Future changes in seasonal climate predictability

Dillon J. Amaya[1], Nicola Maher[2,3,4], Clara Deser[5], Michael G. Jacox[1,6], Michael A. Alexander[1], Matthew Newman[1], Juliana Dias[1], and Jiale Lou[7]

[1]Physical Science Laboratory, National Oceanic and Atmospheric Administration, 325 Broadway, Boulder, CO 80305, USA

[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, 216 UCB, University of Colorado Boulder campus, Boulder, CO 80309, USA

[3]Department of Atmospheric and Oceanic Sciences (ATOC), University of Colorado at Boulder, Boulder, CO 80309, USA

[4]Australian National University and ARC Centre of Excellence for Climate Extremes, Canberra, ACT 0200, Australia

[5]National Center for Atmospheric Research, 1850 Table Mesa Dr, Boulder, CO 80305, USA

[6]Environmental Research Division, Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, 99 Pacific St #255A, Monterey, CA 93940, USA

[7]Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, NJ, USA

Corresponding author: Dillon J. Amaya, dillon.amaya@noaa.gov, 816-916-8348

Abstract

Seasonal forecasts provide critical decision support tools for managing important socioeconomically-relevant resources. As the result of continued model development, the skill of such tools has improved over the years. However, further advancements are hampered by the climate's "potential predictability", an upper limit for how accurately we can predict different parameters that is intrinsic to the chaotic nature of the climate system. Recent studies have shown that potential predictability and actual forecast skill have varied throughout the historical record, primarily as a result of natural decadal variability. In this study, we explore whether potential predictability will change in the future as a distinct response to anthropogenic climate change. We quantify the potential predictability limits of the El Niño-Southern Oscillation (ENSO) as well as global surface temperature, precipitation, and upper atmospheric circulation anomalies from 1921-2100 by applying a perfect model framework to five coupled model large ensembles. We find that the sign, magnitude, and timing of predictability changes are highly model dependent, with some producing a robust increase or decrease in potential predictability by 2100, and others producing no significant change. While there is large intermodel uncertainty in future predictability changes, a common physical mechanism emerges that allows us to anticipate how real-world predictability may change in the coming decades. In particular, predictability changes in each model are strongly linked to their projected change in ENSO amplitude. Therefore, historical forecast skill relationships that depend on ENSO and its teleconnections may be altered as the climate continues to change.

1. Introduction

Seasonal climate forecasts provide important decision support tools to help stakeholders manage a variety of socioeconomically-relevant resources. For example, initialized dynamical forecasts are routinely used to provide seasonal outlooks of regional precipitation and surface temperature, tropical cyclone activity, and climate modes such as the El Niño-Southern Oscillation (ENSO). While recent advances in model physics, resolution, ensemble sizes, and data assimilation schemes have led to increases in seasonal forecast skill (Barnston et al. 2012; Barnston and Tippett 2017), prediction systems are still limited by the so-called "potential predictability" of different climate parameters. Potential predictability is a hard predictability limit intrinsic to the chaotic nature of the climate system (Sardeshmukh et al. 2000), a limit that most traditional dynamical forecasts often fail to reach due to the presence of model errors. As a result of this ceiling, further reduction of model biases may yield only incremental increases in forecast skill as predictability limits are reached for different aspects of the climate system (Newman & Sardeshmukh, 2017).

However, there may still be opportunities to improve seasonal forecast systems. Recent studies have shown that potential predictability limits are not stationary or fixed in time (Newman & Sardeshmukh, 2017; Weisheimer et al., 2022; Zhao et al., 2016). As a result, actual forecast skill has also varied substantially in the past (Derome et al., 2005; Kumar, 2009; MacLeod et al., 2018; O'Reilly et al., 2017, 2019; Shi et al., 2015; Weisheimer et al., 2017, 2019). For example, Lou et al. (2023) and Weisheimer et al. (2022) showed that long-lead ENSO forecast skill was higher at the beginning and end of the twentieth century, with a multidecadal period of lower skill from the 1930s-1950s. Further, Weisheimer et al. (2020) found that past seasonal predictability of extratropical atmospheric circulation patterns such as the Pacific-North American (PNA) pattern and the North Atlantic Oscillation (NAO) have also experienced pronounced decadal variations. While these past changes in prediction skill may result from varied model performance relative to historical observations (e.g., Weisheimer et al., 2022), these skill changes may also be driven by changes in the intrinsic predictability of the climate system itself (Becker et al. 2014; Newman and Sardeshmukh 2017).

Given these historical changes, it is reasonable to expect that potential predictability and actual prediction skill may similarly vary in the future, whether as a result of natural decadal variability (Weisheimer et al., 2020), a possible response to anthropogenic climate change (Zheng et al. 2022), or some combination of both. In particular, some general circulation models (GCMs) project that ENSO and its remote impacts may change in response to an increase in greenhouse gasses (e.g., Cai et al., 2021). For example, some models project significant changes in ENSO variability (Maher et al. 2023; Heede and Fedorov 2023), frequency (Berner et al. 2020), flavor (i.e., central vs eastern Pacific; Capotondi et al., 2015), and teleconnection strength/position (Gan et al. 2017; McGregor et al. 2022; O'Brien and Deser 2023; Zhou et al. 2014). Although, there is substantial uncertainty in the sign and intensity of these changes across models. Nevertheless, due to its far-reaching teleconnections (e.g., Horel & Wallace, 1981), ENSO is the single most important source of predictability on seasonal timescales for much of the globe (e.g., Barnett & Preisendorfer, 1987; Jacox et al., 2019; Quan et al., 2006). Therefore, any future changes to ENSO's strength and/or its connectivity to the rest of the climate system could significantly impact the potential predictability of many socioeconomically-relevant climate parameters.

It is crucial to assess how potential predictability may evolve as climate continues to change. Many previous studies have used hindcast systems to estimate potential predictability in the past (e.g., Shi et al., 2015; Weisheimer et al., 2019, 2020, 2022). However, model hindcasts are not useful for quantifying possible future changes in predictability as they are by definition retrospective and depend on past observations for their initialization. A different technique that can overcome these limitations and assess time-varying climate predictability in the past and the future is the "model-analog" approach. In the traditional analog framework, past observed climate states are found that closely match the current state and their subsequent evolution are treated as forecasts (Lorenz 1969). Alternatively, coupled GCMs allow for analogs to be drawn from climate simulations (often pre-industrial control runs; Ding et al., 2018), with the model evolution of these analogs then treated as the forecast. This method increases the "library" of possible climate states to compare against the current observed state, resulting in closer analog matches and allowing for the generation of forecast ensembles. Such model-analog forecasts have been shown to be as skillful as initialized dynamical forecasts (Ding et al. 2018, 2019), with the added benefit of being more computationally efficient.

The "perfect model-analog" technique utilizes these same methods, but whereas the goal of a traditional model-analog is to leverage climate simulations to forecast the real world, the goal of the perfect model framework is to instead forecast the climate simulation itself. This is accomplished by treating a portion of a climate simulation as "observations", and then drawing the analog forecasts from a different, independent portion of the same climate simulation. The resulting ensemble forecast is "perfect" in that it has no unconditional or conditional biases (von Storch & Zwiers, 1999). Thus, the forecast skill in a perfect model framework is a measure of the potential predictability (or equivalently, "potential skill") in the climate system. Since the perfect model framework does not depend on real world observations, it can be readily applied to past and future climate simulations to explore how these predictability limits change over time.

In this study, we quantify seasonal climate predictability limits from 1921-2100 by applying the perfect model framework to five coupled model initial condition large ensembles (LEs) that are each forced with time-varying radiative forcing. Model LEs have been widely used in climate science studies to separate the response to external forcing from internal climate variations (see review by Maher et al., 2021). In our analysis, the large number of ensemble members provided by each model LE (ranging from 30-100 depending on model) allows us to generate hundreds of thousands of perfect model forecasts with which to assess any future changes in potential predictability. In particular, we generate 24-month forecasts of global surface temperature, precipitation, and upper atmospheric circulation anomalies as well as for ENSO. The forecasts are then verified against independent portions of the same large ensembles using anomaly correlation coefficient (ACC) and reliability categories—a probabilistic measure of forecast skill. Finally, we relate future changes in potential predictability to future ENSO changes in each model.

2. Data and Methods

(a) Climate model simulations and observations

We apply the perfect model framework to five coupled model initial condition LEs that span the Coupled Model Intercomparison Project Phase 5 (CMIP5) and CMIP6 eras (Table 1). Such a comparison across models allows us to test the sensitivity of our results to inter-model uncertainty found in the climate response to increased radiative forcing. For efficiency, all model data output was first interpolated to a common 2.5˚ x 2.5˚ grid.

The models used in our analysis include: the Community Earth System Model version 1.2 LE (CESM1-LE; 40 members; (Kay et al. 2015), CESM version 2 LE (CESM2-LE; 100 members; Rodgers et al., 2021), the Geophysical Fluid Dynamics Laboratory Seamless System for Prediction and Earth System Research Medium Resolution Simulation (GFDL-SPEAR; 30 members; Delworth et al., 2020), the GFDL Earth System Modeling version 2M (GFDL-ESM2M; 30

members; Burger et al., 2022), and the Max-Planck Institute Grand Ensemble (MPI-GE; 100 members; Maher et al., 2019). The analysis period is 1921-2100, during which each model uses a specified external forcing scenario: (1) historical + retrospective emissions pathway 8.5 (RCP8.5), (2) historical + shared socioeconomic pathway 3-7.0 (SSP3-7.0), or (3) historical + SSP5-8.5.

Dataset	**Forcing (ens. size)**	$\boldsymbol{\sigma_{3.4}}$ **trend (˚C dec^{-1}) 1950-2022**	$\boldsymbol{\sigma_{3.4}}$ **trend (˚C dec^{-1}) 1950-2100**	**Reference**
CESM1-LE	HIST+RCP8.5 (40)	0.04 $\pm$ 0.03	0.02 $\pm$ 0.02	Kay et al. (2015)
CESM2-LE	HIST+SSP3-7.0 (100)	0.03 $\pm$ 0.04	0.00 $\pm$ 0.02	Rodgers et al. (2021)
GFDL-SPEAR	HIST+SSP5-8.5 (30)	0.02 $\pm$ 0.03	0.03 $\pm$ 0.01	Delworth et al. (2020)
GFDL-ESM2M	HIST+RCP8.5 (30)	0.02 $\pm$ 0.05	-0.02 $\pm$ 0.02	Burger et al. (2022)
MPI-GE	HIST+RCP8.5 (100)	0.00 $\pm$ 0.04	0.00 $\pm$ 0.01	Maher et al. (2019)
ERSSTv5		0.03		Huang et al. (2017)

Table 1 Observational and model datasets used in this study. First column: radiation forcing scenario used by each model. The number of ensemble members available in each model is in parentheses. Second column: December-February averaged Nino3.4 standard deviation ($\boldsymbol{\sigma_{3.4}}$) trend (˚C decade^{-1}) in 30-year running windows (i.e., Figure 1) for the period 1950-2022. For climate models, the ensemble mean trend is reported along with +/- one standard deviation. Third column: As in the second column, but for the period 1950-2100. Fourth column: Dataset references.

