Benchmarking Prediction Skill in Binary El Niño Forecasts

Xinjia Hu (✉ xinjia@pks.mpg.de)

Max-Planck-Institut für Physik komplexer Systeme <https://orcid.org/0000-0003-2030-0185>

Jan Eichner

Munich Re

Eberhard Faust

Munich Re

Holger Kantz

Max-Planck-Institute for the Physics of Complex Systems: Max-Planck-Institut für Physik komplexer Systeme

Research Article

Keywords: El Niño prediction, benchmark, climate network analysis, prediction skill

Posted Date: July 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-717862/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Climate Dynamics on September 7th, 2021. See the published version at <https://doi.org/10.1007/s00382-021-05950-2>.

Benchmarking prediction skill in binary El Niño forecasts

Xinjia Hu · Jan Eichner · Eberhard Faust · Holger Kantz

Received: date / Accepted: date

Abstract Reliable El Niño Southern Oscillation (ENSO) prediction at seasonal-to-interannual lead times would be critical for different stakeholders to conduct suitable management. In recent years, new methods combining climate network analysis with El Niño prediction claim that they can predict El Niño up to 1 year in advance by overcoming the spring barrier problem (SPB). Usually this kind of method develops an index representing the relationship between different nodes in El Niño related basins, and the index crossing a certain threshold is taken as the warning of an El Niño event in the next few months. How well the prediction performs should be measured in order to estimate any improvements. However, the amount of El Niño recordings in the available data is limited, therefore it is difficult to validate whether these methods are truly predictive or their success is merely a result of chance. We propose a benchmarking method by new surrogate data for a quantitative forecast validation for small data sets. We apply this method to a naïve prediction of El Niño events based on the Oscillation Niño Index (ONI) time series, where we build a data-based prediction scheme using the index series itself as input. In order to assess the network-based El Niño prediction method, we reproduce two different climate network-based forecasts and apply our method to compare the prediction skill of all these. Our benchmark shows that using the ONI itself as input to the forecast does not work for moderate lead times, while at least one of the two climate

network-based methods has predictive skill well above chance at lead times of about one year.

Keywords El Niño prediction · benchmark · climate network analysis · prediction skill

1 Introduction

The El Niño-Southern Oscillation (ENSO) is the dominant seasonal to inter-annual climate variation generated through coupled interaction between atmosphere and ocean in the tropical Pacific (Bjerknes, 1969). El Niño is the warm phase of ENSO and can exert its influence on large-scale changes in climate regionally as well as remotely through atmospheric teleconnections, affecting extreme weather events worldwide (Diaz et al., 2001; Alexander et al., 2002; McPhaden et al., 2006). There is ample evidence for the close relationship between El Niño and climate anomalies across the globe (Wallace et al., 1998; White et al., 2014; Luo and Lau, 2019). As El Niño evolves at a slower rate compared to the weather system and is one of the most predictable climate fluctuation on the earth (on lead times of up to several months) (Goddard et al., 2001; Wang et al., 2020b), it can be used in statistical prediction models to make seasonal forecasts which are unfeasible for dynamical prediction models. One example is the seasonal forecast of tropical cyclones (Wahiduzzaman et al., 2020). Therefore, a long-time lead and skillful prediction of El Niño is crucial and necessary for decision makers when estimating climate anomalies (Capotondi and Sardeshmukh, 2015; Wang et al., 2020a). During the past decades, there were numerous efforts to improve El Niño prediction (Latif et al., 1998; Mason and Mimmack, 2002; Y. and W., 2002; Luo et al., 2008; Clarke, 2014; Tao et al., 2020). The most

Xinjia Hu, Holger Kantz
Max Planck Institute for the Physics of Complex Systems,
Nöthnitzer Str. 38, D 01187 Dresden, Germany
E-mail: xinjia@pks.mpg.de

Jan Eichner, Eberhard Faust
Munich Reinsurance company, Königinstr. 107, 80802
München, Germany

severe restriction to long-lead time El Niño prediction is the so-called spring prediction barrier (SPB), referring to a rapid decrease in prediction skill in boreal spring (Zheng and Zhu, 2010; Duan and Wei, 2013; Duan and Hu, 2016). Prediction schemes including complex climate models and statistical approaches have relatively limited prediction power for long-lead (over 6 months) forecasting due to the SPB (Collins et al., 2002). In order to solve the SPB problem, statistical El Niño prediction approaches based on complex network analysis have been introduced, which extend El Niño prediction to 1 year in advance (Ludescher et al., 2013, 2014; Meng et al., 2018; Ludescher et al., 2019). Since climate fields such as temperature can be interpreted as a climate network (Yamasaki et al., 2008; Gozolchiani et al., 2011), predictions have benefited from network representations of El Niño-related climate variables. In the established network, different regions of the world are regarded as nodes, which interact with each other by exchanging heat and material, etc., and these interactions are the links of the whole climate network (Steinhaeuser et al., 2010). After the pioneering work of Ludescher (Ludescher et al., 2013), similar approaches were conducted aiming at further improvements of El Niño prediction using climate network (Fan et al., 2017, 2018; Meng et al., 2020). All of these predictions are formed in a similar way: If a prediction index crosses a certain threshold from below, it predicts that during some time interval in the future, there will be an El Niño event. More precisely: based on meteorological input data at many nodes, a scalar predictive index is calculated as an indication of link strength. If this index overcomes a threshold value with other conditions satisfied, the El Niño index is expected to become positive within a certain time window. If so, this is counted as a successful prediction, otherwise, it is a false alarm. If a positive ENSO phase occurs without prior threshold crossing of the predictive index within the given time window, this is regarded as a missed hit. The differences between these methods lie in the choice of input climate variables, area of nodes for establishing network, and how link strength is measured, which one could call “technical details”, but also might be decisive for prediction skill. According to the reported hit rates and false alarm rates, these methods are very promising and seem to have solved the SPB problem.

However, due to the length of the time window during which the positive ENSO phase might occur after the threshold crossing of the predictive index, and the quasi-periodic occurrence of ENSO itself (Gauchere, 2010; Zhang and Zhao, 2015), it is not easy to see whether these predictions have real predictive skill or are just a consequence of the rather generous interpreta-

tion of success together with the regularity of the ENSO phenomenon. In the original publications, the authors provided performance measures to evaluate prediction skill, for example, by reporting the receiver operating characteristic (ROC) curve or splitting their data into training and test sets (Ludescher et al., 2013). However, there is no common performance measurement to compare these works, and in particular no comparison to a benchmark predictor. Besides, it is not evident how to provide a benchmark prediction in these studies. Therefore, in this paper, we develop a benchmarking method that can estimate and compare prediction skills, aiming at providing a reference for improving El Niño predictions in the future.

