npj climate and atmospheric science

www.nature.com/npjclimatsci

ARTICLE OPEN

Check for updates

Multi-decadal variation of ENSO forecast skill since the late 1800s

Jiale Lou[1,2,3]✉, Matthew Newman[2] and Andrew Hoell[2]

Diagnosing El Niño-Southern Oscillation (ENSO) predictability within operational forecast models is hindered by computational expense and the need for initialization with three-dimensional fields generated by global data assimilation. We instead examine multi-year ENSO predictability since the late 1800s using the model-analog technique, which has neither limitation. We first draw global coupled model states from pre-industrial control simulations, from the Coupled Model Intercomparison Project Phase 6, that are chosen to initially match observed monthly sea surface temperature and height anomalies in the Tropics. Their subsequent 36-month model evolution are the hindcasts, whose 20th century ENSO skill is comparable to twice-yearly hindcasts generated by a state-of-the-art European operational forecasting system. Despite the so-called spring predictability barrier, present throughout the record, there is substantial second-year ENSO skill, especially after 1960. Overall, ENSO exhibited notably high values of both amplitude and skill towards the end of the 19th century, and again in recent decades.

npj Climate and Atmospheric Science (2023)6:89 ; https://doi.org/10.1038/s41612-023-00417-z

INTRODUCTION

Seasonal-to-interannual prediction skill and potential predictability of El Niño-Southern Oscillation (ENSO) have varied over the past several decades[1–6]. Higher forecast skill appears evident during periods of larger amplitude ENSO events[2,7], but whether such changes in ENSO forecast skill represent inherent changes in potential predictability remains an active research question[8–10]. For example, multi-decadal changes in the climate base states might modulate ENSO characteristics, making its evolution less predictable in some decades than in others[8–11]. How internal and externally forced processes could combine to drive such base state changes is itself unclear[7]. Alternatively, event-to-event variations in ENSO amplitude and pattern might occur by chance, driven by unpredictable atmospheric weather, with related long-term – but similarly unpredictable – variations in ENSO forecast skill[12–16]. Some studies have found that ENSO predictions might be skillful for forecast leads of at least two years, but again with substantial event-to-event variation in skill[7,17–21].

Characterizing decadal variations in ENSO skill requires sufficiently long hindcast records, that is, retrospective forecasts initialized using only those observations potentially available at the forecast time. Unfortunately, most hindcast datasets generated by coupled climate models typically cover only 20–30 years[22], due to computational cost and the need for three-dimensional global analyses for their initialization. Some longer hindcast records have more recently been developed[1,20,22]; however, these datasets are limited, either by restricting the frequency of initialization to only a few times a year, and/or by capping the forecast lead times at one year.

Here, the model-analog approach[23,24] is used to extend ENSO hindcasts back to the late 1800s and to investigate the long-term variation of ENSO forecast skill, including its dependence upon the seasonal cycle. In traditional analog forecasting, observed states are found that are (in some sense) close matches to the current initial state, and their evolution in the past is used to make a current forecast[25]. This approach may not be too skillful if the available library of previous climate states is limited[26]. However, climate simulations in excess of 500 years could provide sufficient data for analog forecasting[23]. That is, their more extensive databases can be used to generate ensembles of suitable initial states. Then, how these states evolve within the model simulations provides the forecast ensemble, with no additional model integration needed, making the model-analog climate predictions computationally efficient. Such model-analog ensembles have been shown to yield ENSO prediction skill comparable to traditional assimilation-initialized hindcasts made by the same operational climate models[23]. The model-analog technique has been successfully applied to, for example, seasonal forecasts[23,24], decadal predictions[27], and climate projections[28].

There are a few additional advantages of using model-analog ensembles to make hindcasts for an extended historical period. First, both the initial conditions (i.e., model-analog states) and the subsequent model-analog ensemble lie entirely within the model space, which may avoid initialization shock[29] commonly seen in initialized forecast systems, arising primarily from an imbalance between the initial analyzed state and all model states. Moreover, the selection of these model-analog states based upon only a few monthly-averaged variables, such as sea surface temperature anomalies (SSTA) and sea surface height anomalies (SSHA), nevertheless appears sufficient to generate hindcasts that capture monthly tropical Indo-Pacific skill found in operational model hindcasts. This appears true even for the skill of some variables not used to choose the initial model-analog ensemble, notably including precipitation, since any variable output as part of the model simulations is also immediately part of the output model-analog forecast ensemble[23]. Moreover, the large number of pre-existing climate model simulations (for example, the CMIP6 archive) means that multi-model ensembles, which typically improve overall prediction skill[30], may be easily constructed[23]. These advantages make the model-analog technique suitable for investigating the long-term variations of ENSO forecast skill.

[1]Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA. [2]NOAA Physical Sciences Laboratory, Boulder, CO, USA. [3]Present address: Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, NJ, USA. ✉email: jiale.lou@noaa.gov

In this study, we construct model-analog hindcasts for leads of 1–36 months based on over two dozen different CMIP6 models, with seven different sea surface datasets used both for monthly model-analog initialization and for hindcast verification, allowing us to make hindcasts starting in the late 1800s, and (for one dataset) as far back as 1854. ENSO prediction skill is determined from both SSTA and sea level pressure anomalies (SLPA), measured with the NINO3.4 and equatorial Southern Oscillation (eqSOI) indices (whose definitions are in Methods). Monthly anomalies used both to determine the model-analogs and to verify the subsequent hindcasts are computed by removing a fair-sliding[31] climatology, in which a 30-year sliding window behind the forecast is used as the reference period (see Methods). This approach avoids introducing information not known at the time of the hindcast; also, it largely acts to detrend the anomalies. Complete details of our model-analog approach, as well as the datasets used, skill metrics, and how the multi-model model-analog hindcast ensemble was constructed, are in the Methods section.

RESULTS

Comparison of model-analog and dynamical model hindcast skill

We begin by comparing the seasonal forecast skill of model-analog hindcasts for the years 1901–2009, initialized with observed SSTA and SSHA drawn from the CERA-20C reanalysis, to the skill of the SEAS5-20C hindcasts[1], created with a lower-resolution version of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational seasonal prediction system SEAS5[32] and initialized with the CERA-20C reanalysis on November 1 and May 1 of each year, in Fig. 1. Variations of the seasonal mean skill, measured by anomaly correlation (AC) of the model-analog NINO3.4 predictions with the CERA-20C SSTA verifications, are shown in Fig. 1a and b for October and April initializations. Note that because model-analogs are initialized with monthly anomalies, a fair skill comparison to traditional operational models requires using model-analogs starting with the prior month; e.g., we compare October-initialized model-analog hindcasts with November 1-initialized SEAS5 hindcasts. AC skill was computed over 30-year sliding windows, advancing by increments of one year within the period of 1901–2009, with skill plotted at the central year of the corresponding 30-year window[1]. The October initialization (Fig. 1a) generally has better skill at shorter time leads than the April initialization (Fig. 1b), consistent with ENSO phase locking[33] and the ENSO spring predictability barrier[34]. Also, for both initializations there is a return of skill for hindcasts verifying during SON/DJF of the following year (that is, when predicting the second-year ENSO). There also appears to be a notable increase in skill starting around 1960. These variations of model-analog hindcast skill (Fig. 1a and b) are largely insensitive to the SST dataset used as verification (Supplementary Fig. 1).

Both the SEAS5-20C (Fig. 1c and d) and model-analog hindcasts show generally comparable skill with similar notable features, including the return to skill for second year ENSO and the increased forecast skill starting around 1960. We also examined the root mean square skill score (RMSSS) of the NINO3.4 index for both sets of hindcasts (Supplementary Fig. 3) and found similar evolution of skill. Compared to the November 1 SEAS5-20C initializations, the model-analogs had higher skill – often significantly so – for hindcasts with leads beyond about 6 months that verified during seasons ranging from JJA to DJF (also see skill difference in Supplementary Fig. 2), although these differences are much greater prior to 1960. In contrast, for both initialization months, the model-analog hindcast ensemble had more pronounced springtime skill minima than did the SEAS5-20C hindcasts (hatching area in Fig. 1; see also Supplementary Fig. 2), although this difference was also diminished in the most recent few decades.

Some of the higher ENSO skill seen comparing Fig. 1a and b to Fig. 1c and d could be due to aspects of the model-analog technique. As noted earlier, model-analogs do not suffer from initialization shock. Additionally, initialization errors in variables other than SSTA and SSHA could degrade the hindcast skill of SEAS5-20C but are irrelevant to the model-analogs. This might be particularly consequential prior to about 1960, when the three-dimensional fields of the CERA-20C are constrained by fewer observations. However, at least some of the skill improvement appears due to the use of a multi-model ensemble; for example, skill computed from the CESM2 model-analogs alone (Supplementary Fig. 4) shows more modest improvement relative to the SEAS5-20C November 1 initializations, and slightly lower overall skill relative to the SEAS5-20C May 1 initializations. Of course, the boost in ENSO skill from the use of a multi-model ensemble has been described previously[18,30], but it is a notable advantage of model-analogs since they can be constructed from many different available model simulations. On the other hand, poorer springtime skill in the model-analogs prior to the satellite era might reveal a deficiency of the technique and/or be a consequence of missing important information that would otherwise constrain the choice of the analogs; for example, choosing analogs only within the tropical domain could miss extratropical forcing[35] of meridional modes also relevant to ENSO evolution[36–38]. In summary, model-analog hindcasts based upon CMIP6 models, initialized only with SSTA and SSHA, can largely capture the decadal variations in lead-dependent ENSO forecast skill found by a traditionally-assimilated full-field initialization of an operational forecast model, supporting our extension of the earlier post-1960 model-analog analysis of ENSO skill variations[24] to the entire 20th century.

Seasonal variations of ENSO skill

ENSO forecast skill is known to have a strong seasonal dependence[16] and Fig. 1 seems to suggest this dependence may itself vary over the 20th century. However, the SEAS5-20C hindcasts were only initialized twice a year and therefore have limited utility to fully evaluate the seasonal dependence of ENSO skill (see also in Ref. [20]). On the other hand, because of their minimal computational cost and comparable skill, model-analogs can be used for this purpose. Also, given the apparent return of second-year ENSO skill in Fig. 1, and since there is no additional cost, we can extend the model-analog ENSO hindcasts to leads of up to 36 months.

