Check for updates

CLIMATOLOGY

Using initial states to identify successful El Niño–Southern Oscillation forecasts 1 year in advance

Nandini Ramesh[1,2]*, Mandy B. Freund[3,4], Damien B. Irving[5], Dougal T. Squire[6], Carly R. Tozer[5], Didier P. Monselesan[5]†, Tess Parker[5]†, Annette Stellema[5]†, James S. Risbey[5]

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

El Niño–Southern Oscillation (ENSO) is the primary source of climate predictability on seasonal and longer timescales over much of the globe, but predicting its phase beyond 9 months has remained challenging. Here, we leverage an underappreciated factor in ENSO forecasting: the substantial variation in forecast skill based on the ENSO phase at the time a forecast is issued. Our analysis of transitions between ENSO phases in a dynamical hindcast dataset demonstrates that this variation markedly exceeds intermodel differences in predictive capability. Forecasts initialized during La Niña conditions perform best, with three models correctly predicting transitions to El Niño 1 year in advance over 70% of the time. We present our findings in a simple framework that allows decision-makers to assess the likelihood of a forecast being correct on the basis of the current ENSO state. Applied effectively, these findings could improve long-lead ENSO forecasting strategies, enhancing global risk management.

INTRODUCTION

El Niño–Southern Oscillation (ENSO) stands out as the most prominent modes of interannual climate variability, arising from air-sea interactions in the tropical Pacific. Through its far-reaching impacts on global weather patterns, it affects ecosystems, agriculture, water resources, public health, and disaster management worldwide.

ENSO is unique among climate phenomena in being predictable months in advance (*1*), offering a rare opportunity to prepare well ahead for a range of potential consequences (*2*). This predictability stems predominantly from the slow timescale of the oceanic processes involved (*3*–*6*). Consequently, early ENSO forecasts play a crucial role in decision-making across a wide range of industries and sectors. The economic significance of these predictions is substantial, with seasonal forecasts influencing an estimated market capitalization of 13 trillion USD (*7*). Forecasting agencies typically issue declarations about the presence or absence of El Niño (warm phase) or La Niña (cool phase) events, which can trigger a cascade of responses globally. These declarations may prompt drought warnings for farmers and water management authorities, lead to fluctuations in agricultural commodity prices, initiate or intensify disaster mitigation efforts, and inform the planning and implementation of public health interventions (*8*, *9*).

Skillful forecasting of ENSO is understood to be limited by conditions during March to May [known as the "spring predictability barrier" (SPB) (*10*–*12*) for its occurrence during boreal spring months] that are highly sensitive to small perturbations, hampering models' ability to forecast the evolution of ENSO. However, several studies have suggested that ENSO may be predictable at longer lead times under specific circumstances (*6*, *13*–*18*). Many have highlighted particular types of events, such as multiyear La Niñas (*19*–*22*) or large El Niños (*13*, *23*–*25*) as being predictable well in advance. Others have pointed out that, as in a chaotic system, conditions when the forecast is issued are key determinants of forecast skill, whether on seasonal (*10*–*12*), interannual (*17*, *18*, *26*, *27*), or decadal (*12*, *28*–*31*) timescales.

Many of these studies of long-lead ENSO prediction have been conducted on individual models or datasets, using different techniques and benchmarks, making comparison challenging and rendering it virtually impossible to integrate their many promising findings into operational settings. In this work, we apply a simple interpretable framework that we generalize across multiple dynamical (physics-based) models to facilitate a fair comparison of model skill and bias when predicting and making predictions from the different phases of ENSO.

Here, we reduce the target of prediction to solely a single quantity per year: the phase of ENSO (i.e., either La Niña, neutral, or El Niño) during the expected event peak [October to February (ONDJF)] for each year (Fig. 1, A to C). Some prior work has examined dynamical model forecast skill in terms of three categories at the seasonal timescale, using metrics such as the ranked probability skill score (RPSS) (*32*, *33*). Rather than assessing model skill in predicting these categories from 1 month or season to the next, we examine skill in predicting the transition from one ENSO event peak to the next on the timescale of 1 year. Since there is generally one event per year (when years are defined from July to June, centered on December to January), this matches the timescale of ENSO events and is consistent with the objective of declaring the presence or absence of an event as early as possible each year. This information can be of use to stakeholders whose decisions are triggered by the absence or presence of ENSO events, for whom details of magnitude or timing of the sea surface temperature (SST) anomaly peak may be a secondary concern.

We begin by examining the sequence of ENSO events in the historical record by categorizing each year as one of the three ENSO phases and assessing the likelihood of transition between each pair of phases ("Constructing a baseline for categorical forecasts: The historical pattern of transition probabilities" section). While the historical record is too short for high confidence, we find that all transitions are not equally probable, indicating some inherent year-to-year

[1]Natural Systems Modelling Group, Technology, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney 2145, Australia. [2]School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia. [3]School of Geography, Earth and Atmospheric Sciences, University of Melbourne, Melbourne 3053, Australia. [4]ARC Centre of Excellence for the Weather of the 21st Century, University of Melbourne, Melbourne 3053, Australia. [5]Climate Variability and Hazards, Environment, CSIRO, Hobart 7004, Australia. [6]ACCESS-NRI, Australian National University, Canberra 2601, Australia.
*Corresponding author. Email: nandini.ramesh@csiro.au
†These authors contributed equally to this work.

Downloaded from https://www.science.org on October 24, 2026

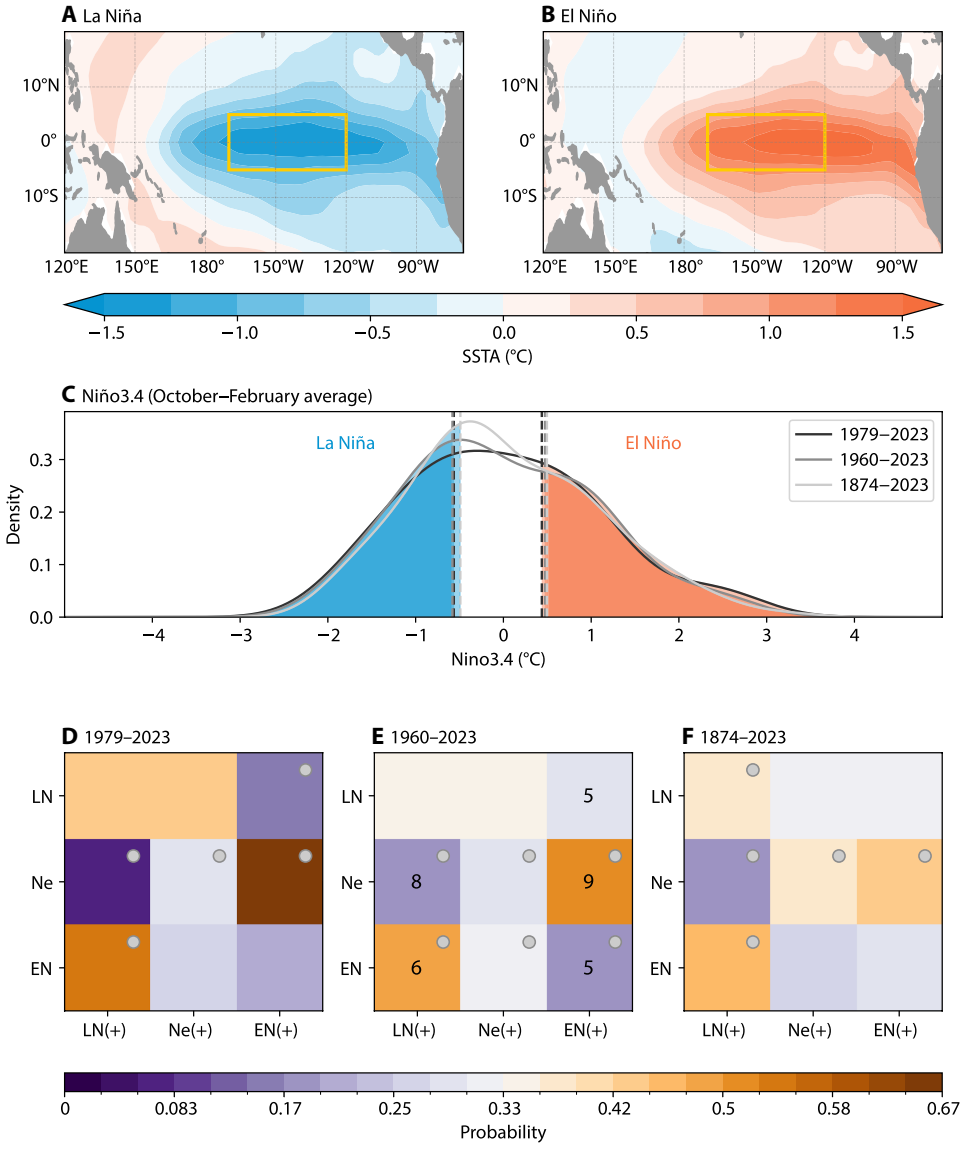


Fig. 1. La Niña and El Niño events and the transition probability matrix of year-to-year transitions between La Niña, Neutral, and El Niño conditions. (**A** and **B**) SST anomalies (°C) averaged over La Niña and El Niño events respectively during the event peak (October to February) for the 1960 to 2023 period, with the Niño3.4 region indicated by the yellow box. (**C**) Distribution of Niño3.4 values, averaged over October to February for three different periods, with the terciles used to determine the ENSO phase for each period indicated by dashed vertical lines in the corresponding shades. (**D** to **F**) TPMs for the satellite period, the overlapping period with the models used, and a 150-year period, respectively. Each row represents an initial state [La Niña (LN), Neutral (Ne), or El Niño (EN)], and columns represent the state in the subsequent year [indicated by (+)]. Colors indicate the historical empirical probabilities of these occurring. The color scale is centered at one-third such that lighter colors correspond to probabilities close to random chance. Gray circles indicate the transitions that are distinct from others in their row on the basis of their 95% confidence intervals. Numbers in (E) indicate the number of models out of nine that agree on the ranking of the most or least likely transition, where a majority of models agree.

predictability in the natural system. To make forecasts with meaningful skill on this timescale, a model must outperform this inherent predictability. For a larger sample size, we turn to a dataset of decadal hindcasts from nine dynamical coupled models, where we find some agreement with the patterns seen in the observations. However, initialized forecasts from these models display a substantially altered pattern of probabilities, indicative of model biases arising from initialization that can be large enough to erase this inherent predictability.

We then examine forecast skill at the 1-year lead time in this multimodel dynamical hindcast dataset as a function of the phase from which a forecast is issued and the phase being predicted ("Assessing categorical forecast skill: Dynamical model forecasts of ENSO categories at the 1-year lead time" section). We find that forecast skill is sensitive to these factors, with models generally performing best when predicting from La Niña conditions.

To systematically assess these variations in model skill in a manner applicable to the interpretation of real-world forecasts, we calculate model *precision*, a quantity constructed using the information available at the time of a forecast, for each possible transition type ("Assessing categorical forecasts in real time: Precision based on

Downloaded from https://www.science.org on October 24, 2026

ENSO transition type" section). We propose that this metric could be used to communicate forecast confidence to stakeholders. We find that confidence in model forecasts is highest when predicting the transitions from La Niña to El Niño and vice versa and that the transition type being predicted has a stronger influence on precision than intermodel differences.