2 Existing climate network-based ENSO prediction schemes

In this paper, we take two existing climate network-based El Niño prediction models from separate studies as examples and create benchmarks to estimate and compare their predictability. One scheme is from Ludescher’s pioneering work (Ludescher et al., 2014) and the other one is from Meng’s work (Meng et al., 2018). Both papers develop a prediction index based on climate network analysis and give an El Niño warning for the next year or next 18 months when the index crosses a given threshold from below. However, due to relatively complicated conversion of the generated predictive index series into the forecasts of events, it is not perfectly transparent how much of the performance is due to skill of the predictive index, and how much is due to suppressing spurious forecasts by the conversion into event prediction.

In order to make our paper self-contained, we describe in this section in some detail previous work. Both studies define an index to represent the degree of cooperative resp. disorder between nodes in certain basins. This index is based on the time evolution of the link strength between the nodes i and j . In Ludescher’s work, links are established between nodes of two basins: the El Niño basin (i.e., the equatorial Pacific corridor) and the rest of the Pacific Ocean. In the other study, only nodes from the Niño3.4 area (i.e., 5°S–5°N, 120°W–170°W) are used (see Fig. 1). The procedures to calculate these finite time-window cross-correlations between different nodes are identical in both studies, as the time-delayed cross-covariance function is defined:

$$C_{i,j}^{(t)}(-\tau) = \frac{\langle T_i(t)T_j(t-\tau) \rangle - \langle T_i(t) \rangle \langle T_j(t-\tau) \rangle}{\sqrt{\langle (T_i(t) - \langle T_i(t) \rangle)^2 \rangle} \sqrt{\langle (T_j(t-\tau) - \langle T_j(t-\tau) \rangle)^2 \rangle}} \quad (1)$$

$$C_{i,j}^{(t)}(\tau) = \frac{\langle T_i(t-\tau)T_j(t) \rangle - \langle T_i(t-\tau) \rangle \langle T_j(t) \rangle}{\sqrt{\langle (T_i(t-\tau) - \langle T_i(t-\tau) \rangle)^2 \rangle} \sqrt{\langle (T_j(t) - \langle T_j(t) \rangle)^2 \rangle}} \quad (2)$$

This time-delayed cross-covariance is calculated for each 10th day t during the whole time span in Ludescher's work. $T_k(t)$ in the formula [1] and [2] is the daily atmospheric temperature anomaly at node k . In Ludescher's work, daily surface atmospheric temperatures (1950-2013) are used, which are divided into two parts, i.e. the learning phase (1950.01-1980.12) and the prediction phase (1981.01-2013.12). The data is obtained from the National Centers for Environmental Prediction/National Center for Atmospheric Research Reanalysis I project. They use Niño 3.4 index as the El Niño index. For Meng's study, they calculate this $C_{i,j}^{(t)}$ for each month t (the first day where the months starts) in the selected time span (1981.01-2017.12) and the daily mean near surface (1000hPa) air temperature fields are applied for analyzing the links between nodes in the Niño3.4 basin. In Meng's work, they used 4 different data-sets separately to establish this link strength. In our paper, we only take the dataset which gives the best result in their paper, i.e. the daily near surface (1000hPa) air temperature (1979-2018) from the ERA-Interim re-analysis. They use the Oceanic Niño Index (ONI) to represent El Niño's variation (obtained from https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php). In formula [1] and [2], time lag τ is between 0 and 200 days in both models, which can guarantee a reliable estimate of the background noise level (Ludescher et al., 2013; Meng et al., 2018). The brackets denote an average over the past 365 days, according to

$$\langle f(t) \rangle = \frac{1}{365} \sum_{a=1}^{365} f(t-a) \quad (3)$$

The two studies then convert the set of these $C_{i,j}^{(t)}(\tau)$ into an index representing the link strength in two different ways. Ludescher et. al define the link strength $S_{ij}(t)$ as the difference between the maximum and the mean value of the cross-correlations $C_{i,j}^{(t)}(\tau)$ in the time window, divided by its standard deviation as:

$$S_{ij}(t) = \frac{\max(C_{i,j}^{(t)}(\tau)) - \text{mean}(C_{i,j}^{(t)}(\tau))}{\text{Std}(C_{i,j}^{(t)}(\tau))} \quad (4)$$

The average of $S_{ij}(t)$ over all pairs of nodes (i, j) in the whole network forms the index $S(t)$, and its threshold crossing from below is interpreted as a precursor of

an El Niño event. The optimal value for this threshold is determined by optimizing the prediction skill on the first half of data (training set, years 1950-1980, as mentioned above). The performance is then tested in the second part of the data (test set, 1981-2013), which is called the prediction phase, and the values of thresholds are further modified to adjust both parts of the data. Prediction skill is evaluated by the hit rate (HR) and false alarm rate (FAR).

In Meng's study, a forecasting index (FI) is proposed which presents the coherence between the different nodes in the Niño3.4 area. $C_{i,j}^{(t)}(\theta)$ is the highest peak value of the cross-correlation function, where θ indicates the time lag when the function comes to the highest (Meng et al., 2018). The degree of coherence/disorder of the nodes in the Niño 3.4 basin is qualified by the average value of all links at their peak:

$$C(t) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N C_{i,j}^{(t)}(\theta) \quad (5)$$

where N is the number of nodes in the Niño 3.4 basin, $N = 105$. Therefore, higher values of $C(t)$ represent higher coherence in the Niño 3.4 basin. The forecasting index FI is defined as:

$$FI(t) = \frac{1}{m+1} \sum_{a=0}^m \ln(C(t-a)) - \ln(C(t)) \quad (6)$$

where m is the total number of months preceding t since the first month of the whole data (i.e. January 1981), so that equation [6] denotes the negative anomaly of $\ln C(t)$. This minus sign was used so that peaks of FI correspond to peaks in the ONI. FI is evaluated from January 1984 for each month (Meng et al., 2018). The threshold set for this model is 0.

Actually, the rules for making predictions are slightly more involved than just the threshold crossing from below. In Ludescher's work, it is defined that when the link strength $S(t)$ crosses the threshold from below and simultaneously the Niño 3.4 index is below 0.5°C , a warning will be given and an El Niño is expected to start in the following calendar year. In Meng's work, one needs to track both ONI and FI from the beginning of the previous El Niño. If $FI(t)$ rises from a local minimum for at least two continuous months, the time at which $FI(t)$ crosses 0 (if it is not during El Niño/La Niña period, i.e. $-0.5^\circ\text{C} < \text{ONI} < 0.5^\circ\text{C}$) is counted as a potential warning of either El Niño or La Niña event within approximately the next 18 months. Additionally, if there occurs a La Niña event within these 18 months, another El Niño is predicted to occur within

Fig. 1: The climate network area of the two studies which we compared. The network of Ludescher’s work consists of 14 grid points in the El Niño basin (red symbols, including the NINO1, NINO2, NINO3, and NINO3.4 regions and one grid point south of the NINO3.4 region) and 193 grid points outside of this basin (blue symbols). The network of Meng’s study consists of 105 grid points in the NINO3.4 region (yellow symbols)

18 months after the end of La Niña. Following these, a true hit of El Niño is counted if within 18 months after the potential signal an El Niño occurs or a La Niña followed by an El Niño within the next 18 months occurs, otherwise, it is counted as a false alarm.

