LETTER • OPEN ACCESS

Long-term ENSO prediction with echo-state networks

To cite this article: Forough Hassanibesheli et al 2022 Environ. Res.: Climate 1 011002

View the article online for updates and enhancements.

You may also like

- Monitoring the pendulum between El Niño and La Niña events
  Jingzhi Su, Tao Lian, Renhe Zhang et al.

- Prediction of Indian rainfall during the summer monsoon season on the basis of links with equatorial Pacific and Indian Ocean climate indices
  Sajani Surendran, Sulochana Gadgil, P A Francis et al.

- Causal effects of Indian Ocean Dipole on El Niño–Southern Oscillation during 1950–2014 based on high-resolution models and reanalysis data
  Thanh Le, Kyung-Ja Ha, Deg-Hyo Bae et al.
Long-term ENSO prediction with echo-state networks

Forough Hassanibesheli\textsuperscript{1,2,*}, Jürgen Kurths\textsuperscript{1,2,3} and Niklas Boers\textsuperscript{1,4,5}\textsuperscript{1} Research Domain IV—Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany
\textsuperscript{2} Department of Physics, Humboldt University Berlin, Berlin, Germany
\textsuperscript{3} Saratov State University, 83, Astrakhanskaya Str., 410012 Saratov, Russia
\textsuperscript{4} Earth System Modelling, School of Engineering & Design, Technical University of Munich, Munich, Germany
\textsuperscript{5} Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, United Kingdom

\textsuperscript{*} Author to whom any correspondence should be addressed.
E-mail: forough.hassanibesheli@pik-potsdam.de

Keywords: ENSO, artificial neural networks, echo state network

Abstract

The El Niño-Southern Oscillation (ENSO) is a climate phenomenon that profoundly impacts weather patterns and extreme events worldwide. Here we develop a method based on a recurrent neural network, called echo state network (ESN), which can be trained efficiently to predict different ENSO indices despite their relatively high noise levels. To achieve this, we train the ESN model on the low-frequency variability of ENSO indices and estimate the potential future high-frequency variability from specific samples of its past history. Our method reveals the importance of cross-scale interactions in the mechanisms underlying ENSO and skilfully predicts its variability and especially El Niño events at lead times up to 21 months. This study considers forecasts skilful if the correlation coefficients are above 0.5. Our results show that the low-frequency component of ENSO carries substantial predictive power, which can be exploited by training our model on single scalar time series. The proposed machine learning method for data-driven modeling can be readily applied to other time series, e.g. finance and physiology. However, it should be noted that our approach cannot straightforwardly be turned into a real-time operational forecast because of the decomposition of the original time series into the slow and fast components using low-pass filter techniques.

1. Introduction

The El Niño-Southern Oscillation (ENSO) is the dominant variability mode of the global climate system at interannual time scales [1]. It is related to interannual temperature and pressure anomalies in the tropical Pacific Ocean. ENSO can be classified into three main variability phases, namely the warm (i.e. El Niño) and cool (i.e. La Niña) phases with sea-surface temperature (SST) anomalies substantially above and below average, respectively, as well as the remaining neutral phases [2]. The warm ENSO phases have typical yet irregular return periods between 3 and 7 years, rendering them challenging to predict [3]. In the neutral phase, trade winds blow east to west resulting in a pile-up of warm water masses at the western boundary of the tropical Pacific Ocean. The resulting east–west SST gradient causes air to ascend in the western Pacific and circulate back to the eastern boundary of the tropical Pacific Ocean, where it descends again. This atmospheric circulation system is called the Walker cell. During El Niño (La Niña) phases, this circulation is weakened (strengthened), leading to warm (cool) SST anomalies in the central and eastern parts of the tropical Pacific.

During the last decades, substantial progress has been made in assessing spatial patterns, the temporal dynamics, the underlying physical mechanisms, and possible changes of ENSO in response to global warming [4, 5]. Pronounced anomalies of ENSO, i.e. both El Niño and La Niña events, can trigger droughts, floods, heat waves, and other extreme weather conditions around the globe [6]. Skillful forecasts of ENSO variability provide crucial information for decision makers to reduce the negative socioeconomic impacts of strong ENSO anomalies. The predictability of ENSO at interseasonal and longer time scales has therefore
attracted substantial attention, using process-based general circulation models [7, 8] and statistical approaches [9, 10]. In recent years, also machine learning models have been demonstrated to provide skillful forecasts of ENSO variability [11–13]. Many linear statistical methods such as linear inverse models [14] and model analogs [15] have been developed to forecast ENSO variability. Since ENSO dynamics is arguably nonlinear [16], also nonlinear statistical models have been introduced [10, 17]. In addition to these data-driven inverse modeling approaches, several statistical methods, based on complex network theory, have been proposed to forecast El Niño events. It has been demonstrated that complex networks provide powerful tools for investigating the spatial co-variability patterns of climatic variables [18]. Regarding the prediction of different ENSO phases, this approach suggests that spatial co-variability in the tropical Pacific plays a crucial role in ENSO predictability. Nevertheless, as for the classical statistical approaches above, the forecast horizon of network-based approaches remains limited to 1 year [19–21].