RESULTS

Constructing a baseline for categorical forecasts: The historical pattern of transition probabilities

We quantify the historical pattern of probabilities using a transition probability matrix (TPM), a commonly used structure in the analysis of Markov chains or networks (Fig. 1, D to F). Each row and each column of the matrix corresponds to one of the three categories. Rows (*i*) represent the category from which the system begins (the "source" or initial state), and columns (*j*) represent the category to which the system transitions in the subsequent year (the "sink" or final state in an interyear transition). Each element of the matrix represents the probability $P(A_{i,j})$ that the transition from the corresponding row *i*'s category to column *j*'s category will occur (e.g., the bottom-left cell represents the empirical probability of a La Niña year following an El Niño year). The rows each sum to one and can each individually be interpreted as a histogram.

Constructing this matrix over the observed record (Fig. 1, D to F) shows that the different transitions are not all equally probable. The color scale is centered at the probability of a transition based on random chance (i.e., one-third, since the categories are constructed as terciles); higher values thus indicate transition probabilities more likely than chance. For example, an El Niño year is most likely to be followed by a La Niña (orange shading in Fig. 1D) and least likely to be followed by another El Niño (violet shading). Following a neutral year, La Niña is unlikely but possible. The transition matrices are shown for the satellite era (since 1979; Fig. 1D), the period covered by the hindcast dataset used in the next section (Fig. 1E), and a 150-year period from 1874 (Fig. 1F). The transition matrices are not identical for different periods of time as a result of higher uncertainty in earlier parts of the dataset and the probabilities of some of these transitions changing over time because of both low-frequency variability and long-term trends (*34*, *35*) (see fig. S1). The thresholds used to categorize years are based on the terciles of Niño3.4 over the time period used, meaning that they also vary slightly between the three periods as shown in Fig. 1C.

These TPMs characterize variations in the probabilities of transitions between states that have been highlighted by prior work in other ways, showing, for example, that the transition from El Niño to La Niña is more likely than the reverse (*36*), that the transition from neutral to La Niña is rare (*35*), and that there is a tendency (albeit at the seasonal timescale) for El Niño to emerge from neutral states (*37*). These also show qualitative agreement with another study using a different SST dataset and event thresholds (*27*).

While the observations are more likely to be accurate in recent years, especially during the satellite era, the inclusion of earlier years allows for a longer record and therefore higher confidence from a larger sample size, albeit with some uncertainty due to inaccuracies or long-term trends and variability. We found that the histogram represented by the probabilities following neutral years (middle row) was significantly different from a uniform distribution (i.e., all three categories occurring with equal probability in the year following a given category) at the 95% confidence level for Fig. 1 (D and F); however, this is highly dependent on sample length. The shortness of the record also means that the TPMs may be sensitive to single events; we test this by resampling the time series (see the "Quantifying uncertainty in observed TPMs" section of Materials and Methods and table S1) to estimate 95% confidence intervals of the TPM cells. Gray circles indicate the transitions whose 95% confidence intervals did not overlap with others in their row and may therefore be considered distinct. Following neutral years, all three possible transitions were found to have distinctly different probabilities, irrespective of the period over which they were calculated.

Following La Niña years, the distribution of transition probabilities remains close to a uniform distribution, meaning that all three categories are equally likely. While the transition probabilities (and how distinct they are from each other) vary on the basis of the period used, the relative ranking of probabilities in each row is fairly stable (see also fig. S1).

The stability of the ranking of probabilities in each row over time suggests that the natural system contains some, if small, inherent predictability of what should follow from 1 year to the next. In other words, this implies that knowledge of the current state can be used to issue a "naïve" forecast of the next year's state with some skill greater than random chance. This predictability is highest following neutral years and lowest following La Niña years.

The Decadal Climate Prediction Project (DCPP; see Materials and Methods) hindcast archive provides a much larger sample size with which to construct these matrices. We use the last 6 years (i.e., five transitions) of each ensemble forecast to construct these (a sample size of 37,500 interyear transitions with a minimum of 1500 transitions per model) to eliminate the effects of initialization shock at the start of the forecast and to therefore capture the models' "free-running" state, subject to the same radiative forcings as the observed system from 1960 to 2023.

We find that while the TPMs differ between models (fig. S2), some features of the rankings of probabilities in each row are consistent with the observations (numbers in Fig. 1E) in most models:

1) Following a neutral year, all models agree that the most likely category to occur next is an El Niño event.

2) Following a neutral year, eight of nine models agree that the least likely category to occur next is a La Niña event.

3) Following an El Niño event, six of nine models agree that the most likely category is a La Niña event.

4) Following an El Niño event, five of nine models agree that another El Niño event is the least likely outcome.

5) Following a La Niña event, five of nine models agree that an El Niño event is the least likely outcome (noting that while this was not statistically significantly different from uniform in the observations, it is consistent with the ranking over different historical periods).

Together, these probabilities are consistent with known temporal features of ENSO including El Niño events ending in boreal spring rather than persisting for multiple years (*38*, *39*) and generally occurring at irregular intervals multiple years apart (*40*, *41*, *42*) along with longer, multiyear periods made up of La Niña and neutral years (*20*, *43*). This also aligns with the expected pattern from the recharge-discharge oscillator paradigm (*4*), in which the recharge phase (occurring during La Niña and some neutral years) persists for longer than the short-lived discharge phase (during El Niño years) (*44*, *45*).

Models show the highest agreement with each other and with the observations (fig. S2) regarding the probabilities following a neutral

Downloaded from https://www.science.org on October 24, 2026

state and low agreement on the ranking of probabilities after a La Niña event, which is also close to uniform in the observations. Models are therefore able, at least qualitatively, to reproduce key features of the observed pattern of interyear transitions, implying that their representations of the ENSO system similarly have some inherent interyear predictability, particularly following neutral years.

Assessing categorical forecast skill: Dynamical model forecasts of ENSO categories at the 1-year lead time

Constructing TPMs from the DCPP hindcast using only the first predicted year (i.e., the transition from a known initialized state to the following year) instead shows a substantially altered pattern of transition probabilities (fig. S3), with the only feature in common with those of the free-running models and observations being that La Niñas are least likely after neutral years. This is an indication that the *reliability*, or ability of models to reproduce each possible outcome with a similar frequency to that observed (*46*), is diminished by initialization shock when forecasting from neutral and El Niño states.

For model forecasts of ENSO category to have meaningful skill 1 year in advance, they must at minimum outperform forecasts based on the TPM (Fig. 1E), which represents the baseline or climatological skill inherent in the system. The RPSS (Table 1) compares model performance to climatological reference skill from the observed TPM (see the "Forecast verification" section of Materials and Methods for definition). IPSL-CM6A-LR, CESM1-1-CAM5-CMIP5, and EC-Earth3 stand out as the best performers overall (last column). Two models (CanESM5 and IPSL-CM6A-LR) were initialized in January, while all others were initialized in the preceding November or October (Table 2); we find that models initialized in later months did not systematically perform better than others, suggesting that initialization date within the boreal winter months is a less important influence on skill than intermodel differences. Similarly, the ensemble size does not have a clear relationship with model performance: While the model with the smallest ensemble (MPI-ESM1-2-HR, 5 members) shows the poorest performance, the model with the largest ensemble (CESM1-1-CAM5-CMIP5, 40 members) is not the top-performing model; instead, it is tied for second place after IPSL-CM6A-LR (10 members) with EC-Earth3 (10 members).

When the RPSS is calculated for forecasts initialized from each ENSO phase, we find that forecasts from La Niña outperform forecasts from other states in every model. Meanwhile, forecasts from neutral states show negative skill, indicating that models consistently perform worse than the observed TPM when initialized from these conditions.

The differences within each model when initialization states are separated far exceed the differences between models. This indicates that the state at initialization is a powerful influence on the skill of forecasts at the annual timescale. The high RPSS (multimodel mean = 0.43, SD = 0.12) when predicting from La Niña states at the 1-year lead time is comparable to the skill at the 3- to 4-month lead time found in the North American Multi-Model Ensemble (NMME) (*47*), a hindcast using operational seasonal forecasting models, when predicting three categories (*33*). Three to 4 months is a timescale at which ENSO forecasts are generally considered trustworthy and widely used, suggesting that these forecasts from La Niña are potentially usable. When predicting from El Niño events, the RPSS (mean = 0.22, SD = 0.11) is comparable to that at the 5- to 6-month timescale in NMME. The skill when predicting from neutral states, which is consistently negative, brings down the RPSS when considered across all states (mean = 0.12, SD = 0.1), masking the potentially useful early forecasts from the other states.

The low skill from neutral years and high skill from La Niña are, in part, attributable to the properties of the observed TPM discussed earlier: The inherent predictability after neutral states means that models must perform better than an already-high baseline to add value to the forecast; meanwhile, the virtual lack of information following La Niña states in the TPM means that models are held to a lower standard when assessing skill from La Niña events.

However, the reduced reliability of the models upon initialization (inferred from the fact that the free-running TPMs more closely resemble the observed TPM than the initialized TPMs do) suggests that the baseline observed TPM is not the only factor in these variations in skill: Initialization appears to alter model performance differently

Downloaded from https://www.science.org on October 24, 2026

Table 1. The RPSS for each model's probabilistic ensemble predictions of ENSO phase with 1-year lead time. The first three columns show the RPSS calculated for forecasts initialized during different ENSO phases and the overall RPSS is displayed in the last column. The 95% confidence intervals for each RPSS are indicated in parentheses. The last two rows show the multi-model mean and standard deviation across models. Higher values of RPSS indicate greater skill. The RPSSs of the three best-performing models for each column are highlighted in bold font.

Model	La Niña	Neutral	El Niño	All
CanESM5	0.41 (0.39, 0.42)	−1.30 (−1.38, −1.20)	0.13 (0.10, 0.17)	0.08 (0.07, 0.10)
CESM1-1-CAM5-CMIP5	**0.59** (0.57, 0.59)	−1.27 (−1.35, −1.18)	0.25 (0.23, 0.29)	**0.19** (0.19, 0.22)
CMCC-CM2-SR5	**0.54** (0.52, 0.54)	**−1.12** (−1.19, −1.03)	0.13 (0.10, 0.17)	0.15 (0.14, 0.18)
EC-Earth3	0.49 (0.47, 0.49)	−1.14 (−1.21, −1.05)	**0.29** (0.27, 0.32)	**0.19** (0.18, 0.22)
HadGEM3-GC31-MM	0.42 (0.39, 0.42)	−1.14 (−1.21, −1.05)	**0.31** (0.29, 0.34)	0.17 (0.16, 0.20)
IPSL-CM6A-LR	**0.56** (0.54, 0.56)	−1.35 (−1.43, −1.25)	**0.44** (0.42, 0.47)	**0.23** (0.23, 0.26)
MIROC6	0.19 (0.16, 0.20)	**−1.09** (−1.16, −1.00)	0.08 (0.04, 0.12)	0.01 (−0.00, 0.04)
MPI-ESM1-2-HR	0.31 (0.28, 0.32)	−2.31 (−2.41, −2.16)	0.18 (0.16, 0.22)	−0.10 (−0.11, −0.07)
NorCPM1	0.33 (0.30, 0.33)	**−0.88** (−0.94, −0.80)	0.19 (0.16, 0.23)	0.13 (0.13, 0.16)
Mean	0.43	−1.29	0.22	0.12
SD	0.12	0.38	0.11	0.10

Table 2. The set of models from the DCPP used in this study, their ensemble sizes, and initialization months.