As noted earlier, the model-analog hindcasts reproduce (and, except for the last few decades, may overdo) the so-called "spring predictability barrier", a common feature of both numerical and statistical ENSO forecasts[16,39,40], where predictions made just before and during boreal spring (i.e., initial month within February through April) have notably lower forecast skill than those made during the rest of the year. Nevertheless, there is currently no consensus on the degree to which the spring predictability barrier may limit ENSO prediction skill for forecast leads greater than one year[2]. In model-analog hindcasts, the springtime forecast skill minima are quite clear in Fig. 2a–d, which show the seasonal cycle of skill (as a function of initialization month and forecast lead) for both the entire record and for the 1901–1930, 1931–1970, and 1971–2009 periods. There is a pronounced tilted forecast skill minimum for ENSO predictions verifying during boreal summer for almost all leads (Fig. 2). Interestingly, forecast skill returns for even longer leads; for example, a forecast initialized in late summer will see skill that first slowly declines, then more rapidly declines at leads of ~6–9 months, and then *increases* with increasing lead to plateau at about 15 months, finally decreasing again for longer leads (Fig. 2). The skill minimum might therefore reflect temporarily lower boreal summer signal-to-noise ratios[16] rather

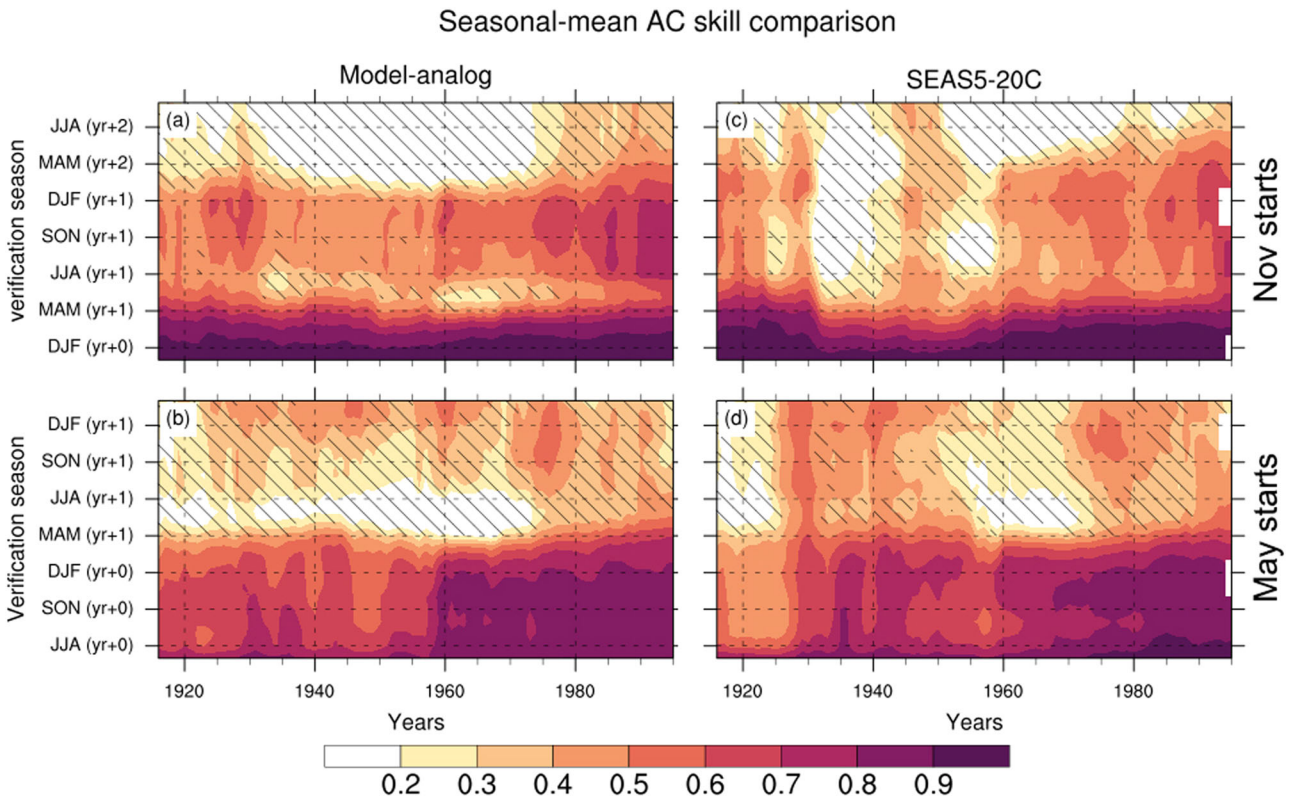

Fig. 1 Comparison of ensemble-mean seasonal-averaged anomaly correlation (AC) skill of NINO3.4 predictions for model-analog and numerical model hindcasts. Ensemble-mean AC skill of NINO3.4 predictions as a function of hindcast period on the horizontal axis and forecast lead time on the vertical axis. **a**, **b** Model-analog hindcasts initialized in October and April and determined using SSTA + SSHA. **c**, **d** SEAS5-20C hindcasts initialized on November 1st and May 1st, respectively; note that these are essentially the same as Fig. 1a and b in Ref. [1]. Correlation coefficients for each 30-year sliding hindcast window are shown at the central year (for example, the value for 1940 represents skill for the 1926–1955 period). Hatching indicates the correlation coefficients that are not statistically significant at the 95% significance level, following Ref. [58]. To make a fair comparison (see Methods), the November 1/May 1 initialized SEAS-20C should be compared to October/April initialized model-analog hindcasts.

than a barrier to skill at longer leads. The main features of the spring predictability barrier appear to be robust over the entire 20th century, since they exist to varying degrees for each of the periods, despite variations in overall forecast skill. However, the forecast skill minimum appears much less pronounced, and more confined to the early summer months, in recent decades; while skill increases for almost all months and leads between the 1931–1970 and 1971–2009 periods, the greatest increase in skill (~0.3–0.4; cf. Fig. 2c and d) occurs for hindcasts verifying in late spring. Moreover, similar features are evident for eqSOI skill (Fig. 2e–h), which also are robust to the SLP dataset used for verification (e.g., HadSLP2 as shown in Supplementary Figure 5).

We also find some third-year skill for both NINO3.4 and eqSOI for hindcasts initialized during the latter half of the year (Fig. 2), although only during the early period (1901–1930) was this skill statistically significant (Fig. 2b, f). Again, this third-year skill occurs after a (second) springtime skill minimum. It is possible that the increased third year ENSO forecast skill found in the early period may be attributed to data sampling issues or to specific ENSO events and ENSO precursors during that time, which warrants further investigation. Notwithstanding, the results here support the conclusion that multi-year ENSO forecast skill exhibits substantially higher values in both the early period and recent decades. Since there is some uncertainty about the robustness of recent third-year skill, we focus on leads up to 24 months for the remainder of this paper.

As Fig. 1b suggested, longer-lead ENSO forecast skill notably increased over the last few decades. For example, in the 1971-2009 period, NINO3.4 (eqSOI) AC values were above 0.6 (0.5) for predictions made in summer for the winter that followed about 18 months later (Fig. 2d, h). This recent second-year winter ENSO forecast skill was especially elevated compared to midcentury skill for the same leads (Fig. 2c, g), with differences that were sometimes statistically significant (at the 95% level), especially for eqSOI (cf. Figure 2g, h). Differences between second-year winter forecast skill for the 1901–1930 period and the two later periods were generally not significant (not shown); still, it does appear that second-year winter skill was at a minimum in midcentury, although even then statistically significant second-year winter skill was found for late summer and fall initializations. Note also that since SLPA were not used to determine the model-analogs, the similar variations in eqSOI skill act to independently validate the above NINO3.4 skill variations.

Multi-decadal variations of ENSO forecast skill

Figure 3a shows that, since 1900, there have been pronounced multi-decadal variations of ENSO skill, with a largely monotonic increase in ENSO skill starting in midcentury that reaches a high plateau in the late 20th century. Interestingly, forecast skill appears to have declined near the start of the 20th century, which raises the possibility that skill might have been even higher in the late 19th century. To investigate this point, we need to extend the hindcast period further back in time, but this cannot be accomplished unless we can use SSTA alone to determine model-analogs. Therefore, we first evaluate how including SSHA in the initialization impacts model-analog skill. We find that model-analog hindcasts based only on SSTA yields hindcast skill

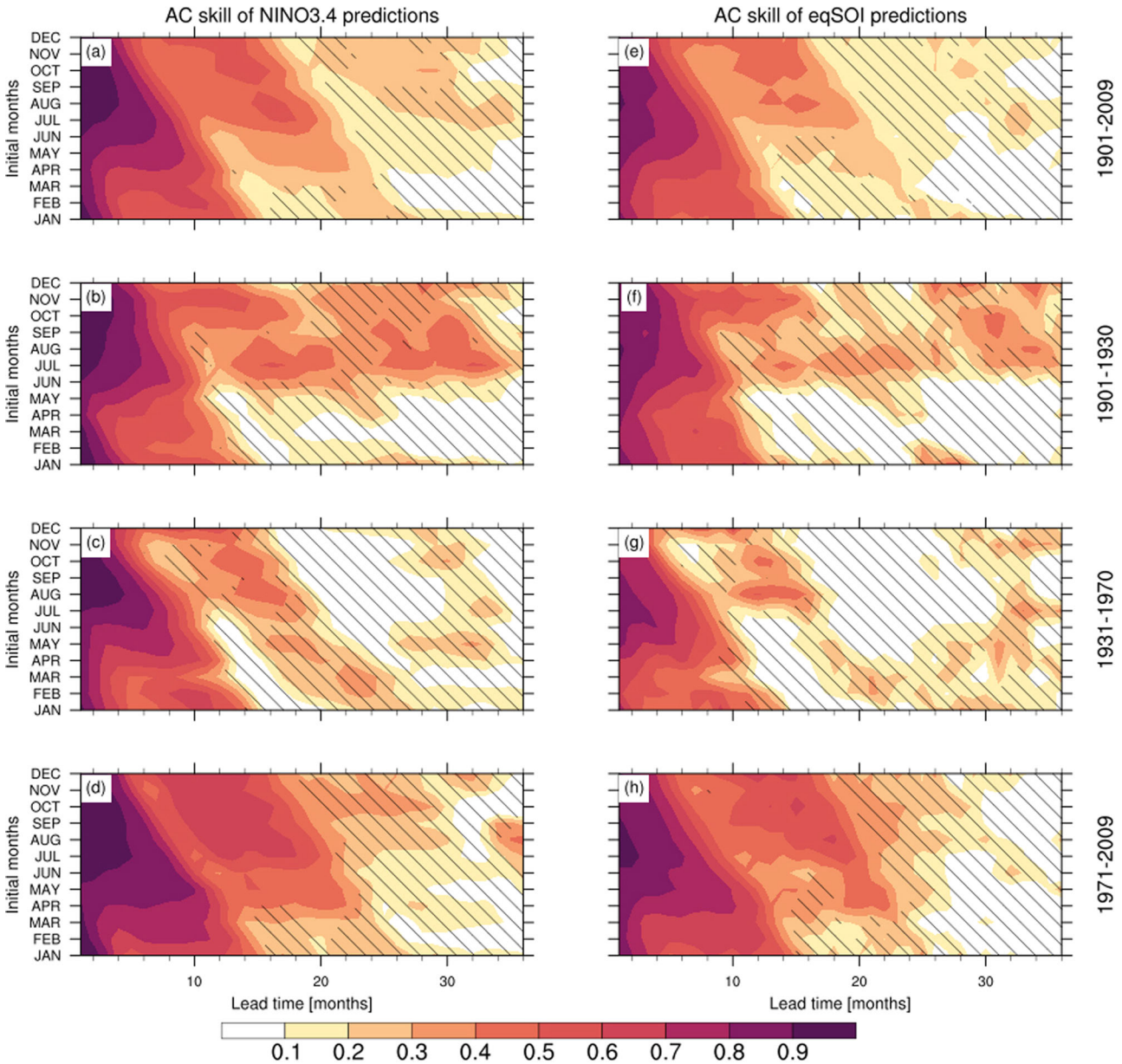

Fig. 2 Ensemble-mean anomaly correlation (AC) skill of ENSO predictions as a function of the initial months. Ensemble-mean AC skill of (**a**–**d**) NINO3.4 and (**e**–**h**) equatorial SOI as a function of initialization month (vertical axis) and the forecast lead times up to 36 months (horizonal axis). The AC skill is computed based on the entire period of (**a**, **e**) 1901–2009, and three sub-periods of (**b**, **f**) 1901–1930, (**c**, **g**) 1931-1970, and (**d**, **h**) 1971-2009, respectively. The verification dataset is taken from CERA-20C. Model-analogs are determined from SSTA + SSHA. Hatching indicates the correlation coefficients that are not statistically significant at the 95% significance level following Ref. [58].