3 Benchmarking methodology

Generally, prediction skill can be estimated by comparing the proximity of both the prediction and a benchmark to the observations (Pappenberger et al., 2015). A forecast can only be regarded as skillful if it has more prediction power than benchmark forecasts (Jolliffe and Stephenson, 2008). Common benchmarks contain naïve forecasts, namely based on either climatology or on persistence (Cattell, 1991). As a benchmark, climatology is useful for phenomena with strong seasonality, while prediction based on persistence works for short range forecast (Harrigan et al., 2018). A forecast scheme based on persistence often shows good predictability for short lead times and the prediction skill breaks down afterwards (Chiera et al., 2009).

El Niño predictions can be roughly divided into 2 types. One is the forecast of SST anomalies or of the index in the Niño basins, spanning the westernmost Niño-4 area to the easternmost Niño1+2 region, in which the Niño indices or results from other models are often used as benchmarks to verify El Niño prediction skill (Jin et al., 2008; Barnston et al., 2015; Tseng et al., 2017). The other one is as what we reported above, namely the two-category outlooks (above or below threshold), acting as a categorical (or binary) forecast for telling

whether there will be an El Niño event or not (Fan et al., 2018; Meng et al., 2020). These climate network-based El Niño predictors are data based, which means that they do not use a physical model of the process but instead make use of correlations and dependencies between different data sets. For such type of forecasts, a classical way is to use surrogate data as a benchmark: one would generate artificial random time series of the predicative index, which share essential statistical properties with the true predictive index series. Conventionally, these properties are the auto-correlation function (or power spectrum), and the marginal distribution (Schreiber and Schmitz, 2000). These surrogate data are otherwise random, thus their predictive skill will provide a benchmark compared with any truly skillful predictor. However, for short data sets, the generation of appropriate surrogates is error prone, so that this benchmark is hard to obtain. In Ludescher’s work (Ludescher et al., 2013), instead, a randomization of blocks of annual data was used which destroys correlations on time scales of more than one year. We propose here a modification of the method of surrogate data as a quantitative assessment of the predictive skill for statistical data based El Niño prediction.

3.1 Typical patterns in ONI time series

Before introducing our surrogates for benchmarking, we firstly extract typical patterns in the Oceanic

Niño Index (ONI) time series from 1950 to 2019 denoted as x in formula [7].

$$E(x(\sigma)|\alpha) = \frac{1}{n} \sum_{t:|x_t-\alpha|<\epsilon} x_{t+\sigma} \quad (7)$$

where α is the value of the condition, ϵ is some tolerance which controls how rarely or how frequently the condition will be satisfied by the data, n is the number of time indices t fulfilling the condition and σ is the lead time for forecasting. Whenever $x_t \in [\alpha - \epsilon, \alpha + \epsilon]$, $x_{t+\sigma}$ contributes to the corresponding conditional value. Hence, we calculate the climatological future of x_t fulfilling the condition. One can also interpret those values $x_{t+\sigma}$ where x_t satisfies the condition as an empirical ensemble, and $E(x(\sigma)|\alpha)$ is the ensemble mean.

A simple El Niño forecast can be achieved by applying this conditional average method, with a slightly more complicated condition: in addition to $x_t \in [\alpha - \epsilon, \alpha + \epsilon]$, $x_t - x_{t-1} > \delta$ should be ensured, i.e., that the signal x_t is rising with a slope larger than δ when crossing the threshold α . One can also interpret this in the sense of the analogues method, see Lorenz (Lorenz, 1969). Fig. 2 (left) shows the conditional expectations of the ONI index over lead time σ for different threshold values α . Negative σ are also considered, i.e., the average signal before it satisfies the condition. The pattern as a whole shows that before and after an El Niño event, the ONI index has the tendency to be negative, which is in agreement with the fact that successive El Niño events are typically separated by several years.

We choose the tolerance $\epsilon=0.3$ in order to find sufficiently many segments of ONI(t) which match the condition. δ is set to be 0.3 to ensure that the time series is increasing within the tolerance of certain x_t . We take 6 different condition values, $\alpha = 0.6, 0.3, 0, -0.3, -0.6$. According to Fig. 2 (left), the conditional average signals of ONI with different α all grow and reach a maximum after about 8 months, which indicates that on average an El Niño is expected to reach a peak after these 8 months. For $x_t = 0.6 \pm \epsilon$ (i.e. $\alpha = 0.6$), it is already during the onset of an El Niño, hence the whole result is not surprising but merely a proof of what we already know. When we lower α to 0, still the conditional average shows the signature of an El Niño, which actually is so merely because in addition we require the slope of the ONI to be positive. So this threshold crossing with a positive slope can be used as an El Niño forecast, where the onset occurs on average already 1-2 months later (at $\sigma = 1$ or 2), and the peak occurs around the month 8. Hence, if we use the analogue forecast to predict El Niño based on the ONI itself, the lead time is much less than 1 year. This simple prediction can be

used as a benchmark to the climate network-based El Niño prediction.

3.2 Conditional average of ONI given by network-based El Niño prediction models

We use the method of conditional averages now to examine the network-based forecasts as well to make a comparison of both models. The statement in Ludescher's (Ludescher et al., 2014) and Meng's work (Meng et al., 2018) is that if their respective predictive index (built on external inputs) crosses a certain threshold from below and not during certain ENSO period, then there will be an El Niño event in the future (where "future" is slightly different in the two forecast schemes). We therefore calculate the conditional expectation value of real ONI, where the condition is that the predictive index of either forecast method crosses its respective threshold from below.

The result is shown in Fig. 2 (right). According to Fig. 2 (right), for the predictions given by the index $S(t)$ of Ludescher et al., El Niño onset (i.e., $\text{ONI}(t) > 0.5$, crossing blue dashed line) occurs around 12 months after the warnings, while from Meng's index, this prediction lead time is shorter (i.e., the time when the red dashed line crosses the blue dashed line). Additionally, the peak value of the conditional average of ONI obtained from Meng's prediction index is much smaller than that from Ludescher's work, indicating that the response of ONI to the threshold crossings of FI is less specific or the prediction lead time is less stable. This difference can be partly explained by Fig. 3, in which we present the temporal distribution of the real El Niño onsets after a warning is given by these two models. In Fig. 3, month 0 on the vertical axis indicates again that the corresponding predictive index crosses its threshold from below with extra conditions satisfied, this is the time of the given warning. The color bar represents the value of the observed ONI in the following 24 months, where the color changing to reddish indicates the ONI values > 0.5 (El Niño starts). The x axis indicates each warning given by these two models (For Ludescher's work, the prediction period is between 1981-2013, during which there are 8 warnings issued in total; for Meng's work, the prediction period covers 1984-2018, during which there gives 11 warnings, 9 are correct hits). For example, the warning 1 in the left panel, real El Niño starts 15 months after this warning (color turns to reddish), while warning 3 is launched 6 months before the El Niño starts. From Fig. 3, we can see that El Niño occurs at least 6 months later when a warning is given by Ludescher's model (reddish appears after month 6). However, the lead time of Meng's model