Due to their distinctive ability to capture nonlinear relationships and their capability of handling large amounts of data, ANNs have recently been employed for a broad range of applications in the geosciences, including pattern recognition, classification, and signal prediction [22, 23]. Despite their need for very large training datasets, deep-learning models [24] have been employed to successfully classify El Niño events [25] and to predict ENSO variability [12, 13]. The most skillful ENSO prediction scheme to date, which outperforms process-based dynamical models at yearly time scales, is based on a convolutional neural network (CNN) architecture [12]. This model has high forecast skill (with correlation to the observed ENSO index above 0.5) for 17 months and thus outperforms operational forecasts using process-based models by 5 months. The need for large amounts of training data for such complex ANN models has been addressed by training on process-based climate model simulations of historical climate variability before training further on reanalysis data [12].

A specific kind of ANNs that has been developed for sequential data such as climate time series is recurrent neural networks (RNNs). Due to their recurrent nature, general RNNs are computationally expensive during training and are potentially prone to vanishing/exploding gradient problems. As a result, echo state networks (ESNs) [26, 27] have been proposed as simplified RNN models with substantially faster training processes in comparison to conventional RNNs. Due to the simplified training process of ESNs, they typically require smaller amounts of training data to learn the dynamics of the underlying complex system compared to other ANN architectures. ESNs have been applied successfully for forecasting deterministically chaotic dynamical systems [26, 28]. Here, we develop an ESN-based method to predict non-deterministic time series with comparably low signal-to-noise ratio, which frequently arises in geosciences.

The key assumption underlying our approach is that ENSO variability can be approximately decomposed into slow and fast modes, where the slow mode is only mildly perturbed by high-frequency forcing such as, e.g. westerly wind bursts or the Madden–Julian oscillation [29]. Several methodologies and techniques have been developed to decompose time series into slow and fast variability components, such as, e.g. singular spectrum analysis (SSA) [30, 31], spectral methods such as the Butterworth (BW) filter [32, 33], or moving average (MA) filters [34]. Upon a careful evaluation of their performance, we focus on the BW filter in the following. In our approach, we employ an ESN architecture to model the low-frequency component (LFC) of ENSO. To address the high-frequency components and perform a forecast of the full variability, we combine the ESN with the past noise forecasting (PNF) [17] method, which samples suitable high-frequency components from past observations to force the low-frequency dynamics.

### 2. Data and method

#### 2.1. Data

We employ monthly SST anomalies for the time span 1891–2019 [35], averaged over the Niño-3 (5 N–5 S, 150 W–90 W) and Niño-3.4 (5 N–5 S, 170 W–120 W) regions in the tropical Pacific (figure 1(a)), which are commonly used to define the Niño-3 (figure 1(b)) and Niño-3.4 indices (figure 1(c)). To obtain the monthly indices, the seasonal cycles were removed by subtracting the 1981–2010 monthly means from the corresponding monthly values. The resulting time series, were then standardized.

#### 2.2. Echo state networks (ESN)

A supervised machine learning method that is suited for temporal and sequential data processing is given by ESNs [27, 37], a specialized type of RNNs with simpler implementation and training process. ESNs are comprised of three parts: an input layer, a so-called reservoir that processes the input, and an output layer that uses the reservoir output for prediction. The reservoir can be considered as a high-dimensional dynamical system with states $r(t)$ that evolve according to the following equation:

$$r(t + \Delta t) = F(W_r r(t) + W_{in} x(t)).$$  \hspace{1cm} (1)
Figure 1. Monthly sea-surface temperature (SST) anomalies in the Pacific Ocean for the 1997 El Niño event, and SST-based ENSO indices. (a) SST anomalies \(^3\) in the Pacific Ocean for December 1997, known as the strongest El Niño on record. The blue and black boxes delineate the Niño 3 (5°N–5°S, 150°W–90°W) and the Niño 3.4 (5°N–5°S, 170°W–120°W) regions from where SST anomalies are used to define the Niño-3 and Niño-3.4 indices \(^3\), respectively. (b) The Niño-3 index, given by a time series of SST anomalies averaged over the Niño-3 region in the eastern tropical Pacific as shown in (a), from 1890 to 2019. (c) SST anomalies averaged over the Niño-3.4 region from 1890 to 2019. Red and blue colors indicate SST anomalies above and below zero, respectively.