Within a given model, each ensemble member starts from a different initial condition. Over time, the ensemble members diverge due to the chaotic nature of the coupled climate system. As a result, once the memory of the initial condition fades, each ensemble member can be treated as an independent sample of the climate that has its own unique sequence of internal variability superimposed on a common forced response. We compare a portion of our model results to monthly mean data from National Oceanic and Atmospheric Administration (NOAA) Extended Reconstructed Sea Surface Temperature (SST) version 5 (ERSSTv5; Huang et al., 2017) from 1921-2022.

(b) Perfect model-analog framework

In each LE, perfect model forecasts are generated and evaluated for different 30-year periods spaced every 10 years from 1921-2100 (e.g., 1921-1950, 1931-1960…2071-2100). The forecasts are produced within each of these 30-year periods separately using the following method. For a given model and 30-year period:

(1) We extract SSTs from each ensemble member for the 30-year period of interest.

(2) We then remove the long-term monthly mean SSTs at each grid point based on the contemporaneous climatology calculated using all ensemble members (i.e., anomalies in 1921-1950 are relative to a 1921-1950 climatology).

(3) We further remove the ensemble mean SST anomaly (SSTA) (i.e., the model-specific externally-forced signal) at each grid point from each of the model's individual ensemble members.

(4) We arbitrarily treat the 1st ensemble member as the "truth" or "observations". Because each ensemble member is independent from one another, a data library of possible analog matches to the "observations" can then be constructed for each month using the remaining ensemble members. For example, the data library for January in CESM1-LE consists of 39 ensemble members x 28 years = 1092 samples. Note that it is only 28 years because we aim to generate 24-month forecasts, so any possible analog matches in the final two years would extend beyond our 30-year window of interest. Thus, the final two years in each 30-year window are excluded from our data libraries.

(5) For a given time step, we choose analogs by minimizing the distance between the climate state in the "observed" ensemble member and those found in the corresponding monthly data library (i.e., by comparing an "observed" January to the January data library). The distance between climate states is estimated by calculating the total root-mean-squared (RMS) difference between the "observed" SSTAs from 60˚S-60˚N and at all longitudes and those from every possible match in the data library. Note that we do not area weight the RMS difference calculation used in our analysis (see following section for more details). The distances are then ranked in descending order. The 10 closest states from the data library and their subsequent 24-month evolution are chosen as the forecast ensemble for that time step.

(6) We repeat (1)-(5) by treating each other model ensemble member as "observations" and constructing the monthly data library using the remaining ensemble members.

This procedure generates a 10-member forecast for every timestep and every ensemble member in a given model LE. For example, applying this perfect model framework to CESM1-LE for a given 30-year period generates 40 (ensemble members) x 12 (months) x 28 (years) x 10 (forecast members) = 134,400 24-month forecasts with which we can estimate seasonal climate predictability. Although we use SSTAs to identify analogs, we are not limited only to SSTA forecasts for analysis. Once the nearest climate states are selected, the evolution of any model variable can be treated as a forecast and subsequently verified against the corresponding variable from “observations” (e.g., Ding et al., 2019). In this way, we assess the forecast anomalies of the following variables from each model, with the CMIP standard variable name shown in parenthesis: SST (tos), 2m temperature over land (tas), precipitation (pr), and the 500mb streamfunction, which was calculated using the U/V wind components at 500mb (ua, va). As previously mentioned for SST, anomalies for all other variables are derived by removing both the long-term monthly means of the contemporaneous 30-year period and each model’s respective ensemble mean.

(c) Perfect model-analog sensitivities

There are several arbitrary choices that must be made when adapting the perfect model-analog technique for LEs. Here, we briefly discuss these decisions and how they might influence our results or conclusions. (1) We remove a given model’s ensemble mean from each of its members in order to isolate the internal component of each parameter, while still allowing for the rectification of the forced response on climate variability. Doing so allows us to focus on possible forced changes in the predictability of climate variations, as opposed to the more trivial exercise of predicting the forced trend. (2) Ding et al. (2018) showed that for data libraries of several hundreds of years, analog forecast ensembles of 10-20 members produced the most accurate forecasts. This is because larger forecast ensembles include increasingly poor analog matches, resulting in lower skill over the length of the forecast. We choose the top 10 analogs for our forecast ensembles for computational efficiency; however, our results and conclusions are not qualitatively impacted when increasing the forecast ensemble size to the top 15 or 20 matches. (3) We do not area weight the RMS difference calculation so as not to overweight the tropics when drawing analogs. We find that this choice increases the overall forecast skill in the mid-latitudes without overly decreasing it in the tropics. We select analogs based on SSTAs from 60˚S-60˚N and at all longitudes for similar reasons (i.e., to improve the representation of the extratropics when selecting analogs). Our results and conclusions are not qualitatively impacted by these decisions.

(d) Potential predictability metrics and signal-to-noise

To assess lead-dependent potential skill in each model, we calculate N_e estimates of the anomaly correlation coefficient (ACC) between each ensemble mean forecast and the corresponding "observations", where N_e is the number of ensemble members in a given LE (i.e., the number of "observed" timeseries used to generate analogs). For example, there are 40 estimates of the ACC for 1921-1950 when evaluating CESM1-LE. We repeat this procedure for each 30-year period separately, and we report the ensemble mean ACC in our results. We test the significance of the ensemble mean ACC using a 95% confidence interval based on two-sample *t*-test. We further determine the robustness of the change in ACC between 30-year periods by indicating where 80% of a given model's ensemble members agree on the sign of the change.

We further evaluate the forecasts using the reliability categories proposed by Weisheimer & Palmer (2014). Reliability categories are advantageous because they provide a highly interpretable measure of whether a forecast system is useful for decision making. Overall, there are five categories. Forecasts that fall into reliability category 5 are considered "Perfect", category 4 = "Very Useful", category 3 = "Marginally Useful", category 2 = "Not Useful", and category 1 = "Dangerously Useless". Reliability categories are defined by the slope of a forecast system's reliability diagram, which simply plots the observed frequency of a given event (say temperatures in the upper tercile) for different forecast probability bins. The slope of the reliability diagram is estimated using a weighted linear regression, where the weights are the number of samples in each probability bin. Using a bootstrapping technique with replacement, the uncertainty around the reliability slope is estimated by randomly resampling the forecasts and recomputing the slope. The reliability category is then determined based on the sign and magnitude of the reliability slope and whether or not the uncertainty intervals encompass the one-to-one perfect reliability line. See Weisheimer & Palmer (2014) for more details.

In our analysis, we assess the reliability categories of surface temperature and precipitation in the upper and lower terciles. We follow Weisheimer & Palmer (2014) with the following exceptions. First, for computational efficiency, we resample our forecasts 500 times when applying the bootstrapping algorithm. Second, we include the full range of reliability slope uncertainty (i.e., a 100% confidence interval) when calculating categories. Finally, because we are able to draw a large number of forecasts from the LEs (>100,000), we have enough data to calculate reliability categories at each grid cell. However, for brevity, we only show the fraction of the global area that falls within each category in our results. This contrasts from Weisheimer & Palmer (2014) and others who, in order to achieve a larger sample size, calculated a single reliability metric for large areas (e.g., all of North America) by aggregating short hindcast records in space. We do not expect any of these methodological differences to qualitatively influence our results or conclusions.

Finally, we assess the lead-dependent signal-to-noise (S2N) ratio in our forecasts following Sardeshmukh et al. (2000). For each model ensemble member e, the S2N ratio at lead l is:

$$S2N(e,l) = \left(\frac{\sum_{i=1}^{n} \overline{x_f}^2}{\frac{1}{K}\sum_{i=1}^{m} {x'_f}^2}\right)^{1/2} \quad (1)$$

Where $x'_f = x_f - \overline{x_f}$ is the deviation of each individual forecast member (x_f) from the ensemble mean forecast ($\overline{x_f}$) at each time step n, and m is n times the number of forecast ensemble members K (in our analysis K = 10). Therefore, for a given 30-year period, n = 12 (months) x 28 (years) = 336 and m = 3360. As with ACC, we calculate N_e estimates of the S2N ratio for each LE (one for each ensemble member), and report the ensemble mean values in our results. A higher S2N ratio indicates that there is a larger ensemble mean anomaly and/or less spread among the forecast ensemble, which results in a more skillful forecast in the perfect model framework (Sardeshmukh et al. 2000).

3. Results

(a) Forced changes in ENSO amplitude

Given ENSO's dominant role in driving seasonal climate predictability, we first assess the simulated response of ENSO amplitude to historical and future radiative forcing in each LE. The CESM1-LE shows a consistent increase in Nino3.4 (i.e., SST anomalies or SSTA, averaged 5˚S-5˚N, 170˚W-120˚W) standard deviation from 1921-2060, after which it levels off (Figure 1 and Table 1). The Nino3.4 amplitude in GFDL-SPEAR is relatively stable from 1921-2020, after which it increases until about 2080 before decreasing slightly. In contrast, the ENSO variability in CESM2-LE rises consistently through 2040 before decreasing consistently through 2100. The positive ENSO amplitude trends from 1921-2022 in the ensemble means of CESM1-LE, CESM2-LE and GFDL-SPEAR compare favorably to observations (Figure 1 black line; Table 1), although CESM1-LE and CESM2-LE show large positive ENSO variability biases. While GFDL-ESM2M also exhibits positive ENSO variability biases, its Nino3.4 standard deviation is relatively stable until about 2040, after which it sharply decreases through the end of the century. In MPI-GE, there is little change in Nino3.4 variability throughout the record. The large inter-model uncertainty in future ENSO variability is consistent with Maher et al. (2023) (see their Figure 4). Based on these results, we primarily focus on CESM1-LE, MPI-GE, and GFDL-ESM2M when evaluating the forecast skill of our perfect model-analogs as these LEs span the range of possible changes in future ENSO amplitude (i.e., increasing, no change, and decreasing, respectively).

