Koopman Theory for Enhanced Pacific SST forecasting

Paula Lorenzo-Sánchez [a,b] , Matthew Newman [c] , John Albers [c] , Aneesh C. Subramanian [f] and Antonio Navarra [a,e]

[a] *Centro Euromediterraneo sui Cambiamenti Climatici, Bologna, Italy*
[b] *Universita' di Bologna, Dipartimento di Fisica e Astronomia, Bologna, Italy*
[c] *National Center for Atmospheric Research, Boulder, Colorado, USA*
[e] *Universita' di Bologna, Dipartimento di Scienze Geologiche, Biologiche e Ambientali, Bologna, Italy*
[f] *Department of Atmospheric and Oceanic Sciences,University of Colorado, Boulder,CO, USA*

ABSTRACT: ENSO is a complex climatic phenomenon that significantly impacts global weather patterns and ecosystems. Improving its prediction is therefore of high societal value. However, Global Circulation Models present severe biases when predicting ENSO, and their skill remains comparable to that of vastly simpler empirical models such as Linear Inverse Models (LIMs). LIMs, however, rely on linear dynamics, and they have inherent limitations in capturing the behavior of non-linear phenomena. In this context, Koopman operator theory has emerged as a powerful mathematical framework, offering a novel perspective for analyzing complex non-linear systems. While previous studies have demonstrated the potential of Koopman methods for ENSO forecasting, a fundamental question remains: how much data is actually needed to obtain robust Koopman operator estimates? In this study, we address this issue by performing a systematic sensitivity analysis on the reliability of Koopman-based forecasts as a function of data length. Using 2000 years of tropical SST pre-industrial CESM2 data, we assess the skill of the Niño 3.4 index forecasts within the nonlinear Koopman framework and compare these results to the benchmark set by LIMs. Our findings reveal nuances in the robustness of Koopman operator estimates, particularly evident when using shorter training periods. However, a notable breakthrough emerges, as we demonstrate the higher skill of the Niño 3.4 Koopman Ensemble Forecasts (KEFs), which showcase consistent improvements over linear models. In addition, the model shows notable improvements in capturing the variability of the western Pacific, as well as in the reliability of the El Niño and La Niña events forecasts.

SIGNIFICANCE STATEMENT: This study aims to improve our ability to predict ENSO variability, whose El Niño and La Niña events have major effects on global weather and ecosystems. We explored a new data-driven approach using Koopman decomposition, which provides a solid dynamical framework to analyse and possibly understand highly nonlinear systems. By applying this method to long records of sea surface temperature data, we found that it can significantly improve the accuracy of El Niño and La Niña forecasts, especially for predictions made well in advance. This advancement could lead to more reliable forecasts, helping communities better prepare for and respond to ENSO events.

1. Introduction

The El Niño–Southern Oscillation (ENSO) is the dominant mode of tropical atmosphere-ocean variability, with its influence extending globally through atmospheric teleconnections. In addition, ENSO affects ecosystems, agriculture, freshwater supplies, and other severe weather events worldwide (Ropelewski and Halpert 1986; Siegert et al. 2001). Understanding ENSO predictability is therefore a matter of high societal relevance, and improving ENSO predictions is an important goal of the seasonal forecasting community (L'Heureux et al. 2020; Zhang et al. 2022). However, many general circulation models (GCMs) struggle in simulating the basic statistical properties of ENSO (Bellenger et al. 2014; Planton et al. 2021), and their prediction skill remains comparable to that of vastly simpler empirical models (Newman and Sardeshmukh 2017; Richter et al. 2020).

Among those, linear inverse models (LIMs; Penland and Sardeshmukh (1995)) have been successfully implemented for decades, demonstrating significant long-lead forecasting skill for Pacific SSTs. LIMs are empirical dynamic models that assume that the temporal evolution of the predictand is described by a multivariate linear Markov process plus noise, representing the rapidly evolving (and unpredictable) nonlinearities (Penland 1996). The mechanism of LIMs represents ENSO as a linear combination of modes, usually obtained through empirical orthogonal function analysis, whose temporal evolution can be inferred through linear regression techniques allowing for the prediction of SST anomalies based on past observations.

The LIM was first implemented for tropical SSTs by Penland and Sardeshmukh (1995), and subsequently extended to capture subsurface and ocean-atmosphere coupling effects (Newman et al. 2011), and through the development of cyclostationary LIMs, which include annual cycles of linear dynamics and stochastic forcing

Corresponding author: Paula Lorenzo-Sánchez, paulalorenzo-sanche2@unibo.it

(Shin et al. 2021). In addition, Newman and Sardeshmukh (2017) showed that the prediction skill of LIMs could meet or exceed that of the GCMs, comprising the North American Multimodel Ensemble (Shin et al. 2021; Newman and Sardeshmukh 2017). However, LIMs assume that predictable dynamics are effectively linear, and they are therefore inherently limited for capturing predictable aspects of nonlinear phenomena.

Within this context, the works of Koopman (1931) and Koopman and Neumann (1932) provide the basis for an alternative to the problem. The Koopman operator framework, rooted in ergodic theory, states that associated to every dynamical system are intrinsic linear operators acting on a potentially infinite-dimensional space of observables (functions on the state space). By representing the evolution of observables within a much higher-dimensional function space, Koopman analysis helps identify coherent structures and dominant modes of variability. These modes can then be used to compute forecasts by tracking their temporal evolution (Navarra et al. 2021; Mezić 2005; Mezić 2013; Rowley et al. 2009; Berry et al. 2015; Wang et al. 2020). The infinite-dimensional nature of the Koopman operator made its practical application computationally challenging, leading to limited use for almost a century. However, recent advances in numerical techniques, such as Extended Dynamic Mode Decomposition (EDMD) and the Variational Approach of Conformation Dynamics (VAC) (Schütte et al. 2016), have renewed interest in the framework due to its potential for analyzing complex, high-dimensional nonlinear systems.

Some recent studies, such as Berry et al. (2015), Wang et al. (2020), and Navarra et al. (2021), have applied Koopman methods to predict the Niño 3.4 index using kernel-based approximations of the Koopman operator. Wang et al. (2020), in particular, employed a Gaussian kernel method in a framework known as Kernel Analog Forecasting (KAF), which produces forecasts mathematically similar to kernel-weighted ensemble forecasts (Wang et al. 2016). In addition, Navarra et al. (2021) compared the performance of the k-EDMD algorithm using linear and Gaussian kernels, following the idea that LIMs are a specific case of the Koopman framework in which linear functions are used for expanding the observable's space. Their results highlight the potential for improved ENSO forecasting using Koopman-based methods, suggesting that nonlinear kernels could enhance ENSO predictions beyond the operational benchmark set by LIMs. However, these studies focus on direct forecasting applications, and do not assess the robustness of Koopman operator estimates with respect to data availability. This gap is crucial, as the reliability of data-driven forecasting approaches depend on its sensitivity to the characteristics of the training data and the parameters involved in the tuning of the algorithm (Navarra et al. 2021).

These are common issues for machine and deep learning methods, which often require extensive data records to outperform simpler linear approaches. Nonlinear methods, in fact, require large amounts of data to fit the nonlinear patterns successfully. Otherwise, they struggle with generalization to unseen samples and overfitting (Zhang 2012). Not only is pre-processing such datasets resource-intensive, but the availability of continuous and reliable data is often challenging, particularly for multidecadal phenomena (Mu et al. 2019; Zhang 2012). These are compelling concerns for the Koopman framework, as it is based on the estimation of a potentially infinite-dimensional underlying operator from finite data records, leading to sampling uncertainty issues.