Here \(W_{\text{in}} \in \mathbb{R}^{N \times L}\) is the input matrix mapping the input of dimension \(L\) to the reservoir space of dimension \(N\). \(W_r \in \mathbb{R}^{N \times N}\) is a sparse, weighted adjacency matrix with elements drawn randomly from a normal distribution with zero means and \(\mathcal{F}\) refers to a nonlinear activation function for which we choose the hyperbolic tangent. The reservoir does not require remembering the entire temporal information of the past; hence, it skips unnecessary information after some time, resulting in extremely fast converging of the training. In this architecture, an input \(x(t)\) of dimension \(L\) is non-linearly embedded in the reservoir through a randomly chosen \((N \times L)\) matrix \(W_{\text{in}}\). Contrary to most conventional NN models, in the ESN setup, the values of \(W_{\text{in}}\) and \(W_r\) are not optimized but fixed at randomly chosen values. Hence, the only trainable connections are given by the reservoir-to-output layer matrix \(W_{\text{out}}\), through which the reservoir states are mapped back to the \(L\)-dimensional outputs \(y(t)\), i.e.

\[
y(t + \Delta t) = (W_{\text{out}}r(t + \Delta t)) \, .
\]

The goal of the system is to approximate the desired outputs \(x_d\), i.e. to find optimal values for \(W_{\text{out}}\) that minimize the loss function \(\mathcal{L}\). This optimization is most conveniently and robustly done using Ridge Regression (also known as Tikhonov Regularization) \(^3\) with a regularization penalty parameter \(\lambda\), which enhances the numerical stability and prevents over-fitting. The task is, thus, to minimize the cost function

\[
\mathcal{L} = \sum_{t=0}^{T} ||W_{\text{out}}r(t) - x_d(t)|| + \lambda||W_{\text{out}}||^2 .
\]

Here \(|.|.|\) is the \(L_2\)-norm of a vector and the regularized optimal output matrix \(W_{\text{out}}\) is determined via:

\[
W_{\text{out}} = (R^TR + \lambda I)^{-1}R^Tx_d
\]

where \(R \in \mathbb{R}^{N \times t}\) and \(I\) denotes the \(N \times N\) identity matrix. The prediction phase shown in figure S5 (corresponding to \(t \geq 0\)) initiates with the information of the reservoir state \(r(t^*)\) at the last time step of the training phase to make the prediction \(y(t^* + 1)\). Further predictions are then iteratively made forward in time, by passing \(y(t^* + 1)\) to the ESN as input to produce a forecast \(y(t^* + 2)\), and so on. Specifically, we train the model between \([t^* - T, t^*]\) with a fixed training data length \(T = 1092\) and then iteratively predict the
subsequent 25 months. This forecasting procedure is first carried out for \( t^* \) equal to January 1982, then repeated for February 1982, and so on until the end of the time series is reached.

As with all ANN models, various hyper-parameters of the ESNs have to be determined. The choice of the reservoir size \( (N) \), the sparsity \( (p) \), and the spectral radius \( (\rho) \) plays an important role for the efficiency of the learning process. In order to find optimal values for the hyper-parameters, we tested for different sets of parameters \( (N, p, \rho) \) (see 8.1). Regarding the reservoir size \( N \), it is generally expected that the larger the reservoir, the better the approximation of the underlying dynamics of the system, although this typically saturates. For our ESN we selected \( N = 600 \). Concerning sparsity, \( p \) determines the distribution of nonzero elements in the reservoir matrix \( W_r \). In this study, we chose \( p = 0.2 \). Another hyper-parameter that can affect the ESN performance is the spectral radius \( \rho \) that controls the echo state properties of the reservoir [37, 39, 40]. To set the spectral radius \( \rho \) of the reservoir matrix \( W_r \), its entries are multiplied with the fraction of the desired spectral radius which is the largest absolute eigenvalue of the matrix \( W_r \) [39]. We choose \( \rho = 0.95 \). We employ these choices for the hyper-parameters consistently to initialise the model to learn the dynamics and predict the target SST time series.