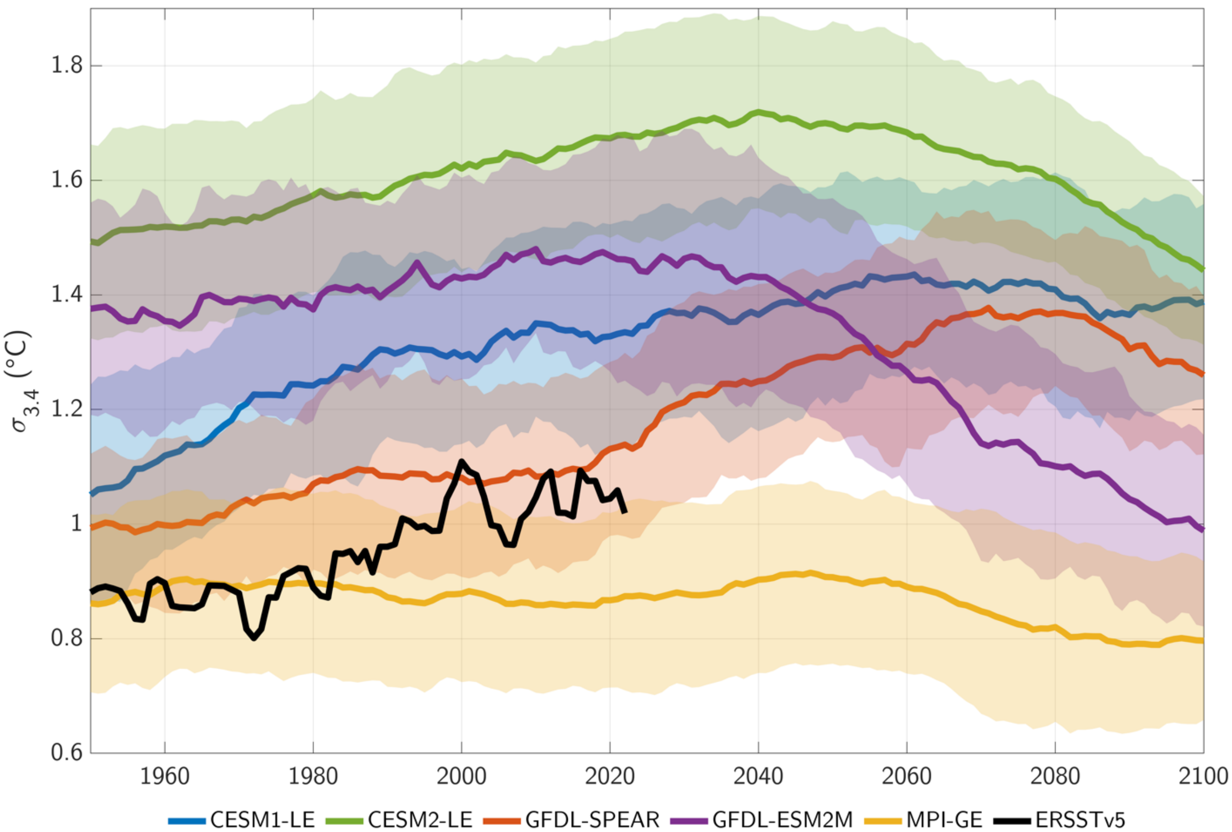

Figure 1 Standard deviation of December-February averaged SSTA in the Nino3.4 region in running 30-year windows from 1921-2100. Years indicate end of the window (e.g., 1960=1931-1960). Colors represent different model large ensembles, with thick curves for ensemble mean values and shading for the one standard deviation spread across the ensemble. Black curve shows the observed values based on ERSSTv5 from 1921-2022.

(b) Potential predictability and future changes, ACC

1) Sea surface temperature and surface air temperature

Perfect model-analog forecasts (hereafter referred to as "forecasts") of SSTA for 1921-1950 in CESM1-LE show significant potential skill (hereafter referred to as "skill") at 0-month lead for most of the globe (globally averaged ACC = 0.62), with the tropical Pacific exhibiting the highest skill (ACCs > 0.9; Figure 2a). There is also significant skill of surface air temperature anomalies over land (SATA) at 0-month lead in most regions. However, SATA skill is generally weaker than for SSTA (global average ACC = 0.48), especially in mid-latitudes. The higher overall SSTA skill or "potential predictability" (hereafter referred to as "predictability") at 0-month lead is expected since our analogs are chosen by minimizing the distance between the "observed" SSTA and the data library. Indeed, the high 0-month lead SSTA skill gives us confidence that the perfect model framework is reliably drawing analogs that closely correspond to the "observed" climate states at each time step. Results are similar for the other LEs (Figures S1-S5).

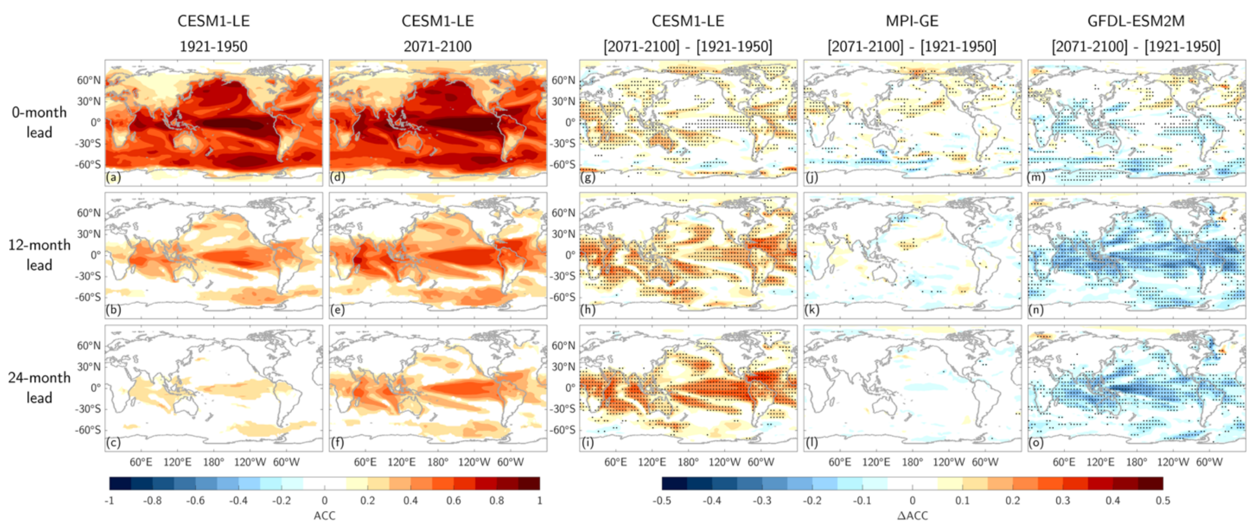

Figure 2 Surface temperature potential predictability. (a)-(c) Ensemble mean skill of surface temperature anomalies in CESM1-LE as measured by ACC calculated across all months in the period 1921-1950. (d)-(f) As in (a)-(c), but for the period 2071-2100. (g)-(o) Change in ACC between past and future periods for (g)-(i) CESM1-LE (j)-(l) MPI-GE (m)-(o) GFDL-ESM2M. Skill values in (a)-(f) are only shown when 95% significant. Stipples in (g)-(o) indicate where 80% of a respective model's ensemble agrees on the sign of the change. See Figures S1-S5 for the full surface temperature anomaly skill in the other large ensembles.

We further assess the predictability at increasing lead times; however, for brevity, we only show the skill at 12-month and 24-month leads (Figure 2b-c; see Figures S1-S5 for skill maps at additional lead times). Skill of surface temperature decreases with increasing lead time, although this reduction is more apparent for SATA than for SSTA. This difference is consistent with the higher thermal capacity of the ocean relative to the atmosphere, which typically leads to higher predictability at longer leads for SSTA than for SATA. In particular, SSTA ACCs at 12-month lead exceed 0.6 in the tropical Pacific, consistent with previous model-analog forecast studies (e.g., Ding et al., 2018). There is also significant SATA predictability over tropical land surfaces, as well as significant SSTA predictability throughout most of the North Pacific, the tropical Atlantic, the tropical Indian Ocean, and the Southern Ocean west of the Drake Passage. These regions are known to be influenced by large-scale ENSO teleconnections (e.g., He et al., 2020; Horel & Wallace, 1981; Mo & Ghil, 1987), suggesting that ENSO is a key source of long-lead predictability in our forecasts. Skill further degrades out to 24-month leads (Figure 2c); however, there is still the significant SATA skill over northern South America and significant SSTA skill in the tropical and South Pacific and the Indian Ocean.

In CESM1-LE, there is a robust increase in SSTA and SATA predictability in the future at all leads, with only a few small regions of decreasing predictability (Figures 2d-i). In particular, the 0-month lead SSTA skill increases in the western tropical Pacific as well as the Indian and Atlantic Oceans (Figure 2i). Similarly, there is a robust increase in future SATA predictability at 0-month lead over much of Africa, portions of eastern Asia, equatorial South America, and all of Australia. An increase in forecast skill at 0-month lead implies that the distance between the "observed" and analog climate states decreases in the future (i.e., the analogs more closely match the "observations"). Further, the widespread ensemble agreement (black stipples) indicates that these predictability changes are a "robust" (defined here as 80% ensemble agreement on the sign of the change) part of the model's forced response and not due to random natural decadal variations.

The CESM1-LE changes in SSTA/SATA predictably are starker at 12 and 24-month leads (Figures 2h-i), with robust increases in ACC throughout the global tropics in an ENSO-like pattern. The increased predictability along the equatorial Pacific, in particular, suggests that ENSO itself is more predictable in the future in CESM1-LE. We will explore ENSO predictability in more detail in Section 3d. There are also robust long-lead increases in SSTA and SATA predictability in the mid-latitudes. For example, there is an increase in SSTA skill in the North Atlantic in a pattern reminiscent of the SSTA footprint generated by the NAO (i.e., a horseshoe shape from southern Greenland to the tropical North Atlantic; Kushnir et al., 2006). There are also pronounced increases in SSTA skill in the North Pacific and along the U.S. west coast and SATA skill in the American Southwest, which may be associated with an eastward shift in ENSO's teleconnections to the Pacific North America region (O'Brien and Deser 2023). Other LEs generally disagree with CESM1-LE on the sign and magnitude of future predictability changes (Figure 2j-o and Figures S1-15). The MPI-GE at 0-month lead shows some isolated regions of increasing and decreasing SSTA/SATA skill, but without a clear pattern. At longer leads, the skill change in MPI-GE is close to zero nearly everywhere and there is little agreement among the ensemble on the sign of the change. In contrast, GFDL-ESM2M shows a robust decrease in SSTA/SATA predictability for most the globe (Figure 2m-o) in a similar ENSO-like pattern as seen in CESM1-LE (pattern correlation = -0.68 at 12-month lead), though with less loading in the Northeast Atlantic. This suggests that ENSO predictability decreases in the future in GFDL-ESM2M.