Fig. 2: Conditional average of ONI (If ONI reaches $\alpha \pm \epsilon$ at any time t , the conditional ONI calculates the average of all ONI values after a time σ (i.e. conditioned on the value of α). Trivially, conditional ONI at $\sigma = 0$ is identical to α (within $\pm \epsilon$); left: based on ONI index itself with $\epsilon = 0.3$, $\delta = 0.3$; right: based on two climate network-based El Niño prediction models; at $\sigma = 0$ the predictive index crosses its respective threshold)

is less stable, some is very long to 2 years (for example, warning 1 in the right panel of Fig. 3), while some even smaller than 4 months (for example, warning 7).

Fig. 2 and Fig. 3 suggest that the prediction lead time for El Niño forecast is more stable in Ludescher's model, because of a more homogeneous El Niño response. The authors of both papers claimed that their climate network-based method would extend El Niño prediction lead time so that the spring barrier problem might be overcome. We verify by Fig. 2 and Fig. 3 that the onset of El Niño starts on average more than 6 months after the respective warning. However, the lead time for prediction is not stable, especially in Meng's model.

3.3 Compare two prediction models in a relatively fair way

In the original papers, the prediction performance was quantified by considering the hit rate HR and false alarms rate FAR under some additional rules, which convert the threshold crossing into a warning. Fig. 3 illustrates that we need additional benchmarks in order to fully understand the power of these two forecasts schemes. In Sec. 4.1 we will introduce a new benchmark method for comparing these two models in their original spirit, which is the main methodological contribution of this paper. Here, however, we compare these two models by a warning scheme for both of them in order to assess their accuracy under the same lead time, and which model can give longer lead times under the same accu-

racy. For that we reproduce these two forecast models (see Fig. 4), i.e., we calculate their respective predictive indices as time series. Our reproduction result is very close to the original paper, although there exist some small differences, due to some technical details not mentioned in the original studies.

The lead time for predictions in both studies are different and also not very exact. For example, Ludescher et al. defined that an El Niño will develop in the following calendar year after their prediction model gives an alarm warning; while the lead time for Meng's prediction model is over 18 months. Here, we study the performance for the same lead times for both models, so that we can compare them under a relatively same standard. For a warning, we still use the same conditions as defined in the two original studies, but we count separately the number of hits in the lead time interval of 0-6 months, and in an interval of 6-12 months. In Fig. 5, month 0 on the y -axis denotes when the prediction model launches a warning. We calculate the accuracy, namely the hit rate HR (the number of correct hits / the number of warnings), for both models in these lead time intervals. Fig. 6 (left panel) shows that Meng's model gives a better hit rate on the 0-6 months interval, while Ludescher's is superior on the 6-12 months lead time. We do this in better time resolution in Fig. 6 (right panel): For every month after the warning, we count in how many cases there is an ongoing El Niño event in this month. We see that around month 15 after the warnings by Ludescher's model, there are close to 80% of ongoing El Niños. So we find that as claimed

Fig. 3: Temporal distribution of predicted El Niño onset given by climate network-based El Niño prediction models and the following real ONI (left: Ludescher's model; right: Meng's model)

Fig. 4: Reproduction of the predictive indices from two cited papers and their 2-year shifted version (left: Ludescher's model; right: Meng's model). The green shades under the Niño index curve are the El Niño period. True forecasts are marked by the green lines; false alarms are marked in the orange dashed lines)

by the authors the onset of El Niño is at earliest a few months after the warnings of both models. Ludescher's prediction shows less spread of the timing of El Niño around its peak and at peak has a higher hit rate, so that we conclude that it indeed serves as a predictor on the annual time scale.

4 A new benchmarking method for El Niño predictions

4.1 Circular time shifted sequence as benchmark

Our analysis has revealed that both climate network-based prediction models can forecast an El Niño event at larger lead times than the simple prediction scheme based on the analogues of ONI values (our benchmark of Sec.3.1). Here we propose a new benchmark method for phenomena which like ENSO show oscillations on some typical time scales but are not periodic, where data sets are small and the actual prediction is a complicated function of the inputs. For the binary predica-

Fig. 5: Temporal distribution of predicted El Niño onset at the exact lead time given by climate network-based El Niño prediction model and the following real ONI (left: Ludescher's model; right: Meng's model. The number inside the bars indicate the true ONI values.)

Fig. 6: Comparison of two models under the same standard (left: the accuracy of both models under the same prediction lead time; right: temporal distribution of accuracy at different prediction lead time)

tion of El Niño (i.e., a yes/no prediction), a benchmark would be a random predictor. If the true predictor gives n warnings, we could generate a time series which contains exactly n level crossings at random time instances. One way to achieve this is to cut the time series of $S(t)$ or $FI(t)$ into segments of 12 months and to randomly re-arrange them (shuffling), although this would introduce a large number of jumps in the new sequences at the cuts. However, it will not be surprising if this shuffling predictor is much less skillful than the original one, since, e.g., two warnings might follow each other with a very short time interval in between, or there might be a gap of many years, while the original predictor has a certain regularity. Hence, we need a method which decouples the time series of predictions from the time series of the true signal, but which conserves relevant properties of the predictions such as modes of oscillation, locking with the seasons, or smoothness.

The El Niño prediction to which we refer contains only 19 El Niño events during the period from 1950 to 2013 (the period used in Ludescher’s pioneering work (Ludescher et al., 2013, 2014)). Moreover, these prediction methods use a time window in which the predicted event should occur. A useful benchmark is required to meet some criteria (Cattell, 1991): the first one is relevance, which indicates the ability to capture the characteristic of the original problem. Portability and scalability are also included, which means the benchmark can be used for testing different systems and systems of different sizes. Besides, the benchmark needs to be comprehensible and therefore credible (Vihinen, 2012). Following these criteria, we here introduce a new approach of creating benchmarks for predictions by the means of shifting the predictive index series in time, relative to the time series. It means that part of the series is shifted into the future and the most recent prediction is wrapped around and will cover the initial time window, where due to the time shift no predictive index is available. So in formula, if the original predictive index is $I(t)$ and is given on the time interval $t \in [t_0, t_1]$, the shifted index reads:

$$\hat{I}(t) = \begin{cases} I(t - \delta t) & t \in [t_0 + \delta t, t_1] \\ I(t_1 - \delta t + t) & t \in [t_0, t_0 + \delta t] \end{cases} \quad (8)$$

By this procedure, the temporal correspondence between predictive index and the signal is destroyed, while the overall properties of the index, in particular its potential temporal regularity and its marginal distribution are conserved exactly. We will use integer multiples of years for the shift δt , because the El Niño phenomenon is tied to the seasons (17 out of 20 El Niño events from 1950 to 2020 started in the first half

year covering MAMJJA and only 3 in SONDJF). If we shifted by intervals less than full years, although there will be more samples in the benchmark surrogate, part of the shifting would destroy the seasonality of the El Niño phenomenon, leading to meaningless results. Therefore, from the 34 years in the two studies we compared, 33 additional values of performance are obtained for the 33 possible shifts. If one of the two signals, forecast or verification, has a slowly decaying auto-correlation function, one should exclude shifts of $|\delta| < \text{the correlation time}$ in order to make sure that the shift really destroys the correlation between forecast and verification. In the case of El Niño, a shift by more than 1 year satisfies this condition.