2.3. Past noise forecasting (PNF)

Generally the pivotal problem of forecasting stochastic time series, such as the ones arising from measuring natural systems like climate, is to understand the dynamics of unresolved, high-frequency variables, since they may have considerable influence on the prediction of the resolved, low-frequency variables. A possible solution strategy is to estimate the time-dependent high-frequency forcing from suitable samples from its past history. Chekroun et al [17] originally developed a particular prediction methodology called PNF to circumvent this problem. The purpose of PNF is to select the best sample of past stochastic forcing to drive the system into the future, using the knowledge of the past noise trajectory. To find such potential noise realizations from past parts of the time series, the stochastic forcing is conditioned on the LFC of the underlying system. Here, we use BW low-pass filter to decompose the time series into an LFC and the corresponding high-frequency part. We then split the LFC into different segments of length \( \Delta \) and search for those segments that resemble the reference LFC segment just preceding time step \( t^* \), at which we initiate the forecast:

\[
\text{RMSE}(\text{LFC}(t_j, t_j + \Delta) - \text{LFC}(t^* - \Delta, t^*)) \leq \alpha
\]

(5)

\[
\text{CORR}(\text{LFC}(t_j, t_j + \Delta) - \text{LFC}(t^* - \Delta, t^*)) \geq \gamma.
\]

(6)

With these selection criteria, one can then identify high-frequency components starting at different \( t_j < t^* - \Delta \), which we then use to force the ESN prediction for times \( t > t^* \).

2.4. ESN implementation and combination with PNF

We train the ESN based on 80% of the original dataset (the training interval thus has a length of 1092 months) prior to \( t^* \), the time that we aim to commence the prediction task (see figure S5). Using a BW low-pass filter, we decompose the training data into low- and high-frequency components, where the latter is treated as noise forcing. We then pass both components as input to the ESN. At each time, we perturb the LFC at time \( t \) by noise at time \( t + 1 \). During the training procedure, the weights of the output matrix \( W_{out} \) are optimized by minimizing the loss function using ridge regression [38]. Using the optimized \( W_{out} \), the trained model starts recursively forecasting 25 consecutive months ahead of \( t^* \).

In the prediction phase, since we do not have any information about the future behavior of the noise, we employ the past-noise forecasting (PNF) method to estimate past noise segments suitable to force the system into the future. As described in section 2.3, first we find those LFCs \( (t_j, t_j + \Delta) \) that resemble the LFC \( (t^* - \Delta, t^*) \) with \( \Delta = 6 \). We call the two LFC segments analogous if \( \alpha \leq 0.5 \) and \( \gamma \geq 0.95 \). In the next step, we select the corresponding high-frequency variability components of size \( T = 25 \). The prediction for the first time step is then, together with corresponding candidate noise analogues, passed to the ESN as inputs to predict the second time step, and so on. We repeat this training procedure for every month \( t_s \) starting from January 1891 with \( r \in \{0, ..., 432\} \) and predict 25 months ahead for each training set. After completing the prediction task for all training sets, we obtain the predicted time series at different lead times ranging from 1 to 25 months running from 1982 to 2019. For every single iteration, the predicted values are not seen by the model during the training phase.

Note that the skill of our ESN-based prediction scheme is slightly affected by varying the cutoff value \( C \) of the filter used to separate the low- from the high-frequency components of the ENSO index. Our results are
Figure 2. Summary of the ESN forecast skill for the Niño-3 and Niño-3.4 indices. Panels (a) and (b) show the prediction skill of the model in terms of the root mean square error (RMSE) and the Pearson correlation coefficient (PCC) between predicted and observed values, respectively. The model is trained with a fixed training data length $T = 1092$ and the prediction task starts from 1982. Results for Niño-3 are shown in blue and results for Niño-3.4 in red color. In panel (b) the magenta curve shows the ENSO correlation skill of a previously introduced CNN together with a comparison to other process-based based ENSO predictions. The ability of our ESN–PNF model to predict El Niño events (i.e. times with Niño-3 index above 1 °C) are assessed by two binary classifiers, the Heidke skill score (HSS) and the probability of detection (POD), in (c) and (d), respectively. Solid lines indicate the average over 100 realizations of the forecast, and shading around the lines represent ±1 σ.

Presented here for the optimal choice in terms of the overall forecast skill based on the above metrics ($C = 0.03$), and compared to alternative choices in (figure S13).

3. Results

Our ESN implementation consists of two main steps; the training phase during which a loss function is minimized to find optimal output weights for the ESN model, and the prediction phase during which the optimized ESN is used to predict unseen data (see figure S5).

