieth century, when observations of the ocean were sparse. Rather, the low-skill period broadly coincides with years when (a) the discrepancies in the SST mean state and seasonal cycle in observations and the model are largest with the model developing a strong warm bias and weakened seasonal cycle (Figure 3), (b) the observed ENSO amplitude is very weak (Figure 4), and (c) the persistence of observed SST anomalies is low (Figure 1h).

The observational coverage of the tropical Pacific saw a drop during the 1940s (Huang et al., 2020) which raises the question whether the increased uncertainty in the ICs impacted the forecast skill during this period. In order to estimate the sensitivity of our results to the specific SST product used, we repeated our analysis using the ERSSTv5 (Huang, Thorne, et al., 2017) data set for verification. A comparison of the ERSSTv5 and CERA-20C (HadISST) data in terms of NINO3.4 SST mean state and variability and the corresponding model performance for biases and skill can be found in Figures S3–S5 in Supporting Information S1. The agreement in forecast skill suggests that these two observational data sets might share similarities in their methodology and in situ data. However, the lack of input data sets over the tropical Pacific precludes being able to definitively rule out related errors. On the other hand, the overall very good agreement of forecast skill and its multi-decadal modulation implies that reduced ENSO skill during the 1930s to 1950s is not purely a result of the increased uncertainties in the observations or a potential temporal data issue in HadISST during the 1930s to 1950s. Rather, the robust findings using two different verification data sets and the atmospheric manifestation of ENSO through the SOI (see Figures 1c and 1d) strongly point toward a genuine attribute of predictive skill arising from changes in the mean state and variability.

The interpretation of the periods of decreased skill as a manifestation of multi-decadal variations in predictability associated with modulations in the strength of the ENSO signal is fully consistent with the delayed oscillator mechanism discussed in Kirtman and Schopf (1998). Capturing these variations in background state and associated impacts on ENSO variability and (perfect model) predictability is still a major challenge for state-of-the-art general circulation models used in seasonal and decadal predictions. Indeed, there are large discrepancies between the decadal modulations of SEAS5-20C real and perfect model skill, which need to be further understood. A simple analysis with detrended time series (Figure S6 in Supporting Information S1) indicates that the increased perfect model skill during the latest period of the record cannot be attributed to the presence of trends in SSTs (L'Heureux et al., 2022).

By contrasting the varying characteristics of ENSO mean state and variability throughout the hindcast period as diagnosed in Figures 3 and 4 with the skill variations reported in Figure 1, we find that the overall higher levels of forecast skill during the second half of the twentieth century found in the November hindcasts concur with periods of strong ENSO amplitudes in both the observations and the shorter forecasts ranges, lending support for the suggested link between ENSO variability and predictability (Capotondi et al., 2015; Chen et al., 2004; Gu & Philander, 1997; Suarez & Schopf, 1988; Zhao et al., 2016) which reflects a reduction of the signal-to-noise ratio (Huang et al., 2021; Kumar, 2009).

The sharp decrease in prediction skill during boreal spring, known as the boreal spring predictability barrier (Liu et al., 2019; Webster & Yang, 1992), is a critical challenge for ENSO predictions on longer time ranges. We find that forecast skill beyond the spring barrier varies between decades of substantially longer skill from the 1980s onward (see also Balmaseda et al., 1995; Fang et al., 2019) and prolonged periods of abrupt and significant skill drops during spring. Once the forecast skill extends beyond the first boreal spring barrier, the predictability limit is much enhanced. For example, Dunstone et al. (2016) reported significant ENSO skill into the second winter and Mayer and Balmaseda (2021) found useful information of SEAS5 in year 2 in the specific case of the strong 1997/98 and 2014–2016 ENSO periods. Our results provide further support for the feasibility of skillful ENSO forecasts up to 18 months.

Questions as to changes in the spatial ENSO characteristics and which forced or unforced mechanisms exactly govern periods with enhanced long-range skill (Timmermann et al., 2018; Wengel et al., 2018) will require more detailed research in the future. We conclude by noting that these extensive reforecast experiments can potentially also become useful data sets to train deep learning algorithms and other machine learning techniques (Ham et al., 2019; LeCun et al., 2015; Rasp et al., 2018) in order to facilitate the development of powerful and reliable dynamical long-range forecasting systems for the future.