4.2 Evaluation measures

In the case of climate network-based El Niño prediction as introduced before, the forecast is a binary one, i.e., it predicts whether or not in a given time window in the future there will be an El Niño event. For a binary forecast, there are two types of errors, often called the issue of sensitivity versus specificity: either an event might be predicted to follow but it will not - a false alarm; or an event that was not predicted finally occurs - a missed hit (Kaiserman et al., 2005; Fenlon et al., 2018). Since both errors might induce very different costs, a representation called Receiver Operating Characteristic (ROC) is in use (Kharin and Zwiers, 2003; Gönen et al., 2006). It reports the fraction of successfully predicted events (hit rate HR) as a function of the false alarm rate FAR, where the latter is tuned by the sensitivity of prediction scheme. In the El Niño predictions which we mentioned above, hit rates and false alarm rates are calculated either for setting the optimal threshold value or as a quantifier of the prediction skill in both two studies (Ludescher et al., 2014; Meng et al., 2018).

In order to derive a benchmark for these predictors, we calculate the false alarm rates (FAR) and hit rates (HR) for every δt shifted predictive series with a fixed threshold. Since we do not adjust the threshold again but use the same one as the original forecast, these can be assumed to be randomized out of sample predictions. These results will be represented as a scatter plot in the plane of hit rate versus false alarm rate. If the hit rate and false alarm rate of the prediction scheme are better than the performances of all the shifted index series, then this is a clear indication that the predictor performs well due to genuine correlations between predictive index and El Niño index. If instead, performance of the true predictor is inside the skill distribution of the shifted versions of the predictive index, we conclude

that it is the similarity of statistical properties of these two time series that give rise to the performance, not the specific correlation between them.

4.3 Benchmarking a naïve El Niño prediction scheme

Before applying the circular time shifted benchmark method to the real predication case, we firstly design a categorical El Niño prediction model based on the Oceanic Niño Index (ONI), which is used as an illustration of how circular time shifted benchmark works. We take ONI from 1950 to 2019 (data from <https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni>). In Sec.3.1, the conditional average of the ONI under the condition $\alpha = 0$ and $\delta = 0.3$ shows a clear signature of an El Niño event, with a lead time of 2 months on average. We hence use this condition here (i.e. threshold $\alpha = 0$ and $\delta = 0.3$, the predictability is decreasing when the threshold comes to negative, therefore we take the positive threshold. When $\alpha = 0.6$ is already at the El Niño period, therefore we choose $\alpha = 0$ as the threshold as it provides the longer prediction lead time compared to $\alpha = 0.3$, see Fig. 2 (left)) to create a naïve predictor. As mentioned before, this can be interpreted in the spirit of an analogues forecast (Lorenz, 1969). Based on this definition, a hit is counted if the ONI is in the positive state of ENSO (i.e., $\text{ONI} > 0.5$ for at least 5 months) at least once within 5 months after lead time, considering that once El Niño starts, it comes to a peak on average after 5 months. Otherwise, it is counted as a false alarm.

Empirically, for a lead time of 6 months, a threshold of value 0 yields optimal predictability, with a hit rate of 0.65 and false alarm rate of 0.24. When we increase the lead time, the predictability decreases, with a smaller hit rate and larger false alarm rate (the threshold remains at 0 for all lead times, since it was not our goal to optimize these forecasts). Therefore, depending on lead time, we have a forecast which we call “good” (lead time 6 months), “mediocre” (lead time 8 months) and “bad” (lead time 10 months), see Table [1].

We then apply the circular shift method to these three forecasts. Indeed, as shown in Fig. 7, the circular shifts affect the performance of the three forecasts in quite different ways: The model with the best prediction skill (lead time = 6 month) loses most of its predictability when the predictions are shifted with respect to the verification, these shifts give less hits and more false alarms compared to the original model. The model with the worst prediction skill (lead time = 10 months) does not lose much of its predictability after shifting, and some shifts can even yield more hits and

less false alarms. This indicates that if the model has a good predictability, after time shifting, it will lose its predictability, while for a bad model does not.

In order to further quantify how the prediction skill changes among the original model and its shifted versions, we represent the skill of each individual forecast (original or time shifted) by a vector in the plane of HR versus FAR. In the scatter plot of skills (FAR vs. HR) of the 69 shifted forecasts (as we take ONI from 1950-2019 to create these analogous models, therefore we get 69 shifted series), we fit a bi-variate normal distribution to these data. The lines of equal probability density (contour lines) are ellipses centered around (μ_x, μ_y) , described by

$$\begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix} \Sigma^{-1} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix} = c^2. \quad (9)$$

where x and y represent the FAR and HR, respectively, (μ_x, μ_y) are the mean values over the 69 performances of the shifted predictions, and Σ^{-1} is the inverse of a co-variance matrix. The outermost ellipses in Fig. 8 and Fig. 10 denote the 95% quantile, i.e., 95% of the probability lies inside. We assume a null hypothesis that the original forecast is a member of the shifted forecasts ensemble, which would mean that it has not more skill than a random forecast. This hypothesis is rejected with 95% confidence if the performance of the unshifted predictions is outside this ellipse. In addition, in order to perform better, its HR should be higher and FA should be lower than the mean over the shifted forecasts, i.e., its vector (FA_original model, HR_original model) should be in the second quadrant. In Fig. 8 such scatter plots together with the 95% confidence ellipse are shown for the three naïve predictions. The good predictor doubtlessly outperforms all its shifted versions, with the original vector (FA_original, HR_original) located outside of the outermost ellipse as well as in the second quadrant, indicating the higher HR and lower FA compared to all of the shifted series. The mediocre predictor is closer to the border of the 95% confidence ellipse, and the bad predictor lies well within the ellipses and it is not in the quadrant which indicates a better prediction skill as well. So for the bad prediction model, we cannot reject the null hypothesis that this is a member of these randomly shifted predictions and hence it has no further forecast skill. The advantage over pure randomization of the predictor lies in the fact that all shifted predictors are identical in all statistical properties, but the specific correlations between predictant and shifted predictors are destroyed.