Addressing these issues, our study provides a structured evaluation of how data availability impacts Koopman-based predictions, offering key insights into the conditions necessary for their successful application to climate forecasting. We seek to explore a systematic evaluation of Koopman prediction skill for ENSO forecasting using extensive simulated pre-industrial data from the CESM2 model, in order to test the ability of Koopman operators to extract the dynamics of a stationary complex system. In addition, the availability of 2000 years of data enables to probe the sensitivity of Koopman skill to data availability. By employing subsamples of varying lengths, we investigate the impact of data record length on the operator's robustness and forecast performance, discerning the optimal conditions for Koopman-based forecasting. Yet, our analysis reveals the presence of significant uncertainty in the estimated Koopman operators, which undermines forecast reliability. To address this issue, we propose a novel approach —Koopman Ensemble Forecasts (KEF) — aimed at enhancing forecast robustness and mitigating the effects of noise. In this approach, several estimates of the Koopman operator are computed from sub-samples of the data and used as independent models to produce an ensemble. By leveraging ensemble forecasting techniques, we seek to harness the collective skill of multiple Koopman operators, offering a more resilient forecasting framework capable of overcoming the inherent limitations of individual operators.

The paper is organized as follows. In Section 2 we review the basics of Koopman theory and spectral decomposition. In Section 3, the data and methodology are described. Section 4 discusses the problem of robustness in the Koopman spectrum. Section 5 presents the results of the KEFs to the Niño 3.4 index, while Sections 6 and 7 focus on Pacific SST forecasts and the reliability of El Niño/La Niña events. In Section 8 we extend our results

for shorter training data lengths, to come closer to a realistic scenario of observed data. Finally, we summarize the results and conclude in Section 9.

2. Koopman spectral decomposition and estimation of the Koopman spectrum from data

a. Basics of Koopman theory

A deterministic dynamical system can be defined by a differential equation of the form:

$$\frac{d\mathbf{z}}{dt} = \mathbf{F}(\mathbf{z}) \tag{1}$$

where $\mathbf{z}$ $(z_0, z_1, \ldots z_n)$ is an $n+1$ dimensional state vector and $\mathbf{F}$ denotes the vector field governing the dynamics of the system. When the vector $\mathbf{z}$ corresponds directly to physical parameters, such as temperature or sea level pressure, we call them "full-system states". In practical applications, however, we often work with reduced representations of the state, denoted as $\mathbf{x}$, which may be obtained via empirical orthogonal function (EOF) analysis or other dimensionality reduction techniques. These reduced representations can be expressed as vector-valued functions of the full system state, i.e., $\mathbf{x} = f(\mathbf{z})$, and are sometimes referred to as "state vectors" in a reduced sense. However, it is important to recognize that $\mathbf{x}$ does not necessarily evolve according to an autonomous deterministic equation like Eq. 1. Instead, the Koopman framework allows us to study the evolution of observables $g(\mathbf{x})$, which are any functions of the state. The value of a general observable $\mathbf{g}(\mathbf{x})$ at time t, starting from $\mathbf{x}_0$ at time zero, is given by

$$\mathbf{g}(t, \mathbf{x}_0) = \mathbf{g}(\mathbf{x}_t). \tag{2}$$

In this context, the Koopman operator K^τ describes the evolution of observables $\mathbf{g}(\mathbf{x})$ over a lag time τ as

$$K^\tau(\mathbf{g}(\mathbf{x}_0)) = g(U^\tau(\mathbf{x}_0)) = \mathbf{x}_t, \tag{3}$$

in the same way that U^τ, the flow map associated with the field $\mathbf{F}$, evolves the state $\mathbf{z}$ for a fixed lag time τ. Note that the Koopman operator (and therefore its eigenvalues and eigenfunctions) implicitly depend on the chosen lag time for its computation. For the sake of simplicity, we will omit this dependency and refer to K^τ as K.

A key property of the Koopman operator is its linearity, which allows the use of spectral decomposition techniques despite the underlying nonlinear dynamics of the system. This framework enables us to represent complex behavior through a set of linear modes, but it comes at the cost of dealing with an infinite-dimensional space. In practical applications, we approximate this infinite-dimensional operator using a finite set of eigenfunctions, whose associated eigenvalues determine the temporal evolution of the system. This spectral decomposition is fundamental for forecasting applications, as it provides a structured way to analyze and predict system dynamics.

Assuming that we can estimate the spectral decomposition of a system's Koopman operator, we can use it to expand its observables (and states, by using the identity observable $\boldsymbol{g(x) = x}$, which returns the state vector itself), and evolve them in time to produce forecasts.

A vector observable $\mathbf{g}(\mathbf{x})$ can be expressed in terms of its Koopman spectral decomposition as follows (Mezić 2005; Mezić 2013; Navarra et al. 2021)

$$K^t(\mathbf{g}(\mathbf{x})) = \overline{\mathbf{g}(\mathbf{x})} + \sum_k \mathbf{v}_k \phi_k(\mathbf{x}) e^{\lambda_k t} + \mathbf{n}(t, \mathbf{x}) \tag{4}$$

where three components can be distinguished: the time mean $\overline{\mathbf{g}(\mathbf{x})}$, an almost-periodic component based on Koopman eigenfunctions and eigenvalues ($\phi_k(\mathbf{x}), \lambda_k = \sigma_k + i\omega_k$), and a completely aperiodic component corresponding to the continuous part of the spectrum ($\mathbf{n}(t, \mathbf{x})$). Note that as the time mean $\overline{\mathbf{g}(\mathbf{x})}$ is included as a separate term in the expansion, the eigenspace corresponding to $\lambda_k = 1$ is excluded from the point spectrum component sum.

The coefficients $\mathbf{v}_k$, known as Koopman modes (see A6 in Appendix A for a detailed formulation), represent the projection of each component of the observables vector on the Koopman eigenfunctions, and portray spatial patterns of variability that evolve with time. On the other hand, the values that the eigenfunctions take at the data points can be interpreted as the magnitude of these spatial patterns at specific time steps, while the Koopman eigenvalues, with their real and imaginary components, tell us about the timescales and frequencies of the modes. The real part of the eigenvalue reflects the growth or decay rate of a mode, which can provide insights into the stability of specific patterns. The imaginary part relates to the oscillatory nature of the mode, giving us information about the period of the oscillations.

It is important to note that the validity of this spectral expansion depends on the properties of the Koopman operator and the function space in which it is defined. In the case of measure-preserving dynamics, where the Koopman operator is unitary on the L^2 space associated with the invariant measure, the expansion holds in an L^2 sense, meaning it describes the evolution of observables in a mean sense over the attractor rather than pointwise (Mezić 2005). In more general cases, it serves as an approximation, especially in finite-dimensional settings. Here, we adopt a data-driven approach, estimating the Koopman operator from time series and analyzing its spectral properties within this framework.

b. Koopman operator spectra from data

The Koopman decomposition can be derived from time series data using various techniques. In this work we use the k-EDMD (Klus et al. 2019) algorithm (see Appendix A for the details of the methodology), which has been previously employed in the context of ENSO forecasting by Navarra et al. (2021).