Model abbreviation	Institutions	Ensemble size	Initialization month
CanESM5	CCCma (Canadian Centre for Climate Modelling and Analysis)	20	January
CESM1-1-CAM5-CMIP5	NCAR (National Centre for Atmospheric Research), USA	40	November
CMCC-CM2-SR5	Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy	10	November
EC-Earth3	EC-Earth European Consortium, Rossby Center, Swedish Meteorological and Hydrological Institute/SMHI	10	November
HadGEM3-GC31-MM	MOHC/ UK Met Office (The UK Meteorological Office)	10	November
IPSL-CM6A-LR	IPSL (Institut Pierre-Simon Laplace), France	10	January
MIROC6	JAMSTEC (Japan Agency for Marine-Earth Science and Technology), AORI (Atmosphere and Ocean Research Institute), University of Tokyo, NIES (National Institute for Environmental Studies), R-CCS (RIKEN Center for Computational Science)	10	November
MPI-ESM1-2-HR	AWI (Alfred Wegener Institute), Helmholtz Centre for Polar and Marine Research, Germany	5	November
NorCPM1	Norwegian Meteorological Institute (MET Norway), NORCE (Norwegian Research Centre), UiB (University of Bergen), UiO (University of Oslo), CICERO (Center for International Climate Research), NILU (Norwegian Institute for Air Research)	10	October

according to initial states. We next explore, using a range of metrics, the contributions of initialization and the observed TPM to this variation in skill and how models perform when ensemble members are grouped in different ways.

Measuring model performance at the 1-year lead time

We assess the performance of the models at the 1-year lead time by using the first year of each decade-long forecast. The baseline against which these should be compared (also used in the RPSS calculation above) is made up of naïve forecasts constructed from the observed TPM. We also include a comparison to forecasts similarly constructed from the models' own TPM using only the last five transitions of the decade-long forecasts: This period is beyond the "memory" of the initialization shock and captures how models would behave in the absence of initialization.

We perform this assessment in three different ways, in order of increasing complexity:

1) Deterministic individual forecasts: Here, we treat each ensemble member of each model as a prediction. This allows us to leverage the large number of samples in the dataset and to use an easily interpretable metric, the Matthews correlation coefficient (MCC or phi coefficient; see Materials and Methods, Eq. 1). Analogous to the Pearson's correlation coefficient but applicable to time series made up of discrete categories rather than continuous data, the MCC ranges from −1 (perfect negative correlation) to 1 (perfect correlation). We compare this with naïve deterministic forecasts constructed from the observed TPM as the most likely category to follow each state. To make a fair and consistent comparison that excludes future information available only after the forecast (*48*), we construct the observed TPM using the 40 years before the initialization date for each initialization year.

2) Deterministic ensemble forecasts: In this case, we consider each ensemble of each model initialized at a single time as a single prediction. The deterministic forecast is the category predicted by the largest number of ensemble members, i.e., the winner of the plurality of votes. This reduces the sample size to the number of initialization years (approximately 60; therefore approximately 20 per category) for each model but more closely corresponds to how model forecasts are made and how stakeholders may interpret them in practice (*49*). To assess performance, we use the ranked probability score (RPS), which takes low values when model performance is good (see Materials and Methods, Eq. 2). We compare this with the naïve deterministic forecasts constructed from the observed TPM as described above.

3) Probabilistic ensemble forecasts: This captures the nuances of the full three-category probability distribution predicted by the models; each ensemble forecast is interpreted as a set of three probabilities representing the predicted likelihood of each category. This corresponds to how ensemble forecasts are generally treated in practice. As above, we use RPS to assess performance. This probabilistic measure captures reliability and resolution of the models' initialized forecasts. We compare this with naïve probabilistic forecasts that use the rows of the observed TPM corresponding to each initial state. This metric was also used in calculating the RPSS in Table 1.

These metrics (Fig. 2) show that initial state has a marked impact on forecast performance at the 1-year lead time, indicative of the varying sensitivity to initial conditions expected in a chaotic system. When all initialization years are considered together, the models generally perform better than the observed TPM when evaluated in deterministic terms (gray bars in Fig. 2, A and B) and similarly to the observed TPM in probabilistic terms (gray bars in Fig. 2C). In every model in all three panels, predictions issued from La Niña perform better than those from neutral states and, with exceptions in the deterministic cases (MIROC6 in Fig. 2A; MPI-ESM1-2-HR and NorCPM1 in Fig. 2B), better than those from El Niño. Across

Downloaded from https://www.science.org on October 24, 2026

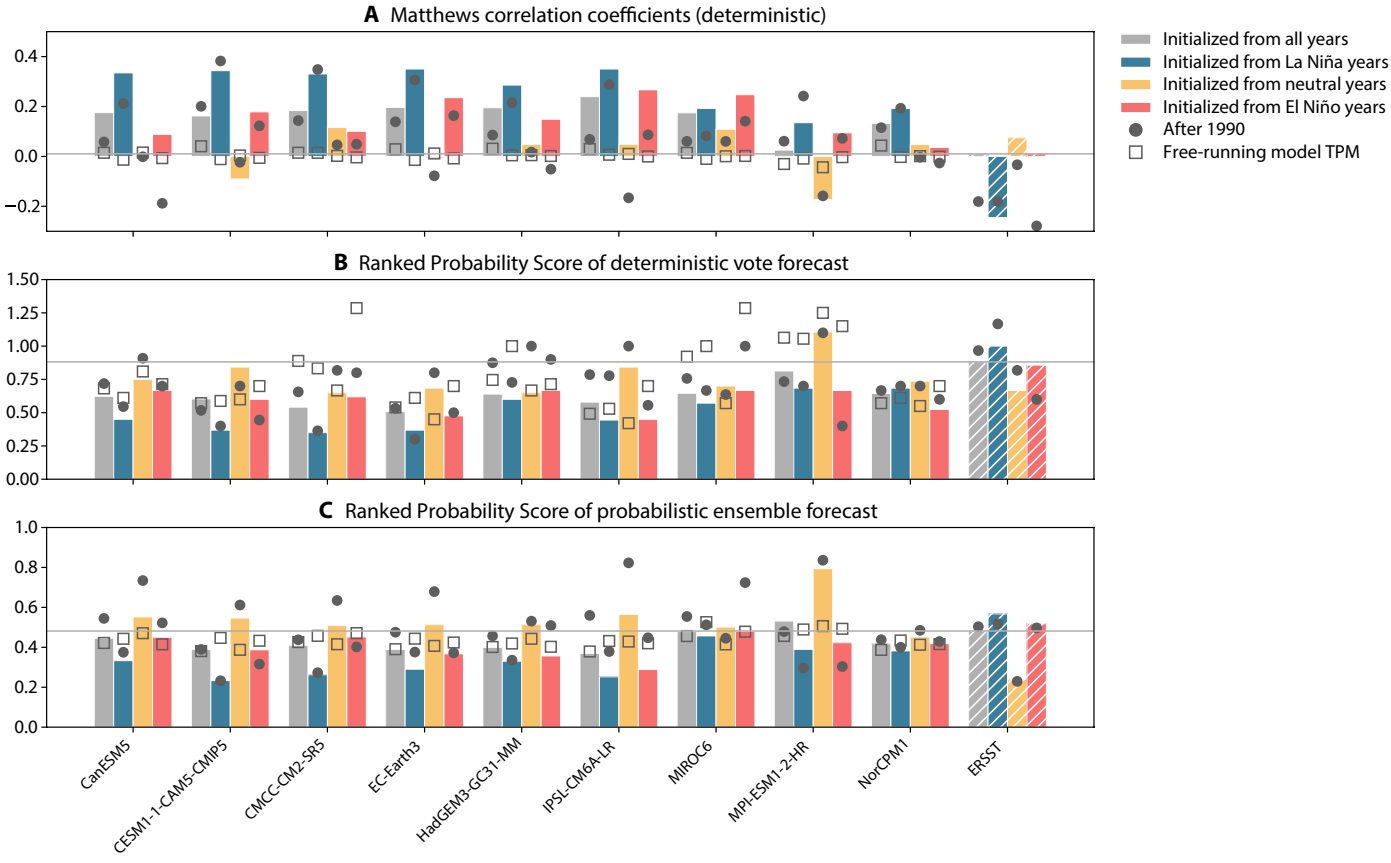


Fig. 2. Model performance when predicting from all states (gray), La Niña (blue), neutral (yellow), and El Niño (red) conditions, measured in three different ways. (**A**) MCC when each ensemble member is assessed individually, (**B**) RPS when the category with the highest number of ensemble members/probability is issued as a deterministic forecast, and (**C**) RPS when ensembles are treated as probabilistic forecasts. Bars with hatching represent the same constructed from the observations, and the horizontal gray line indicates performance of the observed TPM when all initial states are used. Dots indicate the corresponding scores when assessed for the second half of the hindcast dataset (1990 onward) with the climatology constructed from earlier years. Squares indicate the corresponding scores evaluated from the models' free-running TPMs. Larger positive values in (A) and smaller values in (B) and (C) are indicative of better performance.

metrics, the variation in performance based on initial state within each model is larger than that between models when no separation based on initial states is made.

As future information included in the climatology can artificially enhance skill (*48*), we also perform these assessments on the second half of the hindcast dataset (after 1990; gray circles), using only past information to construct the climatology. While this results in an expected reduction in performance in many models (further information in section S4 and table S3), it is still the case that forecasts from La Niña perform best and forecasts for neutral states show the lowest predictive performance.

Predictions from La Niña conditions consistently add skill compared to the observed TPM, which is negatively correlated with the observations. Meanwhile, predictions issued from neutral conditions appear to almost universally perform poorly, with some models even showing a negative correlation with real-world outcomes.

This runs counter to the pattern of inherent predictability implied in the observed and the models' own free-running TPMs, which suggests that there is higher predictability after neutral states than after La Niña. Along with the models' initialized TPMs, these results indicate that the process of initializing these forecasts substantially alters model behavior.

We also examine the impact of initialization by including the scores from the free-running models' TPMs (gray squares) in each case. These forecasts are constructed from TPMs (fig. S2) that represent model behavior when the model is in the same phase as the real world. This has the advantage that the free-running model starts in the correct ENSO phase without needing to be initialized (pulled away) to the observed state, meaning that this version of the model is not subject to initialization shock. When each ensemble member is considered separately (Fig. 2A), the predictive ability of these TPMs is comparable to random chance (zero correlation). In general, the initialized models perform better than their uninitialized TPMs. However, the uninitialized TPMs generally perform better when predicting from neutral states (this is true of every model in Fig. 2C and most models in Fig. 2B), implying that initialization degrades forecasts from neutral states but improves them from others.

Forecast initialization involves two factors: first, adding information about the current state of the real world through data assimilation, which can itself induce shocks that affect skill (*50*); and second, perturbing the system (both deliberately to ensure ensemble spread and unintentionally because of model drift and errors). Figure 2, along with the models' initialized TPMs, shows that the positive effect of the added information is most powerful during La Niña and

Downloaded from https://www.science.org on October 24, 2026

the negative effect of the added perturbation most pronounced during neutral conditions.