2) Precipitation

Forecasts of precipitation anomalies for 1921-1950 in CESM1-LE show peak skill over the tropical oceans (Figure 3a-c; see Figures S6-S10 for other models). For example, 0-month lead precipitation skill is highest over the central equatorial Pacific, with ACCs exceeding 0.9. There is also significant skill at 0-month lead over tropical land surfaces and in the mid-latitudes along the U.S. west coast. Precipitation predictability similarly decreases with increasing lead, with only the equatorial Pacific and Indian Oceans displaying any significant skill at 12-month lead. By 24-month lead, precipitation predictability is generally insignificant, except for isolated regions in the Indo-Pacific warm pool.

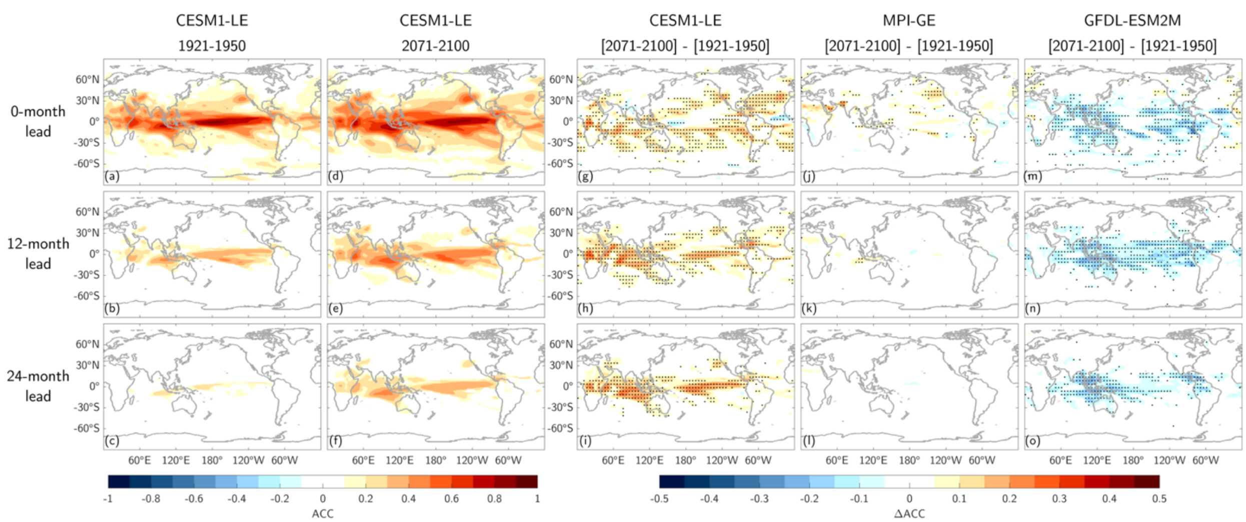

Figure 3 As in Figure 2, but for precipitation predictability.

Similar to SSTA/SATA, there are robust increases in future precipitation predictability at all leads in CESM1-LE (Figure 3d-i), with centers of action in the Indian Ocean, the equatorial Pacific, the Caribbean, and the U.S. west coast. The North Atlantic also shows robust increases in predictably at 0- and 12-month lead. The increase in predictability at 24-month lead is of particular note given that there is virtually no significant skill in the past. In the future, however, there is significant predictability over the equatorial Pacific and Indian Oceans. Additionally, the region of highest skill along the equatorial Pacific shifts eastward from about the dateline in the period 1921-1950 to about 140˚W in the period 2071-2100. This eastward shift may be related to CESM1-LE simulated El Niño events shifting eastward in the future (O'Brien and Deser 2023; Williams and Patricola 2018). The sign and relative magnitude of the skill changes in the other LEs are also consistent with their respective SSTA/SATA predictability changes (Figures 3j-o). Specifically, MPI-GE once again shows isolated regions of robust precipitation skill change at 0-month, but no significant change at longer leads. Similarly, GFDL-ESM2M shows a robust decrease in precipitation predictability at all leads throughout the tropics.

3) Upper atmosphere circulation

Forecasts of 500mb streamfunction anomalies (ψ_{500}) during the period 1921-1950 in CESM1-LE show significant skill at 0-month and 12-month leads (Figure 4a-c; see Figures S11-S15 for other models). In particular, there are regions of high ACC in the subtropical and mid-latitude North and South Pacific as well as over North America. These centers of action are consistent with the locations of the PNA and Pacific-South American (PSA) patterns (Horel and Wallace 1981; Mo and Ghil 1987). Combined, these two results suggest that the forecasts are successfully capturing the upper atmospheric wave train response to tropical heating anomalies associated with ENSO.

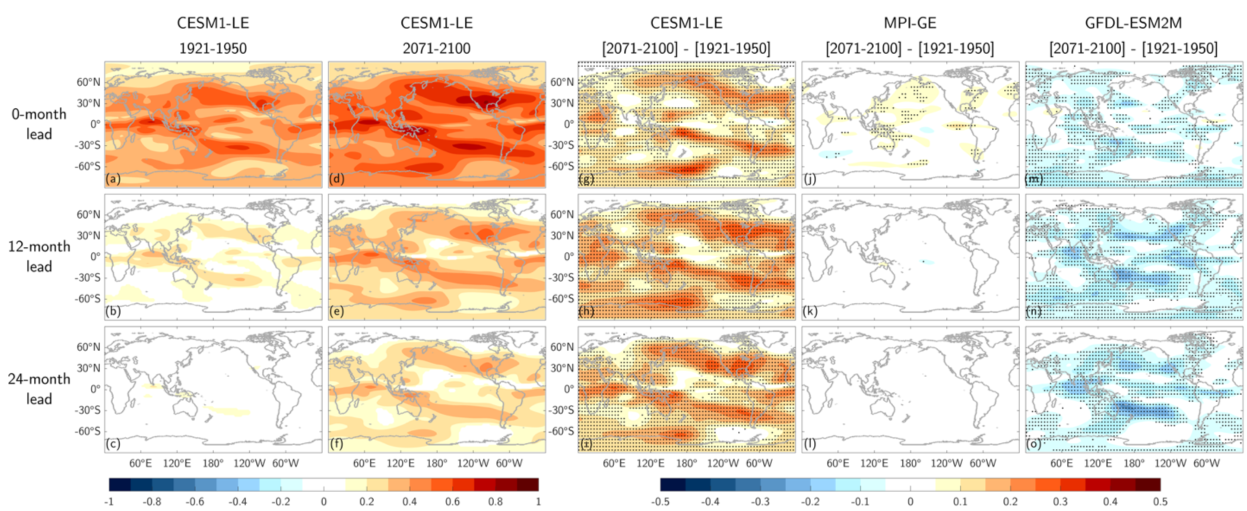

Figure 4 As in Figure 2, but for 500mb streamfunction (ψ_{500}) predictability.

In the future, there is a near-global increase in CESM1-LE ψ_{500} predictability at all leads (Figure 4d-i). Of note are the increases in ψ_{500} ACC in the PNA and PSA regions, respectively, which may be an indication of stronger ENSO teleconnections in CESM1-LE in the future (O'Brien and Deser 2023). The robust predictability increases in the PNA region are also consistent with the increases seen in both SSTA/SATA and precipitation predictability along the U.S. west coast (see Figures 2g-i and 3g-i). Similar to precipitation forecasts, long-lead ψ_{500} predictability is especially impacted in CESM1-LE, with significant increases in predictability nearly everywhere at 24-month lead. As with SSTA, SATA, and precipitation, the other LEs disagree with CESM1-LE on the sign of future ψ_{500} predictability changes (Figure 4j-o). MPI-GE shows no regions of robust predictability changes beyond 0-month lead, and GFDL-ESM2M once again produces a decrease in ψ_{500} skill for most of the globe. In particular, GFDL-ESM2M shows a decrease in predictability in the PNA and PSA regions of the North and South Pacific, which may suggest that ENSO-related teleconnections in this model are weaker in the future.

(c) Potential predictability and future changes, reliability

1) Surface temperature

To test the sensitivity of our results to our choice of skill metric, we further evaluate future changes in climate predictability using probabilistic reliability categories. Upper tercile surface temperature (including SSTA and SATA) forecasts in CESM1-LE show strong reliability during the period 1921-1950 (Figure 5a-c; blue bars). For example, at 0-month lead, 34% and 59% of the globe falls within the category 5 ("perfect") and 4 ("very useful") forecast bins, respectively. The fraction of the globe in these higher categories decreases with increasing lead time, with the majority of forecasts across the globe falling into reliability category 3 ("marginally useful") by 12- (area fraction = 65%) and 24-month (area fraction = 67%) lead. In the future, CESM1-LE surface temperature forecasts become more reliable (Figure 5a-c; red bars), with a clear shift in the distribution towards higher reliability categories at all leads. For example, at 12-month leads, the global area fraction of forecasts that fall into reliability category 4 increases from 25% in 1921-1950 to 43% by 2071-2100, with a corresponding decrease in reliability category 2 ("not useful") and 1 ("dangerously useless") forecasts.

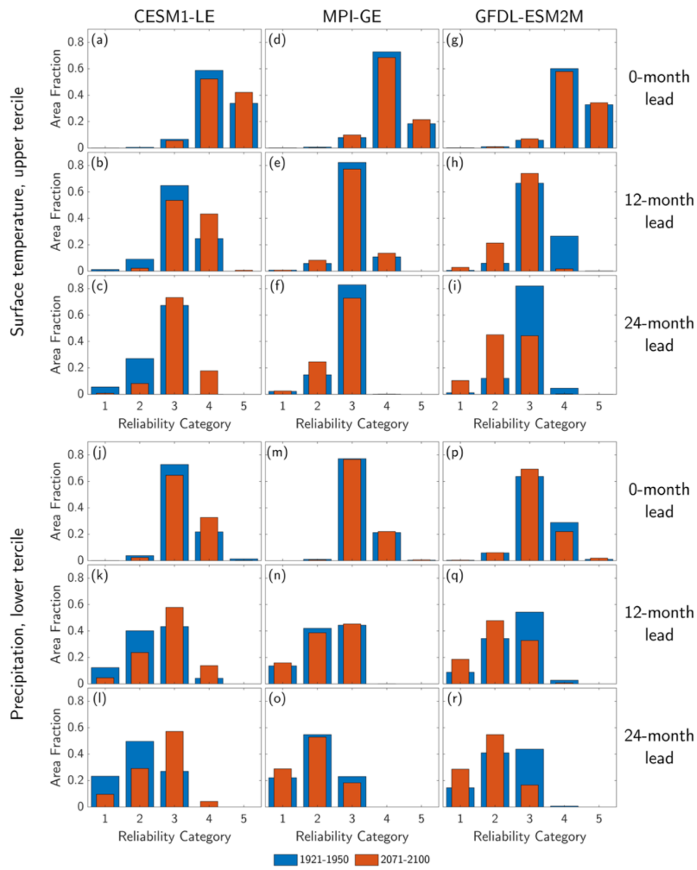

Figure 5 Fraction of global area in each reliability category for (a)-(i) forecasts of upper tercile surface temperature anomalies and (j)-(r) forecasts of lower tercile precipitation anomalies. Values are for 0-, 12-, and 24-month leads in (left column) CESM1-LE, (middle column) MPI-GE, and (right column) GFDL-ESM2M. The reliability categories are 5 = perfect, 4 = very useful, 3 = marginally useful, 2 = not useful, and 1 = dangerously useless. Each category is calculated across all months in the periods (blue) 1921-1950 and (red) 2071-2100.