EDMD relies on a set of nonlinear features $P = [p_1, p_2, \ldots]$ to elevate the data to a higher-dimensional space. These functions can take different forms, such as Gaussians, polynomials or radial functions, and they act as a basis to expand the space where the Koopman operator acts. However, explicit calculations of features can quickly become computationally expensive. Kernels play a pivotal role in mitigating this burden. They provide a Reproducing Kernel Hilbert Space (RKHS), where data can be implicitly expressed in terms of the feature functions. The value of the features in any state vector $\mathbf{x}$ is obtained from the kernel function as $p_i(\mathbf{x}) = k(\mathbf{x}_i, \mathbf{x})$. The *representer theorem* states that for any function g in the RKHS, and for any tolerance $\epsilon > 0$, there exists a finite set of points $x_1, \ldots, x_n$ and scalars α_i such that the residual $g(x) - \sum_{i=1}^{n} \alpha_i p_i(x)$ has an RKHS norm less than ϵ, where the α_i are expansion coefficients of g with respect to the kernel features (Navarra et al. 2021). It is important to note that not all observables necessarily lie in the RKHS induced by a given kernel. For example, when using a Gaussian kernel, the RKHS consists only of bounded functions, meaning that functions of the form $g(\mathbf{x}) = \mathbf{x_j}$, which are unbounded, are generally not part of this space. However, practical applications often rely on approximations within the RKHS, which can still yield meaningful representations of the system's dynamics.

This approach, also known as the *"kernel trick"*, allows to employ richer sets of functions, improving the approximations of the Koopman operator (Scholköpf and Smola 2002). Importantly, the choice of the kernel determines the function space used to approximate the Koopman operator, which can change how well we capture the system's evolution. A well-chosen kernel can help reveal hidden structures in the data by mapping observations into a space where complex behaviors become easier to model. When the standard inner product is used to measure similarity, the function space is restricted to linear functions, yielding the same eigenvalues and eigenfunctions as those obtained in LIM algorithms (Tu 2013; Tu et al. 2014; Navarra et al. 2021).

In our work, we use the Gaussian kernel, which is *characteristic* and *universal*, meaning that it can approximate all continuous functions of the state variables (complete set), and allows for perfect distinction between different distributions (Muandet et al. 2017). Therefore, applicability tests such as the τ- test are not needed, as *universal* kernels guarantee that the model can distinguish between different underlying distributions. In addition, Gaussian kernels are positive definite and limited between zero and one, not requiring further normalization, and they have been used in previous applications for SST data demonstrating good results (Navarra et al. 2021, 2024). The Gaussian kernel

$$p_i(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}s)^{n/2}} \exp(-||\mathbf{x} - \mathbf{x}_i||^2/2s^2), \qquad (5)$$

will be compared to the linear one, used as a benchmark ($p_i(\mathbf{x}) = \langle \mathbf{x_i}, \mathbf{x} \rangle$). In the Gaussian kernel, the bandwidth s determines the influence radius of data points with respect to each other. Ideally, a larger bandwidth is desired in regions with low state density, and a smaller one in regions with higher state density (Froyland et al. 2021). There are different approaches to choose s, including bandwidths that normalize distances to unit standard deviation (Navarra et al. 2021) or "self-tuning" kernels that adjust the bandwidth depending on the density of the region (Berry et al. 2015; Berry and Harlim 2016). In this study, a bandwidth based on the distance between temporal nearest neighbors has been used. Although the detailed values of the eigenvalues might be sensitive to small changes in the bandwidth selection, the properties of the modes remain robust.

3. Data and methodology

In this paper, we use 2000 years of pre-industrial control (pi-Control) runs from the Community Earth System Model (CESM2). CESM2 is a fully-coupled GCM developed by the National Center for Atmospheric Research (NCAR), and it provides state-of-the-art simulations of the Earth's past, present and future climate states (Danabasoglu et al. 2020). In particular, we use SST and SSH (sea surface height) data from the monthly averaged pi-Control runs of the Ocean Post Processed Data, available at *www.cesm.ucar.edu:/models/cesm2/*.

Pre-industrial control runs are simulations where the model is run without changes in external forcings, such as greenhouse gas concentrations, and they are often a baseline to understand the natural variability of the climate system. Due to their relatively constant dynamics, pi-Control runs allow to test the Koopman framework skill to capture and forecast the system's internal variability without the influence of external factors or measurement errors from observational datasets. In addition, pi-Control runs typically span several centuries, providing sufficiently long time-series to test Koopman operator sensitivity to the training record length. Moreover, the CESM2 pre-industrial simulation has demonstrated to realistically simulate several aspects of ENSO, including its dominant timescales, tropical and extratropical precursors, composite evolution of El Niño and La Niña events and ENSO

teleconnections to the extratropics. However, the modeled ENSO amplitude is about 30% larger than observed, with larger variability occurring in the central Pacific rather than the eastern Pacific, limiting the model's ability to represent the full diversity of El Niño spatial patterns (Capotondi et al. 2020).

The original data comprises SST and SSH monthly means featured in a tripolar grid. As a part of the pre-processing, they have been regridded into a 1ºx 1º regular grid. The study region was defined as 32º N to 31º S and 130 ºE to 70 ºW (see Figure 6), resulting in a total number of 9121 grid points (after the land-sea mask). Although the dataset should not present externally forced trends, it has been linearly de-trended point by point to ensure no tendencies are present. Anomaly monthly means have been computed with respect to the monthly climatology and normalized by the total standard deviation.

Prior to the Koopman analysis, the data has been pre-processed through an EOF analysis. Two different subsets of EOFs will be used through the study: a configuration with 30 EOFs (20 SST + 10 SSH, retaining 84% and 65% of the total variance, similar to configurations used for constructing tropical Pacific LIMs from observational data (Penland and Sardeshmukh 1995; Newman et al. 2011; Navarra et al. 2021)) and a configuration retaining all the variability (keeping the rank of the data matrix: 845 EOFs for SST and 1000 for SSH) which is used only for the Gaussian kernel. After this transformation, the state vectors contain the EOF coefficients for every monthly mean anomaly. We can then organize the data according to A1, where m equals the total number of months and the dimension of the vectors x_i depends on the number of retained EOFs.

Additionally, we will subsample the record in chunks of different lengths (1000, 400, 200, 100 and 50 years). Each subsample will be used to compute a single estimation of the Koopman operator, which we label by the length of the data used (e.g., $\mathbf{K}_{100}$ is a Koopman operator estimated using a 100-yr long record). The spectra of the estimations will be compared to study their sampling uncertainty. Subsequently, Koopman Ensemble Forecasts (KEF) will be carried out. To do so, each operator is used to produce forecasts that cover the entire data series *except* the period from which that operator was computed, leading to a cross-validation-like structure. The forecasts are treated as members of the KEF, which is composed of j members, where j equals the number of total available years (2000) divided by the length of the sub-samples minus one (as the operator computed using a certain chunk of data is not included in its forecast). For instance, the 100-yrs KEF will be composed of 19 members, each one produced by one of the 20 possible $\mathbf{K}_{100}$ from the entire dataset. The properties and skill of the KEF mean and members are then compared. The idea behind the KEF is to reduce the sampling uncertainty of the estimated Koopman operators, which presents in the form of noisy departures from the real underlying operator of the system. This approach is closely related to *bagging* techniques, widely used in the machine learning community to improve unstable modelling procedures by combining estimates from different models (Petropoulos et al. 2018; Liu et al. 2023). By averaging forecasts made with different estimates of the Koopman operator, we seek to mitigate the effects of the sampling uncertainty, hopefully leading to more robust forecasts.

Forecast skill of the KEFs is assessed with the Anomaly Correlation Coefficient (ACC), which gives a measure of the correlation between anomalies of the forecasts and those of verifying values:

$$ACC = \frac{\sum_{i=1}^{n}(O_i' - \overline{O}')(F_i' - \overline{F}')}{\sqrt{\sum_{i=1}^{n}(O_i' - \overline{O}')^2}\sqrt{\sum_{i=1}^{n}(F_i' - \overline{F}')^2}} \qquad (6)$$

where O_i' and F_i' are the anomalies of the observed and forecasted values at time step i, $\overline{O}'$ and $\overline{F}'$ are the means of the observed and forecast anomalies, and n is the total number of time steps. Other skill measures, such as the Root Mean Square Error Skill Score (RMSEss), have also been evaluated (*Section 4*); since they give similar results, they are not included in the main text. However, some of the most relevant results can be found in the *Online Supplemental Material*.