Data Availability Statement

CERA-20C data is available through Laloyaux et al., 2018. The SEAS5-20C reforecast experiment data used in this study (ECMWF, 2021) are available under a Creative Commons Attribution 4.0 International license (CC BY 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

References

Balmaseda, M. A., Davey, M. K., & Anderson, D. L. T. (1995). Decadal and seasonal dependence of ENSO prediction skill. *Journal of Climate*, *8*, 2705–2715. https://doi.org/10.1175/1520-0442(1995)008<2705:dasdoe>2.0.co;2

Barnston, A. G., Li, S., Mason, S. J., DeWitt, D. G., Goddard, L., & Gong, X. (2010). Verification of the first 11 years of seasonal climate forecasts. *Journal of Applied Meteorology*, *49*, 493–520. https://doi.org/10.1175/2009jamc2325.1

Barnston, A. G., Tippett, M. K., Ranganathan, M., & L'Heureux, M. L. (2019). Deterministic skill of ENSO predictions from the North American multimodel ensemble. *Climate Dynamics*, *53*, 7215–7234. https://doi.org/10.1007/s00382-017-3603-3

Capotondi, A., Wittenberg, A. T., Newman, M., Di Lorenzo, E., Yu, J.-Y., Braconnot, P., et al. (2015). Understanding ENSO diversity. *Bulletin of the American Meteorological Society*, *96*(6), 921–938. https://doi.org/10.1175/BAMS-D-13-00117.1

Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E., & Huang, D. (2004). Predictability of El Niño over the past 148 years. *Nature*, *428*, 733–736. https://doi.org/10.1038/nature02439

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*, 809–814. https://doi.org/10.1038/ngeo2824

ECMWF. (2021). *Global monthly reforecast sea surface temperatures from 1901 to 2010 with 24-month lead time*. ECMWF. Retrieved from https://apps.ecmwf.int/research-experiments/expver/guxf/. https://doi.org/10.21957/fzf9-te33

Fang, X.-H., Zheng, F., Liu, Z.-Y., & Zhu, J. (2019). Decadal modulation of ENSO spring persistence barrier by thermal damping processes in the observation. *Geophysical Research Letters*, *46*6892. https://doi.org/10.1029/2019gl082921

Gu, D., & Philander, S. G. H. (1997). Interdecadal climate fluctuation that depend on exchanges between the tropics and the extratropics. *Science*, *275*, 805–807. https://doi.org/10.1126/science.275.5301.805

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*, 568–572. https://doi.org/10.1038/s41586-019-1559-7

Huang, B., Menne, M. J., Boyer, T., Freeman, E., Gleason, B. E., Lawrimore, J. H., et al. (2020). Uncertainty estimates for sea surface temperature and land surface air temperature in NOAAGlobalTemp Version 5. *Journal of Climate*, *33*, 1351–1379. https://doi.org/10.1175/jcli-d-19-0395.1

Huang, B., Shin, C.-S., Kumar, A., L'Heureux, M., & Balmaseda, M. (2021). The relative roles of decadal climate variations and changes in the ocean observing system on seasonal prediction skill of tropical Pacific SST. *Climate Dynamics*, *56*, 3045–3063. https://doi.org/10.1007/s00382-021-05630-1

Huang, B., Shin, C.-S., Shukla, J., Marx, L., Balmaseda, M. A., Halder, S., et al. (2017). Reforecasting the ENSO events in the past 57 years (1958–2014). *Journal of Climate*, *30*, 7669–7693. https://doi.org/10.1175/jcli-d-16-0642.1

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended reconstructed sea surface temperature version 5 (ERSSRv5): Upgrades, validations and intercomparisons. *Journal of Climate*, *30*, 8179–8205. https://doi.org/10.1175/jcli-d-16-0836.1

Jin, E. K., Kinter, J. L., III, Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B. P., et al. (2008). Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dynamics*, *31*, 647–664. https://doi.org/10.1007/s00382-008-0397-3

Jin, F.-F., Neelin, J. D., & Ghil, M. (1996). El Niño/Southern oscillation and the annual cycle: Subharmonic frequency locking and aperiodicity. *Physica D*, *98*, 442–465. https://doi.org/10.1016/0167-2789(96)00111-x

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

Kirtman, B. P., & Schopf, P. S. (1998). Decadal variability in ENSO predictability and prediction. *Journal of Climate*, *11*, 2804–2822. https://doi.org/10.1175/1520-0442(1998)011<2804:DVIEPA>2.0.CO;2

Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal predictions. *Monthly Weather Review*, *137*, 2622–2631. https://doi.org/10.1175/2009mwr2814.1

Kumar, A., & Hu, Z.-Z. (2014). How variable is the uncertainty in ENSO sea surface temperature prediction? *Journal of Climate*, *27*, 2779–2788. https://doi.org/10.1175/jcli-d-13-00576.1

Kumar, A., Peng, P., & Chen, M. (2014). Is there a relationship between potential and actual skill? *Monthly Weather Review*, *142*, 2220–2227. https://doi.org/10.1175/MWR-D-13-00287.1

Laloyaux, P., De Boisseson, E., Balmaseda, M. A., Bidlot, J.-R., Broennimann, S., Buizza, R., et al. (2018). CERA-20C: A coupled reanalysis of the twentieth century. *Journal of Advances in Modeling Earth Systems*, *10*, 1172–1195. https://doi.org/10.1029/2018ms001273

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

L'Heureux, M. L., Tippett, M. K., & Wang, W. (2022). Prediction challenges from errors in tropical Pacific sea surface temperature trends. *Frontiers in Climate*. https://doi.org/10.3389/fclim.2022.837483