We focus on forecasting the Niño-3 and Niño-3.4 (figure 2) indices during the time span from 1982 to 2019. To evaluate the skill of our ESN–PNF approach, we consider four different metrics, namely (I) the root mean square error (RMSE) and (II) the Pearson correlation coefficient (PCC) to evaluate the overall performance in predicting the ENSO index, as well as (III) the Heidke-skill score (HSS) and (IV) the probability of detection (POD) to evaluate the binary prediction of El Niño events (see Appendix A). For Niño-3 the RMSE remains almost constant at around 0.5 °C for lead times between one and 14 months, and then increases linearly up to around 1.4 °C at 24 months lead time (figure 2(a)). The PCC between the observed and simulated time series for the prediction phase remains constant around 0.8 also up to lead times of around 14 months and drops below 0.5 after 18 months (figure 2(b)). Note that, following [12] we consider forecasts as skillful if the correlation coefficients is above 0.5. To assess the skill of our model in predicting El Niño events (i.e. time steps with ENSO index larger than one), we applied the HSS (figure 2(c)) and the POD (figure 2(d)), cf appendix for details on these scores. Both event-based metrics also indicate high skill up to forecast lead times of 18 months. For the prediction phase starting at 1992, we obtain even a much higher skill during the first 14 months (figures S6 and S7). Our results remain similar when 5 month running averages of the Niño-3 and Niño-3.4 indices are considered, and when El Niño events are defined as time spans for which these smoothed indices are above 0.5 °C for six consecutive months (see figure S8).

Note that the prediction score of our ESN–PNF model varies moderately over time, depending on the absolute values of the employed ENSO index. For example, we observe higher RMSEs during the 1997–98 and 2015–2016 El Niño events (figure S9). Hence, for El Niño and La Niña events with particularly high or
Figure 3. Correlation skill of the ESN–PNF model as function of lead time and target month for the Niño-3 index. The Pearson correlation coefficient (PCC) between the predicted and observed Niño-3 index targeted to each calendar month at different lead times, averaged over 100 realizations. Here we train the model on a fixed training data length $T = 1092$ and start forecasting from 1982. See (figure S10) for the Niño-3.4 index.

low values of the ENSO index, the RMSE is comparably higher, but the binary prediction of these events is still successful.

In order to further assess the reliability of our ENSO forecast, we investigate the sensitivity of the model’s forecast skill to the target months at different lags (figures 3 and S10). In agreement with previous studies [12], our ESN–PNF model exhibits the longest forecast horizon for target months in boreal winter (the correlation between the observed and simulated ENSO index is above 0.5 for 21 months) and the shortest forecast horizon (18 months) for target months in late boreal spring. This spring predictability barrier [41] may be a result of the comparably weak Walker circulation and susceptibility of the coupled ocean–atmosphere system to the external forcing at that time of the year. Nevertheless, a recent CNN model has a valid forecast horizon (i.e. lead times at which the correlation remains above 0.5) that is substantially longer than for process-based dynamical models (11 months vs. 4 months) for the May–June–July season.

For the May–June–July season, the valid forecast horizon of our ESN–PNF model is even substantially longer compared to the CNN model [12] (18 months vs 11 months). This indicates that our method is less affected by the spring predictability barrier compared to the CNN model and substantially less affected than process-based models, possibly because we estimate the future fast variables that drive the system using the PNF method.

As an example, the predicted SST anomaly between 1983 and 2019, at a lead time of 17 months, shows high visual resemblance to the observational Niño-3 index (figure 4(a)), although this is the longest lead time for which we consider our forecast to exhibit skill. The predicted values are computed as averages over 100 realizations at 17 month lead. The performance of our model is also examined by comparing the statistical properties such as the autocorrelation function (ACF, figure 4(b)) and probability density functions (PDF, figure 4(c)) of the predicted and observational time series. The ACF shows very close resemblance between the observed and simulated indices in terms of their correlation structure, and according to the Kolmogorov–Smirnov test, the hypothesis that the underlying distributions of the observed and simulated indices are identical cannot be rejected ($p = 0.35$). Corresponding results for the Niño-3.4 are very similar (figure S11).