Table 1: Analogous-based El Niño prediction model with different predictability

Threshold	Lead time	Hit rate	False alarm rate
0	6	0.65	0.24
0	8	0.4	0.34
0	10	0.1	0.46

Fig. 7: Frequency distributions of hit rates and false alarm rates given by shifted naïve El Niño forecast model (from left to right: prediction model using 6, 8, 10 months as the lead time; as the prediction model is based on the ONI from 1950-2019, we shift the model by 1 year each time, therefore the maximum shift length is 69)

4.4 Circular time shifted benchmarking for climate network-based models

Eventually, we apply our time-shifting method to the two time series of the predictive indices $S(t)$ and $FI(t)$. In both papers, the time period covers 34 years (in (Ludescher et al., 2014): 1980-2013; in (Meng et al., 2018): 1984-2017). The time series in Fig. 4 (upper panels) are our own reproductions of the two predictive indices using the same dataset as mentioned in the original paper. The lower panels show one the same index time series after shift by 2 years as an example. Our reproductions catch the main characteristics of original results which are published as figures in the corre-

sponding papers, with the same correct hits and false alarms, even though there are some small differences in very details. We shift the predictive index time series by multiples of 1 year and totally acquire 33 time series in each case. Correct predictions and false alarms given by these time series are counted following the rules defined in these 2 papers, and the frequency distributions of hit rates and false alarm rates of these shifted series are presented in Fig. 9 together with the performance of the original forecasts.

Fig. 9 shows how time shifting of the index series $S(t)$ and of $FI(t)$ reduces the number of correct predictions and increases the number of false alarms compared to its original predictive index before shifting.

Fig. 8: Comparison of prediction skills of the original models and their shifted versions by fitting bivariate Gaussians: the outermost ellipses denote the 95% quantiles, and the red disks represent the performances of the unshifted predictor. (from top to bottom: prediction model using 6, 8, 10 months as lead time; (FA_ori,HR_ori) indicates the original FAR and HR before shifting; (FA_shift, HR_shift) denotes the FAR and HR of shifted series)

Fig. 10 shows the 95%-quantiles of the fitted bi-variate Gaussian's in the FAR-HR-plane. While the original (FA_ori,HR_ori) of Meng's model is rather close to the boundary of the corresponding ellipse and also its FAR is larger than the mean FAR over all shifted versions, this model is at the border of our benchmarking test: Its performance is close to being compatible with the performance of a random $FI(t)$ series which has the same statistical properties as Meng's $FI(t)$ series. The result is much more favorable for Ludescher's forecast: While its FAR is close to the mean FAR of all shifted

versions, its HR is definitely much better and sticks out from all those of the shifted versions. We conclude that Ludescher's model passes our benchmark approach much better than Meng's method. This is consistent with the conclusion we obtained in Sec.3.3. We want to stress that the lead time of the two network based models is much larger than that of the naïve ONI predictor used as another benchmark, so that both of them will find their applications.

Fig. 9: Frequency distributions of hits and false alarms given by shifted versions of the two different El Niño prediction models (from left to right: the results of Ludescher’s model and the results of Meng’s model; when the shift length is 0, the corresponding dots indicate the hits and false alarms given by the original series)

5 Conclusion

We compare two climate network based El Niño prediction models in different ways and construct a method for benchmarking predictions on short data sets which relies on a new type of surrogate data. By shifting the forecast time series with respect to the verification time series, we destroy correlations between forecast and verification while conserving all statistical properties of the forecast. We demonstrate the usefulness of this benchmark to assess and compare different prediction schemes for El Niño forecasts.

We first compare two existed prediction models by compare their prediction skills on the same prediction lead times; and compare how long the lead time can be given at their best accuracy. From this initial comparison based on the observational data, we find that Ludescher’s model is more stable and indeed has skill on lead times of about or more than one year. We further compare these two models with the naïve prediction model based on analogous method. Even though the lead times in both network-based models are not very stable, they

outperform the naïve prediction model, since they have much longer prediction lead time. Finally, we compare these two climate network based prediction models by a new benchmark method. Prediction skills of shifted time series in both models decrease compared to their original time series. However, Ludescher’s model shows a more obvious decrease under shifting compared to Meng’s model, which indicates that the former one performs more robustly than the latter one. The reason for causing this difference may come from the choice in different study domains and different indices created in these two studies. In Meng’s work, the nodes only cover the Niño 3.4 basin, without considering nodes outside the basin as Ludescher’s work does.

In future work on El Niño prediction, we may need to think more about the influence from outside of the El Niño basin. Our benchmark method can also be used to assess other forecasts and can be easily generalized to multi-categorical forecasts or real values forecasts, by studying the root mean squared prediction error as a function of the time shift. **As the prediction for La Niña is not covered by these climate network based forecasts**

Fig. 10: Comparison of prediction skill of the two models in the FAR-HR-plane (left: Ludescher's model, right: Meng's model; (FA_ori,HR_ori) indicates the original FAR and HR before shifting; (FA_shift, HR_shift) denotes the FAR and HR of shifted series).

we discussed here, we do not explore our benchmarking on La Niña's prediction. Besides, La Niña and El Niño events are not symmetric (Wu et al., 2010; Im et al., 2015), El Niño usually decays rapidly by next summer, while La Niña can persist and appear twice or even multiple times over consecutive years (Bonsal and Lawford, 1999; Okumura and Deser, 2010), the prediction mechanism behind El Niño and La Niña may be very different. However, if there exists binary La Niña forecasts, our benchmarking methods can also be applied to estimate their prediction skills. In this paper, we do not cover this aspect.

Declarations

Funding This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813844

Conflicts of interest The authors declare that they have no conflict of interest.

Availability of data and material

The data used in this study are publicly available online:

- The ONI used in reproducing Meng's model is from: (<https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni>);
- The ONI used in analogous forecast in Sec.3.1 is from: (<https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni>);
- The Niño 3.4 index used in reproducing Ludescher's model is from: (https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/nino34.long.anom.data);
- The daily surface atmospheric temperature dataset (1950-2013) in Ludescher's model is from: (<https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.surface.html>);
- The daily mean near surface (1000hPa) air temperature dataset (1979-2018) in Meng's model is from the ERA-Interim re-analysis: (<https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=pl/>).

Code availability Not applicable

Acknowledgements This work is part of the Climate Advanced Forecasting of sub-seasonal Extremes (CAFE) which

has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.813844.