It has also been suggested that the type of event being predicted (*14*) and its magnitude [such as a large El Niño (*13*, *23*–*25*)] may affect forecast skill. We found that grouping forecasts on the basis of the phase being predicted (see table S4) did not result in as clear a separation in skill as when they were grouped on the basis of initial state. We examine the influence of the upcoming event's strength in Fig. 3, which shows the relationship between the real-world magnitude of the predicted event (the predictand; Niño3.4 in October to February) and the percentage of ensemble members in each forecast that predicted the correct category. The percentage of correct forecasts shows a statistically significant linear relationship with the real-world magnitude of the predictand (Fig. 3) when the predictand is a La Niña. For El Niño predictands, this relationship is not statistically significant, but poor predictions for events with magnitude greater than 1.5°C are relatively rare. The worst predictions are generally of events with absolute values of Niño3.4 between 0.75° and 1.5°C.

The contrast between forecasts issued from La Niña and neutral states is also evident from Fig. 3 (colors of dots indicating initial state). More than 50% of ensemble members predicted the correct category from most of the La Niña years (12 blue dots higher than the 50% line) with 1-year lead time. This threshold was passed only by six forecasts from El Niño years (red) and five from neutral (yellow). Meanwhile, it is rare for forecasts from La Niña years to perform worse than chance (only three blue dots below 33.33%) while forecasts from nine neutral years performed worse than chance, as did forecasts from six El Niño years. We did not find statistically significant linear relationships between Niño3.4 magnitude during initialization and the percentage of ensemble members that predicted the correct category.

Assessing categorical forecasts in real time: Precision based on ENSO transition type

Our results indicate that forecast performance is influenced by both the initial state from which a forecast is issued and the predicted state, with the former being more influential. To interpret how models perform in the case of each possible combination of initial and final states, we assess performance for each of the nine transitions using a discrimination metric known as precision (see the "Discrimination metrics" section in the Materials and Methods). This allows us to answer the question: What is the probability, based on historical performance, that a current forecast will come true?

The precision is the ratio between the number of times the model correctly predicted a particular transition and the number of times that the model predicted that transition at all, i.e., the fraction of model forecasts of a particular transition that came true. This can be calculated using information available at the time a forecast is issued and aligns with the situation of a stakeholder interpreting a model forecast.

In other words, this is the fraction of forecasts of a given transition that were not false alarms. Note that this metric does not penalize instances where a future category was missed (false negative), only penalizing instances where a forecast of the transition did not come true (false alarm or false positive). Therefore, while the precision

Downloaded from https://www.science.org on October 24, 2026

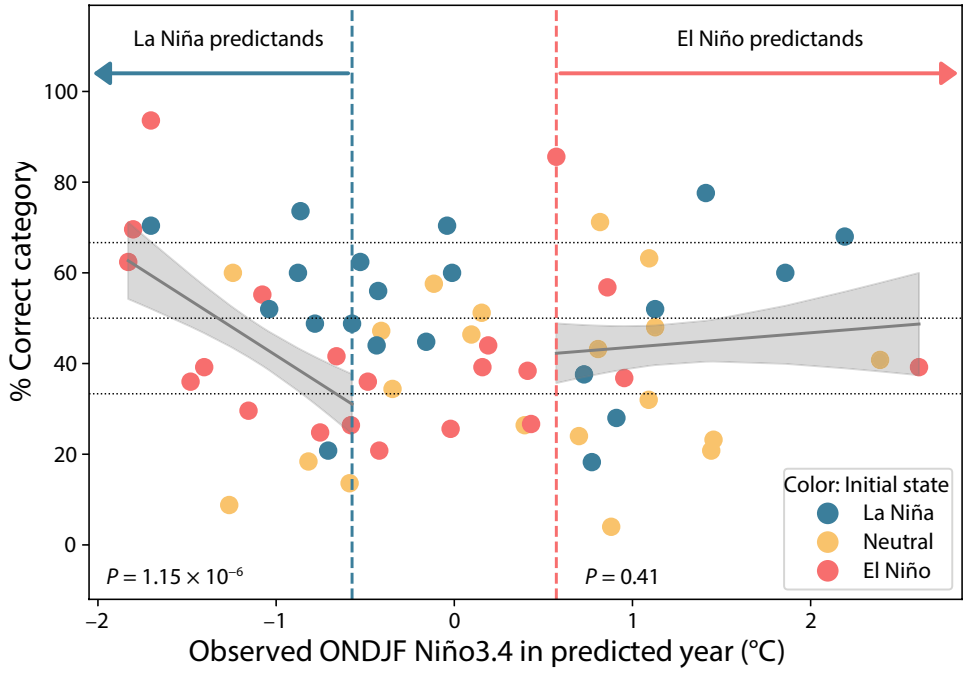


Fig. 3. The percentage of correct forecasts versus the observed magnitude of the predicted event. Each dot represents the aggregated forecast for a single initialization year across all models' ensemble members. Colors indicate the state from which the forecast was issued. The real-world Nino3.4 index of the predicted event (during October to February) is shown on the *x* axis. Vertical dashed lines indicate the thresholds separating La Niña (blue) and El Niño (red) from neutral years for the predictands. The lines of best fit from a linear regression between the forecast performance shown on the *y* axis and the magnitude of the predictand event are shown in gray with shading indicating 95% confidence intervals. Linear regressions were performed for predicted La Niñas (left) and El Niños (right) separately, with the associated *P* values shown in the corresponding bottom corners. Horizontal lines indicate 33.33 (corresponding to random chance), 50, and 66.67%.

presents an incomplete picture of model skill, it captures the information relevant to assessing how much confidence to assign to a current forecast and depends on both the predictive abilities of a model and the historical probability of a particular transition occurring in the real world.

The precision of dynamical forecasts varies markedly within models depending on the type of transition predicted. Figure 4 shows the precision for each transition type in the nine models assessed. The variation in precision within each model based on transition type far exceeds the differences between models overall. For example, in the CESM1-1-CAM5-CMIP5 model (Fig. 4B), forecasts from La Niña of an upcoming El Niño come true 82.21% of the time and are therefore quite trustworthy even despite the 1-year lead time; on the other hand, when the same model predicts La Niña from neutral, it should not be believed, as this is very unlikely to come true (correct only 8.94% of the time). These differences, internal to each model, are much larger than the differences in overall precision (also known as accuracy; indicated in parentheses) between the various models. This means that the transition type predicted by a model provides far more information regarding whether a real-world forecast should be believed than the choice of model does.

Four models show the highest precision when predicting the transition from La Niña to El Niño, with three attaining a precision exceeding 70% a year in advance of the event peak. All models show a higher precision for this transition than could be attained by climatological predictions, i.e., predictions obtained by randomly sampling the TPM from the observations (indicated by stars in Fig. 4). Three models show the highest precision in the El Niño to La Niña transition, with 74% precision seen in the MIROC6 model and eight of nine models attaining a precision greater than that from the observed TPM. The lowest precision is almost universally (eight models) in predicting the transition from neutral years to La Niña; these precision matrices therefore reveal a model bias toward predicting La Niñas arising from neutral conditions when these transitions are, in fact, rare in nature.

The precision attained by model forecasts outperforms the baseline TPM-based forecast in most transitions from La Niña and many

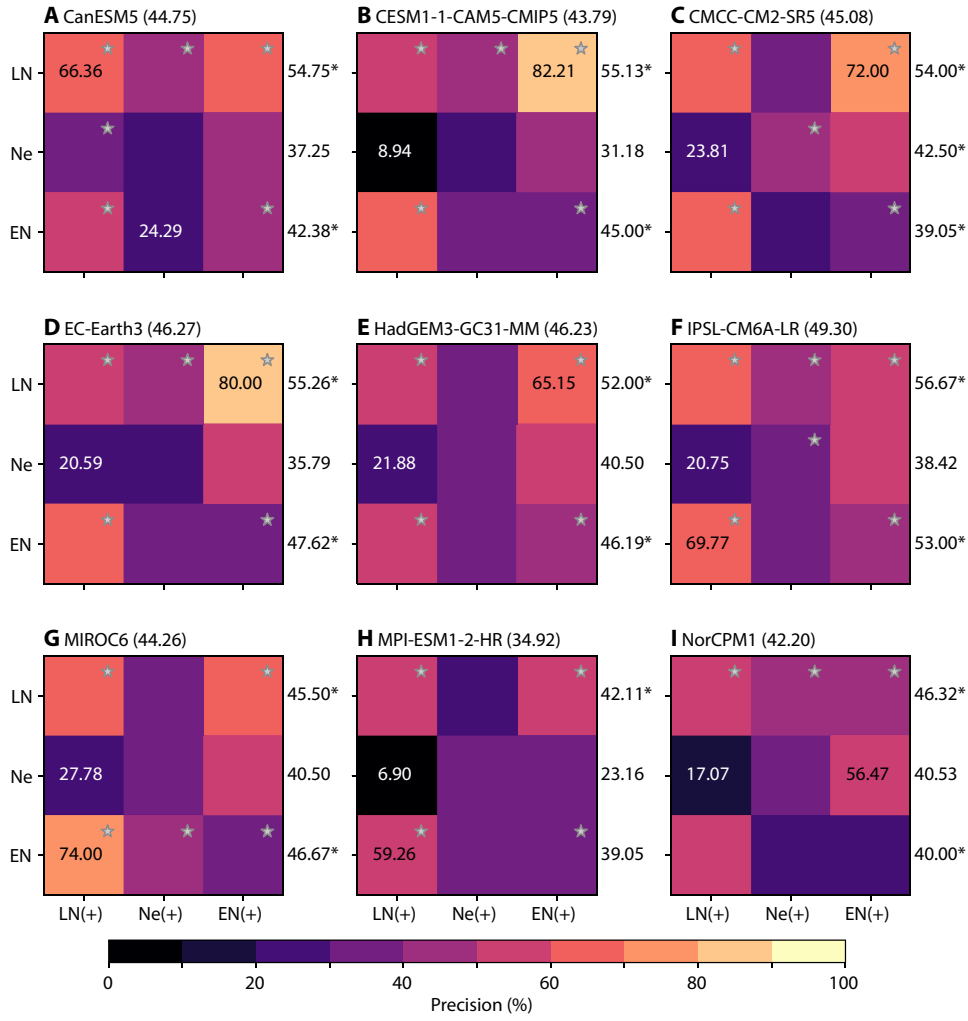


Fig. 4. The precision of predictions of ENSO phase 1 year in advance for each possible transition, for all models in the study. (**A** to **I**) Rows and columns correspond to ENSO phase during initialization and the subsequent year in the forecast, respectively. Colors indicate precision, with the maximum precision within each model indicated in black font and the minimum precision in white. The accuracy for each initial state is shown to the right of the corresponding row. The overall accuracy for each model is indicated in parentheses next to the model name. Stars indicate where the precision is higher than that obtained using the observed TPM.

Downloaded from https://www.science.org on October 24, 2026

transitions from El Niño (indicated by gray stars and asterisks). However, models are entirely unable to outperform the baseline when predicting the transition from neutral to El Niño: In other words, they are more likely to issue a false alarm for El Niño when predicting from neutral states than the observed TPM is. This is consistent with previous findings of a high false alarm rate when predicting El Niño events from boreal spring months (*51*, *52*). Together, these results indicate that forecasts of El Niño should be treated with caution when issued from neutral conditions (as they are likely to be a false alarm) but can be issued with confidence when emerging from a La Niña. Inferences regarding biases for each model and each transition type can be made using the precision along with the sensitivity or hit rate (presented in fig. S4).

When the precision is calculated for each row as a group or aggregated over all transitions, it is equal to simply the fraction of forecasts that were correct, known as the accuracy. We find that models generally have the highest accuracy from La Niña conditions, consistent with our earlier results; they also consistently perform better than the observed TPM from El Niño but not from neutral states.