4. Summary and discussion

There have been considerable advances in machine learning-based approaches to predict ENSO variability. One of the critical obstacles that most deep learning models encounter in predicting climate phenomena such as ENSO is the unavailability of sufficiently long observational time series, hindering training of these models purely on observations. In contrast to deep ANNs, the comparably simple ESNs can learn the dynamics of the underlying system from comparably small amounts of training data. Here, we expanded the ESN approach to predict time series with high noise levels, as they may arise from real-world measurements and experiments. We specifically applied our ESN model to predict the SST-based Niño-3 and Niño-3.4 indices. We decomposed the considered ENSO indices into dominant low- and high-frequency variability by using low-pass filter techniques and trained the ESN model on the slow mode of the system. To model the effect of the high-frequency forcing on the low-frequency variability, we estimated the potential future
high-frequency forcing of the system by relying on PNF method, which substantially enhanced the skill of our ESN–PNF model in predicting El Niño events. This improved predictability shows that interactions across multiple time scales indeed play a crucial role in generating the dynamics of ENSO. At the same time, the long valid time of our statistical prediction, with horizons up to 21 months, reveals that ENSO is in principle predictable for much longer lead times than previously thought, despite the fact that we have trained our ESN–PNF model only on single scalar ENSO indices.

Several studies demonstrated that using preprocessing techniques such as low-pass filtering can enhance forecasting skill of ANNs by providing high-quality inputs [42–44]. In practice, however, filtering methods such as SSA, BW filters, or MA-based methods incorporate information from future values to determine the LFC. As for all approaches involving filtering prior to the predictive modelling, our approach can therefore not directly turn into a real-time forecasting of ENSO variability. Of course, the filtering can in principle be applied solely to the training part, but in that case well-known edge effects, i.e. errors in estimating the LFC arising at the beginning and end of the time series, would hinder a skillful prediction. We outline improved methods to reduce such errors when estimating low-frequency time series components as important subject for future research, which would also improve the skill of the method proposed here in an operational real-time forecasting setting.

While traditional modelling approaches are often challenged by cross-scale interactions, our approach in fact exploits the dependencies between low- and high-frequency variability for a more accurate modelling and statistical prediction. Our methods should therefore also prove valuable in other fields where cross-scale interactions are relevant, such as Neuroscience or Finance. Our approach indeed outperforms existing statistical, process-based, and deep-learning based approaches at lead times beyond 1 year, with the caveat that for the reasons explained above, we have not made our predictions operational. Nevertheless, our results reveal remarkable long-term predictability of ENSO far beyond the spring barrier [45]. For future studies, it might be of interest to combine process-based models with machine learning models such as the one presented here, to combine the strengths of both approaches.
Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Mickael Checkroun for helpful discussions. This work has been financially supported by the German Academic Exchange Service (DAAD) under Project No. 57299294. N B acknowledges funding from the VolkswagenStiftung, the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No. 820970 and under the Marie Sklodowska-Curie Grant Agreement No. 956170, as well as from the Federal Ministry of Education and Research under Grant No. 01LS2001A.

Authors contribution

FH and NB conceived and designed the study. FH performed the numerical analysis and wrote the manuscript together with NB. All authors discussed results and contributed to editing the manuscript.

Appendix A. Scores for forecast validation

A.1. RMSE and correlation coefficient

To quantify the skill of our time series forecast, we use two well-established metrics, namely the root-mean-squared error (RMSE) and the Pearson’s correlation coefficient (PCC). RMSE measures the average error magnitude with quadratic weight, which gives a quantitative evaluation of the closeness of the predicted variables $x^p$ compared to the reference variables $x^r$ (see figures 2(a), S9 and S13(a)).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x^r_i - x^p_i)^2}. \quad (S1)$$

PCC measures the linear relationship between the original data and corresponding forecasts. PCC can take values between $-1$ and 1, where the positive and negative signs correspond to the direction of the linear relationship.

$$\text{PCC} = \frac{\sum_{i=1}^{n} (x^r_i - \bar{x}^r)(x^p_i - \bar{x}^p)}{\sigma_{x^r} \cdot \sigma_{x^p}}. \quad (S2)$$

where $\sigma_{x^r}$ and $\sigma_{x^p}$ are the standard deviation of the reference and predicted variables, respectively.

A.2. Heidke skill score (HSS)

To verify the skill of our model for the binary forecast of El Niño events (i.e. months with ENSO index larger than 1°C), we employ the HSS [46], which is widely applied in binary classification problems. This metric is a measure of accuracy of a forecast with respect to a randomly generated forecast, adjusted to predictions that are correct by chance:

$$\text{HSS} = \frac{TP + TN - CRF}{N - CRF}. \quad (S3)$$

where TP and TN stand for true positives and true negatives, respectively, $N$ is the total number of possible events, and CRF indicates the number of correct random forecasts, which can be calculated as follows:

$$\text{CRF} = \frac{(TP + FN)(TP + FP) + (TN + FN)(TN + FP)}{N}. \quad (S4)$$

Negative HSS values imply that the forecast skill of the model is worse than a random forecast, HSS = 0 indicates that the forecast is just as good as the random forecast, and HSS = 1 would indicate a perfect binary forecast. We also consider the probability of detection (POD) as a second metric to evaluate our prediction of El Niño events, which as defined as

$$\text{POD} = \frac{TP}{TP + FN}. \quad (S5)$$
Figure S1. Echo state network performance with variations of reservoir size and spectral radius at fixed sparsity = 0.2.