References

- Alexander MA, Blad I, Newman M, Lanzante JR, Lau NC, Scott JD (2002) The atmospheric bridge: The influence of enso teleconnections on air-sea interaction over the global oceans. *Journal of climate* 15(16):2205–2231, doi:[https://doi.org/10.1175/1520-0442\(2002\)015<2205:TABTIO>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2)
- Barnston AG, Tippett MK, van den Dool HM, Unger DA (2015) Toward an improved multimodel enso prediction. *Journal of Applied Meteorology and Climatology* 54(7):1579–1595, doi:<https://doi.org/10.1175/JAMC-D-14-0188.1>
- Bjerknes J (1969) Atmospheric teleconnections from the equatorial pacific. *Monthly weather review* 97(3):163–172, doi:[https://doi.org/10.1175/1520-0493\(1969\)097<0163:ATFTEP>2.3.CO;2](https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2)
- Bonsal B, Lawford R (1999) Teleconnections between el niño and la niña events and summer extended dry spells on the canadian prairies. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 19(13):1445–1458, doi:[https://doi.org/10.1002/\(SICI\)1097-0088\(19991115\)19:13<1445::AID-JOC431>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1097-0088(19991115)19:13<1445::AID-JOC431>3.0.CO;2-7)
- Capotondi A, Sardeshmukh PD (2015) Optimal precursors of different types of enso events. *Geophysical Research Letters* 42(22):9952–9960, doi:<https://doi.org/10.1002/2015GL066171>
- Cattell R (1991) The Benchmark Handbook for Database and Transaction Processing Systems, chapter 6
- Chiera BA, Filar JA, Zachary DS, Gordon AH (2009) Comparative forecasting and a test for persistence in the el nino southern oscillation. vol 138, pp 255–273, doi:https://doi.org/10.1007/978-1-4419-1129-2_9
- Clarke AJ (2014) El niño physics and el niño predictability. *Annual review of marine science* 6(1):79–99, doi:[10.1146/annurev-marine-010213-135026](https://doi.org/10.1146/annurev-marine-010213-135026), URL <https://doi.org/10.1146/annurev-marine-010213-135026>
- Collins M, Frame D, Sinha B, Wilson C (2002) How far ahead could we predict El Niño? *Geophysical Research Letters* 29(10):130–1–130–4, doi:<https://doi.org/10.1029/2001GL013919>
- Diaz HF, Hoerling MP, Eischeid JK (2001) Enso variability, teleconnections and climate change. *International Journal of Climatology: A Journal of*

- the Royal Meteorological Society* 21(15):1845–1862, doi:<https://doi.org/10.1002/joc.631>
- Duan W, Hu J (2016) The initial errors that induce a significant spring predictability barrier for el niño events and their implications for target observation: results from an earth system model. *Climate Dynamics* 46(11):3599–3615, doi:<https://doi.org/10.1007/s00382-015-2789-5>
- Duan W, Wei C (2013) The spring predictability barrier for enso predictions and its possible mechanism: results from a fully coupled model. *International Journal of Climatology* 33(5):1280–1292, doi:<https://doi.org/10.1002/joc.3513>
- Fan J, Meng J, Ashkenazy Y, Havlin S, Schellnhuber HJ (2017) Network analysis reveals strongly localized impacts of el niño. *Proceedings of the National Academy of Sciences* 114(29):7543–7548, doi:<https://doi.org/10.1073/pnas.1701214114>
- Fan J, Meng J, Ashkenazy Y, Havlin S, Schellnhuber HJ (2018) Climate network percolation reveals the expansion and weakening of the tropical component under global warming. *Proceedings of the National Academy of Sciences* 115(52):E12128–E12134, doi:<https://doi.org/10.1073/pnas.1811068115>
- Fenlon C, OGrady L, Doherty ML, Dunnion J (2018) A discussion of calibration techniques for evaluating binary and categorical predictive models. *Preventive veterinary medicine* 149:107–114, doi:<https://doi.org/10.1016/j.prevetmed.2017.11.018>
- Gaucherel C (2010) Analysis of enso interannual oscillations using non-stationary quasi-periodic statistics: A study of enso memory. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 30(6):926–934, doi:<https://doi.org/10.1002/joc.1937>
- Goddard L, Mason SJ, Zebiak SE, Ropelewski CF, Basher R, Cane MA (2001) Current approaches to seasonal to interannual climate predictions. *International Journal of Climatology* 21(9):1111–1152, doi:<https://doi.org/10.1002/joc.636>
- Gönen M, et al. (2006) Receiver operating characteristic (roc) curves. *SAS Users Group International (SUGI)* 31:210–231, URL <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.2918&rep=rep1&type=pdf>
- Gozolchiani A, Havlin S, Yamasaki K (2011) Emergence of el niño as an autonomous component in the climate network. *Physical review letters* 107(14):148501, doi:<https://doi.org/10.1103/PhysRevLett.107.148501>
- Harrigan S, Prudhomme C, Parry S, Smith K, Tanguy M (2018) Benchmarking ensemble streamflow prediction skill in the uk. *Hydrology and Earth System Sciences* 22(3):2023–2039,