(Fig. 3b) only slightly lower than when both SSTA and SSHA were used (see difference plot in Fig. 3c). SSHA primarily improves the forecast skill for longer-lead hindcasts prior to about 1960, which could mean that CERA-20C SSHA from this period provides some additional information, beyond SSTA, to constrain the selection of initial model-analogs. Although the inclusion of SSHA can modulate the selection of the initial model states and result in slightly better long-lead ENSO forecast skill (Fig. 3c), the initial model-analog reconstruction (Supplementary Fig. 6) as well as the relatively small difference in skill suggest that using SSTA alone may be largely sufficient to find good model-analog ocean states whose subsequent evolution matches the observed evolution of SSTA, at least within the tropical Pacific region. This might be expected, for example, in the "fast wave limit" where tropical Pacific monthly upper oceanic anomalies respond relatively quickly to surface wind anomalies and SSTAs[41,42].

However, even early in the 20th century these forecast skill differences are not statistically significant, which also suggests that using only SSTA may be sufficient not only to select skillful model-analogs but also to estimate long-term variations of ENSO hindcast skill (Fig. 4), albeit with some slightly underestimated values at long leads prior to about 1960. Using the other six SST datasets to repeat the SSTA + SSHA model-analog analysis for the common period of 1901–2009 also yielded no significant differences (Supplementary Fig. 7). While maximum skill is clearly obtained from the CERA-20C SSTA + SSHA model-analogs, we find that using either SSTA-only or SSTA + SSHA model-analogs still yields the same qualitative picture of variation in ENSO skill as a function of year, lead time, and season (not shown).

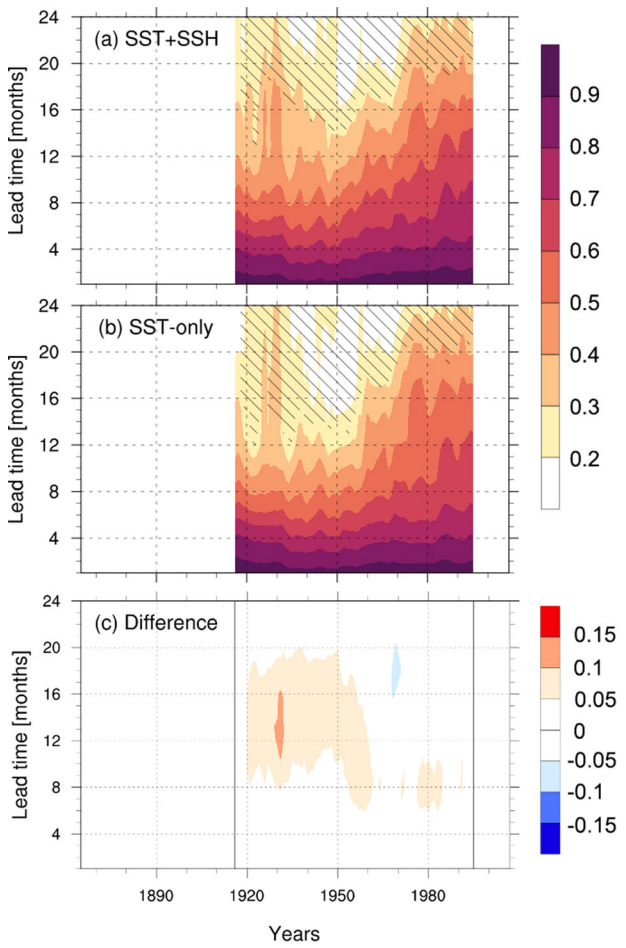

Fig. 3 Impact of including SSHA in model-analogs upon the anomaly correlation (AC) skill variation of NINO3.4 predictions. Ensemble-mean AC skill variation of NINO3.4 predictions as a function of lead months up to 24 months on the vertical axis and hindcast period on the horizonal axis. **a**, **b** AC skill based on model-analogs determined from SSTA + SSHA and SSTA-only, respectively. **c** AC skill difference between SST + SSH model-analog experiment and SST-only experiment. The verification dataset is taken from CERA-20C. All correlation coefficients are shown in the central year of the corresponding 30-year period, while anomalies are based on the sliding 30-year window that ends the year prior to the hindcast. Hatching indicates correlation coefficients that are not statistically significant at the 95% significance level, following Ref. [58].

By extending the hindcast period back into the late 19th century and forward into the 2010s, based on model-analogs conditioned by SSTA alone, it becomes clearer that ENSO hindcast skill has been roughly U-shaped (Fig. 4), starting from high values in the late 19th century, declining to a minimum in the mid-20th century, and strongly increasing in the late 20th century (Fig. 4). This result is robust to the choice of SST dataset, as shown in Fig. 4, both for selecting the initial model-analog ensembles and verifying the resulting hindcasts. However, there are also some clear quantitative differences amongst different model-analog hindcast datasets for forecast skill during the early period, which is not surprising given the uncertainty in SSTA then (see Supplementary Figure 7 and Methods for details). Comparison to Fig. 3b also suggests that there is somewhat greater skill in the first half of the 20th century when using the CERA20C SSTAs to both determine and verify model-analog hindcasts. The increase of second-year ENSO skill during the late 20th century, however, is quantitatively robust to the choice of SST dataset. Also, using hindcasts that now extend to near-present day, it appears that the second-year ENSO forecast skill increase may have begun to reverse, with some reduction since the beginning of the 21st century. Note, however, that for shorter leads, on the order of a few seasons, skill has plateaued and has not recently decreased.

We find a generally similar multi-decadal evolution of skill for El Niño and La Niña separately, such as in the probabilistic relative operating characteristic (ROC) scores shown in Fig. 5, for hindcasts based on ERSSTv5 SSTA and verified on both ERSSTv5-based NINO3.4 and 20CRv3-based eqSOI. Note that all the oceanic NINO3.4 (Fig. 5a, b) and atmospheric eqSOI (Fig. 5c, d) forecast skill estimates are qualitatively quite similar, showing multi-decadal skill variations as found above for the deterministic AC skill (for example, Figs. 3, 4). It is also clear, however, that there are some differences in forecast skill between El Niño and La Niña, especially for NINO3.4, and for some periods these differences appear statistically significant ($p < 0.05$ comparing Fig. 5a with b, and c with d), as indicated by the hatching in Fig. 5 (see Methods for details on statistical testing). We find that in the last few decades, La Niña predictions have been generally more skillful than El Niño, at both short and long leads, but that this difference may be a recent phenomenon. In fact, there are also a few periods, both in the 1970s and in the beginning of the 1900s, where predictions of El Niño were significantly more skillful at longer leads (over 12 months), both for NINO34 and eqSOI. Whether the asymmetry in El Niño and La Niña prediction skill is associated with underlying dynamical changes or is the result of randomness[12,43] is beyond the scope of this analysis. To address these questions might require applying model-analogs within a perfect model experiment aimed at investigating potential causes of forecast skill asymmetry.

Relating variations in ENSO forecast skill to observational estimates of intrinsic ENSO variability

Finally, we examine whether some aspects of either the model-analog technique or the observational datasets might spuriously impact our estimates of the multi-decadal variation of ENSO forecast skill. Specifically, we investigate whether observational uncertainties impacting model-analog 'initializations' and/or hindcast verifications, as well as intrinsic ENSO variations, might affect both the selection of initial model-analogs and the verification process. Note that this analysis is another advantage of the model-analog technique, for which evaluation of the impact of product uncertainty requires only a few fields.

We earlier found (see Fig. 4) that the greater observational uncertainty in the early part of the record (prior to 1950) leads to some minor quantitative, but not qualitative, uncertainty in the multi-decadal variations of ENSO forecast skill. Still, it is worth further exploring this issue, since it is possible that observational uncertainty might impact the hindcasts, through the initial selection of the model-analogs, differently than the verifications. Therefore, in Fig. 6 we show the predicted AC skill of NINO3.4 and eqSOI, determined for hindcasts initialized with each SST dataset and verified against all other datasets, at some chosen forecast leads. The grey shading shows the range of all the cross-verifications, where the initial model-analogs are determined from one particular SST dataset but verified against a different dataset, providing additional evaluation of uncertainty in the variation of skill. We can see that prior to about 1960 the uncertainties in skill are relatively large, with the spread subsequently converging considerably, consistent with the observational uncertainty. Given this, we still find that all the hindcasts display broadly similar skill evolution, including the pronounced midcentury ENSO prediction skill minimum. It is interesting to note that, in general, the highest skill for model-analog hindcasts using any given initialization dataset occurs when they are verified against ERSSTv5 SSTAs, especially in the early period (not shown); also, the highest skill overall occurs for hindcasts that are initialized using ERSSTv5, both