While model accuracy when predicting from neutral conditions is in the range of 23 to 43%, using the observed TPM to issue a naïve climatological forecast after each neutral year results in an accuracy of 61%, highlighting models' low skill at forecasting the correct category from neutral conditions: A decision-maker would be better off believing the naïve climatological forecast than the models when predicting the next year's conditions from a neutral state.

Multiyear forecasts

Although some studies have detected 2-year or longer skill in ENSO forecasts in particular models (*17*) or for specific types of events (*13*, *18*, *27*), we find that the precision is reduced below 50% for every transition type in the second year.

To further simplify the target of prediction, we assess models' ability to predict the absence of each ENSO phase for the first 4 years from initialization (Fig. 5). In this setting, a model forecast is correct if it predicts, e.g., the absence of any La Niñas for a number of consecutive years and the real outcomes for those years are either El Niño or neutral. This information, though less specific than annual ENSO phase, can still be of use to stakeholders who may be concerned about the occurrence of one ENSO phase but not the others or of multiyear periods with weakly warm or cool conditions that can lead to longer-term impacts such as multiyear droughts (*53*).

We find that in these terms, the only case where models display consistently high performance for the second year is when the absence of El Niño is predicted from an El Niño. Beyond the first year, even with the highly simplified target for prediction, models perform relatively poorly. At the 2-year lead time, only when predicting the absence of La Niña from a neutral state (Fig. 5A) and the absence of El Niño from another El Niño (Fig. 5C) do models consistently attain a precision greater than 50%, i.e., the threshold above which a forecast is more likely to be correct than wrong. In the case of the former, the multimodel mean precision is still lower than that achieved by a climatological forecast.

The multimodel mean precision is higher than that of the observed TPM-based forecast (indicated by gray circles) for some years following a La Niña and in some cases after El Niño; however, after the second year, both the models and observed TPM are more likely to be wrong than correct as the precision falls below 50%.

There is only a single situation where all models perform better than the baseline forecast (indicated by the gray star) while also predicting with precision greater than 50%: when they predict the absence of El Niño for 2 years following an El Niño event. This is also the case for the sensitivity or hit rate (fig. S4), indicating that all of the models are both likely to be correct when they issue such a forecast (based on the precision) and unlikely to miss this situation when it occurs (based on the sensitivity).

This is in agreement with prior work using a single model that has shown the presence of skill at the 2-year lead time when predicting from the peak of an El Niño event, attributable to the strong discharge of upper-ocean heat content (*17*), but further demonstrates that this skill is consistently present across nine models. This forecast skill when predicting from El Niño may be contributed by the very strongest El Niño events, which have been shown to lead to forecast skill for the subsequent 2 years (*18*). While weaker El Niño events may not facilitate such long-range predictability, here, we

Downloaded from https://www.science.org on October 24, 2026

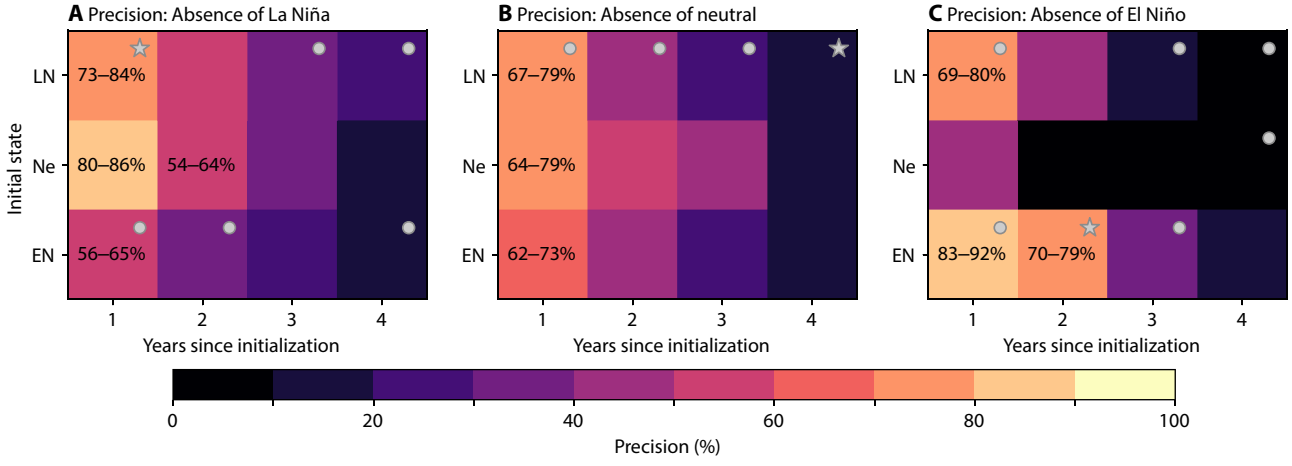


Fig. 5. The precision of predictions of the absence of each ENSO phase in ONDJF 1 to 4 years in advance for each initial state (rows), averaged over all models in the study. The precision (color) is displayed for the absence of La Niña (**A**), neutral (**B**), and El Niño (**C**) conditions respectively. Here, the columns each indicate the number of years since initialization. The minimum and maximum values across models are indicated on cells where the metric exceeded 50% for all models. Gray circles indicate where the multimodel mean exceeds the performance of the observed TPM, and stars indicate where all models perform better than the observed TPM.

have also coarsened the target of prediction to a simple absence of events, not discriminating between La Niña and neutral states, which may boost models' ability to detect persistent cool states (*20*).

DISCUSSION

Our analysis of observed year-to-year transition probabilities between ENSO phases shows that there is predictability inherent to the ENSO system from 1 year to the next, despite the well-known SPB. This is supported by the qualitative agreement between several large model ensembles and the observations regarding the probabilities of transition between the three main states of El Niño, La Niña, and neutral. The examination of model forecasts in these terms has produced two key results that can be applied to interpreting forecasts of ENSO:

1) La Niña events present a window of opportunity for skillful forecasts as early as 1 year before the ENSO peak season of October to February. Forecasts of the ENSO phase made from La Niña events show skill comparable to operational forecasts at a much shorter lead time and can therefore be issued with higher confidence before the boreal spring. Forecasts of the transition from La Niña to El Niño in particular are likely to be correct.

2) Initialized model forecasts made from the same season during neutral conditions perform poorly despite the inherent predictability in the system. This suggests that methods that do not rely on initializing models (such as statistical forecasts) may be more successful in this setting and that dynamical forecasts from neutral states should not be used until later seasons when confidence may be higher.

These results can be used to inform decisions regarding when to communicate forecasts and warnings regarding ENSO events and the level of confidence to assign to them. During the peak of an ongoing La Niña, forecasts of the next year's event could be communicated to the public; during neutral states, it would be more prudent to wait until later in the year. The transition-specific precision metric is straightforward to calculate for any model and can be used to support these decisions. The second result above motivates further research into the performance of statistical forecasting methods for ENSO that leverage the system's predictability based on past information, which hold promise for predicting from neutral states. These differences between initial states hold true across nine dynamical models developed by different organizations. The differences in performance between the various models are far smaller than the differences arising from the ENSO phase at initialization.

The relatively high skill when predicting from La Niña, particularly when predicting an upcoming El Niño, indicates that these forecasts are trustworthy earlier than the SPB. Boreal spring is the time of year when the east Pacific is at its warmest, the zonal SST gradient across the Pacific at its smallest, and the intertropical convergence zone at its southernmost position. These conditions are thought to contribute to the system's characteristic unpredictability at this time of year, though the exact mechanisms are still an area of active investigation. It may be the case that SST anomalies that linger beyond the La Niña event peak into boreal spring [since La Niñas, unlike El Niños, do not abruptly terminate around February to March (*39*)] ameliorate these conditions by cooling the east Pacific and strengthening the zonal SST gradient. It has also been pointed out that atmospheric noise is less prevalent during La Niña conditions (*54*, *55*), perhaps leading to higher predictability. Another possibility is that the degree to which the warm water volume in the upper tropical Pacific (an early indicator of ENSO activity) is replenished by the recharge phase during La Niña is a key determinant of what will follow in the next year, meaning that the information added through initialization has a strong influence on skill. Potential explanations such as these for the increased skill when predicting from La Niña will be the focus of future work.

The finding that initializing dynamical forecasts diminishes skill when predicting from neutral states is consistent with other work that found low skill when predicting from neutral states (*27*). The dynamical models used in DCPP are not necessarily identical to those used in operational seasonal forecasting and may use differing initialization protocols. However, the consistency of the lowered skill from neutral states across different models' hindcasts created using different modeling centers' data assimilation, ensemble generation, and initial perturbation techniques suggests that this finding likely applies widely across different initialization methods. Our method demonstrates that the loss of skill from neutral states arises from forecast initialization by comparing the free-running and initialized behavior of the models. This is a key difference from methods that use analog or other statistical techniques that eliminate the initialization process (*18*, *26*), meaning that they might not detect this variation in skill.

While initialized ensemble forecasts perform poorly from neutral conditions, the historical TPM (both in observations and models) displays inherent predictability which is erased by initialization. Methods that do not require the initialization of a perturbed ensemble, such as analog (*56*) or statistical machine learning (*15*, *57*) techniques, may therefore be better strategies to add skill when predicting from neutral states as they rely on knowledge of historical patterns.

Although it has been suggested that specific types of events may be better predicted at long lead times (*6*, *13*, *14*), we find that this is less of a factor than the phase during initialization. Stronger La Niña events are predicted better than weaker ones; this effect is less clear for El Niño.

Our results show that the variation in skill within a single dynamical model, based on the transition type, surpasses the differences in overall performance between different dynamical models. This suggests that weighting or assessing confidence on the basis of the transition type may be a more effective way to interpret and enhance probabilistic forecasts than model selection or weighting schemes (*58*, *59*). The framework we apply, of assessing trustworthiness on the basis of forecast precision by transition type, is a simple, intuitive method that can be used to support forecast interpretation and communication in real time.

Over the 2-year timescale, higher skill in forecasts initialized from El Niño is consistent with other skill assessments (*17*, *18*, *27*). We find that this is primarily attributable to the ability to detect a lack of El Niño events for a 2-year period, potentially due to the time necessary for the warm water volume to be recharged. The enhanced predictability from La Niña conditions may explain the predictability into a second year found both after strong El Niño events (*17*, *18*) (that are frequently followed by La Niña) and of multiyear La Niña events (*19*–*22*). While the lack of El Niño events for 2 years following an El Niño is the only situation predicted well at the 2-year lead time, this is valuable information for stakeholders: Such persistent cool states of the tropical Pacific have been shown to produce multiyear droughts across the global subtropics (*53*).

Downloaded from https://www.science.org on October 24, 2026

Our analysis provides a straightforward, interpretable approach to assess ENSO predictability by reducing the target of prediction to only the ENSO phase during the event peak and categorizing forecasts on the basis of the ENSO phase at initialization 1 year in advance. We use only annual lead times to be able to characterize the relationship between one ENSO event and the next. A large sample size enables robust statistical interpretability; however, complexities such as background trends, decadal variability, and ENSO diversity are neglected here. Subsequent research could build upon our findings by progressively reintroducing these complexities, to assess how they modulate patterns of predictability.