For the period 1921-1950, upper tercile surface temperature forecasts from MPI-GE and GFDL-ESM2M produce a similar distribution of reliability categories as CESM1-LE (Figure 5d-i; blue bars). Both LEs have mostly category 4 and 5 forecasts at 0-month lead, with the distribution shifting towards lower reliabilities at longer leads. By 12-month lead, forecasts for 82% of the global area fall within category 3 for MPI-GE, while forecasts for 67% of the global area fall within the same category for GFDL-ESM2M. In the future period, the global area fraction within each reliability category for MPI-GE remains relatively stable at all leads (Figure 5d-f), with only a small decrease in category 3 forecasts (from 83% to 73%) and corresponding increase in category 2 forecasts at 24-month lead (from 15% to 25%). While the reliability distribution for GFDL-ESM2M forecasts do not change much at 0-month, there is a noticeable shift towards lower categories at 12- and 24-month lead going from the period 1921-1950 to 2071-2100 (Figure 5h-i).

At 24-month lead, the global area fraction with category 3 forecasts in GFDL-ESM2M decreases from 82% to 44% and the global area fraction with category 2 forecasts increases from 12% to 45%. Therefore, forecasts of upper tercile surface temperature in GFDL-ESM2M become less reliable in the future for most of the globe, consistent with the decreasing ACCs shown previously.

2) Precipitation

Repeating this analysis for lower tercile precipitation forecasts, we find that the CESM1-LE precipitation forecasts are overall less reliable than the surface temperature forecasts, as indicated by skew of the reliability distribution towards category 1-3 forecasts at all leads (Figure 5j-l). However, the future change in lower tercile precipitation forecast reliability in CESM1-LE is consistent with that seen in upper tercile surface temperature, with a clear shift in the distribution towards higher categories. For example, at 24-month lead, the global area fraction with category 3 forecasts increases from 27% to 57% between 1921-1950 and 2071-2100 with a corresponding decrease from 50% to 29% for category 2 forecasts. The future changes in lower tercile precipitation reliability in MPI-GE and GFDL-ESM2M are also consistent with their respective surface temperature reliability changes, with MPI-GE forecasts showing little change in the reliability distribution (Figure 5m-o), and GFDL-ESM2M showing a clear shift towards categories 1-2 (Figure 5p-r). In particular, at 24-month lead in GFDL-ESM2M, the area fraction with category 3 forecasts decreases in the future from 44% to 17%, with a corresponding increase in category 1 and 2 forecasts. The above results are consistent for lower and upper tercile forecasts of surface temperature and precipitation, respectively (Figure S16).

(d) Linking future predictability changes and ENSO amplitude

1) Signal-to-Noise

To briefly summarize the above results, seasonal climate predictability in the future generally increases in CESM1-LE, does not change in MPI-GE, and decreases in GFDL-ESM2M, as measured by different forecast skill metrics (ACC and reliability) across multiple variables (SSTA, SATA, precipitation, and ψ_{500}). While the models disagree on the sign of future predictability changes, they are each self-consistent with their projected change in future ENSO amplitude (i.e., Figure 1). The link between future climate predictability and future ENSO amplitude may be related to ENSO's role as the dominant internal climate mode, allowing one to detect its influence across much of the globe despite the presence of other forms of variability (e.g., weather or other climate modes). For example, if ENSO amplitude increases in the future (e.g., as projected by CESM1-LE), then that may lead to an increase in the signal-to-noise (S2N) ratio of ENSO and its teleconnections, which would tend to contribute to an overall more deterministic climate system and more skillful forecasts (e.g., Sardeshmukh et al., 2000). To test this hypothesis, we calculate changes in the S2N ratio (Eq. 1) for surface temperature as a function of lead time in each of the two time periods (Figure 6). During the period 1921-1950, the S2N ratios in CESM1-LE forecasts at 0-month lead follow an ENSO-like pattern, with the highest values in the equatorial Pacific (maximum value = 1.94). Weaker (but still elevated) values are seen in the Indian Ocean, the South Pacific, the Northeast Pacific along the U.S. west coast, the North Atlantic, and over the tropical African and South American land surfaces (Figure 6a). The S2N decreases with increasing lead time (Figure 6b-c); however, the ENSO-like pattern of elevated S2N persists at 12-month lead before mostly dissipating at 24-month lead.

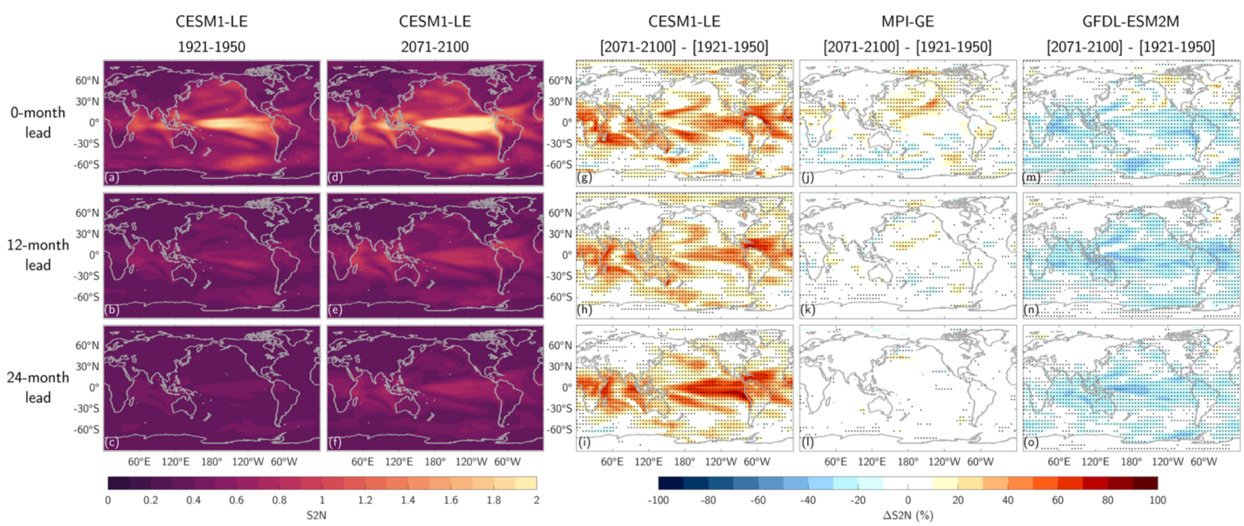

Figure 6 Signal-to-noise (S2N) ratios for surface temperature anomaly forecasts. (a)-(c) Ensemble mean S2N of surface temperature forecasts in CESM1-LE calculated across all months in the period 1921-1950. (d)-(f) As in (a)-(c), but for the period 2071-2100. (g)-(o) Percent change in S2N between past and future periods for (g)-(i) CESM1-LE (j)-(l) MPI-GE (m)-(o) GFDL-ESM2M. Stipples in (g)-(o) indicate where 80% of a respective model's ensemble agrees on the sign of the change.

The patterns of future S2N change in each of the LEs are remarkably similar to the surface temperature ACC changes seen in Figure 2 (Figure 6d-o), with pattern correlations between the ACC and S2N maps at 0-, 12-, and 24-month lead of 0.86, 0.97, and 0.98 for CESM1-LE, 0.76, 0.90, and 0.83 for MPI-GE, and 0.69, 0.95, and 0.95 for GFDL-ESM2M, respectively.

Decomposing the S2N equation into a signal and noise component (i.e., the numerator and denominator of Eq. 1, respectively), we find that the changes in the signal are over five times larger than changes in the noise for much of the globe (Figures S17-S18). For example, the signal change averaged 60˚S-60˚N at 12-month lead in CESM1-LE is 27%, compared to just a 4.7% change in the noise. In the case of CESM1-LE, this indicates that the amplitude of a typical ensemble mean forecast anomaly is larger in the future without a substantial increase in the average forecast spread (i.e., the forecast uncertainty). These results are consistent with previous studies linking ENSO amplitude to S2N and/or climate predictability (Capotondi et al., 2015; Chen et al., 2004; Gu & Philander, 1997; Sardeshmukh et al., 2000; Suarez & Schopf, 1988; Weisheimer et al., 2022; Zhao et al., 2016).

2) Time-varying potential predictability changes, Nino3.4

To further relate changes in ENSO amplitude to global predictability, we explore skill changes as a function of time. A time-varying perspective of predictability is important given the non-monotonic changes in ENSO amplitude seen in most LEs (e.g., Figure 1). Such variability in each model's forced ENSO response may give rise to periods of predictability that differ not only from the historical period, but also from the total changes seen at the end of the 21st century (i.e., Figures 2-4). Further, by evaluating whether time-varying skill changes are robust across a given model's ensemble, we can quantitatively estimate the "time of emergence" for forced changes in predictability within each model.

To illustrate, we show the forecast skill of SSTAs averaged in the Nino3.4 region for six different 30-year periods from 1921-2100 (Figure 7). In addition to CESM1-LE, MPI-GE, and GFDL-ESM2M, we also include CESM2-LE and GFDL-SPEAR in this analysis as ENSO amplitude changes in these models are particularly varied, with prolonged periods of increasing and decreasing variability. Treating 1921-1950 as the base period, Nino3.4 skill tends to be highest (exceeding 0.8) at leads of less than ~6 months and for forecasts initialized in boreal fall and winter (Figure 7; left column). For boreal spring and summer initializations, predictability tends to be similarly elevated at leads that encompass boreal winter in the forecast. For example, June initialized forecasts in GFDL-ESM2M show a peak in Nino3.4 skill at 2-10 month leads, and then again at 16-22 month leads (i.e., October-April of the following year).