In addition, the sharpness and reliability of the KEFs are assessed using reliability diagrams that evaluate the categorical prediction of El Niño and La Niña events, defined as the months lying in the upper and lower terciles of the Niño 3.4 model observations. Reliability diagrams are graphical tools used to assess the calibration of probabilistic forecasts by comparing predicted probabilities against observed frequencies, typically grouped into bins. In this study, the forecast probability of an event (El Niño or La Niña) is computed as the fraction of ensemble members predicting that event at a given lead time. Specifically, for each forecast initialization, we count the number of ensemble members classifying the event as occurring and normalize by the total number of ensemble members, yielding a probability between 0 and 1. The observed frequency, in contrast, represents the fraction of times the event actually occurs in the verification dataset when the forecast probability falls within a given bin.

A well-calibrated model will show observed frequencies closely matching predicted probabilities, and therefore the reliability will lay close to the diagonal. In other words, if the forecast model predicts a 60% chance of rain, the

observed frequency of rain should be close to 60% for instances where the model made similar predictions. If the observed frequencies consistently deviate from the predicted probabilities across different bins, it suggests that the forecast model may be poorly calibrated. This tool is very useful for detecting systematic biases, helping to improve the accuracy and reliability of forecast models.

4. Spectrum of Tropical SST data

To assess the stability of the Koopman operator estimates, we start by analyzing the eigenvalue spectra obtained from two independent 1000-year subsamples of the dataset ($\mathbf{K}_{1000}$). Figure 1 presents the spectral distributions for both subsamples, where the first and second rows refer to the first and second sub-samples of the data. The first column shows the spectra retrieved using the Gaussian kernel and all EOFs (Gaussian-ALL from now on). The coloring of the dots is proportional to the norm of the corresponding mode, which gives a measure of their contribution to the explained variance. These distributions are compatible with a chaotic system with a partial continuous spectrum, where the continuous part is often associated with chaotic behavior. However, we cannot rigorously determine whether the observed spectrum is genuinely continuous or consists of a dense set of discrete eigenvalues, as our numerical methods do not allow us to distinguish between these cases. The eigenvalues inside the unit circle (module <1) represent modes with decaying amplitudes over time, which in the context of ENSO may be associated with transient dynamics. Some eigenvalues lie close to the unit circle but not exactly on it, indicating quasi-periodic behavior. Note that there are as many non-zero eigenvalues as time-steps used for its computation, independently of the number of spatial degrees of freedom. The second column shows the eigenvalues obtained using the linear kernel. Due to the EOF truncation, only 30 of them are different from zero, corresponding to the number of degrees of freedom of the covariance matrix. Had all the EOFs been retained, we would have retrieved the same number of eigenvalues as in the Gaussian example (Xu et al. 2024).

The third column of Figure 1 shows the density of eigenvalues retrieved from the Gaussian kernel (blue line) and the discrete linear eigenvalues (red lines) for each sub-sample. The eigenvalue density is estimated using Kernel Density Estimation (KDE) with a Gaussian kernel. Specifically, the imaginary parts of the logarithm of the eigenvalues are used to estimate the probability density function of the eigenvalue distribution along the frequency axis. The resulting density function represents the number of eigenvalues per frequency band. Several local maxima in the density function can be identified in Figure 1, which highlight dominant frequency components (Navarra et al. 2024). While both sub-samples share similarities, such as peaks around the 3-4 and 1.5-year periods, they also exhibit significant differences in the number and position of density peaks. In particular, the number and position of peaks vary significantly between sub-samples, indicating a degree of uncertainty in the spectral estimation. This variability is also reflected in the Gaussian eigenvalues (first column), which show distinct distributions depending on the data subset used. As the eigenvalues do not appear consistently for different samples of the data, which should be expected for data coming from the same dynamical system (see more in Figure 2), we define them as non-robust. In contrast, the shape and distribution of the linear spectrum is more stable, suggesting that 1000 years are sufficient for robust linear estimations of the Koopman operator.

Shorter data records make this robustness issue worse, as shown in Figure 2, which presents composites of the eigenvalues for Koopman operator estimates computed using different record lengths, i.e. two $\mathbf{K}_{1000}$, five $\mathbf{K}_{200}$, etc. The number of eigenvalues in each panel is equal to 24.000 (the total number of time-steps in the time series, but note that in the case of the linear kernel just 30 of them per operator are different from zero). However, there is increased spread across sub-samples for shorter data records, with the eigenvalues showing weaker robustness and convergence properties. When the linear kernel is used (lower row) the spread becomes notable only for shorter data spans. Further analysis including all lengths (1000, 400, 200, 100, and 50 years; see Supplemental Material Figure 1) confirms that the uncertainty of the eigenvalues increases when reducing the operator's length. Also the number of damped Gaussian eigenvalues increases as the record length shortens. As we know that the eigenvalues of a bounded system should lie close to the unit circle, this can be seen as the uncertainty of the estimated eigenvalues increasing for shorter data records. However, non-negligible errors (intended as variations in the computed eigenvalues as the training interval is varied) are retrieved even for very long record lengths (1000 years).

5. Koopman Ensemble Forecasts

To address the lack of robustness in estimated Koopman operators, we propose the Koopman Ensemble Forecasting (KEF) approach. This method combines forecasts from multiple operator estimates into an ensemble, reducing sampling uncertainty.

To evaluate the effectiveness of the KEFs, we apply the $\mathbf{K}_{200}$ operators to predict the Niño 3.4 index. Figure 3 shows the KEF forecast for the first 100 years of the Niño 3.4 CESM2 data (Gaussian kernel in the left column, linear on the right). Each of the 9 thin red lines represents one of the KEF members; the thick blue line is the KEF mean and the gray one is the Niño 3.4 model observation. The ACC

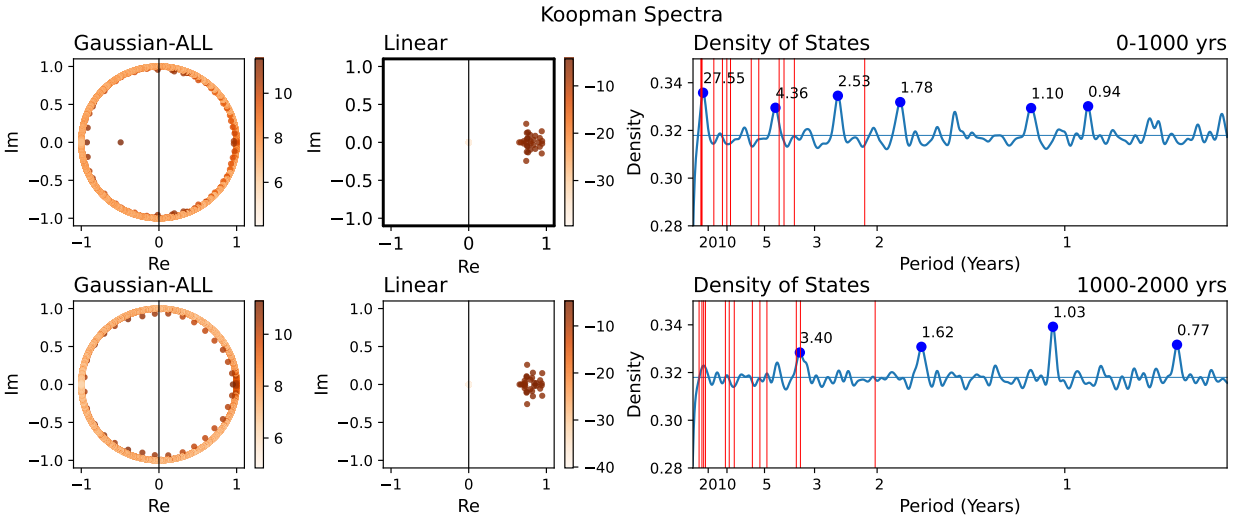

FIG. 1. Koopman spectra of the two possible 1000-yrs subsamples. The first two columns show the eigenvalues using the Gaussian (all EOFs) and the linear kernels respectively. The coloring of the dots is proportional to the norm of the corresponding mode of the Koopman operator. The third column shows the eigenvalue density retrieved from the spentrum of the Gaussian kernel (blue line) and the discrete eigenvalues of the linear kernel (red vertical lines). The units of the density are number of eigenvalues per frequency band.

between the KEF mean and the model observations, and the averaged ACC between each one of the KEF members and the observations, are shown in the lower part of each panel.