Moreover, strategies for skill improvement that may be used operationally, including model calibration or weighting of different models on the basis of past skill could be applied in the future. The skill that we detect may be a lower estimate than what is possible in real forecasting when information such as ENSO flavors, warm-water volume, and interbasin interactions may be used as indicative precursors. We have also not sought to explore the many recent developments in analog methods and statistical machine learning (*15*, *56*, *57*, *60*, *61*) that may enhance skill when predicting from states that are sensitive to initialization-induced perturbations and biases. Overall, our study demonstrates that the initial ENSO phase is a critical determinant of forecast skill that presents an opportunity to assess confidence in early forecasts. Leveraging this insight strategically could substantially enhance long-lead ENSO prediction and forecast communication, with implications for global risk management and adaptation.

MATERIALS AND METHODS

Data

In this study, we use the Extended Reconstructed SST version 5 (*62*) produced by the US National Oceanic and Atmospheric Administration (NOAA). Our analysis of predictability in models uses hindcasts from the DCPP archive (*63*) of the Coupled Model Intercomparison Project version 6 (CMIP6). Table 2 lists the nine models used, the timing of their ensemble initialization, and the size of the ensemble. Different modeling centers around the world developed these nine models, and each applied their ensemble generation, initial perturbation, and data assimilation techniques to generate the DCPP hindcasts, ensuring that a range of initialization techniques are represented in this study.

This multimodel hindcast archive consists of decade-long simulations, initialized in each year from 1961 to approximately 2020. Although some modeling centers initialize forecasts multiple times each year, we make use of only a single initialization per year, using those initialized between the months of October and January to align with the typical peak period of ENSO events. The ensemble size for the different models ranges from 5 to 40. This is therefore an extensive collection of interyear transitions: Each 10-year simulation contains 9 annual transitions, we use a total of 125 ensemble members across models, and the simulations are initialized in 60 different years, giving a total of 67,500 annual transitions. Restricting our dataset to only those simulations that were initiated between October and January allows us to clearly differentiate between ENSO phases during initialization and to also examine the predictability of events at a lead time of ~1 year in advance of an event peak, at least one season earlier than the boreal SPB.

Preprocessing hindcast and historical data

We use the Niño3.4 index, averaged over ONDJF to assess the occurrence of ENSO events in the models and observations, as in other work seeking to classify years into La Niña, neutral, and El Niño in large ensembles (*64*). This index is used with some preprocessing, described below, applied.

To account for the bias induced by initialization shock and model drift at the start of each 10-year segment, within each model, we calculated the average SST in the Niño3.4 region for each timestep across all such 10-year segments (giving a "decadal climatology" consisting of 120 timesteps) that was subtracted from the total Niño3.4 SST to obtain the Niño3.4 index, as in other studies using decadal hindcasts (*65*, *66*). Then, to remove the long-term trend, the decadal mean from each 10-year segment (across all ensemble members) was subtracted. No further calibration or bias correction was applied to these data.

To construct the TPMs for each model's internal representation of ENSO (fig. S2), we used only the last 5 years of each 10-year hindcast. As in previous work using large ensembles (*67*, *68*), this is done to ensure independence, i.e., that the frequency of occurrence of these transitions is not biased by the information at the initialization of each hindcast. This, however, retains the external forcings (e.g., radiative forcing from atmospheric composition changes) imposed on the models, making it comparable to the observed record. For assessing forecast skill, we used only the transitions at the beginning of each hindcast, i.e., the transition from initial state to the first forecast ONDJF for single-year skill and to subsequent years for multiyear skill. To facilitate a fair comparison with the DCPP models processed in this manner, we subtracted the 10-year running mean from the Niño3.4 anomalies in the historical SST data instead of detrending.

Defining transition probabilities

We apply a simple definition of the three event categories (La Niña, neutral, and El Niño) that can be applied across models and observational datasets: We calculate the terciles of the empirical distribution of the ONDJF Niño3.4 index as calculated above and assign each year to the category of La Niña if the ONDJF Niño3.4 for that year is below or equal to the lower tercile, neutral if it is greater than the lower tercile but less than the upper tercile, and El Niño if it is equal to or greater than the upper tercile (Fig. 1C). This is applied to each model or dataset on the basis of its own ONDJF Niño3.4 distribution so that by construction, within each model, the three categories are of equal probability. The advantages of this simple definition are that:

1) This accounts for each model's inherent biases in event magnitude, asymmetry, and the shape of the Niño3.4 distribution, capturing ENSO events consistent with the models' internal representation of ENSO; allowing for a comparison of equivalent phenomena across models and observations.

2) Patterns of predictability in interyear transitions can be compared against a baseline of a one-third probability of transition from any given category to another, which represents the probability expected if the sequence of ENSO events was truly random and drawn from a uniform distribution (i.e., a fair "coin toss" of a three-sided coin). Transitions with probability far from this threshold imply inherent predictability.

Examining the ENSO state annually rather than monthly introduces an important difference from some prior studies of models' categorical forecast skill (*18*, *32*, *69*): When transitions are assessed

Downloaded from https://www.science.org on October 24, 2026

at a seasonal or monthly frequency, the system must necessarily pass through a neutral state to move from one extreme to another. For example, if the system is at an El Niño peak in December, over the course of the year, it must pass through a neutral state, even if briefly, to reach a La Niña peak in the following December. Skill assessment metrics must therefore carefully account for this inherent difference in transition probabilities. Assessing the ENSO state annually does not impose this requirement, meaning that the "no information" baseline can simply be set at the intuitive threshold of one-third for each transition's probability, and the TPM can be used to characterize the "climatological" or "damped persistence" forecast (*69*).

In addition to allowing intermodel comparisons and establishing a clear, easy-to-understand baseline, we find that in the historical datasets, this definition does not result in substantial deviation from the ENSO events identified by major forecasting agencies. For example, the NOAA Climate Prediction Center (CPC) categorizes 3-month periods into La Niña, neutral, or El Niño on the basis of whether the seasonal-mean Niño3.4 exceeds a threshold of ±0.5 and persists for five or more overlapping 3-month periods (*70*). In the period from 1960 to the present (65 years), our categorization deviates from theirs in only 4 years (1966, 1967, 1996, and 2008), of which one (1996) lies close to the ±0.5 Niño3.4 threshold and is classified as neutral in their list but has been considered a La Niña event in some detailed studies of the 1995 to 2000 period (*71*, *72*). Meanwhile, the Australian Bureau of Meteorology categorizes years on the basis of more complex criteria involving the Southern Oscillation Index, Trade wind activity, and Niño3.4 SSTs (*73*) and disagrees with NOAA CPC's DJF categorizations in 12 years since 1960 (1968, 1976, 1983, 1984, 1986, 1993, 1995, 2004, 2005, 2014, 2017, and 2018) (*74*), indicating that our simple tercile-based definition of the three categories produces results well within the amount of variation between different agencies' definitions. Further, another study (*27*) that constructed annual transition matrices on the basis of different categorization criteria arrived at a qualitatively similar transition matrix to that in Fig. 1, suggesting that the sensitivity of our results to this definition is low.

Detection of difference from random chance in TPMs

To assess whether the observed time series of ENSO events has a pattern significantly different from random chance, we apply the exact multinomial test to each row of the TPM, with the null hypothesis that the three-category distribution is not different from a uniform distribution of ($^1/_3$, $^1/_3$, $^1/_3$). The test returns a *P* value. This is the discrete equivalent of a *t* test (as is the chi-square test) and performs well in settings with small sample sizes (*75*).

Quantifying uncertainty in observed TPMs

The TPMs are constructed from a small sample because of the shortness of the observational record of SSTs in the Pacific (and the lower confidence in observations preceding the satellite era). This means that they may be sensitive to small errors in the record or single events. To assess this uncertainty, we use a leave-one-out jackknife resampling approach: The TPMs are recalculated repeatedly, each time excluding one event (year). This results in a distribution of TPMs from which we estimate a 95% confidence interval for each transition. These are indicated in table S1. We also use these resampled TPMs to estimate the uncertainty in the climatological baseline used in calculating the RPSS (see the "Forecast verification" section).

Forecast verification

We use a range of verification metrics in this study to confirm that the results are not sensitive to the metric chosen. To assess the performance of deterministic individual forecasts in Fig. 2A, we use the MCC, also known as the phi coefficient, which for *K* categories is defined as

$$\mathrm{MCC} = \frac{c \times s - \sum_{k=1}^{K} p_k \times t_k}{\sqrt{\left(s^2 - \sum_{k=1}^{K} p_k^2\right) \times \left(s^2 - \sum_{k=1}^{K} t_k^2\right)}} \quad (1)$$

where c is the total number of samples correctly predicted, s is the total number of samples, p_k is the number of times category k was predicted, and t_k is the number of times category k truly occurred (*76*). This quantity behaves analogously to the Pearson's correlation coefficient.

To quantify the ensemble forecasts' performance in Fig. 2 (B and C), we have used the RPS. RPS is defined as

$$\mathrm{RPS} = \sum_{k=1}^{K} \left[\sum_{j=1}^{k} \left(p_j - o_j\right) \right]^2 \quad (2)$$

where K is the number of categories, i.e., 3; p_j is the predicted probability of the j^{th} category, and o_j is the observed probability of the same category, expressed as 0 or 1. We present this quantity averaged over multiple forecasts in Fig. 2.

The RPS can then be used to calculate the RPSS

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{\mathrm{ref}} \rangle} \quad (3)$$

where $\mathrm{RPS}_{\mathrm{ref}}$ is the RPS obtained by using the TPM constructed from the observations to issue naïve forecasts, and the angled brackets represent the averaging over many forecasts.

The RPSS, unlike the commonly used Brier skill score, does not penalize all incorrect forecasts equally. Instead, it penalizes a forecast that is two categories away from the correct outcome more harshly than a forecast in the next category (*77*). This means that e.g., when the correct outcome is El Niño, a forecast of La Niña is treated as "more incorrect" than a forecast of neutral. The results in Fig. 2 were found to be similar when Brier skill scores were used instead. Where we have reported RPSS (Table 1 and tables S3 and S4), we also assessed 95% confidence intervals representing plausible variations in $\mathrm{RPS}_{\mathrm{ref}}$ arising from the uncertainty in the TPMs derived from the short observational record. These intervals were estimated from a distribution of RPSS constructed by repeating the calculation of $\mathrm{RPS}_{\mathrm{ref}}$ using each resampled TPM (as described in the "Quantifying uncertainty in observed TPMs" section).

We also used the percentage of ensemble members in the correct category across models for each initialization year in Fig. 3. This is similar to the treatment in calculating the MCC in Fig. 2A and the precision and sensitivity in Figs. 4, fig. S4, 5, and fig. S5. We do not find a qualitative difference in the relative performance between ENSO phases when forecasts are grouped in this manner compared to the probabilistic forecasts.

We aggregate all forecasts in assessing the precision, accuracy, and sensitivity. This ensures sufficiently large samples for each type of transition to estimate these metrics separately.

Downloaded from https://www.science.org on October 24, 2026

Discrimination metrics

We use three discrimination metrics in our analysis: precision, sensitivity, and accuracy, with a focus on precision in the "Assessing categorical forecasts in real time: precision based on ENSO transition type" section. Discrimination metrics have been developed to assess model performance in classification tasks (particularly in machine learning (*78*) and meteorological classification (*46*, *77*, *79*) applications) and are designed to carefully account for the contributions to skill from true positives and true negatives, and the errors introduced by false positives (also known as false alarms) and false negatives, in various combinations. The first two are specific to each transition type. Precision is used to assess the probability of a forecast being correct. Sensitivity is used to examine model bias by showing which transitions are most likely to be missed by forecasts. Accuracy is measured over the entire set of transitions, or for particular initial states, and indicates the fraction of forecasts that were correct.