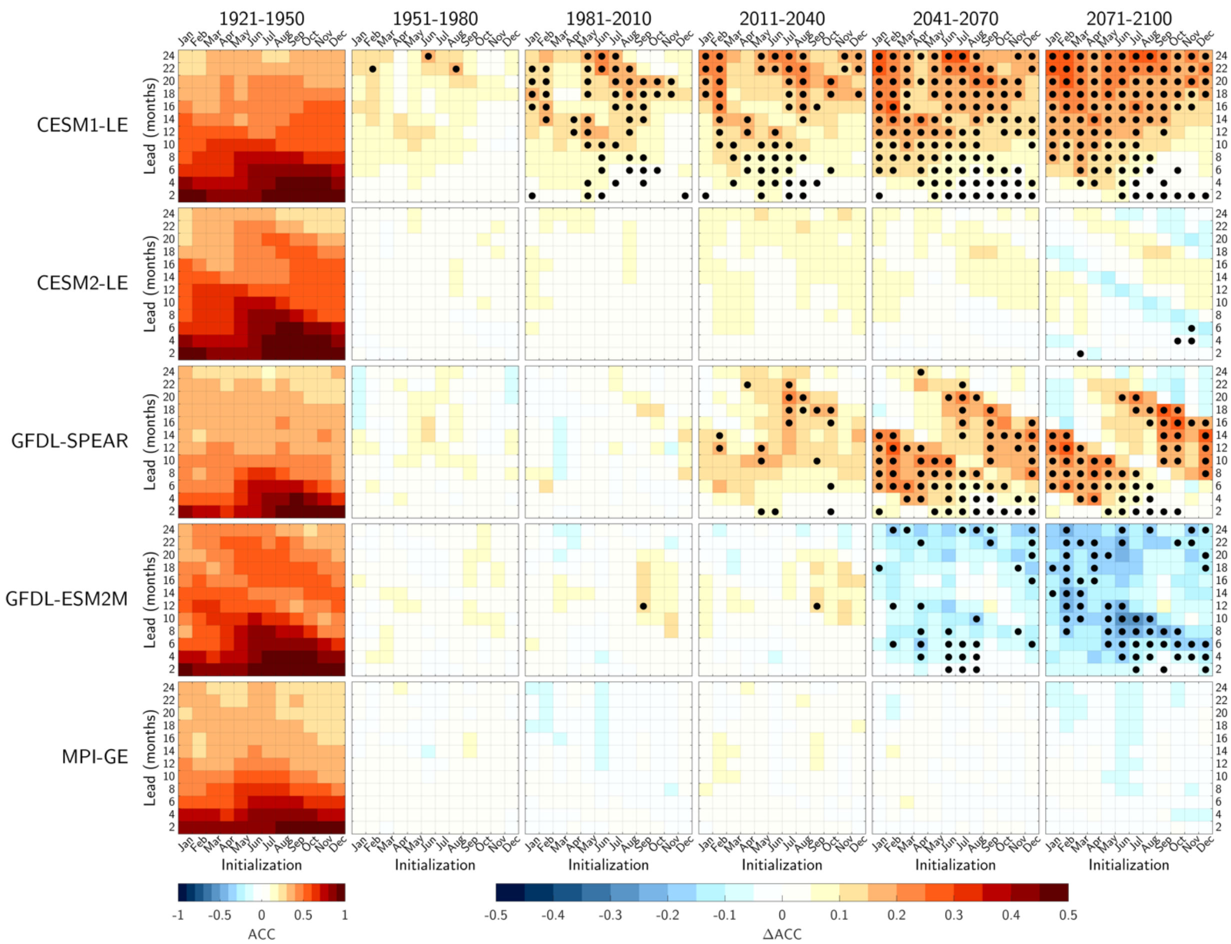

Figure 7 First column: Ensemble mean Nino3.4 potential forecast skill (ACC) as a function of initialization month (x-axis) and lead time (y-axis) for each model large ensemble. Second-fifth columns: Difference in Nino3.4 skill between the base period 1921-1950 and different 30-year periods. For example, the second column shows the difference in skill between the periods 1951-1980 and 1921-1950. Stipples indicate that 80% of the respective model ensemble agrees on the sign of the change.

There is little change in Nino3.4 skill in any of the models for the adjacent 30-year period (1951-1980). However, by the period 1981-2010, CESM1-LE shows a robust increase in Nino3.4 predictability at short leads for May-September initializations and at longer leads for much of the year. This suggests that forced changes in CESM1-LE ENSO predictability begin to emerge above the internal noise inherent to each ensemble member during this period. In 2011-2040, CESM1-LE Nino3.4 skill continues to increase, while GFDL-SPEAR begins to show some robust increases in predictability. Forecast skill in CESM2 also increases slightly during this period, but there is not widespread agreement among its ensemble on the sign of this change. We see the largest period-to-period changes in Nino3.4 skill between 2011-2040 and 2041-2070 (Figure 7; fifth column). For example, CESM1-LE shows robust increases in predictability for leads less than 8 months when initialized in boreal summer to winter and at nearly all initializations beyond 16-month lead. Forced changes to ENSO forecast skill in GFDL-SPEAR also fully emerge during this period, with diagonal bands of increased predictability associated with forecasts that verify in boreal summer to winter. In GFDL-ESM2M, robust decreases in predictability begin to emerge, but without a clear pattern. Finally, by the period 2071-2100, CESM1-LE and GFDL-SPEAR largely maintain the increases in ENSO predictability observed in the previous epoch, while forced decreases in Nino3.4 forecast skill are now fully evident in GFDL-ESM2M.

3) Time-varying potential predictability changes, global

There is clear model diversity in the simulated change of ENSO predictability, both in the sign and intensity of end-of-21st century changes and in the apparent time of emergence for each model's forced response (i.e., Figure 7 black dots). However, similar to our previous results (e.g., Figures 2-5), the sign and timing of ENSO predictability changes in each of the LEs is consistent with their respective time-varying ENSO amplitudes (Figure 1). For example, there are no robust changes in Nino3.4 forecast skill in GFDL-SPEAR until the period 2011-2040, which closely corresponds to the timing of the strongest increasing trend in this model's ENSO amplitude (comparing third row of Figure 7 to orange line in Figure 1). Similarly, ENSO predictability in GFDL-ESM2M remains relatively stable until the period 2041-2070, at which point both the forecast skill and GFDL-ESM2M's ENSO amplitude start to sharply decrease (comparing fourth row of Figure 7 to purple line in Figure 1). The ensemble mean Nino3.4 skill in CESM2-LE also shows hints of a close link to its time-varying ENSO amplitude, with a slight increase in skill/amplitude through 2040 followed by a decrease through the end of the century, though these predictability changes are not robust across the CESM2 ensemble.

The relationship between time-varying ENSO amplitude and climate predictability extends beyond the Nino3.4 region, manifesting on global scales via ENSO-driven changes in the S2N ratio (as previously suggested in Figure 6). Indeed, we find a high correspondence in each LE between their respective time-evolving Nino3.4 amplitude, globally averaged ACC, and globally averaged S2N ratio (Figure 8). For example, at 6-month lead, the globally averaged SSTA skill in CESM1-LE increases roughly linearly over time with increasing ENSO amplitude (Figure 8a circles; $R = 0.95$; Table 2), with over 80% of the model ensemble agreeing on the sign of both the ENSO amplitude and global predictability changes beginning in the period 1981-2010 (i.e., circles with thick black outline). A similar linear relationship is seen in other LEs with different ENSO amplitude trends. For example, in GFDL-ESM2M, there is a decrease in skill over time that closely corresponds to this model's decrease in ENSO amplitude (Figure 8a triangles; R = 0.97), although its forced changes in predictability are not apparent until the period 2041-2070.