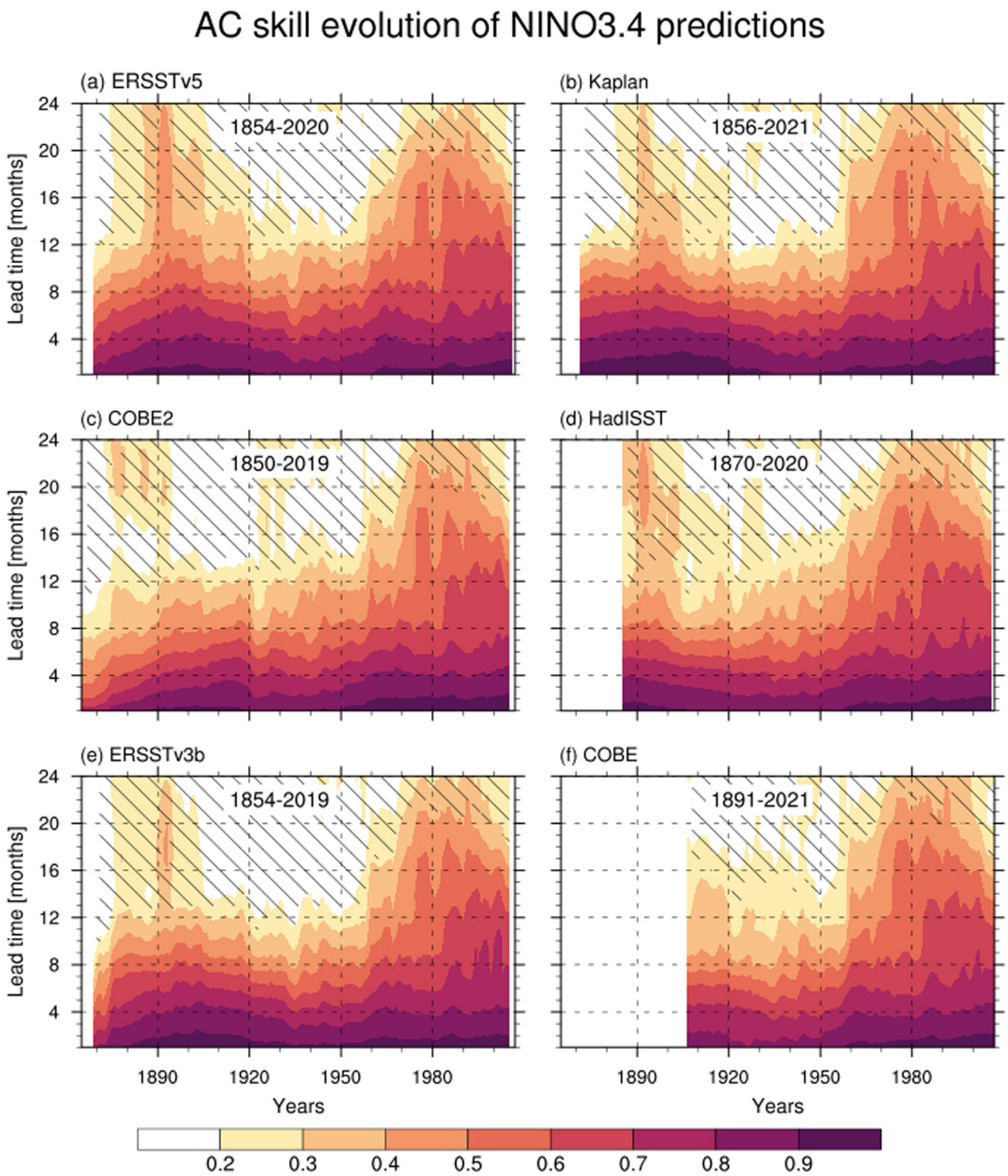

Fig. 4 Ensemble-mean anomaly correlation (AC) skill evolution of predicted NINO3.4 time series since the late 1800's. Predictive AC skill evolution of NINO3.4 time series over the 30-year hindcast windows for lead times up to 24 months. The AC skills are both initialized with and verified against six different SST datasets, (**a**) ERSSTv5, (**b**) Kaplan, (**c**) COBE2, (**d**) HadISST, (**e**) ERSSTv3b, and (**f**) COBE, respectively. Model-analogs are determined from SSTA-only. All correlation coefficients are shown in the central year of the corresponding 30-year period, while anomalies are based on the sliding 30-yr window that ends the year prior to the hindcast. Hatching indicates correlation coefficients that are not statistically significant at the 95% significance level, following Ref. [58].

for NINO3.4 and eqSOI in the early period (e.g., light green curve in Fig. 6), where the ERSSTv5 seems to have the largest variance (Fig. 6d). This may be related to the use of additional observations in the early record in the construction of the ERSSTv5 dataset, which may explain its higher ENSO variance in the early period relative to the other SST datasets[44].

Figure 6d and h show the variance evolution of the observed NINO3.4 and eqSOI over the 30-year sliding windows. We can see

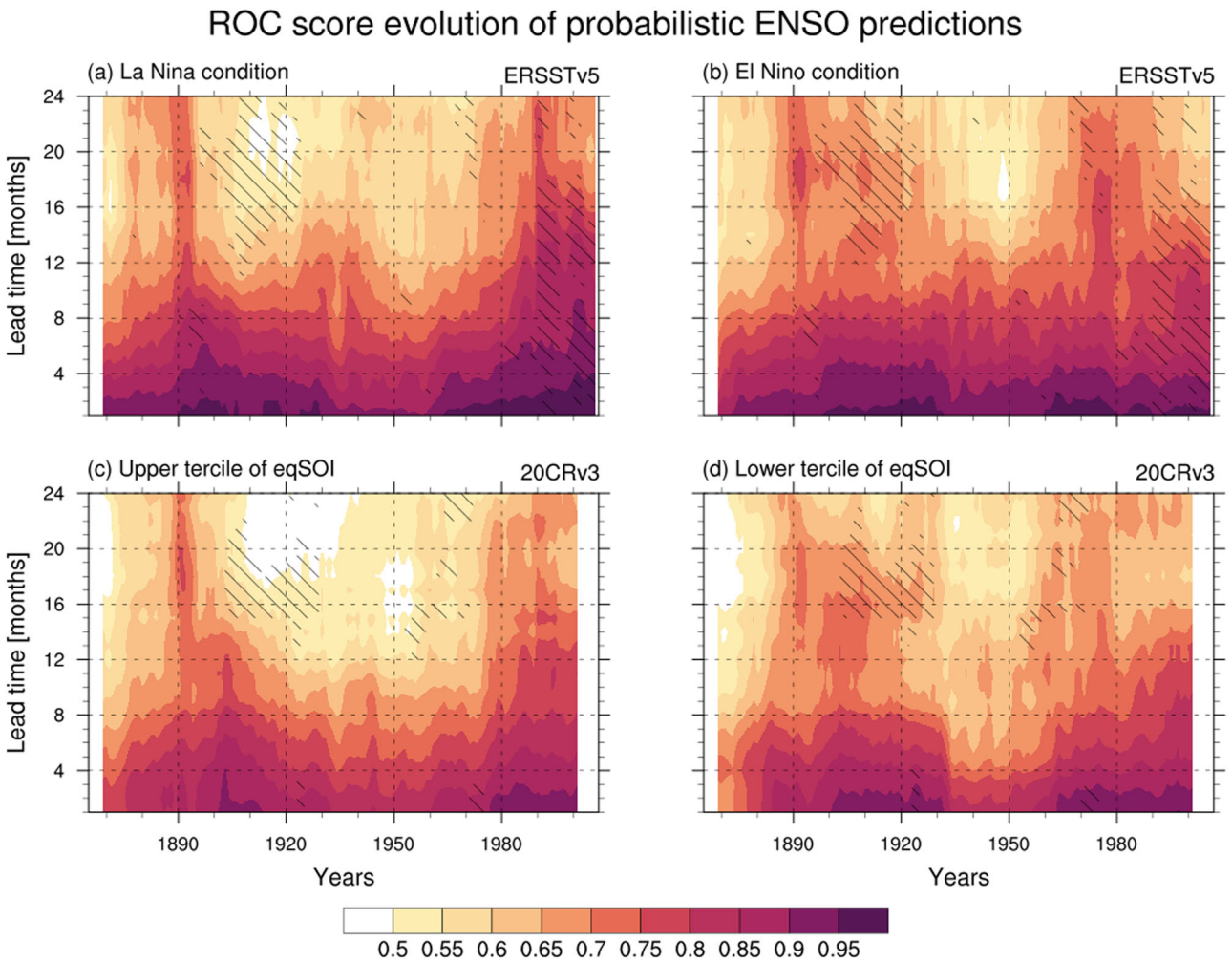

Fig. 5 Relative operating characteristic (ROC) score evolution of probabilistic ENSO predictions. Predictive ROC score evolution for (**a**) La Niña conditions (NINO3.4 < −0.5 °C) and (**b**) El Niño conditions (NINO3.4 > 0.5 °C), and (**c**) upper tercile and (**d**) lower tercile of the equatorial SOI, obtained from the 20 CRv3 reanalysis for the period of 1854–2015, over the 30-year sliding hindcast windows. The model-analog hindcasts are determined from ERSSTv5 SSTA only. All ROC scores are centered in the corresponding 30-year period, while anomalies are based on the sliding 30-year window that ends the year prior to the hindcast. Hatching indicates where forecast skill for one category is significantly different from the other (p value < 0.05), based upon a bootstrapping (resampling size of 2000) performed to test the differences between the upper tercile and lower tercile ROC scores. Note that the categorical forecasts generally outperform reference forecasts derived from climatology (i.e., ROC score > 0.5), indicating that these probabilistic ENSO predictions are generally skillful.

that when ENSO variance is relatively high, the corresponding predictive skill tends to be high as well (also see Fig. 5), suggesting some relationship between decadal ENSO variability and prediction skill, similar to what has been previously suggested on interannual time scales[16,45]. Of note is that, relative to the other SST datasets, ERSSTv5 SSTAs had both the greatest hindcast skill and the greatest variance in the late 19th century. The variance evolution of the traditional Southern Oscillation Index (SOI[46,47]), determined from weather station data obtained from Tahiti and Darwin, is also shown in Fig. 6h. Although coarse resolution of CMIP6 models cannot properly represent the changes in these two weather stations, the traditional SOI here nevertheless provides an independent measurement to qualify the variations of the atmospheric Southern Oscillation. In particular, it shows a pronounced maximum in the late 19th century, which is most consistent with the NINO3.4 amplitude variation in the ERSSTv5 dataset. The eqSOI skill is also relatively high in the late 19th century. Some studies[46,48] pointed out that the 1877/78 ENSO event in the late 19th century is closest in magnitude and temporal variation to the 1982/83 ENSO event, and it has also been suggested that the periods of 1876-1895 and 1976-1995 were dominated by strong ENSO activity and therefore had similar ENSO forecast skill[7]. Finally, it appears that the increase of ENSO variance may have leveled off in the past few decades (despite the large 1997/98 event), and likewise the increase in skill has leveled off (or even declined, as is the case for 18-month leads), although recent 12-month skill remains quite high.

A related issue is whether both the decadal variability of ENSO, and the greater uncertainty in our observational estimate of it during the early part of the record, might impact the ability of the model-analogs to capture the initial states, and if so whether that might impact skill variations. We address this by comparing the initial model-analog states (i.e., reconstructions[23]) with the observations (Fig. 7). The observed NINO3.4 time series is well-represented by the model-analogs throughout the entire record (Fig. 7a), with an overall reconstruction skill, or temporal correlation between the initial model-analog reconstruction and observations, of 0.98 (Fig. 7c). The reconstruction of eqSOI (Fig. 7b) is not quite as good, with a correlation of 0.87 (Fig. 7d), and the initial spread of the eqSOI ensemble is also larger than that of NINO3.4 (indicated in grey in Fig. 7).