Mathematically, they are defined as follows. The accuracy is assessed over the full set of predictions of $k = 3$ categories

$$\text{Accuracy} = \frac{c}{p} = \sum_{k=1}^{K} \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{4}$$

where c and p represent the number of correct predictions and the total number of predictions, respectively, and TP and FP stand for the number of true positives and false positives.

The precision and sensitivity are assessed for each transition type, from state i to state j

$$\text{Precision} = \frac{c_{i,j}}{p_{i,j}} = \frac{\text{TP}_{i,j}}{\text{TP}_{i,j} + \text{FP}_{i,j}} \tag{5}$$

When used to assess the confidence of a current forecast, the precision is to be interpreted as a conditional probability: Given that the model has made a prediction, the precision is the empirical probability that the prediction will come true, i.e.

$$\text{Precision} = P\left(A_j | B_j\right) = \frac{P\left(A_j \cap B_j\right)}{P\left(B_j\right)} \tag{6}$$

where $P\left(A_j\right)$ is the probability of category j occurring in the real world and $P\left(B_j\right)$ is the probability of the model forecasting an outcome of category j. This uses only quantities known at the time the forecast is issued.

The precision does not account for false negatives and therefore cannot be used as a complete measure of performance. For example, a low-skill model that fails to forecast 99 of 100 instances of real-world transitions from El Niño to La Niña (EN-LN) can still have perfect precision if it correctly forecasts the single remaining EN-LN transition and never produces another forecast of EN-LN; on the other hand, high skill always implies high precision. The precision is equal to one minus the false alarm rate.

The sensitivity (also known as the hit rate) is a complementary metric that assesses how many real-world instances of a transition were correctly predicted by the model.

$$\text{Sensitivity} = \frac{c_{i,j}}{o_{i,j}} = \frac{\text{TP}_{i,j}}{\text{TP}_{i,j} + \text{FN}_{i,j}} \tag{7}$$

where $o_{i,j}$ is the number of times the transition from i to j occurred in the real world. $\text{FN}_{i,j}$ represents the number of false negatives, i.e., the number of real-world occurrences of the transition from i to j that were missed by the model forecasts.

Like precision, sensitivity provides an incomplete picture of performance. For example, a low-skill model that only ever issues forecasts of EN-LN in every instance will have perfect sensitivity as it never misses this transition when it occurs but issues a large number of false positives. Note that the difference between precision and sensitivity is in the denominator: The shared numerator of the number of correct predictions is compared against the properties of the forecasts for precision and against the properties of the observations for the sensitivity. The sensitivity, since it relies on knowledge of the real-world outcome, cannot be estimated at the time of issuing a forecast but can be used to assess model bias.

These three discrimination metrics are compared against climatological forecasts which are constructed by randomly sampling the appropriate row of the historical TPM computed from observational data before the forecast date and drawing the same number of samples as the model ensemble size. This is repeated using the jackknife resampling method to then construct a distribution of the metric for each cell of the matrix from which 95% confidence intervals are estimated. Cells or rows of the models' precision and sensitivity matrices in Figs. 4 and 5, and figs. S4 and S5, are marked using stars or asterisks if their performance metric is higher than the upper limit of the confidence interval estimated from the resampled observed TPMs, indicating skill relative to the observations. In Fig. 5, where the multimodel mean is shown, a gray circle is displayed where the multimodel mean exceeds this threshold and a gray star where the metric from all models exceeds the threshold.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S5
Tables S1 to S4
References

REFERENCES

1. M. A. Cane, S. E. Zebiak, S. C. Dolan, Experimental forecasts of El Niño. *Nature* **321**, 827–832 (1986).
2. L. Goddard, M. Dilley, El Niño: Catastrophe or opportunity. *J. Climate* **18**, 651–665 (2005).
3. K. Wyrtki, Fluctuations of the dynamic topography in the Pacific Ocean. *J. Phys. Oceanogr.* **5**, 450–459 (1975).
4. F.-F. Jin, An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model. *J. Atmos. Sci.* **54**, 811–829 (1997).
5. C. S. Meinen, M. J. McPhaden, Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña. *J. Climate* **13**, 3551–3559 (2000).
6. N. Ramesh, R. Murtugudde, All flavours of El Niño have similar early subsurface origins. *Nat. Clim. Change* **3**, 42–46 (2013).
7. D. Lemoine, S. Kapnick, Financial markets value skillful forecasts of seasonal climate. *Nat. Commun.* **15**, 4059 (2024).
8. M. Bouma, R. Kovats, S. Goubet, J. S. H. Cox, A. Haines, Global assessment of El Niño's disaster burden. *Lancet* **350**, 1435–1438 (1997).
9. Y. Jeon, S. Kim, S.-J. Sohn, S. Han, S. Moon, Toward a sustainable and resilient society through enhanced ENSO response and preparedness. *Bull. Am. Meteorol. Soc.* **106**, E571–E575 (2025).
10. M. Latif, M. Flügel, An investigation of short-range climate predictability in the tropical Pacific. *J. Geophys. Res. Oceans* **96**, 2661–2673 (1991).
11. P. Webster, The annual cycle and the predictability of the tropical coupled ocean-atmosphere system. *Meteorol. Atmos. Phys.* **56**, 33–55 (1995).
12. M. A. Balmaseda, M. K. Davey, D. L. Anderson, Decadal and seasonal dependence of ENSO prediction skill. *J. Climate* **8**, 2705–2715 (1995).