In the linear case, the KEF members remain notably robust even for long lead times (as the spread among them is small), while the Gaussian members present a wider spread starting from early leads. Even if the averaged ACC of the KEF members are almost identical for the linear and Gaussian kernels, the KEF mean of the Gaussian kernel has notably higher ACC. This suggests that there is non-linear predictability that can be captured by the Gaussian Koopman operators, despite their inherent lack of robustness.

The Gaussian KEF mean consistently outperforms individual ensemble members, whereas the improvement in the linear case is less pronounced. This is illustrated in Figure 4 (left and middle panels), which shows the Niño 3.4 ACC of the KEF mean (thick lines) and the spread of the KEF members (light shaded area) for the Gaussian-ALL and linear kernels using the $\mathbf{K}_{200}$ operators. The increase in skill for the Gaussian KEF mean is greater for operators trained on shorter data records (for example the $\mathbf{K}_{50}$ operators, see Supplemental Material Figure 2), as the uncertainty in their eigenvalues is greater and there is a better chance for improvement than in the case of operators trained on longer records. In contrast, the linear kernel retrieves robust spectra even when trained on shorter lengths, so its skill improvement from the ensemble approach is less notable. Further analysis regarding the comparison between the Gaussian kernel using all EOFs (Gaussian-ALL) versus 30 EOFs (Gaussian-30), demonstrates an overall better performance for the former one, suggesting that the Gaussian kernel yields better results when applied upon the full, rather than truncated state space. This result aligns with expectations, as non-linear kernels are designed to capture the non-linearities inherent in higher-dimensional subspaces of the dynamical system.

The KEF mean and member *ACC* results for the $\mathbf{K}_{200}$ are shown in the right panel of Figure 4. Solid lines feature the lead-dependent skill of the KEF mean, while the dotted ones represent the averaged *ACC* of the KEF members. Following the previous results, the Gaussian-ALL KEF members *ACC* outperforms both the linear and Gaussian-30 approaches. However, the most interesting results appear for the KEF mean's performance, where the Gaussian-ALL kernel remains over the 0.6 *ACC* mark almost 4 months longer than the linear counterpart, with greatest improvement between 11 and 18 leading months, presenting gains over 0.15.

A comprehensive analysis including all the different lengths of Koopman operators has been carried out to investigate the behaviour of the ensembles depending on the operator's length. The results demonstrate that linear KEF means perform very similarly among them irrespectively of their length, with only the $\mathbf{K}_{50}$ forecasts performing slightly better than the others after the 12 month lead-time mark. This marginal improvement of the operators computed over very short periods might be dependent on the

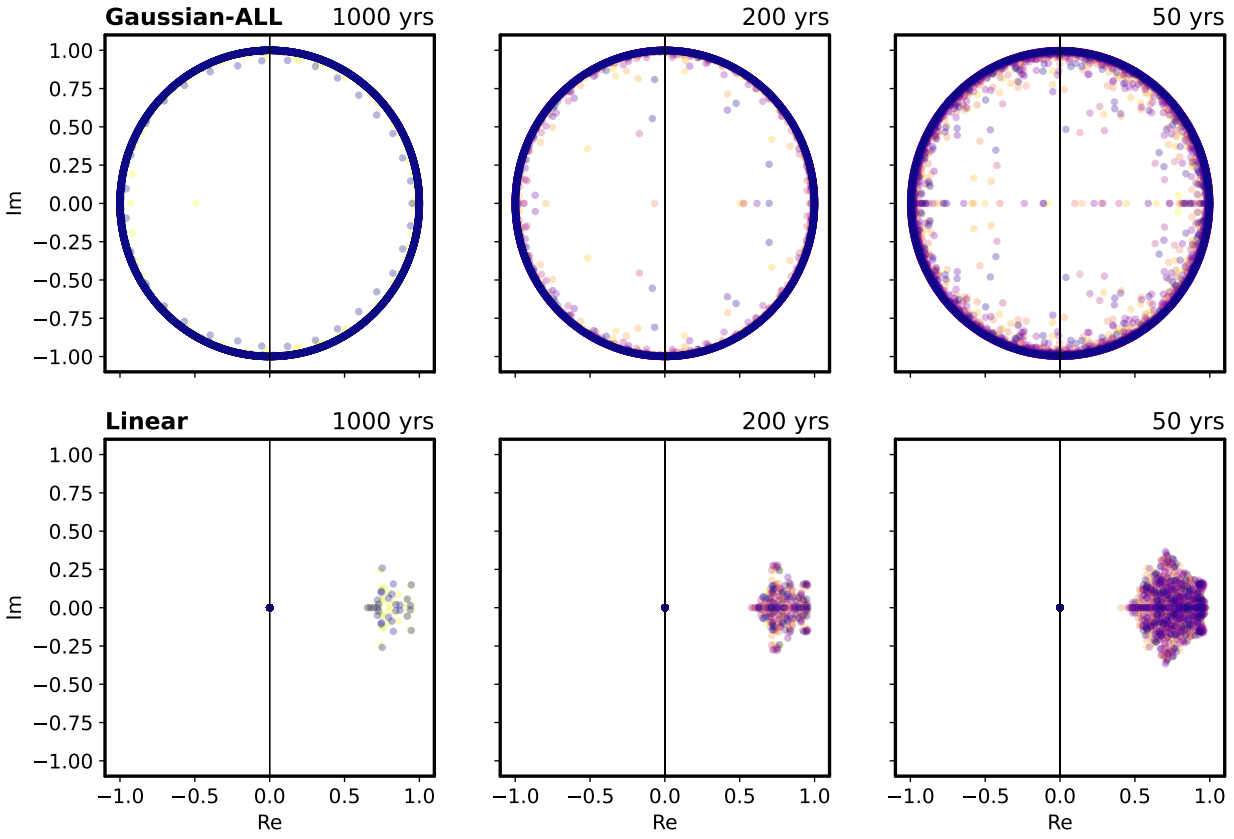

FIG. 2. Composite eigenvalue spectrum for different lengths of the operators (1000, 200 and 50 years). The first row shows the eigenvalues retrieved using the Gaussian kernel and all EOFs (Gaussian-ALL), while the second the ones obtained using the linear kernel. Each color represents the spectrum of a different sub-sample (i.e. yellow and purple for each one of the two 1000 years samples, first column). Note that total number of dots (and therefore eigenvalues) in each panel is the same. However, in the case of the linear kernel, just 30 of them per operator are nontrivial (different from zero).

mean differences across the different chunks of data, as the used EOFs have been computed for the total data series.