Figure S2. Comparison between the statistical properties of the predicted and observed Niño-3 indices. (a) Time series of SST anomalies of the Niño 3 index (red) and the predicted time series for ($\rho = 0.95$, sparsity = 0.2) (magenta), and ($\rho = 0.7$, sparsity = 0.2) (green), at 15 months lead time. (b) and (c) compare the statistics (ACF and PDF) of the original and predicted time series.

Appendix B. Supplementary figures

B.1. Hyper-parameters of the ESN–PNF
The key hyper-parameter that determines the echo state property of the ESN model is spectral radius ($\rho$). Figure S1 displays the performance of the ESN–PNF in terms of RMSE for different choices of reservoir size $N$ and $\rho$ values. Even though we observed better performance at $\rho = 0.7$ and $N = 600$, reducing the $\rho$ value from 0.95 to 0.7 decreases the capability of the ESN model to produce the statistical properties of the underlying dynamics. To illustrate that, we compared the ACF and PDF of the predicted time series and original Niño-3 index at a 15 month lead time for $\rho = 0.95$ (magenta) and $\rho = 0.7$ (green) (see figure S2). According to our results, the reservoir scaled by $\rho = 0.95$ exhibits better performance in predicting the target signal.
Figure S3. Echo state network performance with variations of reservoirs size with different sparsity.

Figure S4. Comparison between the predicted and observed Niño-3 indices. Panels (a) and (b) show the prediction skill of the model in terms of the RMSE and the PCC between predicted and observed values for sparsity $= 0.02$ (blue), and sparsity $= 0.2$ (magenta) with $\rho = 0.95$. Panels (c) and (d) compare the statistics (ACF and PDF) of the original and predicted time series, respectively.

Further, we examined the role of reservoir size and connectivity on the model’s performance (see figure S3) at the fixed $\rho = 0.95$. Our results demonstrate that increasing the reservoir size enhances the prediction performance of the model. We also observed that the model is not much sensitive to the sparsity of the reservoir. To have a better insight, we also compared the model’s outputs with sparsity setting equal $p = 0.2$ and $p = 0.02$ in terms of different metrics (RMSE, PCC, ACF and PDF). As it can be observed from figures S4(a) and S4(b), the sparsity does not affect the performance of the ESN–PNF significantly.
Figure S5. Schematic diagram of echo-state network setup.

Figure S6. Summary of the ESN forecast skill for the Niño-3 index. Panels (a) and (b) respectively show the prediction skill of the model in terms of RMSE and PCC between predicted and observed values. The ability of the model to detect El Niño events (i.e. times with Niño-3 index above \(1 \degree C\)) are assessed by two binary classifiers, the Heidke skill score (HSS) and the probability of detection (POD), in (c) and (d), respectively. Solid lines indicate the average over 100 realizations of the forecast, and shading around lines represent \(\pm 1 \sigma\). The results correspond to the reservoir setup trained with a fixed training data length \(T = 1212\) and the prediction starts from 1992.
Figure S7. Dependency of the forecast skill on the training length. Panel (a) illustrates the RMSE and PCC values between the predicted and absolute values of Niño-3 index at a lead of 12 months with respect to different starting points. Panel (b) shows HSS and POD for the corresponding lead time. Our ESN model exhibits overall better performance the longer the training data. Shading around lines represents ±1 σ.

Figure S8. Detecting El Niño events for Niño-3 and Niño-3.4 indices with 5 month running mean average. The ability of the model to detect El Niño events (i.e., SST index above 0.5 °C) are assessed by two binary classifiers, the HSS and the POD. Panels (a) and (c) show HSS of the 5 month smoothed version of Niño-3 and Niño-3.4 indices respectively. The corresponding POD scores can be found in panels (b) and (d), where colored shadings around lines correspond to ±1 σ.
**Figure S9.** Dependence of the forecast skill on the target year for the Niño-3 index. (a) A 12 month running mean of the Niño-3 index from 1983 to 2018, here the training length is equal to 1092. (b) Forecast skill of our ESN model from 1983 to 2018 in terms of the RMSE. Each yearly value is the average RMSE over all months of the corresponding year, calculated on 12 month lead. We note that in some years with exceptionally strong El Niño events, the prediction accuracy of the model is affected slightly.