Liu, T., Song, X., Tang, Y., Shen, Z., & Tan, X. (2021). ENSO predictability over the past 137 years based on a CESM ensemble prediction system. *Journal of Climate*, *35*, 763–777. https://doi.org/10.1175/JCLI-D-21-0450.1

Liu, Z., Jin, Y., & Rong, X. (2019). A theory for the seasonal predictability barrier: Threshold, Timing and Intensity. *Journal of Climate*, *32*, 423–443. https://doi.org/10.1175/jcli-d-18-0383.1

Mayer, M., & Balmaseda, M. A. (2021). Indian Ocean impact on ENSO evolution 2014–2016 in a set of seasonal forecasting experiments. *Climate Dynamics*, *56*, 2631–2649. https://doi.org/10.1007/s00382-020-05607-6

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*, 9684–9689. https://doi.org/10.1073/pnas.1810286115

Shukla, J., Anderson, J., Baumhefner, D., Brankovic, C., Chang, Y., Kalnay, E., et al. (2000). Dynamical seasonal prediction. *Bulletin of the American Meteorological Society*, *81*, 2593–2606. https://doi.org/10.1175/1520-0477(2000)081<2593:dsp>2.3.co;2

Suarez, M. J., & Schopf, P. S. (1988). A delayed action oscillator for ENSO. *Journal of Atmospheric Science*, *45*, 3283–3287. https://doi.org/10.1175/1520-0469(1988)045<3283:adaofe>2.0.co;2

Tang, Y., Zhang, R.-H., Liu, T., Duan, W., Yang, D., Zheng, F., et al. (2018). Progress in ENSO prediction and predictability study. *National Science Review*, *5*, 826–839. https://doi.org/10.1093/nsr/nwy105

Timmermann, A., An, S.-I., Kug, J.-S., Jin, F.-F., Cai, W., Capotondi, A., et al. (2018). El Niño-Southern oscillation complexity. *Nature*, *559*(7715), 535–545. https://doi.org/10.1038/s41586-018-0252-6

Tippett, M. K., Ranganathan, M., L'Heureaux, M., Barnston, A. G., & DelSole, T. (2019). Assessing probabilistic predictions of ENSO phase and intensity from the North American multimodel ensemble. *Climate Dynamics*, *53*, 7497–7518. https://doi.org/10.1007/s00382-017-3721-y

Titchner, H. A., & Rayner, N. A. (2014). The Met office hadley centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research*, *119*, 2864–2889. https://doi.org/10.1002/2013jd020316

Trenberth, K. E. (1997). The definition of El Niño. *Bulletin of the American Meteorological Society*, *78*, 2771–2777. https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., & Ropelewski, C. (1998). Progress during TOGA in understanding and modelling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research*, *103*, 14291–14324. https://doi.org/10.1029/97jc01444

Wang, B., Luo, X., Yang, Y.-M., Sun, W., Cane, M. A., Cai, W., et al. (2019). Historical change of El Nino properties sheds light on future changes of extreme El Nino. *Proceedings of the National Academy of Sciences*, *116*, 22512–22517. https://doi.org/10.1073/pnas.1911130116

Webster, P. J., & Yang, S. (1992). Monsoon and ENSO: Selectively interactive systems. *Quarterly Journal of the Royal Meteorological Society*, *118*, 877–926. https://doi.org/10.1002/qj.49711850705

Weisheimer, A., Balmaseda, M., Stockdale, T., Mayer, M.,De Boisseson, E., Senan, R., & Johnson, S. (2021). Retrospective two-year ENSO predictions during the 20th century. *ECMWF Newsletter*, *169*, 6–7. https://www.ecmwf.int/en/newsletter/169/news/retrospective-two-year-enso-predictions-during-20th-century

Weisheimer, A., Befort, D., MacLeod, D., Palmer, T. N., O'Reilly, C., & Strommen, K. (2020). Seasonal forecasts of the 20th century. *Bulletin of the American Meteorological Society*, *101*(8), E1413–E1426. https://doi.org/10.1175/BAMS-D-19-0019.1

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., et al. (2009). Ensembles: A new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, *36*, L21711. https://doi.org/10.1029/2009gl040896

Wengel, C., Dommenget, D., Latif, M., Bayr, T., & Vijayeta, A. (2018). What controls ENSO amplitude diversity in climate models? *Geophysical Research Letters*, *45*, 1989–1996. https://doi.org/10.1002/2017GL076849

Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A., & Zeng, F. R. (2014). ENSO modulation: Is it decadally predictable? *Journal of Climate*, *27*, 2667–2681. https://doi.org/10.1175/jcli-d-13-00577.1

Zebiak, S., & Cane, M. (1987). A model El Nino – Southern oscillation. *Monthly Weather Review*, *115*, 2262–2278. https://doi.org/10.1175/1520-0493(1987)115<2262:ameno>2.0.co;2

Zhao, M., Hendon, H. H., Alves, O., Liu, G., & Wang, G. (2016). Weakened eastern Pacific El Niño predictability in the early twenty-first century. *Journal of Climate*, *29*, 6805–6822. https://doi.org/10.1175/JCLI-D-15-0876.1