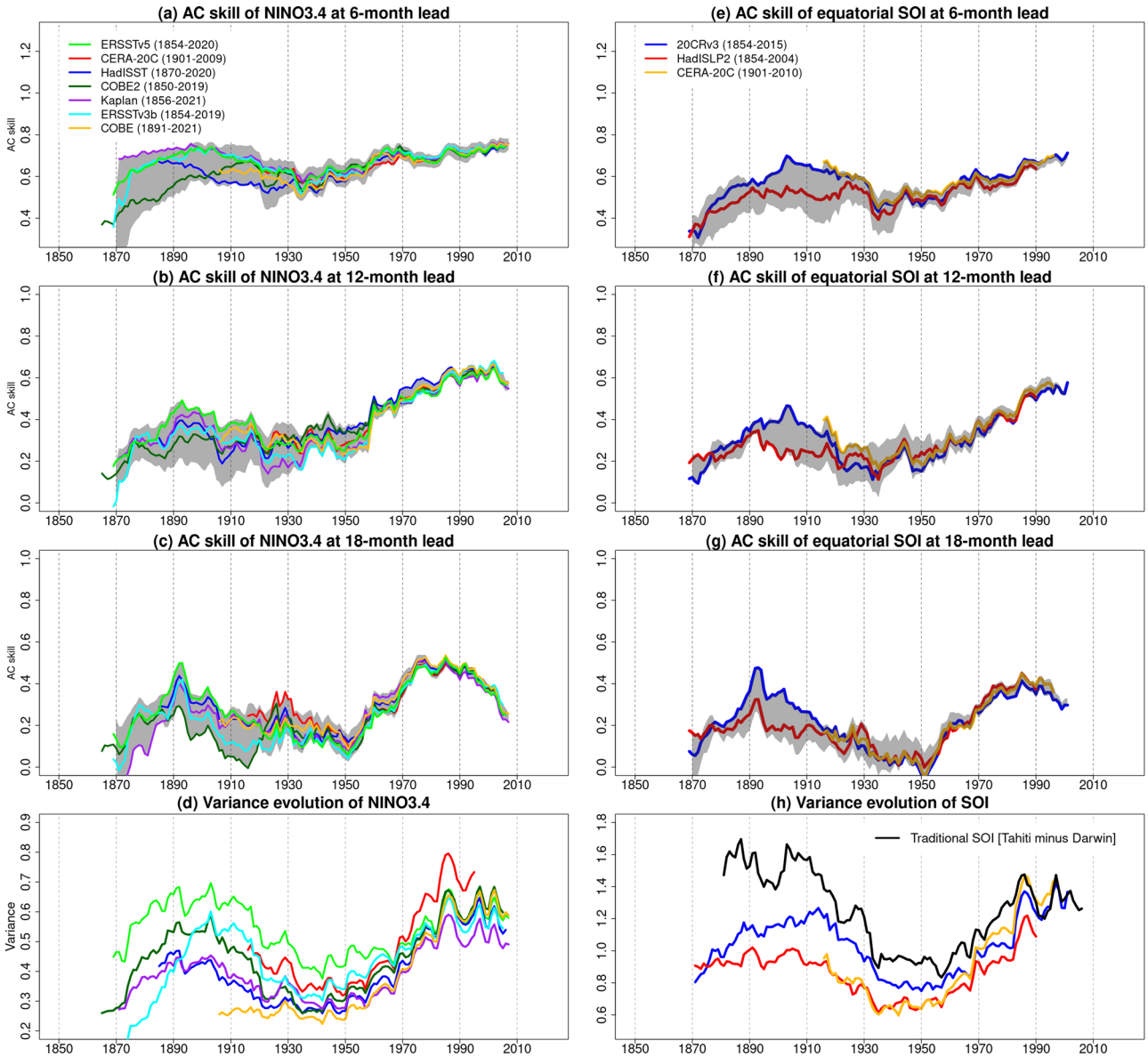

Fig. 6 Cross-verification of the ensemble-mean anomaly correlation (AC) skill evolution of ENSO predictions. Predictive AC skill evolution of ENSO over the 30-year sliding hindcast windows for (**a**–**c**) NINO3.4 time series obtained from seven SST datasets, and (**e**–**g**) equatorial SOI time series obtained from three SLP datasets at prediction lead times of 6-, 12-, and 18 months, respectively. **d**, **h** variance evolution over the 30-year sliding windows. The model-analogs are determined from SSTA-only. Correlation coefficients and variances are shown in the central year of the corresponding 30-year period. The grey shading indicates the spread of the cross verifications, where model-analogs are determined from each SSTA dataset and then verified against other SST and SLP datasets.

The quality of the NINO3.4 reconstruction appears independent of time, which is seen by recomputing the reconstruction skill for each index within a 30-year sliding window, where the value for each year is centered within the window, as shown in Fig. 7c and d. The initial model-analogs also generally capture the amplitude of individual ENSO events for both NINO3.4 and eqSOI (Fig. 7), as well as the observed decadal variation in NINO3.4 amplitude (Supplementary Figure 8) including the minimum in ENSO variance during midcentury. Note that this reduced variance did not lead to a reduction in the NINO3.4 reconstruction skill (cf Fig. 7c). Similar results are obtained when extending back to the late 1800s using ERSSTv5 (Supplementary Figure 9) despite the greater observational uncertainties in the late nineteenth century. Overall, model-analogs reproduce both interannual and interdecadal variations in NINO3.4 observations, suggesting that periods of lower hindcast skill do not necessarily represent times of relatively poorer model representation of nature.

However, the quality of the eqSOI reconstruction is not as consistent. During the mid-20th century, the model-analogs capture the eqSOI decadal variance minimum but with notably reduced reconstruction quality (Fig. 7d). Also, the initial spread of eqSOI (Fig. 7d) in this period is larger compared with other periods, suggesting there might not be sufficient good SST/SSH analogs to determine the initial SLPA model states (which are not chosen using observed SLPA). Whether this represents a change in the ENSO-related SLPA or an overall reduction in ENSO amplitude (making model-analog reconstruction of the atmospheric response more difficult in the presence of noise) remains to be determined.

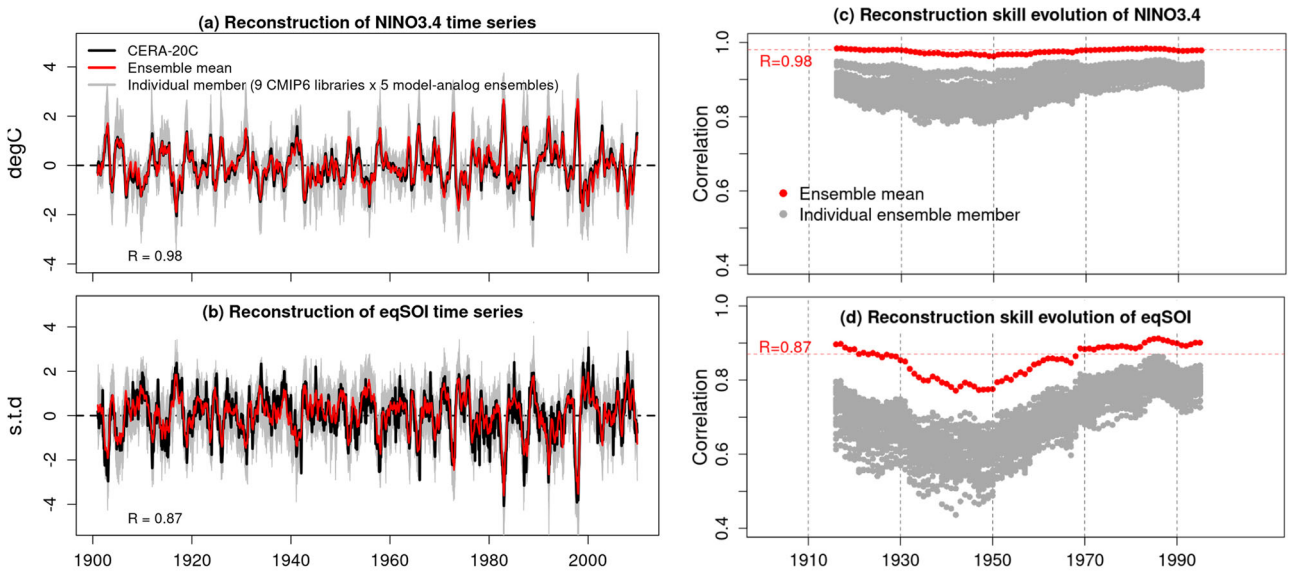

Fig. 7 Model-analog reconstruction of ENSO time series. Time series of (**a**) NINO3.4 and (**b**) the equatorial SOI, and the reconstruction skill evolution of (**c**) NINO3.4 and (**d**) equatorial SOI at $t = 0$ over the 30-year sliding windows. The observed time series are shown in black. The ensemble-mean reconstructed time series at $t = 0$ are shown in red. Arbitrary individual ensemble members from 9 CMIP6 libraries and 5 model-analog ensembles are shown in grey. The time series are verified over CERA-20C based on the common period of 1901-2009. The correlation coefficients between the observed (black) and ensemble-mean reconstructed time series (red) are labelled, which are statistically significant at the 95% level, following the method described by Ref. [58]. The ensemble-mean correlations in (**c**) and (**d**) at each 30-year period are shown in red. Individual ensemble members are shown in grey. The horizontal red line in (**c**) and (**d**) indicates the ensemble-mean correlation throughout the whole period. The correlation coefficients are all statistically significant at the 95% significance level. AC skill for each 30-year window is shown at the central year. These model-analogs are determined from SSTA-only.

DISCUSSION

In this study, using the model-analog technique[23,24] we constructed a multi-model hindcast ensemble of the tropics, for monthly start dates and for leads of up to 36 months, extending from the late 1800s into the present day. We used this hindcast ensemble to examine the robustness of multi-decadal variations of ENSO prediction skill, finding lower skill in the middle of the 20th century and higher skill in the late 19th and 20th centuries (Figs. 2–6), verifying with both SSTA and SLPA tropical indices. There also appeared to be a notable increase in prediction skill of second-year ENSO events in the late-twentieth century (Figs. 2–5), which was not impeded by the so-called spring predictability barrier. These results were consistent across all observed SST datasets used either to identify model-analogs and/or to verify hindcasts, although some additional skill in the first half of the 20th century was obtained for model-analogs based on CERA-20C SSTA and SSHA. The similar results seen for predictions of the SLPA component of ENSO (e.g., eqSOI) provided independent validation of these results, since SLPA was not used to choose the model-analogs. Moreover, while it has been suggested that La Niña conditions are more predictable than El Niño conditions[2], we found that this has not been generally true since the late 1800s, raising the possibility that this aspect of La Niña prediction also may be recent and transient.

ENSO forecast skill evolution exhibits a *U*-shape (i.e., substantially higher forecast skill towards the end of the 19th century and in recent decades) over the past century and a half, although the extent to which this is true depends somewhat upon the SSTA dataset used to identify the model-analogs and could also be impacted if we have slightly underestimated skill prior to 1900, when only SSTA were available for model-analog selection. Notably, the increased second-year ENSO skill in the late-twentieth century was not unique, since similarly high ENSO skill occurred in the late-nineteenth century (Figs 4–6), and it may not be permanent, since second-year ENSO skill may have recently begun declining (Figs. 5 and 6), although skill at leads of up to about a year remains high. Overall, our results (Fig. 6) are consistent with previous studies[16,45] suggesting a robust relationship between ENSO amplitude and forecast skill.

Inhomogeneous observations, particularly prior to 1960, may also contribute to some apparent variations in ENSO forecast skill. For example, to the extent that reduced data sampling leads to underestimated SST variance, we might also expect that the resulting smaller initial SSTA amplitudes would lead to model-analogs whose initial amplitudes likewise represent underestimates, thereby leading to weaker predicted anomalies and some reduction in skill. [Of course, this might be a concern for traditional forecast approaches and empirical techniques as well.] This could explain why using model-analogs based upon the ERSSTv5 dataset, whose relatively higher variance prior to 1960 may be attributable to the incorporation of newer and more diverse data sources[44], yields higher ENSO skill than model-analogs based upon the other datasets. On the other hand, the relationship between dataset variance and hindcast skill (e.g., Fig. 6) is not monotonic, and how dataset uncertainties impact hindcasts and their skill estimates is likely more complex and deserving of a more thorough analysis in the future.

In addition to the well-known challenges posed by data quality issues in ENSO predictions and predictability, another related question arises: how is a relatively simple model-analog technique able to produce forecast skill that is ever greater than that generated by a much more sophisticated prediction system such as SEAS5-20C (cf. Fig. 1)? One possible explanation is that traditional assimilation-based initialization forecasts rely heavily on reliable assimilated full-field initial states, which can be challenging in periods with large uncertainties in observations, potentially leading to less skillful forecasts in the early period. In contrast, the model-analog technique does not require full field variables and can instead utilize a few, but relatively more accurate, key variables (e.g., SSTA/SSHA in our study) to efficiently match the observed states. This advantage of the model-analog

approach may help explain why it yields comparable forecast skill to SEAS5-20C.

Still, the past few decades do appear to represent a particularly sustained period of comparably high ENSO forecast skill, both deterministic and probabilistic, which might be due to natural decadal variability[9,14,43] but also could be related to either the enhanced modern observational network[31] or to anthropogenic climate change[11,49]. A much sparser SST observational network in the earlier part of the record means greater uncertainty in the SST reconstructions, which has led to quantitative uncertainty in the early ENSO forecast skill. On the other hand, it is interesting to note that ERSSTv5 SSTA[44] both generally yields skill that is the most *U*-shaped and has relatively higher late-1800s NINO3.4 variance. This is also consistent with the pronounced *U*-shape of SOI variance, which is based only on two long station records of surface pressure observations. Of course, sparser data coverage during the early period (~1880–1930) also coexisted with higher ENSO forecast skill, and the monotonic increase in ENSO skill since mid-century now may be plateauing and even decreasing for second-year ENSO.