**Figure S10.** Same as (figure 3) but for Niño 3.4 index. The Pearson correlation coefficient (PCC) between the predicted and observed Niño-3.4 index targeted to each calendar month from January to December, at different lead times. Hatches indicate combinations target months and lead times for which the correlation of observed and predicted ENSO index is above 0.5. The prediction period is between 1984 and 2017.
Figure S11. Same as (figure 4) of the main text, but for the Niño 3.4 index. (a) Time series of SST anomalies of the Niño 3 index (red) and the predicted time series (blue) at 17 months lead time during 1984–2017. (b) and (c) compare the statistics (ACF and PDF) of the original time series and predicted time series; for the predictions averages over 100 realizations at 17 months lead time are taken.

Figure S12. Enhancement of the ESN model skill using past-noise forecasting (PNF) method. The magenta and red lines refer to the performance of the model trained on the original data (Niño-3 index) and smoothed data (using a Butterworth low-pass filter) respectively, and colored shadings correspond to ±1 σ. On the other hand, the blue line corresponds to a setup in which we tried to estimate the potential future high-frequency forcing of the system using PNF method. We used a fixed training data length $T = 1092$ to train the model, and then start prediction task from 1982.
Figure S13. Summary of the ESN forecast skill for Niño-3 index for different cutoff choices to decompose the index into slow and fast components. Panels (a) and (b) show the prediction skill of the model in terms of RMSE and PCC between predicted and observed values. The ability of the model to detect El Niño events (i.e. months with Niño-3 index above 1 °C) is assessed by two binary classifiers, the HSS and the POD, panels (c) and (d), respectively. The model is trained with a fixed training data length $T = 1092$ and the prediction task starts from 1982. Here the solid lines indicate the average over 100 different realizations and colored shadings around lines correspond to ±1 σ. Different colors, as indicated in the legend of panel (a), correspond to results for different cutoff thresholds to decompose the ENSO index. In order to determine the optimal cutoff value, we additionally investigated the dependency of the binary forecast skill of predicting El Niño events, using HSS and POD metrics. The ESN model exhibits overall best performance at a cutoff value of $C = 0.03$ (months $^{-1}$). Even though a cutoff at $C = 0.02$ (months $^{-1}$) displays a higher correlation skill for very long lead times, the forecast of El Niño events, and also the reproduction of statistical properties such as the ACF and PDF is worse in this case.

Figure S14. Difference between the trajectory of low-pass-filter of the training signal generated in two different settings. Here red color indicates a signal obtained from the low-pass filter applied on the entire original time series. Blue color manifests the same training set (with the same start and endpoints) created from the corresponding non-smoothed training subset with a length 1092 months.
Figure S15. Absolute average difference. Values illustrate differences between the last 12 points of the decomposed training signals generated in two settings, using (I) low pass filter of the entire original time series and (II) low pass filter of the selected period. Numbers on the x-axis are related to different starting points of the training set with a length of 1092 months.

Figure S16. Summary of the ESN forecast skill for the Niño-3 index. Here we split the whole time series into two subsets (i.e. the training and test set) and filter only the training set as the input of the ESN model. Panels (a) and (b) show the prediction skill of the model in terms of the RMSE and the PCC between predicted and observed values, respectively. The model is trained with a fixed training data length $T = 1092$, and the prediction task starts from 1982. Different colors indicate different low-pass filter cutoffs.
Figure S17. Correlation skill of the ESN trained on a subset preprocessed by low-pass filter, as a function of lead time and target month for the Niño-3 index. The PCC between the predicted and observed data targeted to each calendar month at different lead times. Here we train the model on a fixed training data length $T = 1092$ and start forecasting from 1982.

Figure S18. Comparison between the statistical properties of the predicted and observed Niño-3 indices. (a) Time series of SST anomalies of the Niño 3 index (red) and the predicted time series (blue) at 6 months lead time. The model is trained with a fixed training data length $T = 1092$ and the prediction task starts from 1982. (b) and (c) compare the statistics (ACF and PDF) of the original and predicted time series; for the predictions.

ORCID iDs

Forough Hassanibesheli  
https://orcid.org/0000-0001-6919-4358

Niklas Boers  
https://orcid.org/0000-0002-1239-9034

References