The Gaussian KEFs show particularly notable improvements, with consistent enhancements over the linear case across all lead times and operator lengths. For the best-performing KEFs ($\mathbf{K}_{200}$ and $\mathbf{K}_{400}$), *ACC* gains of up to 0.15 are observed around the 15-month lead time. Additionally, the distribution of *ACC* behavior varies depending on operator length, with $\mathbf{K}_{400}$, $\mathbf{K}_{200}$, and $\mathbf{K}_{100}$ KEFs showing similar *ACC* values across all lead times and outperforming the $\mathbf{K}_{1000}$ and $\mathbf{K}_{50}$ KEFs. This suggests the presence of an optimal-skill pseudo plateau for certain operator lengths, where performance remains nearly constant.

To determine how operator length affects forecast skill, we compare the *ACC* of the KEF mean and ensemble members at a 12-month lead time. Figure 5 displays the ACC of the KEF mean and members at the 12 month lead time for different operator lengths (50, 100, 200, 400, and 1000 years). The orange lines represent the skill of the KEF members, which increase with the length of the operators. Namely, the Gaussian KEF members computed over 1000 years of data perform better than those computed with 400, followed by the 200-years ones and so on, indicating that single operators perform better when longer records are used for their computation. On the other hand, the Gaussian KEF means (solid and dashed blue lines) show a fundamentally different behaviour, reflecting an optimal point surrounded by a pseudo-plateau between the 100-400 years of operator length (for the Gaussian-ALL case). This behaviour is similarly observed for other lead times (9 and 15 months, see Supplemental Material Figure 4), determining slightly better results for the $\mathbf{K}_{200}$ KEFs. Therefore, the $\mathbf{K}_{200}$ forecasts will be used as reference from now on (if not stated differently). It is important to note, though, that the location of the optimal record length will depend on the length of the total available data. For the 2000 years that CESM2 provides, it is located around the 100-400 years.

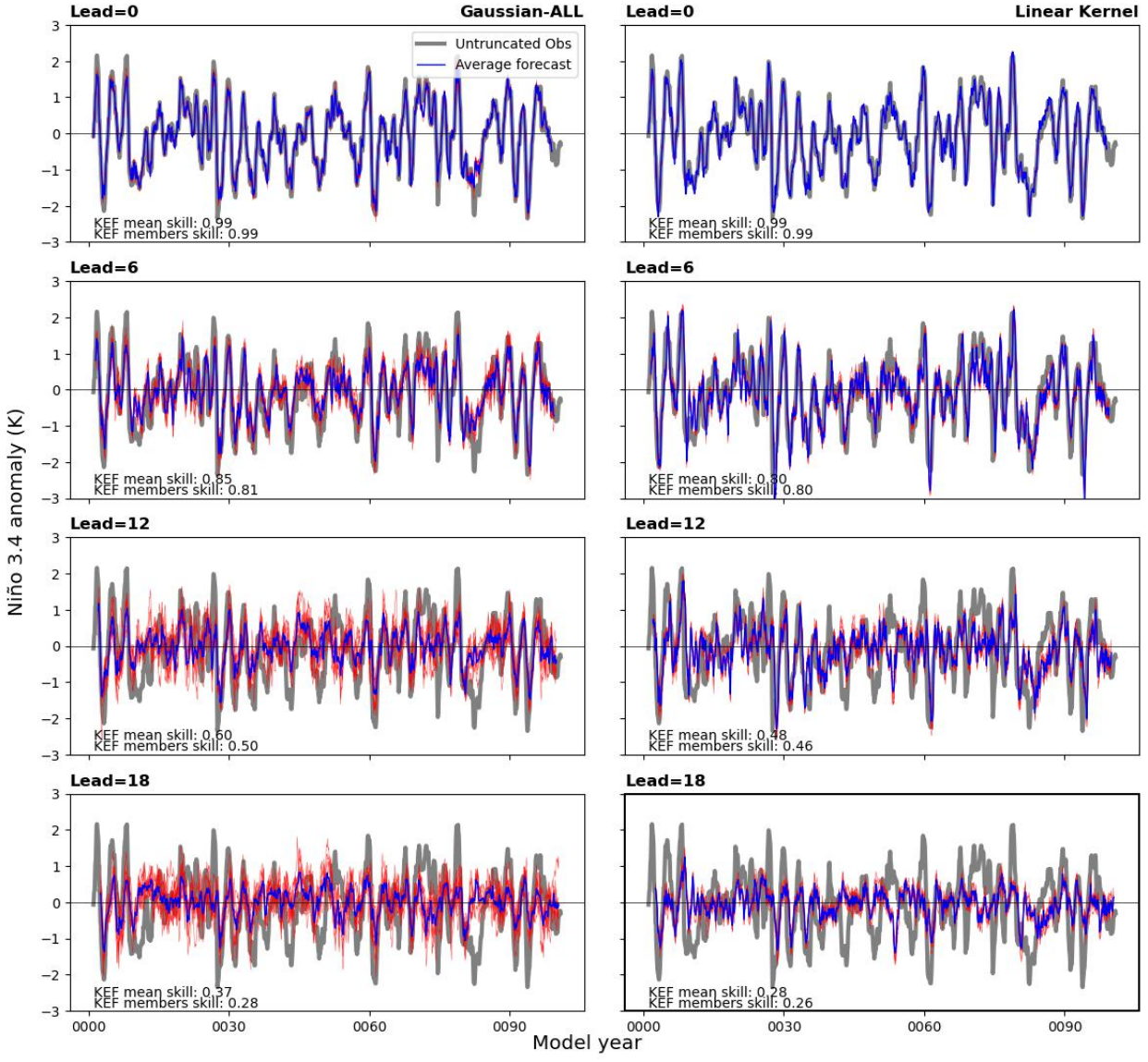

FIG. 3. Niño 3.4 index example of the linear and Gaussian-ALL KEF made with the $\mathbf{K}_{200}$ (200-years operators) for the first 100 years of the CESM data. The grey lines are the model observations, while the red thin lines represent each one of the ensemble members and the blue thick line is the ensemble mean (namely the average of all the red lines). The KEF mean and members ACC, where the KEF member ACC is the averaged ACC of each one of the KEF members, are present in the lower left corner of each panel.

The location of that optimal length for shorter training records is be investigated in Section 8.

6. SST spatial correlation patterns

While the ensemble forecasting approach improves the skill of Niño 3.4 predictions, it is also important to assess the spatial performance of the forecasts across the tropical Pacific. To understand how the choice of kernel affects regional forecast skill, we compare the spatial distributions of *ACC* between the Gaussian-ALL and linear KEF means.

The left panel of Figure 6 presents the spatial distribution of the Gaussian-ALL KEF means for the $\mathbf{K}_{200}$ operator, while the right one shows the *ACC* difference between the Gaussian-ALL and linear KEF means. The solid and dashed gray contours represent the 0.99 and 0.95 confidence levels. The skill gain of the Gaussian-ALL kernel is consistent and widespread across the spatial domain, with large regions exceeding 0.1. Strong skill gains are present around the equatorial belt, especially