Finally, it is noteworthy that the model-analog hindcast ensemble was drawn from 9 different CMIP6 model simulations that all used fixed, pre-industrial radiative conditions. This does not exclude the possibility that climate change is playing a role in changes in recent skill; for example, one suggested impact of external forcing is to change the relative frequency of internal climate states[2,50] rather than to change the states themselves, in which case it could be sufficient even in a changing climate to use a fixed climate simulation to make model-analog predictions. Further model-analog exploration of this point, however, would likely require the use of large ensembles of historical climate simulations. This topic will be further addressed in a future study.

METHODS

Model-analog method

Analogs are chosen at each time t by minimizing a distance between a target state vector $\mathbf{x}(t)$ and each library state vector $\mathbf{y}(t')$, where the target state is defined as the observed anomalous state at the initialization time, and the library consists of all anomalous states drawn from a long climate model simulation[23]. The library states are restricted to the same calendar month as the target states, to take account of seasonality. As in previous studies[23,24], to measure the distance between the target state and each library state, we compute the root-mean-square (RMS) difference between two variables chosen from the full state vectors $\mathbf{x}$ and $\mathbf{y}$, spatial maps of SSTA and SSHA, although we define this distance within the entire tropics (30 °S-30 °N) rather than the tropical Indo-Pacific. Note that each variable here is normalized by its own domain-averaged standard deviation to equally weigh the variables. Then, these RMS distances are sorted in ascending order, and the K best simulated states with lowest RMS distance are chosen as the ensemble of initial states, indicated by the set $\mathbf{y}(t'_1), \mathbf{y}(t'_2), \ldots, \mathbf{y}(t'_k), \ldots, \mathbf{y}(t'_K)$ with k the analog index and t'_k the time of this analog in the library. The subsequent evolution of this ensemble within the control simulation, $\mathbf{y}(t'_1+\tau), \mathbf{y}(t'_2+\tau), \ldots, \mathbf{y}(t'_k+\tau), \ldots, \mathbf{y}(t'_K+\tau)$, is the model-analog forecast ensemble for $\mathbf{x}(t+\tau)$ at lead time τ months. Results are somewhat sensitive to the model-analog ensemble size K, with deterministic skill maximizing for a value of K that logarithmically increases with library size, in contrast with traditional assimilation-initialized ensembles[23]. For data libraries on the order of several hundred years in length, an ensemble size of $K \sim 10$–20 was found to give the best results[23]. In this study, because of the larger number of model simulations we examined, we found that $K = 5$ was sufficient. For more details of the technique, including perfect model experiments, and comparison to traditional assimilation-initialized hindcasts made by operational numerical models, see Ref. [23].

Model datasets

The library datasets consisted of monthly mean output from the pre-industrial control (piControl) experiments conducted using 25 CMIP6 climate models (see details listed in Table 1) whose simulations were at least 500 years in length. The pre-industrial CMIP6 forcings constitute repeating seasonal cycles[51], including CO_2 and other well-mixed greenhouse gases, solar irradiance, ozone, aerosols, and land use[52]. All datasets are remapped onto a regular 2° longitude by 2° latitude grid prior to our analysis. For the CMIP6 model simulations, we first linearly detrend the data to remove potential long-term climate drifts in these fixed climate simulations. Then, the fixed monthly climatology defined through the full length of individual simulations was removed to compute the monthly anomalies, which effectively removes mean biases relative to observations.

Observations

We use SSTA and SSHA taken from the CERA-20C reanalysis[53] to select the model-analogs for the period 1901–2009, and to verify their subsequent skill. We additionally examine the sensitivity of our results both to the dataset used to initialize (that is, used to choose the initial model-analog ensemble) hindcasts and to subsequently verify the hindcasts, using six other SST and three SLP reanalysis datasets listed in Table 2. Since these SST datasets cover years prior to 1900, we also extend our hindcast record, but in this case by using only SST to determine the model-analogs; how much additional skill the SSH information gives for the common period is examined in the 'multi-decadal variations of ENSO forecast skill' section. Note that model-analogs make forecasts of not only the subset of variables used to define the analogs (e.g., tropical SSHA and SSTA), but also any other model quantity, such as SLPA, associated with the library states and their subsequent evolution within the control simulation.

Initialized numerical hindcasts

We compared skill of the model-analog hindcasts to skill of the SEAS5-20C hindcasts[1], which were created with a lower-resolution version of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational seasonal prediction system SEAS5[32]. These hindcasts were initialized using the full three-dimensional CERA-20C fields, twice yearly (May 1st and November 1st), and were run for monthly lead times of up to two years. The SSTA of the SEAS5-20C hindcasts was then bias-corrected, by removing the lead-time dependent forecast climatology. Note that we compared these hindcasts to model-analog hindcasts initialized with the *prior* month's anomaly (i.e., October and April), so that there would be no extra information in the model-analog initialization. Then, as in Ref. [23], the lead 1 forecast for the model-analogs is the November and May monthly mean, which is compared to the corresponding SEAS5-20C predicted monthly mean (November 1–30 and May 1–31, respectively).

Determining anomalies

Defining the climatology for anomaly calculations is particularly important for lengthy records[54] since some methods (e.g., using full length of record as climatology) might spuriously enhance hindcast skill and thus could be considered "unfair"[31]. For example, if we use a reference climatology that contains information that would have been unknown at the forecast initialization time, we might create artificial skill in our hindcasts due to the inclusion of a long-term trend when comparing to equivalent real-time forecasting. Therefore, for the observations, the monthly anomalies are computed by removing a 30-year

Table 1. The 25 CMIP6 pre-industrial control simulations used in this study.

Model name	Modeling center	Length (years)	Ensemble	Key reference
ACCESS-ESM1.5	CSIRO	900	r1i1p1f1	Ziehn et al. (2020)[59]
CAMS-CSM1.0	CAMS	500	r1i1p1f1	Rong (2019)[60]
CanESM5	CCCma	1000	r1i1p1f1	Swart et al. (2019)[61]
CESM2*	NCAR	1200	r1i1p1f1	Danabasoglu et al. (2020)[62]
CIESM*	DESS-THU	500	r1i1p1f1	Lin et al. (2020)[63]
E3SM-1.0	LLNL	500	r1i1p1f1	Bader et al. (2019)[64]
EC-Earth3*	EC-Earth Consortium	501	r1i1p1f1	Döscher et al. (2021)[65]
FGOALS-g3	CAS	700	r1i1p1f1	Pu et al. (2020)[66]
GFDL-ESM4*	GFDL-NOAA	500	r1i1p1f1	Krasting et al. (2018)[67]
GISS-E2.1-G	NASA	851	r1i1p1f1	Kelley et al. (2020)[68]
GISS-E2.1-H	NASA	801	r1i1p1f1	Kelley et al. (2020)[68]
HadGEM3-GC31-LL*	MOHC	500	r1i1p1f1	Kuhlbrodt et al. (2018)[69]
HadGEM3-GC31-MM*	MOHC	500	r1i1p1f1	Senior et al. (2020)[70]
INM-CM5.0	INM-RAS	1201	r1i1p1f1	Volodin et al. (2019)[71]
IPSL-CM6A-LR	IPSL-CMC	1200	r1i1p1f1	Boucher et al. (2020)[72]
KIOST-ESM	KIOST	500	r1i1p1f1	Kim et al. (2019)[73]
MIROC-ES2L	MIROC Consortium	500	r1i1p1f2	Hajima et al. (2020)[74]
MIROC6	MIROC Consortium	800	r1i1p1f1	Tatebe et al. (2019)[75]
MPI-ESM-1.2-HAM	MPI	780	r1i1p1f1	Gutjahr et al. (2019)[76]
MPI-ESM1.2-LR	MPI	1000	r1i1p1f1	Gutjahr et al. (2019)[76]
MRI-ESM2.0	MRI	701	r1i1p1f1	Yukimoto et al. (2019)[77]
NESM3	NUIST	500	r1i1p1f1	Cao et al. (2018)[78]
NorESM2-LM*	NCC	501	r1i1p1f1	Bentsen et al. (2013)[79]
SAM0-UNICON*	SNU	700	r1i1p1f1	Park et al. (2019)[80]
UKESM1.0-LL*	MOHC	1880	r1i1p1f2	Sellar et al. (2019)[81]

*The nine CMIP6 models used in the multi-model model-analog ensemble are marked (see text for more details).

Table 2. The verification (target) datasets used in this study.

Variable	Dataset	Period	Reference
SSH	CERA-20C	1901-2009	Laloyaux et al. (2018)[53]
SST	CERA-20C	1901-2009	Laloyaux et al. (2018)[53]
	COBE	1891-2021	Ishii et al. (2005)[82]
	COBE2	1850-2019	Hirahara et al. (2014)[83]
	ERSSTv5	1854-2021	Huang et al. (2017)[84]
	ERSSTv3b	1854-2019	Smith et al. (2008)[85]
	HadISST	1870-2020	Rayner et al. (2003)[86]
	Kaplan SST	1856-2021	Kaplan et al. (1998)[87]
SLP	20 CRv3	1836-2015	Slivinski et al. (2019)[88]
	CERA-20C	1901-2010	Laloyaux et al. (2018)[53]
	HadISLP2	1850-2004	Allan and Ansell (2006)[89]

sliding climatology. For example, for a prediction made in 1972, the corresponding climatology is defined based on the 1942–1971 period, which avoids the use of any future information that would have not been available at the forecast initialization time. This anomaly calculation is referred to as a "fair-sliding" method[31], where a sliding window behind the forecast is used as the reference climatology period. The resulting anomalies are used both to determine the model-analogs and to verify the subsequent hindcasts. Note that this calculation also acts to largely detrend the anomalies.

As in previous studies[24], we draw from the library states of long pre-industrial control simulations, which means the model-analog hindcasts do not include the effects of external radiative forcing. This issue may be addressed by adding an estimate of the externally-forced component, determined from the CMIP5 historical simulations, to such model-analog hindcast ensemble members[24], although this externally-forced trend does not appear to significantly impact ENSO skill, at least over the 1961–2015 period when it was found that there was no significant secular trend in ENSO skill[24]. When we performed a similar analysis using the CMIP6 historical simulations, we found that the skill results were similar to simply using the 30-year sliding climatology to compute anomalies. Therefore, we used that approach in this study.

ENSO indices

To evaluate hindcast skill, we make use of both SSTA- and SLPA-related ENSO indices. For SSTA, we use the NINO3.4 index, derived from the averaged SSTA in the Niño 3.4 region (5°S-5°N, 170°W-120°W; Supplementary Fig. 10). To evaluate the skill of the model-analog SLPA hindcasts, we use a measure of the Southern Oscillation index (SOI). This serves as an independent metric of skill since SLPA are not used to choose the model-analogs, as opposed to SSTA and SSHA. Ideally, we would use the traditional SOI, the standardized SLPA difference obtained from Tahiti and Darwin weather station data (Supplementary Fig. 10). However, due to the relatively coarse spatial resolution of the CMIP6 models, model output cannot entirely represent the SLP see-saw between these two stations. Therefore, we instead use the equatorial SOI (eqSOI[10,55]), derived from the standardized difference between the standardized SLPA in areas over the eastern equatorial Pacific

(5°S-5°N, 80°W-130°W) and over Indonesia (5°S-5°N, 90°E-140°E; Supplementary Fig. 10).