- doi:<https://doi.org/10.5194/hess-22-2023-2018>
- Im SH, An SI, Kim ST, Jin FF (2015) Feedback processes responsible for el niño-la niña amplitude asymmetry. *Geophysical Research Letters* 42(13):5556–5563, doi:<https://doi.org/10.1002/2015GL064853>
- Jin EK, Kinter JL, Wang B, Park CK, Kang IS, Kirtman BP, Kug JS, Kumar A, Luo JJ, Schemm J, Shukla J, Yamagata T (2008) Current status of enso prediction skill in coupled ocean–atmosphere models. *Climate Dynamics* 31(6):647–664, doi:<https://doi.org/10.1007/s00382-008-0397-3>
- Jolliffe IT, Stephenson DB (2008) Proper scores for probability forecasts can never be equitable. *Monthly weather review* 136(4):1505–1510, doi:<https://doi.org/10.1175/2007MWR2194.1>
- Kaiserman I, Rosner M, Peer J (2005) Forecasting the prognosis of choroidal melanoma with an artificial neural network. *Ophthalmology* 112(9):1608–e1, doi:<https://doi.org/10.1016/j.ophtha.2005.04.008>
- Kharin VV, Zwiers FW (2003) On the roc score of probability forecasts. *Journal of Climate* 16(24):4145–4150, doi:[https://doi.org/10.1175/1520-0442\(2003\)016<4145:OTRSOP>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2)
- Latif M, Anderson D, Barnett T, Cane M, Kleeman R, Leetmaa A, O'Brien J, Rosati A, Schneider E (1998) A review of the predictability and prediction of enso. *Journal of Geophysical Research: Oceans* 103(C7):14375–14393, doi:<https://doi.org/10.1029/97JC03413>
- Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences* 26(4):636–646, doi:[https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2)
- Ludescher J, Gozolchiani A, Bogachev MI, Bunde A, Havlin S, Schellnhuber HJ (2013) Improved el niño forecasting by cooperativity detection. *Proceedings of the National Academy of Sciences* 110(29):11742–11745, doi:[10.1073/pnas.1309353110](https://doi.org/10.1073/pnas.1309353110)
- Ludescher J, Gozolchiani A, Bogachev MI, Bunde A, Havlin S, Schellnhuber HJ (2014) Very early warning of next el niño. *Proceedings of the National Academy of Sciences* p 201323058, doi:[10.1073/pnas.1323058111](https://doi.org/10.1073/pnas.1323058111)
- Ludescher J, Bunde A, Havlin S, Schellnhuber HJ (2019) Very early warning signal for el niño in 2020 with a 4 in 5 likelihood. *arXiv preprint arXiv:191014642*
- Luo JJ, Masson S, Behera SK, Yamagata T (2008) Extended enso predictions using a fully coupled ocean–atmosphere model. *Journal of Climate* 21(1):84–93, doi:[10.1175/2007JCLI1412.1](https://doi.org/10.1175/2007JCLI1412.1)
- Luo M, Lau NC (2019) Amplifying effect of enso on heat waves in china. *Climate Dynamics* 52(5):3277–3289, doi:[10.1007/s00382-018-4322-0](https://doi.org/10.1007/s00382-018-4322-0)
- Mason SJ, Mimmack GM (2002) Comparison of some statistical methods of probabilistic forecasting of enso. *Journal of Climate* 15(1):8–29
- McPhaden MJ, Zebiak SE, Glantz MH (2006) Enso as an integrating concept in earth science. *Science* 314(5806):1740–1745, doi:[10.1126/science.1132588](https://doi.org/10.1126/science.1132588)
- Meng J, Fan J, Ashkenazy Y, Bunde A, Havlin S (2018) Forecasting the magnitude and onset of el niño based on climate network. *New Journal of Physics* 20(4):043036, doi:[10.1088/1367-2630/aabb25](https://doi.org/10.1088/1367-2630/aabb25)
- Meng J, Fan J, Ludescher J, Agarwal A, Chen X, Bunde A, Kurths J, Schellnhuber HJ (2020) Complexity-based approach for el niño magnitude forecasting before the spring predictability barrier. *Proceedings of the National Academy of Sciences* 117(1):177–183, doi:[10.1073/pnas.1917007117](https://doi.org/10.1073/pnas.1917007117)
- Okumura YM, Deser C (2010) Asymmetry in the duration of el niño and la niña. *Journal of Climate* 23(21):5826–5843
- Pappenberger F, Ramos M, Cloke H, Wetterhall F, Alfieri L, Bogner K, Mueller A, Salamon P (2015) How do i know if my forecasts are better? using benchmarks in hydrological ensemble prediction. *Journal of Hydrology* 522:697–713, doi:[10.1016/j.jhydrol.2015.01.024](https://doi.org/10.1016/j.jhydrol.2015.01.024)
- Schreiber T, Schmitz A (2000) Surrogate time series. *Physica D: Nonlinear Phenomena* 142(3–4):346–382, doi:[10.1016/S0167-2789\(00\)00043-9](https://doi.org/10.1016/S0167-2789(00)00043-9)
- Steinhaeuser K, Chawla NV, Ganguly AR (2010) Complex networks in climate science: Progress, opportunities and challenges. In: CIDU, pp 16–26, URL <http://c3.ndc.nasa.gov/dashlink/resources/224/>
- Tao L, Duan W, Vannitsem S (2020) Improving forecasts of el niño diversity: a nonlinear forcing singular vector approach. *Climate Dynamics* 55(3):739–754, doi:[10.1007/s00382-020-05292-5](https://doi.org/10.1007/s00382-020-05292-5)
- Tseng Yh, Hu ZZ, Ding R, Chen Hc (2017) An enso prediction approach based on ocean conditions and ocean–atmosphere coupling. *Climate Dynamics* 48(5–6):2025–2044, doi:[10.1007/s00382-016-3188-2](https://doi.org/10.1007/s00382-016-3188-2)
- Vihinen M (2012) How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis 13(4):1–10, doi:[10.1186/1471-2164-13-S4-S2](https://doi.org/10.1186/1471-2164-13-S4-S2)
- Wahiduzzaman M, Oliver EC, Wotherspoon SJ, Luo JJ (2020) Seasonal forecasting of tropical cyclones in the north indian ocean region: The role of el niño–southern oscillation. *Climate Dynamics* 54(3):1571–1589, doi:[10.1007/s00382-019-05075-7](https://doi.org/10.1007/s00382-019-05075-7)

- Wallace JM, Rasmusson EM, Mitchell TP, Kousky VE, Sarachik ES, von Storch H (1998) On the structure and evolution of enso-related climate variability in the tropical pacific: Lessons from toga. *Journal of Geophysical Research: Oceans* 103(C7):14241–14259, doi:[10.1029/97JC02905](https://doi.org/10.1029/97JC02905)
- Wang L, Ren HL, Zhu J, Huang B (2020a) Improving prediction of two enso types using a multi-model ensemble based on stepwise pattern projection model. *Climate Dynamics* pp 1–15, doi:<https://doi.org/10.1007/s00382-020-05160-2>
- Wang X, Slawinska J, Giannakis D (2020b) Extended-range statistical enso prediction through operator-theoretic techniques for nonlinear dynamics. *Scientific reports* 10(1):1–15, doi:[10.1038/s41598-020-59128-7](https://doi.org/10.1038/s41598-020-59128-7)
- White CJ, Hudson D, Alves O (2014) Enso, the iod and the intraseasonal prediction of heat extremes across australia using poama-2. *Climate dynamics* 43(7-8):1791–1810, doi:[10.1007/s00382-013-2007-2](https://doi.org/10.1007/s00382-013-2007-2)
- Wu B, Li T, Zhou T (2010) Asymmetry of atmospheric circulation anomalies over the western north pacific between el niño and la niña. *Journal of Climate* 23(18):4807–4822
- Y T, W H (2002) Hybrid coupled models of the tropical pacific–ii enso prediction. *Climate dynamics* 19(3):343–353, doi:[10.1007/s00382-002-0231-2](https://doi.org/10.1007/s00382-002-0231-2)
- Yamasaki K, Gozolchiani A, Havlin S (2008) Climate networks around the globe are significantly affected by el nino. *Physical review letters* 100(22):228501, doi:[10.1103/PhysRevLett.100.228501](https://doi.org/10.1103/PhysRevLett.100.228501)
- Zhang WF, Zhao Q (2015) The quasi-periodicity of the annual-cycle forced enso recharge oscillator model. *Communications in Nonlinear Science and Numerical Simulation* 22(1-3):472–477, doi:<https://doi.org/10.1016/j.cnsns.2014.08.013>
- Zheng F, Zhu J (2010) Spring predictability barrier of enso events from the perspective of an ensemble prediction system. *Global and Planetary Change* 72(3):108–117, doi:[10.1016/j.gloplacha.2010.01.021](https://doi.org/10.1016/j.gloplacha.2010.01.021)