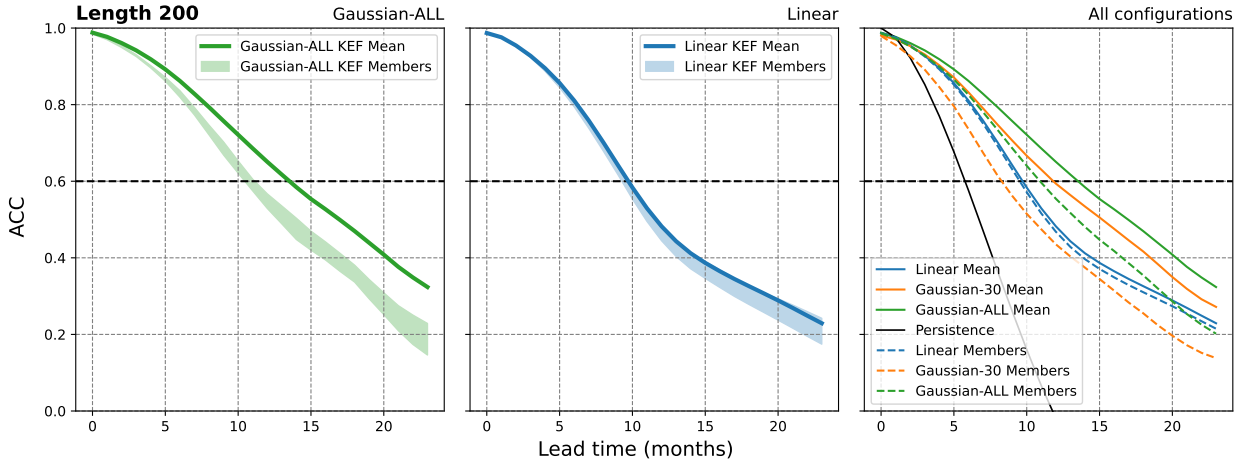

FIG. 4. Left and middle panels: spread of the Niño 3.4 index ACC of the cross-validated ensemble members (light shading) and the ensemble mean of the forecasts (thick lines) for the $\mathbf{K}_{200}$ operators using the Gaussian-ALL (left) and linear (middle) configurations. Right panel: Niño 3.4 index ACC of the Koopman ensemble mean and members for the three studied set ups (linear, Gaussian-30 and Gaussian-ALL) using the $\mathbf{K}_{200}$ operators. The plotted skill for the ensemble members is the mean skill of all the ensemble member forecasts.

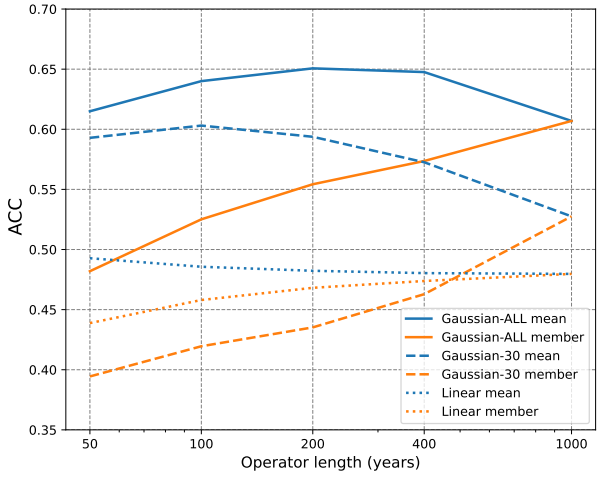

FIG. 5. ACC of the KEF mean and members at the 12 month lead time for the different lengths of the operators used (50, 100, 200, 400 and 1000 yrs). The plotted skill for the KEF members is the mean skill of all the KEF member forecasts.

at longer lead times, where *ACC* gains of up to 0.3 are present. In addition, arrow-shaped gains of up to 0.2 are present in the western Pacific area including the Western Pacific Warm Pool (WPWP), typically known for the poor performance of both GCMs and data-driven traditional methods (Lin et al. 2023). These features hold for the $\mathbf{K}_{100}$ and $\mathbf{K}_{400}$ forecasts, while the $\mathbf{K}_{1000}$ and $\mathbf{K}_{50}$ ones present much poorer improvements (not shown).

Regarding the skill of the entire SST anomaly, Figure 7 shows the spatial pattern correlation as a function of lead time. It has been computed by flattening the spatial SST anomaly fields into vectors and then calculating the Pearson correlation coefficient between the forecasted and observed vectors at each lead time. This provides a measure of the similarity between the spatial structures of the forecasted and observed SST anomalies over time. The solid lines represent the KEF mean *ACC*, while the dashed ones feature the averaged *ACC* of the KEF members. The improvements of the Gaussian KEF mean over the linear counterpart are consistent over the lead time range, but particularly relevant for long lead times. For instance, the Gaussian KEF mean skill exceeds 0.4 correlation for three months longer than the linear one, demonstrating that our approach also improves SST pattern predictions.

7. El Niño/La Niña events

There are numerous long term ENSO events whose evolution is better captured by the Gaussian kernel. Notably, many of them are among the strongest El Niño/La Niña events through the model data. Figure 8 shows composite Hovmoller diagrams for the development of the 1% strongest La Niña events (peak taken as the 240 strongest negative monthly Niño 3.4 index cases). The left panel shows the model observations starting 12 months before the peak of the events. The middle panel show the Gaussian-ALL KEF mean initialized 12 months before the peak of the events, and the right one the difference between the Gaussian-ALL and linear KEF means. The color shadings represent SST anomalies, and the contours SSH ones. Note that two different temperature and altitude scales have been chosen (for the observations and forecast, respectively) to better display the long-lead forecasts

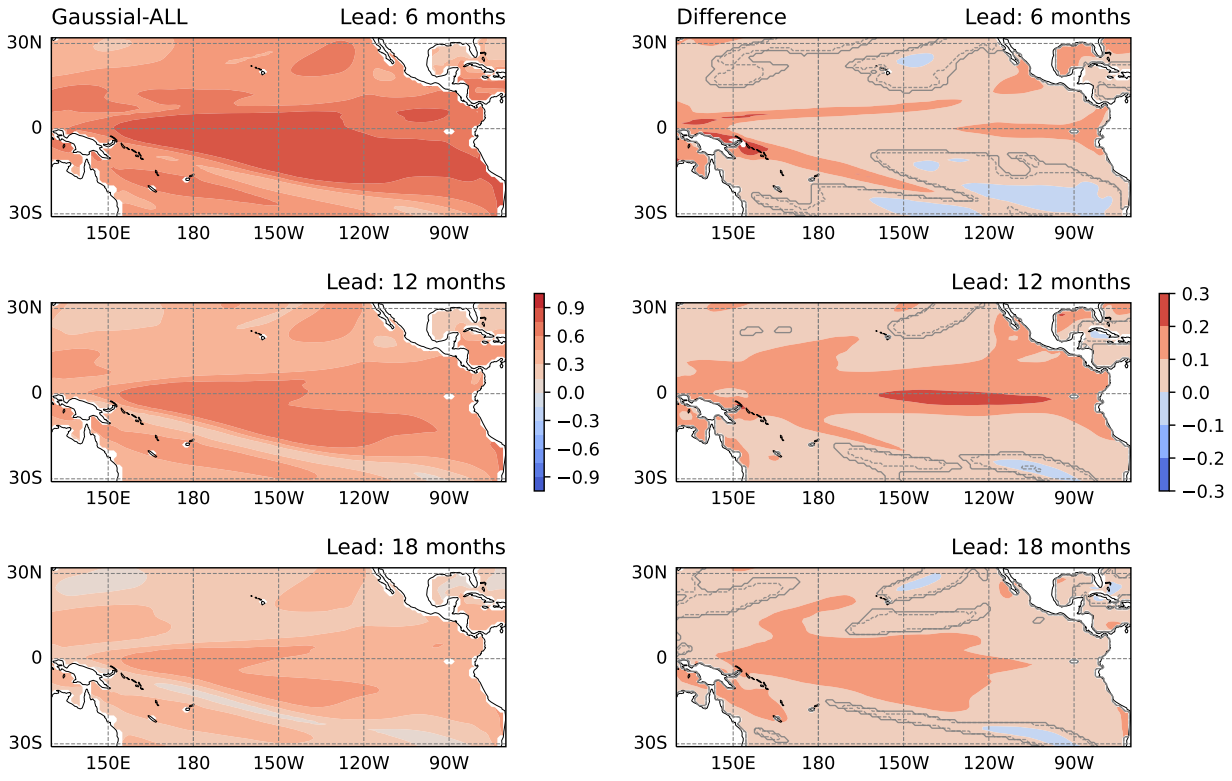

FIG. 6. Left panel: Spatial distribution of the Gaussian-ALL KEF mean ACC for the $\mathbf{K}_{200}$ operators. Right panel: Same as the left panel but for the difference between the Gaussian-ALL and linear KEF means. Solid gray contours represent the 0.99 confidence level, while the lighter dashed contour represents the 0.95 one.

which present lower amplitudes than the observations.