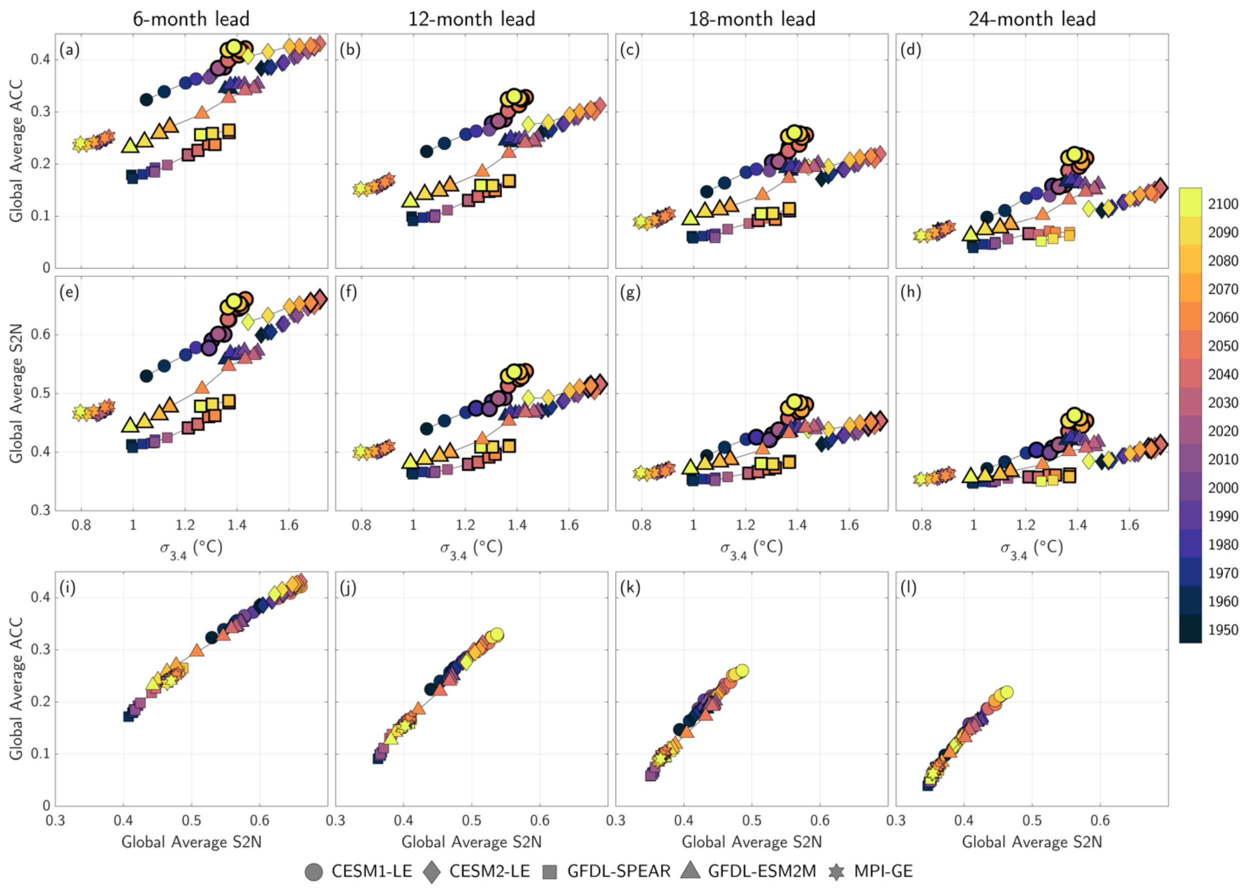

Figure 8 (a)-(d) Global average ensemble mean potential skill at different leads (y-axis) versus December-February averaged Nino3.4 standard deviation (x-axis) in different 30-year periods. (e)-(h) As in (a)-(d), but for global average forecast S2N ratio versus Nino3.4 standard deviation. (i)-(l) As in (a)-(d), but for global average ACC versus global average S2N ratio. All ACC and S2N values are based on ensemble mean SSTA forecasts from each model (i.e., different shapes). Shading of each shape indicates the 30-year window over which the forecast skill, S2N ratio or Nino3.4 standard deviation are calculated, with the year indicating the end of the window. For example, the shading for 1950 corresponds to 1921-1950. Markers with bold outlines in (a)-(h) indicate 30-year windows in which 80% of a given model's ensemble agree on the sign of the change (relative to 1921-1950) for both the ACC/S2N and Nino3.4 standard deviation.

Globally averaged S2N is highly correlated in time with each model's projected ENSO amplitude (Figure 8e-h and Table 2), consistent with the S2N maps discussed earlier. There is also a near-perfect linear relationship between globally averaged S2N and ACC (Figure 8i-l and Table 2), consistent with previous studies relating perfect model skill to S2N (Sardeshmukh et al. 2000). Combined, these results further support our hypothesis that time-varying changes in predictability are driven by same-sign changes in global S2N ratios, which in turn are driven by each respective LE's projected change in ENSO amplitude. The close link between ENSO amplitude, S2N, and forecast skill is consistent across lead times, models (different marker types in Figure 8), and variables (Figures S19-S20). However, the estimated time of emergence for each model's forced response in predictability varies widely from model-to-model, ranging from as early as 1981-2010 in CESM1-LE to as late as 2041-2070 in GFDL-ESM2M at 6-month lead (Table 2).

Dataset	**R(ACC, $\sigma_{3.4}$)**	**R(S2N, $\sigma_{3.4}$)**	**R(ACC, S2N)**	**ToE (ACC, $\sigma_{3.4}$)**	**ToE (S2N, $\sigma_{3.4}$)**
CESM1-LE	0.95	0.95	1.0	2010	1970
CESM2-LE	0.72	0.83	0.98	Not robust	Not robust
GFDL-SPEAR	0.97	0.96	1.0	2030	2030
GFDL-ESM2M	0.97	0.98	1.0	2070	2060
MPI-GE	0.85	0.58	0.90	Not robust	Not robust
All models	0.82	0.82	1.0		

Table 2 Potential skill (ACC), signal-to-noise (S2N), and ENSO amplitude relationships at 6-month lead. First column: Correlation between globally averaged SSTA potential skill and December-February averaged Nino3.4 standard deviation ($\sigma_{3.4}$) for different 30-year windows spanning 1921-2100 (i.e., Figure 8a). Second column: As in the first column, but for globally averaged SSTA signal-to-noise (S2N) ratios (i.e. Figure 8b). Third column: As in the first column, but for globally averaged potential skill and S2N ratios. Fourth column: Time of emergence (ToE) of a given model's forced change in globally averaged SSTA predictability and ENSO amplitude. The ToE is estimated as the first 30-year period in which 80% of a given model's ensemble agrees on the sign of both the potential skill change and Nino3.4 amplitude change. Values reported only if the model ensemble continues to agree on the sign of change through the end of record. The year indicates the end of the 30-year window (e.g., 2010 = 1981-2010). Fifth column: As in the fourth column, but for globally averaged S2N ratios. Results are consistent for other leads.

4. Summary and Discussion

In this study, we investigated future changes in seasonal potential predictability across five coupled GCM LEs. Using a perfect model-analog technique, we generated hundreds of thousands of synthetic seasonal forecasts to estimate predictability changes from 1921-2100. CESM1-LE consistently showed a robust increase in predictability in the future, while predictability in GFDL-ESM2M consistently decreased (e.g., Figures 2-5). These predictability changes were largest at longer leads. In contrast, seasonal predictability in MPI-GE did not exhibit significant changes.

While there was large inter-model uncertainty in the sign, magnitude, and timing of future climate predictability changes, we showed that a common physical mechanism emerges that allows us to anticipate how real-world predictability may change in the coming decades. In particular, the predictability changes in each model were driven by a same-sign change in their respective ENSO amplitude. For example, forecasts from models with increasing ENSO amplitude trends (e.g., CESM1, GFDL-SPEAR, and CESM2 until ~2040) were associated with a higher S2N ratio in the future, which led to an overall more deterministic climate system and increased potential for significant forecast skill. The higher S2N ratio resulted from a larger ensemble mean forecast anomaly (i.e., signal), owing to ENSO's role as a bigger "hammer" to the climate system. The opposite was true for models with decreasing ENSO trends (e.g., GFDL-ESM2M and CESM2 after ~2040).

While previous studies have highlighted natural variations in climate predictability in the past (e.g., Weisheimer et al., 2020), our finding that changes to potential predictability limits are a key component of the response to increased radiative forcing has important implications for future seasonal forecasting systems. Whereas natural variations in climate predictability are random in time and include periods of both high and low predictability, our model results indicate that forced changes in climate predictability are often associated with a long-term shift towards either higher or lower predictability without a prolonged return to historical baselines. This suggests that any future deviations from historical forecast skill relationships may represent a shift in the climate system towards a new predictability regime, rather than a temporary excursion driven by internal variability. Although, non-monotonic forced changes in predictability back towards historical predictability limits are also possible (e.g., as in CESM2-LE).

The climate models analyzed here do not agree on the direction of future predictability changes, but the close link between skill and each model's ENSO amplitude allows us consider the future direction of predictability based on recent observations. Since 1970, the observed trend in ENSO amplitude is positive (Figure 1). Should this trend persist into the future, we might also expect seasonal forecast skill to increase alongside predictability in regions strongly influenced by ENSO and its teleconnections as these portions of the climate system become more deterministic. Of course, this assumes that perfect model predictability is a reasonable proxy for "actual" skill (e.g., skill derived from a dynamical forecast system or traditional model-analog methods), which may not always be the case (e.g., Kumar et al., 2014; Weisheimer et al., 2022). Indeed, actual skill can sometimes exceed potential skill, giving rise to a signal-to-noise paradox (Scaife and Smith 2018).

While our analysis takes an important first step towards understanding future climate predictability changes, there a number of important questions that remain. First, is there a strong seasonality to future global predictability changes? Our study focused primarily on potential skill computed across all months; however, there were some seasonal differences in ENSO predictability changes (Figure 7). For example, ENSO skill changes in GFDL-SPEAR were largest for forecasts initialized (or including) boreal spring to boreal fall (Figure 7; third row). Additionally, Maher et al. (2023) showed that ENSO amplitude changes in the LEs analyzed here are stronger in some seasons (typically boreal winter) than others (see their Figure 4). Therefore, it is possible that ENSO's impact on future predictability may be seasonally dependent. Next, what other ENSO-related factors impact future climate predictability? Many studies have shown that ENSO frequency (e.g., Berner et al., 2020), flavor (i.e., central vs eastern Pacific; Capotondi et al., 2015), and asymmetry (i.e., the duration of El Niño versus La Niña events; Maher et al., 2023) may change in the future. Changes to these characteristics may alter ENSO's influence on the rest of the climate system and thereby climate predictability. Additionally, there may be changes in the background mean state (e.g., the strength of the east-west temperature gradient in the equatorial Pacific) that impact the overall climate response to ENSO (Cai et al. 2021). While we did not find a significant relationship between predictability in our forecasts and each LE's time-varying ENSO frequency or flavor preference (not shown), we encourage future studies to investigate these mechanisms in more detail.

Finally, although ENSO is a dominant driver of seasonal forecast skill for much of the globe, there are likely other mechanisms that contribute to the predictability limits of different regions and variables. For example, Shi et al., (2022) showed that long-term shoaling of the mixed layer in the future may reduce the thermal inertia of the ocean, thereby decreasing ocean memory and year-to-year SST persistence, especially in the mid-latitudes. Similarly, Kumar et al., (2023) found that global warming decreases soil moisture memory over North America due to an increase in potential evapotranspiration. In both cases, the reduction in climate memory increases variability at less predictable high frequencies (e.g., weather timescales) while decreasing variability at lower frequencies (e.g., seasonal and longer), thus "whitening" the power spectrum and contributing to a decrease in persistence-related predictability. However, it is still unclear to what extent these changes may be offset by dynamical drivers of predictability change related to ENSO. More research is needed to unpack the dynamic versus thermodynamic contributions to future climate predictability change.

Acknowledgements

We thank Friedrich Burger and Thomas Frölicher for providing us with the GFDL-ESM2M data used in this study.

Data Availability Statement

Large ensemble datasets are available as follows:

CESM1-LE: https://www.cesm.ucar.edu/projects/community-projects/MMLEA/

CESM2-LE: https://www.cesm.ucar.edu/projects/community-projects/LENS2/

GFDL-SPEAR: https://www.gfdl.noaa.gov/spear_large_ensembles/

GFDL-ESM2M: Provided by Friedrich Burger and Thomas Frölicher at the University of Bern.

MPI-GE: https://esgf-data.dkrz.de/projects/mpi-ge/

References

Barnett, T. P., and R. Preisendorfer, 1987: Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis. *Monthly Weather Review*, **115**, 1825–1850, https://doi.org/10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2.

Barnston, A. G., and M. K. Tippett, 2017: Do Statistical Pattern Corrections Improve Seasonal Climate Predictions in the North American Multimodel Ensemble Models? *Journal of Climate*, **30**, 8335–8355, https://doi.org/10.1175/JCLI-D-17-0054.1.

——, ——, M. L. L’Heureux, S. Li, and D. G. DeWitt, 2012: Skill of Real-Time Seasonal ENSO Model Predictions during 2002–11: Is Our Capability Increasing? *Bulletin of the American Meteorological Society*, **93**, 631–651, https://doi.org/10.1175/BAMS-D-11-00111.1.