One concern with extending the hindcast dataset back into the 19th century is that uncertainty in observations/verifications might impact our evaluation of variations in ENSO prediction skill. The time series of both NINO3.4 and eqSOI (Supplementary Fig. 10) have considerable uncertainties across multiple datasets in the early part of the record, eventually beginning to converge around 1950. In addition, due to the pronounced disagreement in the late 19th century and the early 20th century, the temporal correlation of the eqSOI between 20CRv3 and HadISLP2 datasets dropped to 0.63 for their overlapping period of 1850–2004, while CERA-20C and 20CRv3 has the highest correlation of 0.91 thoughout the common period of 1901–2010. It is worth noting that all the correlations are considerably larger when only the later periods are considered; for example, the mutual correlations exceed 0.92 from 1950 onwards.

Skill metrics

To measure deterministic ensemble-mean forecast skill, we use AC skill and RMS error (RMSE), which is defined as

$$ACC = \frac{\sum_{i=1}^{N} \overline{F_i}' O_i'}{\sqrt{\sum_{i=1}^{N} (\overline{F_i}')^2 \sum_{i=1}^{N} (O_i')^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (O_i' - \overline{F_i}')^2} \quad (2)$$

where $\overline{F_i}'$ is the ensemble-mean anomaly forecast, and O_i' is the observed anomaly with $i = 1 \ldots N$ representing the verification months or years. RMSSS is defined as $RMSSS = 1 - RMSE / \sqrt{\frac{1}{N}\sum_{i=1}^{N} (O_i')^2}$.

We also examined ROC score, which measures the quality of probabilistic forecasts that relate the hit rate to the corresponding false-alarm rate[56,57]. The ROC score is defined as the area under the ROC curve. The ROC score of a no-skill forecast is 0.5, and for a perfect forecast is 1. In this work, oceanic La Niña and El Niño conditions are defined as NINO3.4 < −0.5 °C and NINO3.4 > 0.5 °C, and the upper and lower terciles of the eqSOI are used to compute the probabilistic metrics.

Significance testing of AC skill was conducted following the approach of Ref. [58], which takes account of the effective number of degrees of freedom due to serial auto-correlations. Then, a simple t statistic was applied to assess whether the AC skill is significant. Fisher Z-transformation was applied to the corresponding AC skill. Then, t test was applied to assess whether the differences between AC skill are significant. To test if the differences between two ROC curves are significant, a bootstrapping method with 2000 perturbations was applied. Each stratified replicate contains the same number of controls and cases as in the original sample. For each bootstrap replicate, the ROC scores of the two ROC curves (i.e., La Niña vs. El Niño condition) were computed and the difference was stored. The difference between the original two ROC scores was then compared to the standard deviation of the bootstrap differences following $D = (\mathrm{ROC1} - \mathrm{ROC2})/s$, where s is the standard deviation of the bootstrap differences, and ROC1 and ROC2 are the ROC scores of the two original ROC curves. Finally, D was compared to the normal distribution and p-value is computed. When $p<0.05$, we rejected the null hypothesis (i.e., there was no difference between the two ROC scores).

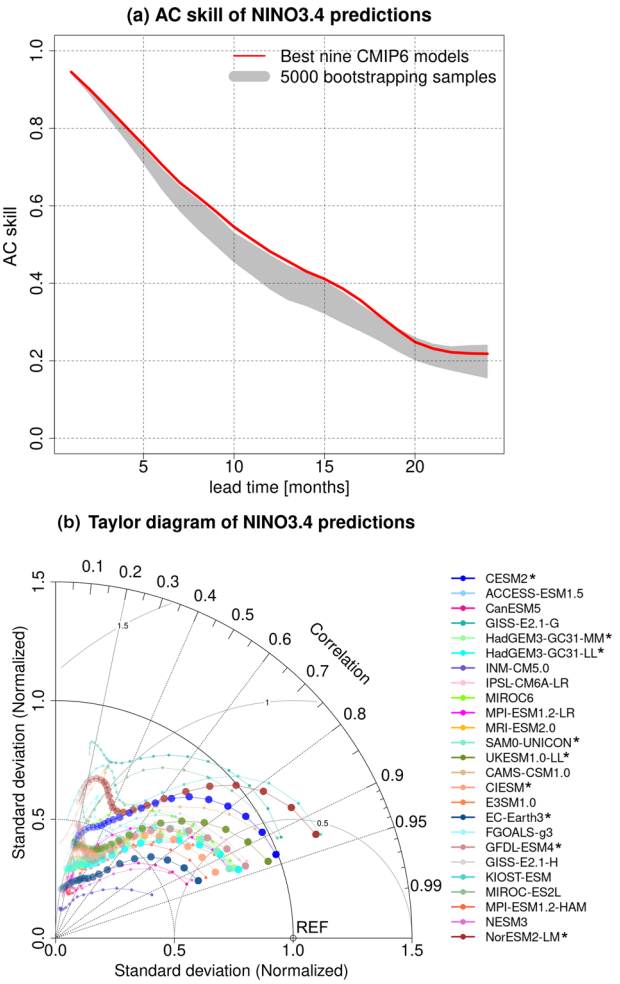

Fig. 8 Testing the choice of CMIP6 library model combinations. **a** AC skill of NINO3.4 predictions verified over CERA-20C for the period of 1901-2009. The red curve represents the ensemble mean AC skill derived from the nine best CMIP6 models used in the text. The grey shading indicates the AC skill range of 5000 bootstrapping samples. **b** Taylor diagram of model-analog forecast skill at lead times of 1–24 months for NINO3.4 index. Each curve represents a skill trajectory from one CMIP6 model, with one dot per forecast lead month ranging from 1 month (opaque dots) to 24 months (transparent dots) in advance. The nine 'best' models are highlighted using bigger dots and marked in the legend.

Constructing the multi-model model-analog ensemble

For this study, we show results for a multi-model ensemble constructed from a subset of nine CMIP6 models (marked in Table 1) that we found yielded the most skillful combination (Fig. 8). We identified these models by first applying a bootstrapping method to randomly choose different nine-model combinations from the CMIP6 library, which was replicated 5000 times. The AC skill of multi-model ensemble-mean ENSO forecasts was then computed. The spread of predictive ENSO skill based on different CMIP6 combinations is reasonably small at all the forecast lead times (Fig. 8). We then selected the highest-skill nine-model CMIP6 multi-model ensemble (marked in Table 1) to form our base of library datasets, whose ensemble-mean skill is generally higher compared to any other combination. It is worth noting that other numbers of CMIP6 library models are also tested, which yields indistinguishable results as presented in Fig. 8. These nine models are also individually the most skillful of the CMIP6 models, as indicated in the Taylor diagram (Fig. 8).

DATA AVAILABILITY

SST datasets taken from CERA-20C, HadSLP2, and HadISST are accessible through https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/cera-20c, https://www.metoffice.gov.uk/hadobs/hadslp2/, and https://www.metoffice.gov.uk/hadobs/hadisst/, respectively. CERA-20C SSH data is accessible upon request to ECMWF. The rest of the reanalysis datasets can be downloaded from https://psl.noaa.gov/data/gridded/. CMIP6 datasets can be downloaded from https://esgf-node.llnl.gov/projects/cmip6/.

CODE AVAILABILITY

The R code used to generate model-analog hindcast investigated in this study is available at the Zenodo open data repository (https://doi.org/10.5281/zenodo.8070768). The corresponding ENSO hindcast products (NINO3.4 and eqSOI) are available at https://downloads.psl.noaa.gov/Projects/LIM/Realtime/Realtime/webData/MA_CERA20C/, which is also the child directory of https://psl.noaa.gov/forecasts/seasonal/, where the deterministic and probabilistic hindcasts used in this study are also accessible in an image browser.

Received: 2 February 2023; Accepted: 30 June 2023;
Published online: 14 July 2023