Downloaded from https://www.science.org on October 24, 2026

13. D. Chen, M. A. Cane, A. Kaplan, S. E. Zebiak, D. Huang, Predictability of El Niño over the past 148 years. *Nature* **428**, 733–736 (2004).
14. P. L. Gonzalez, L. Goddard, Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Climate Dynam.* **46**, 3127–3147 (2016).
15. Y.-G. Ham, J.-H. Kim, J.-J. Luo, Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568–572 (2019).
16. X. Wu, Y. M. Okumura, C. Deser, P. N. DiNezio, Two-year dynamical predictions of ENSO event duration during 1954–2015. *J. Climate* **34**, 4069–4087 (2021).
17. S. Sharmila, H. Hendon, O. Alves, A. Weisheimer, M. Balmaseda, Contrasting El Niño–La Niña predictability and prediction skill in 2-year reforecasts of the twentieth century. *J. Climate* **36**, 1269–1285 (2023).
18. N. Lenssen, P. DiNezio, L. Goddard, C. Deser, Y. Kushnir, S. J. Mason, M. Newman, Y. Okumura, Strong El Niño events lead to robust multi-year ENSO predictability. *Geophys. Res. Lett.* **51**, e2023GL106988 (2024).
19. P. N. DiNezio, C. Deser, Y. Okumura, A. Karspeck, Predictability of 2-year La Niña events in a coupled general circulation model. *Climate Dynam.* **49**, 4237–4261 (2017).
20. N. Ramesh, M. A. Cane, R. Seager, D. E. Lee, Predictability and prediction of persistent cool states of the tropical Pacific Ocean. *Climate Dynam.* **49**, 2291–2307 (2017).
21. P. N. DiNezio, C. Deser, A. Karspeck, S. Yeager, Y. Okumura, G. Danabasoglu, N. Rosenbloom, J. Caron, G. A. Meehl, A 2 year forecast for a 60–80% chance of La Niña in 2017–2018. *Geophys. Res. Lett.* **44**, 11–624 (2017).
22. J.-J. Luo, G. Liu, H. Hendon, O. Alves, T. Yamagata, Inter-basin sources for two-year predictability of the multi-year La Niña event in 2010–2012. *Sci. Rep.* **7**, 2276 (2017).
23. H. Astudillo, F. Borotto, R. Abarca-del Rio, Embedding reconstruction methodology for short time series–application to large El Niño events. *Nonlinear Processes Geophys.* **17**, 753–764 (2010).
24. M. L. L'Heureux, K. Takahashi, A. B. Watkins, A. G. Barnston, E. J. Becker, T. E. di Liberto, F. Gamble, J. Gottschalck, M. S. Halpert, B. Huang, K. Mosquera-Vásquez, A. T. Wittenberg, Observing and predicting the 2015/16 El Niño. *Bull. Am. Meteorol. Soc.* **98**, 1363–1382 (2017).
25. M. A. Ehsan, M. L. L'Heureux, M. K. Tippett, A. W. Robertson, J. Turmelle, Real-time ENSO forecast skill evaluated over the last two decades, with focus on the onset of ENSO events. *npj Clim. Atmos. Sci.* **7**, 301 (2024).
26. Z. Hou, J. Li, R. Ding, C. Karamperidou, W. Duan, T. Liu, J. Feng, Asymmetry of the predictability limit of the warm ENSO phase. *Geophys. Res. Lett.* **45**, 7646–7653 (2018).
27. Y. Liu, M. G. Donat, M. H. England, L. V. Alexander, A. L. Hirsch, C. Delgado-Torres, Enhanced multi-year predictability after El Niño and La Niña events. *Nat. Commun.* **14**, 6387 (2023).
28. B. P. Kirtman, P. S. Schopf, Decadal variability in ENSO predictability and prediction. *J. Climate* **11**, 2804–2822 (1998).
29. C. Karamperidou, M. A. Cane, U. Lall, A. T. Wittenberg, Intrinsic modulation of ENSO predictability viewed through a local Lyapunov lens. *Climate Dynam.* **42**, 253–270 (2014).
30. N. Ramesh, M. A. Cane, The predictability of tropical Pacific decadal variability: Insights from attractor reconstruction. *J. Atmos. Sci.* **76**, 801–819 (2019).
31. J. Lou, M. Newman, A. Hoell, Multi-decadal variation of ENSO forecast skill since the late 1800s. *NPJ Clim. Atmos. Sci.* **6**, 89 (2023).
32. M. K. Tippett, A. G. Barnston, Skill of multimodel ENSO probability forecasts. *Mon. Weather Rev.* **136**, 3933–3946 (2008).
33. M. K. Tippett, M. Ranganathan, M. L'Heureux, A. G. Barnston, T. DelSole, Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Climate Dynam.* **53**, 7497–7518 (2019).
34. T. Geng, F. Jia, W. Cai, L. Wu, B. Gan, Z. Jing, S. Li, M. J. McPhaden, Increased occurrences of consecutive La Niña events under global warming. *Nature* **619**, 774–781 (2023).
35. M. B. Freund, J. R. Brown, A. G. Marshall, C. R. Tozer, B. J. Henley, J. S. Risbey, N. Ramesh, R. Lieber, S. Sharmila, Interannual ENSO diversity, transitions, and projected changes in observations and climate models. *Environ. Res. Lett.* **19**, 114005 (2024).
36. S.-I. An, J.-W. Kim, ENSO transition asymmetry: Internal and external causes and intermodel diversity. *Geophys. Res. Lett.* **45**, 5095–5104 (2018).
37. G. Conti, A. Navarra, J. Tribbia, The ENSO transition probabilities. *J. Climate* **30**, 4951–4964 (2017).
38. Y. M. Okumura, M. Ohba, C. Deser, H. Ueda, A proposed mechanism for the asymmetric duration of El Niño and La Niña. *J. Climate* **24**, 3822–3829 (2011).
39. S. McGregor, N. Ramesh, P. Spence, M. H. England, M. J. McPhaden, A. Santoso, Meridional movement of wind anomalies during ENSO events and their role in event termination. *Geophys. Res. Lett.* **40**, 749–754 (2013).
40. N. K. Larkin, D. Harrison, ENSO warm (El Niño) and cold (La Niña) event life cycles: Ocean surface anomaly patterns, their symmetries, asymmetries, and implications. *J. Climate* **15**, 1118–1140 (2002).
41. S. G. Philander, A. Fedorov, Is El Niño sporadic or cyclic? *Annu. Rev. Earth Planet. Sci.* **31**, 579–594 (2003).
42. X. Wu, Y. M. Okumura, P. N. DiNezio, What controls the duration of El Niño and La Niña events? *J. Climate* **32**, 5941–5965 (2019).
43. R. Seager, N. Henderson, M. Cane, H. Zhang, J. Nakamura, Atmosphere–ocean dynamics of persistent cold states of the tropical Pacific Ocean. *J. Climate* **34**, 5195–5214 (2021).
44. W. S. Kessler, Is ENSO a cycle or a series of events? *Geophys. Res. Lett.* **29**, 40–1–40–4 (2002).
45. S.-I. An, E. Tziperman, Y. M. Okumura, T. Li, "ENSO irregularity and asymmetry" in *El Niño Southern Oscillation in a Changing Climate*, Geophysical Monograph 253, M. J. McPhaden, A. Santoso, W. Cai, Eds. (American Geophysical Union; John Wiley & Sons Inc., 2021), chap. 7, pp. 153–172.
46. G. W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
47. B. P. Kirtman, D. Min, J. M. Infanti, J. L. Kinter III, D. A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M. P. Mendez, E. Becker, P. Peng, P. Tripp, J. Huang, D. G. DeWitt, M. K. Tippett, A. G. Barnston, S. Li, A. Rosati, S. D. Schubert, M. Rienecker, M. Suarez, Z. E. Li, J. Marshak, Y. K. Lim, J. Tribbia, K. Pegion, W. J. Merryfield, B. Denis, E. F. Wood, 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).
48. J. S. Risbey, D. T. Squire, A. S. Black, T. DelSole, C. Lepore, R. J. Matear, D. P. Monselesan, T. S. Moore, D. Richardson, A. Schepen, M. K. Tippett, C. R. Tozer, Standard assessments of climate forecast skill can be misleading. *Nat. Commun.* **12**, 4346 (2021).
49. N. J. Lenssen, L. Goddard, S. Mason, Seasonal forecast skill of ENSO teleconnection maps. *Weather Forecast.* **35**, 2387–2406 (2020).
50. D. P. Mulholland, P. Laloyaux, K. Haines, M. A. Balmaseda, Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Mon. Weather Rev.* **143**, 4631–4644 (2015).
51. M. K. Tippett, M. L. L'Heureux, E. J. Becker, A. Kumar, Excessive momentum and false alarms in late-spring ENSO forecasts. *Geophys. Res. Lett.* **47**, e2020GL087008 (2020).
52. A. F. Levine, M. L'Heureux, C. Wen, Understanding spring forecast El Niño false alarms in the North American Multi-Model Ensemble. *NPJ Clim. Atmos. Sci.* **8**, 94 (2025).
53. C. Herweijer, R. Seager, The global footprint of persistent extra-tropical drought in the instrumental era. *Int. J. Climatol.* **28**, 1761–1774 (2008).
54. I. Eisenman, L. Yu, E. Tziperman, Westerly wind bursts: ENSO's tail rather than the dog? *J. Climate* **18**, 5224–5238 (2005).
55. A. Levine, F. F. Jin, M. J. McPhaden, Extreme noise–extreme El Niño: How state-dependent noise forcing creates El Niño–La Niña asymmetry. *J. Climate* **29**, 5483–5499 (2016).
56. H. Ding, M. Newman, M. A. Alexander, A. T. Wittenberg, Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *J. Climate* **31**, 5437–5459 (2018).
57. P. J. Petersik, H. A. Dijkstra, Probabilistic forecasting of El Niño using neural network models. *Geophys. Res. Lett.* **47**, e2019GL086423 (2020).
58. S. Casanova, B. Ahrens, On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Mon. Weather Rev.* **137**, 3811–3822 (2009).
59. T. DelSole, X. Yang, M. K. Tippett, Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Q. J. Roy. Meteorol. Soc.* **139**, 176–183 (2013).
60. M. Sun, L. Chen, T. Li, J.-J. Luo, CNN-based ENSO forecasts with a focus on SSTA zonal pattern and physical interpretation. *Geophys. Res. Lett.* **50**, e2023GL105175 (2023).
61. Y. Chen, Y. Jin, Z. Liu, X. Shen, X. Chen, X. Lin, R. H. Zhang, J. J. Luo, W. Zhang, W. Duan, F. Zheng, M. J. McPhaden, L. Zhou, Combined dynamical-deep learning ENSO forecasts. *Nat. Commun.* **16**, 3845 (2025).
62. B. Huang, P. W. Thorne, V. F. Banzon, T. Boyer, G. Chepurin, J. H. Lawrimore, M. J. Menne, T. M. Smith, R. S. Vose, H. M. Zhang, Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate* **30**, 8179–8205 (2017).
63. G. J. Boer, D. M. Smith, C. Cassou, F. Doblas-Reyes, G. Danabasoglu, B. Kirtman, Y. Kushnir, M. Kimoto, G. A. Meehl, R. Msadek, W. A. Mueller, K. E. Taylor, F. Zwiers, M. Rixen, Y. Ruprich-Robert, R. Eade, The decadal climate prediction project (DCPP) contribution to CMIP6. *Geosci. Model Dev.* **9**, 3751–3777 (2016).
64. B. Wang, W. Sun, C. Jin, X. Luo, Y. M. Yang, T. Li, B. Xiang, M. J. McPhaden, M. A. Cane, F. Jin, F. Liu, J. Liu, Understanding the recent increase in multiyear La Niñas. *Nat. Clim. Change* **13**, 1075–1081 (2023).
65. D. MacLeod, C. Caminade, A. Morse, Useful decadal climate prediction at regional scales? A look at the ENSEMBLES stream 2 decadal hindcasts. *Environ. Res. Lett.* **7**, 044012 (2012).
66. L. Magnusson, M. Alonso-Balmaseda, S. Corti, F. Molteni, T. Stockdale, Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Climate Dynam.* **41**, 2393–2409 (2013).
67. T. Kelder, M. Müller, L. J. Slater, T. I. Marjoribanks, R. L. Wilby, C. Prudhomme, P. Bohlinger, L. Ferranti, T. Nipen, Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes. *NPJ Clim. Atmos. Sci.* **3**, 47 (2020).
68. D. B. Irving, J. S. Risbey, D. T. Squire, R. Matear, C. Tozer, D. P. Monselesan, N. Ramesh, P. J. Reddy, M. Freund, A multi-model likelihood analysis of unprecedented extreme rainfall along the east coast of Australia. *Meteorol. Appl.* **31**, e2217 (2024).

Downloaded from https://www.science.org on October 24, 2026

69. S. J. Mason, G. M. Mimmack, Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate* **15**, 8–29 (2002).
70. National Weather Service Climate Prediction Center, Cold & Warm Episodes by Season, https://web.archive.org/web/20260922044023/https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso/oni/v5/ [accessed 21 September 2026].
71. G. C. Johnson, M. J. McPhaden, G. D. Rowe, K. E. McTaggart, Upper equatorial Pacific Ocean current and salinity variability during the 1996–1998 El Niño–La Niña cycle. *J. Geophys. Res. Oceans* **105**, 1037–1053 (2000).
72. F. B. Schwing, T. Murphree, L. deWitt, P. M. Green, The evolution of oceanic and atmospheric anomalies in the northeast Pacific during the El Niño and La Niña events of 1995–2001. *Prog. Oceanogr.* **54**, 459–491 (2002).
73. C. Lucas, M. C. Wheeler, H. Nguyen, A. B. Watkins, Z. W. Chua, C. Ganter, "Revising the Bureau's ENSO Alert system: Adapting indices in a changing climate" (Tech. Rep. 083, Australian Bureau of Meteorology, 2023).
74. Australian Bureau of Meteorology, Australian rainfall during El Niño and La Niña events, https://web.archive.org/web/20251023030512/https://reg.bom.gov.au/climate/history/enso/ [accessed 23 October 2025].
75. P. Kroonenberg, A. Verbeek, The tale of Cochran's rule: My contingency table has so many expected values smaller than 5, what am I to do? *Am. Stat.* **72**, 175–183 (2018).
76. J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **28**, 367–374 (2004).
77. E. S. Epstein, A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**, 985–987 (1969).
78. M. Hardt, B. Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* (Princeton Univ. Press, USA, 2022), chap. 2: Fundamentals of Prediction.
79. I. Mason, A model for assessment of weather forecasts. *Aust. Met. Mag.* **30**, 291–303 (1982).
80. S. J. Mason, C. A. Ferro, W. A. Landman, Forecasts of "normal". *Q. J. Roy. Meteorol. Soc.* **147**, 1225–1236 (2021).

Acknowledgments: This study was funded by a grant awarded to N.R. from the Drought Resilience Mission of the Commonwealth Scientific and Industrial Research Organisation. We also acknowledge the Coupled Model Intercomparison Project of the World Climate Research Programme and the Australian National Computing Infrastructure for making climate model outputs available for use. **Funding:** This study was funded by a grant provided to N.R. by the Commonwealth Scientific and Industrial Research Organisation's Drought Resilience Mission.
Author contributions: Conceptualization: N.R., M.B.F., D.B.I., D.T.S., C.R.T., D.P.M., and J.S.R. Methodology: N.R., M.B.F., D.B.I., D.T.S., and J.S.R. Data curation: D.T.S. Software: N.R., D.B.I., and D.T.S. Investigation: N.R. and M.B.F. Formal analysis: N.R. and D.T.S. Validation: N.R., M.B.F., and J.S.R. Writing—original draft: N.R., M.B.F., and J.S.R. Writing—reviewing and editing: N.R., M.B.F., D.B.I., D.T.S., C.R.T., D.P.M., T.P., A.S., and J.S.R. Funding acquisition, supervision, and project administration: N.R. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials. This study did not generate new materials. Data used in this study are publicly available from the Earth System Grid Federation by following the instructions at https://esgf.github.io/ for CMIP6 DCPP and from the National Centers for Environmental Information by following the instructions at www.ncei.noaa.gov/products/extended-reconstructed-sst for ERSST. The code used to perform the analysis is available in a Zenodo repository at DOI:10.5281/zenodo.18333876.

Submitted 2 March 2026
Accepted 22 September 2026
Published 23 October 2026
10.1126/sciadv.aeg8654

Downloaded from https://www.science.org on October 24, 2026

ScienceAdvances

Using initial states to identify successful El Niño–Southern Oscillation forecasts 1 year in advance

Nandini Ramesh, Mandy B. Freund, Damien B. Irving, Dougal T. Squire, Carly R. Tozer, Didier P. Monselesan, Tess Parker, Annette Stellema, and James S. Risbey

Sci. Adv. **12** (43), eaeg8654. DOI: 10.1126/sciadv.aeg8654

View the article online
https://www.science.org/doi/10.1126/sciadv.aeg8654
Permissions
https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org on October 24, 2026

Use of this article is subject to the Terms of service

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).