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and Forecast Skill in NMME. *Journal of Climate*, **27**, 5891–5906, https://doi.org/10.1175/JCLI-D-13-00597.1.

Berner, J., H. M. Christensen, and P. D. Sardeshmukh, 2020: Does ENSO Regularity Increase in a Warming Climate? *Journal of Climate*, **33**, 1247–1259, https://doi.org/10.1175/JCLI-D-19-0545.1.

Burger, F. A., J. Terhaar, and T. L. Frölicher, 2022: Compound marine heatwaves and ocean acidity extremes. *Nat Commun*, **13**, 4722, https://doi.org/10.1038/s41467-022-32120-7.

Cai, W., and Coauthors, 2021: Changing El Niño–Southern Oscillation in a warming climate. *Nat Rev Earth Environ*, **2**, 628–644, https://doi.org/10.1038/s43017-021-00199-z.

Capotondi, A., and Coauthors, 2015: Understanding ENSO diversity. *Bulletin of the American Meteorological Society*, **96**, 921–938, https://doi.org/10.1175/BAMS-D-13-00117.1.

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

Delworth, T. L., and Coauthors, 2020: SPEAR: The Next Generation GFDL Modeling System for Seasonal to Multidecadal Prediction and Projection. *Journal of Advances in Modeling Earth Systems*, **12**, e2019MS001895, https://doi.org/10.1029/2019MS001895.

Derome, J., H. Lin, and G. Brunet, 2005: Seasonal Forecasting with a Simple General Circulation Model: Predictive Skill in the AO and PNA. *Journal of Climate*, **18**, 597–609, https://doi.org/10.1175/JCLI-3289.1.

Ding, H., M. Newman, M. A. Alexander, and A. T. Wittenberg, 2018: Skillful Climate Forecasts of the Tropical Indo-Pacific Ocean Using Model-Analogs. *Journal of Climate*, **31**, 5437–5459, https://doi.org/10.1175/JCLI-D-17-0661.1.

——, ——, ——, and ——, 2019: Diagnosing Secular Variations in Retrospective ENSO Seasonal Forecast Skill Using CMIP5 Model-Analogs. *Geophysical Research Letters*, **46**, 1721–1730, https://doi.org/10.1029/2018GL080598.

Gan, B., L. Wu, F. Jia, S. Li, W. Cai, H. Nakamura, M. A. Alexander, and A. J. Miller, 2017: On the Response of the Aleutian Low to Greenhouse Warming. *Journal of Climate*, **30**, 3907–3925, https://doi.org/10.1175/JCLI-D-15-0789.1.

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

He, S., J.-Y. Yu, S. Yang, and S.-W. Fang, 2020: ENSO's impacts on the tropical Indian and Atlantic Oceans via tropical atmospheric processes: observations versus CMIP5 simulations. *Clim Dyn*, **54**, 4627–4640, https://doi.org/10.1007/s00382-020-05247-w.

Heede, U. K., and A. V. Fedorov, 2023: Towards understanding the robust strengthening of ENSO and more frequent extreme El Niño events in CMIP6 global warming simulations. *Clim Dyn*, **61**, 3047–3060, https://doi.org/10.1007/s00382-023-06856-x.

Horel, J. D., and J. M. Wallace, 1981: Planetary-scale phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813–829, https://doi.org/10.1175/1520-0493(1981)109<0813:PSAPAW>2.0.CO;2.

Huang, B., and Coauthors, 2017: Extended reconstructed Sea surface temperature, Version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Jacox, M. G., M. A. Alexander, C. A. Stock, and G. Hervieux, 2019: On the skill of seasonal sea surface temperature forecasts in the California Current System and its connection to ENSO variability. *Climate Dynamics*, **53**, 7519–7533, https://doi.org/10.1007/s00382-017-3608-y.

Kay, J. E., and Coauthors, 2015: The community earth system model (CESM) large ensemble project : A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kirtman, B. P., and P. S. Schopf, 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 Prediction. *Monthly Weather Review*, **137**, 2622–2631, https://doi.org/10.1175/2009MWR2814.1.

——, P. Peng, and M. Chen, 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.

Kumar, S., C. F. Dewes, M. Newman, and Y. Duan, 2023: Robust Changes in North America's Hydroclimate Variability and Predictability. *Earth's Future*, **11**, e2022EF003239, https://doi.org/10.1029/2022EF003239.

Kushnir, Y., W. A. Robinson, P. Chang, and A. W. Robertson, 2006: The Physical Basis for Predicting Atlantic Sector Seasonal-to-Interannual Climate Variability. *Journal of Climate*, **19**, 5949–5970, https://doi.org/10.1175/JCLI3943.1.

Lorenz, E. N., 1969: Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Lou, J., M. Newman, and A. Hoell, 2023: Multi-decadal variation of ENSO forecast skill since the late 1800s. *npj Clim Atmos Sci*, **6**, 1–14, https://doi.org/10.1038/s41612-023-00417-z.

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

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble - Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems*, 2019MS001639, https://doi.org/10.1029/2019MS001639.

——, S. Milinski, and R. Ludwig, 2021: Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble. *Earth System Dynamics*, **12**, 401–418, https://doi.org/10.5194/esd-12-401-2021.

——, and Coauthors, 2023: The future of the El Niño–Southern Oscillation: using large ensembles to illuminate time-varying responses and inter-model differences. *Earth System Dynamics*, **14**, 413–431, https://doi.org/10.5194/esd-14-413-2023.

McGregor, S., C. Cassou, Y. Kosaka, and A. S. Phillips, 2022: Projected ENSO Teleconnection Changes in CMIP6. *Geophysical Research Letters*, **49**, e2021GL097511, https://doi.org/10.1029/2021GL097511.

Mo, K. C., and M. Ghil, 1987: Statistics and Dynamics of Persistent Anomalies. *Journal of the Atmospheric Sciences*, **44**, 877–902, https://doi.org/10.1175/1520-0469(1987)044<0877:SADOPA>2.0.CO;2.

Newman, M., and P. D. Sardeshmukh, 2017: Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophysical Research Letters*, **44**, 8520–8529, https://doi.org/10.1002/2017GL074088.

O'Brien, J. P., and C. Deser, 2023: Quantifying and Understanding Forced Changes to Unforced Modes of Atmospheric Circulation Variability over the North Pacific in a Coupled Model Large Ensemble. *Journal of Climate*, **36**, 19–37, https://doi.org/10.1175/JCLI-D-22-0101.1.

O'Reilly, C. H., J. Heatley, D. MacLeod, A. Weisheimer, T. N. Palmer, N. Schaller, and T. Woollings, 2017: Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, **44**, 5729–5738, https://doi.org/10.1002/2017GL073736.

——, T. Woollings, L. Zanna, and A. Weisheimer, 2019: An Interdecadal Shift of the Extratropical Teleconnection From the Tropical Pacific During Boreal Summer. *Geophysical Research Letters*, **46**, 13379–13388, https://doi.org/10.1029/2019GL084079.

Power, S., and Coauthors, 2021: Decadal climate variability in the tropical Pacific: Characteristics, causes, predictability, and prospects. *Science*, **374**, eaay9165, https://doi.org/10.1126/science.aay9165.

Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing Sources of U.S. Seasonal Forecast Skill. *Journal of Climate*, **19**, 3279–3293, https://doi.org/10.1175/JCLI3789.1.

Rodgers, K. B., and Coauthors, 2021: Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, **12**, 1393–1411, https://doi.org/10.5194/esd-12-1393-2021.

Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of Probability Associated with El Niño. *Journal of Climate*, **13**, 4268–4286, https://doi.org/10.1175/1520-0442(2000)013<4268:COPAWE>2.0.CO;2.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Clim Atmos Sci*, **1**, 1–8, https://doi.org/10.1038/s41612-018-0038-4.

Shi, H., and Coauthors, 2022: Global decline in ocean memory over the 21st century. *Science Advances*, **8**, eabm3468, https://doi.org/10.1126/sciadv.abm3468.

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters*, **42**, 1554–1559, https://doi.org/10.1002/2014GL062829.

Storch, H. von, and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press,.

Suarez, M. J., and P. S. Schopf, 1988: A Delayed Action Oscillator for ENSO. *Journal of the Atmospheric Sciences*, **45**, 3283–3287, https://doi.org/10.1175/1520-0469(1988)045<3283:ADAOFE>2.0.CO;2.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J R Soc Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

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

——, D. Decremer, D. MacLeod, C. O'Reilly, T. N. Stockdale, S. Johnson, and T. N. Palmer, 2019: How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society*, **145**, 140–159, https://doi.org/10.1002/qj.3446.

——, D. J. Befort, D. MacLeod, T. Palmer, C. O'Reilly, and K. Strømmen, 2020: Seasonal Forecasts of the Twentieth Century. *Bulletin of the American Meteorological Society*, **101**, E1413–E1426, https://doi.org/10.1175/BAMS-D-19-0019.1.

——, M. A. Balmaseda, T. N. Stockdale, M. Mayer, S. Sharmila, H. Hendon, and O. Alves, 2022: Variability of ENSO Forecast Skill in 2-Year Global Reforecasts Over the 20th Century. *Geophysical Research Letters*, **49**, e2022GL097885, https://doi.org/10.1029/2022GL097885.

Williams, I. N., and C. M. Patricola, 2018: Diversity of ENSO Events Unified by Convective Threshold Sea Surface Temperature: A Nonlinear ENSO Index. *Geophysical Research Letters*, **45**, 9236–9244, https://doi.org/10.1029/2018GL079203.

Wittenberg, A. T., A. Rosati, T. L. Delworth, G. A. Vecchi, and F. Zeng, 2014: ENSO Modulation: Is It Decadally Predictable? *Journal of Climate*, **27**, 2667–2681, https://doi.org/10.1175/JCLI-D-13-00577.1.

Zhao, M., H. H. Hendon, O. Alves, G. Liu, and G. Wang, 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.

Zheng, Y., M. Rugenstein, P. Pieper, G. Beobide-Arsuaga, and J. Baehr, 2022: El Niño–Southern Oscillation (ENSO) predictability in equilibrated warmer climates. *Earth System Dynamics*, **13**, 1611–1623, https://doi.org/10.5194/esd-13-1611-2022.

Zhou, Z.-Q., S.-P. Xie, X.-T. Zheng, Q. Liu, and H. Wang, 2014: Global Warming–Induced Changes in El Niño Teleconnections over the North Pacific and North America. *Journal of Climate*, **27**, 9050–9064, https://doi.org/10.1175/JCLI-D-14-00254.1.