REFERENCES

1. Weisheimer, A. et al. Variability of ENSO forecast skill in 2-year global reforecasts over the 20th Century. *Geophys. Res. Lett.* **49**, e2022GL097885 (2022).
2. L'Heureux, M. L. et al. ENSO prediction. In *El Niño Southern Oscillation in a changing climate*. 227–246 (Wiley, 2020).
3. Weisheimer, A. et al. Seasonal forecasts of the twentieth century. *Bull. Am. Meteorol. Soc.* **101**, E1413–E1426 (2020).
4. Chen, D. & Cane, M. A. El Niño prediction and predictability. *J. Comput. Phys.* **227**, 3625–3640 (2008).
5. Balmaseda, M. A., Davey, M. K. & Anderson, D. L. T. Decadal and seasonal dependence of ENSO prediction skill. *J. Clim.* **8**, 2705–2715 (1995).
6. Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S. & DeWitt, D. G. Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? *Bull. Am. Meteorol. Soc.* **93**, ES48–ES50 (2012).
7. Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E. & Huang, D. Predictability of El Niño over the past 148 years. *Nature* **428**, 733–736 (2004).
8. Trenberth, K. E. & Hoar, T. J. El Niño and climate change. *Geophys. Res. Lett.* **24**, 3057–3060 (1997).
9. Power, S. et al. Decadal climate variability in the tropical Pacific: characteristics, causes, predictability, and prospects. *Science* **374**, eaay9165 (2021).
10. Power, S. B. & Kociuba, G. The impact of global warming on the Southern Oscillation Index. *Clim. Dyn.* **37**, 1745–1754 (2011).
11. Zhao, M., Hendon, H. H., Alves, O., Liu, G. & Wang, G. Weakened eastern Pacific El Niño predictability in the early twenty-first century. *J. Clim.* **29**, 6805–6822 (2016).
12. Wittenberg, A. T. Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.* **36**, L12702 (2009).
13. Newman, M., Shin, S.-I. & Alexander, M. A. Natural variation in ENSO flavors. *Geophys. Res. Lett.* **38**, L14705 (2011).
14. Kirtman, B. P. & Schopf, P. S. Decadal variability in ENSO predictability and prediction. *J. Clim.* **11**, 2804–2822 (1998).
15. Flügel, M., Chang, P. & Penland, C. The role of stochastic forcing in modulating ENSO predictability. *J. Clim.* **17**, 3125–3140 (2004).
16. Newman, M. & Sardeshmukh, P. D. Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophys. Res. Lett.* **44**, 8520–8529 (2017).
17. Collins, M., Frame, D., Sinha, B. & Wilson, C. How far ahead could we predict El Niño? *Geophys. Res. Lett.* **29**, 130-1–130-4 (2002).
18. DelSole, T., Nattala, J. & Tippett, M. K. Skill improvement from increased ensemble size and model diversity. *Geophys. Res. Lett.* **41**, 7331–7342 (2014).
19. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568–572 (2019).
20. Sharmila, S., Hendon, H., Alves, O., Weisheimer, A. & Balmaseda, M. Contrasting El Niño–La Niña predictability and prediction skill in 2-year reforecasts of the twentieth century. *J. Clim.* **36**, 1269–1285 (2023).
21. Gonzalez, P. L. M. & Goddard, L. Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Clim. Dyn.* **46**, 3127–3147 (2016).
22. Liu, T., Song, X., Tang, Y., Shen, Z. & Tan, X. ENSO predictability over the past 137 years based on a CESM ensemble prediction system. *J. Clim.* **35**, 763–777 (2022).
23. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *J. Clim.* **31**, 5437–5459 (2018).
24. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs. *Geophys. Res. Lett.* **46**, 1721–1730 (2019).
25. Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **26**, 636–646 (1969).
26. Van Den Dool, H. M. Searching for analogues, how long must we wait? *Tellus Dyn. Meteorol. Oceanogr.* **46**, 314–324 (1994).
27. Menary, M. B., Mignot, J. & Robson, J. Skilful decadal predictions of subpolar North Atlantic SSTs using CMIP model-analogues. *Environ. Res. Lett.* **16**, 064090 (2021).
28. Mahmood, R. et al. Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales – a poor man's initialized prediction system. *Earth Syst. Dyn.* **13**, 1437–1450 (2022).
29. Mulholland, D. P., Laloyaux, P., Haines, K. & Balmaseda, M. A. Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts*. *Mon. Weather Rev.* **143**, 4631–4644 (2015).
30. Kirtman, B. P. et al. The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014).
31. Risbey, J. S. et al. Standard assessments of climate forecast skill can be misleading. *Nat. Commun.* **12**, 4346 (2021).
32. Johnson, S. J. et al. SEAS5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.* **12**, 1087–1117 (2019).
33. Kim, S.-K. & An, S.-I. Seasonal gap theory for ENSO phase locking. *J. Clim.* **34**, 5621–5634 (2021).
34. Clarke, A. J. El Niño physics and El Niño predictability. *Annu. Rev. Mar. Sci.* **6**, 79–99 (2014).
35. Vimont, D. J., Wallace, J. M. & Battisti, D. S. The seasonal footprinting mechanism in the Pacific: implications for ENSO*. *J. Clim.* **16**, 2668–2675 (2003).
36. Chang, P. et al. Pacific meridional mode and El Niño-Southern Oscillation. *Geophys. Res. Lett.* **34**, L16608 (2007).
37. Zhang, H. & Clement, A. & Di Nezio, P. The South Pacific meridional mode: a mechanism for ENSO-like variability. *J. Clim.* **27**, 769–783 (2014).
38. Lou, J., O'Kane, T. J. & Holbrook, N. J. Linking the atmospheric Pacific-South American mode with oceanic variability and predictability. *Commun. Earth Environ.* **2**, 223 (2021).
39. McPhaden, M. J. Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophys. Res. Lett.* **30**, 1480 (2003).
40. Tippett, M. K. & L'Heureux, M. L. Low-dimensional representations of Niño 3.4 evolution and the spring persistence barrier. *NPJ Clim. Atmos. Sci.* **3**, 24 (2020).
41. Penland, C. & Sardeshmukh, P. D. The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **8**, 1999–2024 (1995).
42. Kumar, A., Wen, C., Xue, Y. & Wang, H. Sensitivity of subsurface ocean temperature variability to specification of surface observations in the context of ENSO. *Mon. Weather Rev.* **145**, 1437–1446 (2017).
43. Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A. & Zeng, F. ENSO modulation: Is it decadally predictable? *J. Clim.* **27**, 2667–2681 (2014).
44. Newman, M., Wittenberg, A. T., Cheng, L., Compo, G. P. & Smith, C. A. The extreme 2015/16 El Niño, in the context of historical climate variability and change. *Bull. Am. Meteorol. Soc.* **99**, S16–S20 (2018).
45. Kumar, A., Chen, M., Xue, Y. & Behringer, D. An analysis of the temporal evolution of ENSO prediction skill in the context of the equatorial Pacific Ocean observing system. *Mon. Weather Rev.* **143**, 3204–3213 (2015).
46. Allan, R. J., Nicholls, N., Jones, P. D. & Butterworth, I. J. A further extension of the Tahiti-Darwin SOI, early ENSO events and Darwin pressure. *J. Clim.* **4**, 743–749 (1991).
47. Können, G. P., Jones, P. D., Kaltofen, M. H. & Allan, R. J. Pre-1866 extensions of the Southern Oscillation Index using early Indonesian and Tahitian meteorological readings. *J. Clim.* **11**, 2325–2339 (1998).
48. Kiladis, G. N. & Diaz, H. F. An analysis of the 1877–78 ENSO episode and comparison with 1982–83. *Mon. Weather Rev.* **114**, 1035–1047 (1986).
49. Hu, Z.-Z. et al. The interdecadal shift of ENSO properties in 1999/2000: a review. *J. Clim.* **33**, 4441–4462 (2020).
50. Palmer, T. N. A nonlinear dynamical perspective on climate prediction. *J. Clim.* **12**, 575–591 (1999).
51. Meinshausen, M. et al. Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geosci. Model Dev.* **10**, 2057–2116 (2017).
52. Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).

53. Laloyaux, P., Balmaseda, M., Dee, D., Mogensen, K. & Janssen, P. A coupled data assimilation system for climate reanalysis: coupled data assimilation for climate reanalysis. *Q. J. R. Meteorol. Soc.* **142**, 65–78 (2016).
54. Meehl, G. A. et al. The effects of bias, drift, and trends in calculating anomalies for evaluating skill of seasonal-to-decadal initialized climate predictions. *Clim. Dyn.* **59**, 3373–3389 (2022).
55. Vecchi, G. A. et al. Weakening of tropical Pacific atmospheric circulation due to anthropogenic forcing. *Nature* **441**, 73–76 (2006).
56. Kharin, V. V. & Zwiers, F. W. On the ROC score of probability forecasts. *J. Clim.* **16**, 4145–4150 (2003).
57. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
58. Davis, R. E. Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.* **6**, 249–266 (1976).
59. Ziehn, T. et al. The Australian earth system model: ACCESS-ESM1.5. *J. South. Hemisph. Earth Syst. Sci.* **70**, 193–214 (2020).
60. Rong, X. CAMS CAMS-CSM1.0 model output prepared for CMIP6 ScenarioMIP. https://doi.org/10.22033/ESGF/CMIP6.11004 (2019).
61. Swart, N. C. et al. The Canadian earth system model version 5 (CanESM5.0.3). *Geosci. Model Dev.* **12**, 4823–4873 (2019).
62. Danabasoglu, G. et al. The Community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).
63. Lin, Y. et al. Community integrated earth system model (CIESM): description and evaluation. *J. Adv. Model. Earth Syst.* **12**, e2019MS002036 (2020).
64. Bader, D. C., Leung, R., Taylor, M. & McCoy, R. B. E3SM-Project E3SM1.0 model output prepared for CMIP6 CMIP. https://doi.org/10.22033/ESGF/CMIP6.2294 (2019).
65. Döscher, R. et al. The EC-Earth3 Earth system model for the coupled model intercomparison project 6. *Geosci. Model Dev.* **15**, 2973–3020 (2022).
66. Pu, Y. et al. CAS FGOALS-g3 model datasets for the CMIP6 scenario model intercomparison project (ScenarioMIP). *Adv. Atmos. Sci.* **37**, 1081–1092 (2020).
67. Krasting, J. P. et al. NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP. *Earth Syst. Grid Fed*. https://doi.org/10.22033/ESGF/CMIP6.1407 (2018).
68. Kelley, M. et al. GISS-E2.1: configurations and climatology. *J. Adv. Model. Earth Syst.* **12**, e2019MS002025 (2020).
69. Kuhlbrodt, T. et al. The low-resolution version of HadGEM3 GC3.1: development and evaluation for global climate. *J. Adv. Model. Earth Syst.* **10**, 2865–2888 (2018).
70. Senior, C. A. et al. U.K. community earth system modeling for CMIP6. *J. Adv. Model. Earth Syst.* **12**, e2019MS002004 (2020).
71. Volodin, E. et al. INM INM-CM5-0 model output prepared for CMIP6 CMIP piControl. *Earth Syst. Grid Fed*. https://doi.org/10.22033/ESGF/CMIP6.5081 (2019)
72. Boucher, O. et al. Presentation and evaluation of the IPSL-CM6A-LR climate model. *J. Adv. Model. Earth Syst.* **12**, e2019MS002010 (2020).
73. Kim, Y. et al. KIOST KIOST-ESM model output prepared for CMIP6 CMIP. *Earth Syst. Grid Fed*. https://doi.org/10.22033/ESGF/CMIP6.1922 (2019).
74. Hajima, T. et al. Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geosci. Model Dev.* **13**, 2197–2244 (2020).
75. Tatebe, H. et al. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geosci. Model Dev.* **12**, 2727–2765 (2019).
76. Gutjahr, O. et al. Max planck institute earth system model (MPI-ESM1.2) for the high-resolution model intercomparison project (HighResMIP). *Geosci. Model Dev.* **12**, 3241–3281 (2019).
77. Yukimoto, S. et al. MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP. *Earth Syst Grid Fed*. https://doi.org/10.22033/ESGF/CMIP6.621 (2019).
78. Cao, J. et al. The NUIST earth system model (NESM) version 3: description and preliminary evaluation. *Geosci. Model Dev.* **11**, 2975–2993 (2018).
79. Bentsen, M. et al. The Norwegian earth system model, NorESM1-M – Part 1: description and basic evaluation of the physical climate. *Geosci. Model Dev.* **6**, 687–720 (2013).
80. Park, S., Shin, J., Kim, S., Oh, E. & Kim, Y. Global climate simulated by the Seoul National University atmosphere model version 0 with a unified convection scheme (SAM0-UNICON). *J. Clim.* **32**, 2917–2949 (2019).
81. Sellar, A. A. et al. UKESM1: description and evaluation of the U.K. Earth System Model. *J. Adv. Model. Earth Syst.* **11**, 4513–4558 (2019).
82. Ishii, M., Shouji, A., Sugimoto, S. & Matsumoto, T. Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection. *Int. J. Climatol.* **25**, 865–879 (2005).
83. Hirahara, S., Ishii, M. & Fukuda, Y. Centennial-scale sea surface temperature analysis and Its uncertainty. *J. Clim.* **27**, 57–75 (2014).
84. Huang, B. et al. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
85. Smith, T. M., Reynolds, R. W., Peterson, T. C. & Lawrimore, J. Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Clim.* **21**, 2283–2296 (2008).
86. Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.* **108**, 4407 (2003).
87. Kaplan, A. et al. Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res. Oceans* **103**, 18567–18589 (1998).
88. Slivinski, L. C. et al. Towards a more reliable historical reanalysis: improvements for version 3 of the twentieth century reanalysis system. *Q. J. R. Meteorol. Soc.* **145**, 2876–2908 (2019).
89. Allan, R. & Ansell, T. A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *J. Clim.* **19**, 5816–5842 (2006).

ACKNOWLEDGEMENTS

The authors would like to thank Andrew Wittenberg, and Antje Weisheimer for their helpful comments and discussions that helped to improve this manuscript. We also thank Magdalena Balmaseda for supplying the CERA-20C SSH dataset. We thank Yan Wang for preparing and downloading all the datasets used in this study. We also thank Yan Wang for maintaining the model-analog seasonal forecast webpage (https://psl.noaa.gov/forecasts/seasonal/), where the hindcasts investigated in this study are stored. This work was primarily supported by the Famine Early Warning Systems Network, AID-OFDA-T-17-00002, with additional funding from the U.S. Department of Energy, DE-SC0019418, under NOAA cooperative agreements NA17OAR4320101 and NA22OAR4320151.

AUTHOR CONTRIBUTIONS

J.L. conducted the analyses and led the writing, with contributions from all the authors. M.N. helped shape the scientific ideas, and actively contributed to the discussions, writing, and manuscript polishing. A.H. secured funding sources of this project, and actively participated in the discussions and assisted in polishing the results and writing.

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/s41612-023-00417-z.

Correspondence and requests for materials should be addressed to Jiale Lou.

Reprints and permission 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.