The Gaussian KEF in Figure 8 demonstrates better skill representing the dynamics of extreme La Niña cases, capturing key features that the linear one struggles with. In particular, the Gaussian KEF represents better the La Niña peak, including the spatial distribution of stronger negative anomalies over the western Pacific. A more detailed analysis considering the entire domain (the Hovmoller diagrams comprise only the 5ºN-5ºS band) shows pattern correlations between the predicted and observed SST composites above 0.6 for leads up to 15 months, as well as a better representations of the extension of La Niña into the tropical regions (see Supplemental Material Figure 5). Also the amplitude of the event is larger in the Gaussian KEF, better matching the model observations. The linear KEF reproduces much weaker and very equatorially confined La Niña events (also in Supplemental Material Figure 5), with no temperature differences between the eastern and western Pacific. The evolution of the SST anomalies is also better represented with the Gaussian KEF, with negative anomalies starting 9 months before the peak (lead time 3) that progressively grow into a western peak, matching the model observations. In contrast, the linear forecast struggles with timely depiction of the SST negative anomalies, as well as with the ENSO discharge. The difference patterns indicate that the Gaussian approach better captures the strengthening and decay of positive SST anomalies, suggesting an improved representation of the state-dependent growth and decay processes characteristic of ENSO.

In contrast, both approaches have more difficulty predicting the 1% strongest El Niño events (Figure 9). Although the full composite SST fields still show some improvements of the Gaussian kernel over the linear one (0.61/0.50 pattern ACC for the Gaussian and linear kernels at the peak, respectively), the overall ACCs are much lower than for La Niña cases (see Supplemental Material Figure 6). Figure 9 shows that both kernels tend to predict the peak of ENSO events too early, a common limitation in statistical and dynamical forecasts, and they are shifted well to the west. However, the Gaussian model exhibits a reduced bias, delaying the peak less compared to the linear model and better aligning with observations. Additionally, the Gaussian-based model shows noticeable improvements in representing SST anomalies in the western Pacific warm pool region, where nonlinear interactions between oceanic and atmospheric processes play a crucial role in ENSO dynamics. The enhanced representation of

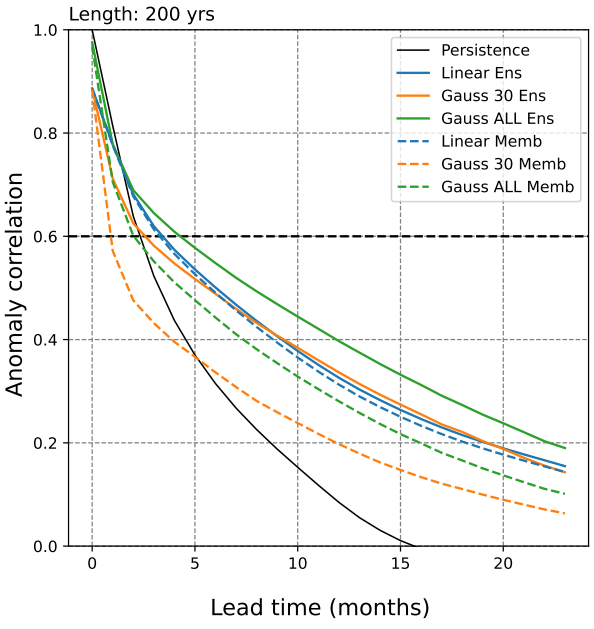

FIG. 7. Spatial anomaly correlation of the forecasted Pacific SST using the linear and Gaussian-ALL configurations for the $\mathbf{K}_{200}$ operators. Solid lines represent the KEF mean correlations, while dotted ones the average ACC of the KEF members.

these processes suggests that the Gaussian kernel captures important nonlinear feedbacks that are relevant for accurately modeling ENSO evolution. These errors are considerably smaller when the forecasts are initialized closer to the peak, especially for the Gaussian kernel. Gaussian forecasts initialized 6 months before the peak show a much better depiction of the development of ENSO, as well as a noticeable improvement of the skill over the WPWP, with reduced overestimation of SSTs during warm events.

Overall, the Gaussian KEFs represent better the evolution of extreme ENSO events, but the amount of improvement depends on the sign of the anomaly and the lead time of the event. Long lead improvements (up to 15 months) are present for La Niña extreme events, in agreement with the findings that long lead La Niña events coming out of El Niño are better credited than other kind of ENSO events (Lenssen et al. 2024). In contrast, the horizon for improved predictions is much shorter for El Niño events. The underlying reasons for this asymmetry require further investigation.

In order to assess the accuracy of our El Niño and La Niña forecasts we use reliability diagrams (Fig.10), which are graphical tools used to evaluate probabilistic forecasting systems by comparing predicted probabilities to observed frequencies. The orange and blue dots represent the observed frequencies of El Niño and La Niña events depending on its forecasted probability (from zero to one, divided in 0.2 bins). The thresholds used for the classification of the observed frequency and forecast's probability have been chosen as the top and bottom terciles of the Niño 3.4 model observations. The diagonal line from the bottom-left to the top-right represents perfect reliability. Points above the diagonal indicate underconfidence, where the predicted probability is lower than the observed frequency, while points below the diagonal indicate overconfidence, with predicted probabilities higher than the observed frequencies. A histogram showing the distribution of the forecast probabilities is included at the bottom, helping to visualize the number of cases used for each reliability point. The vertical error bars represent the margin of error (MoE), and they indicate the range within which the true observed frequency is likely to fall given the sample size and the data variability. Bins with very small sample sizes present large margins of error, while they become negligible in bins with large populations. Finally, the horizontal red and blue dotted lines represent the climatological frequency of El Niño and La Niña events. It is important to note that in our application, these diagrams evaluate the probabilistic forecasts coming from KEFs, meaning that they represent the uncertainty arising from the operator's estimations, and not from the initial condition uncertainty.

The diagrams have been evaluated at the 0, 6, 12 and 18 month lead times, with both kernels showing decreasing reliability with increasing lead times. The reliability points become increasingly apart from the diagonal line as the lead time increases, with low probability forecasts becoming underconfident and high probability ones becoming increasingly overconfident. However, the Gaussian kernel consistently outperforms the linear one in terms of reliability, as evidenced by the closer alignment of its forecast probabilities with observed frequencies starting from the 6 month lead time. This is particularly notable at longer lead times, where the linear kernel approaches the climatology lines much faster than the Gaussian ones. In fact, the Gaussian kernel yields very reliable forecasts, particularly for the 6 and 12 months. These results hold both for the El Niño and La Niña events, but La Niña events present slightly higher reliability all across the board, both for the linear and Gaussian kernels, following the results shown in Figures 8 and 9, but also state-of-the-art studies stating that the transitions to La Niña are more predictable than those to El Niño (Sharmila 2023).

The histograms accompanying the reliability diagrams illustrate the distribution of forecast probabilities. The Gaussian KEF exhibits a broader spread of forecast probabilities, whereas the linear kernel produces a more concentrated distribution, with most cases belonging to the lowest (0-0.2) and highest (0.8-1) bins. This can be explained in terms of sharpness and reliability. Following the precedent discussions, the forecasts made with the linear kernel are

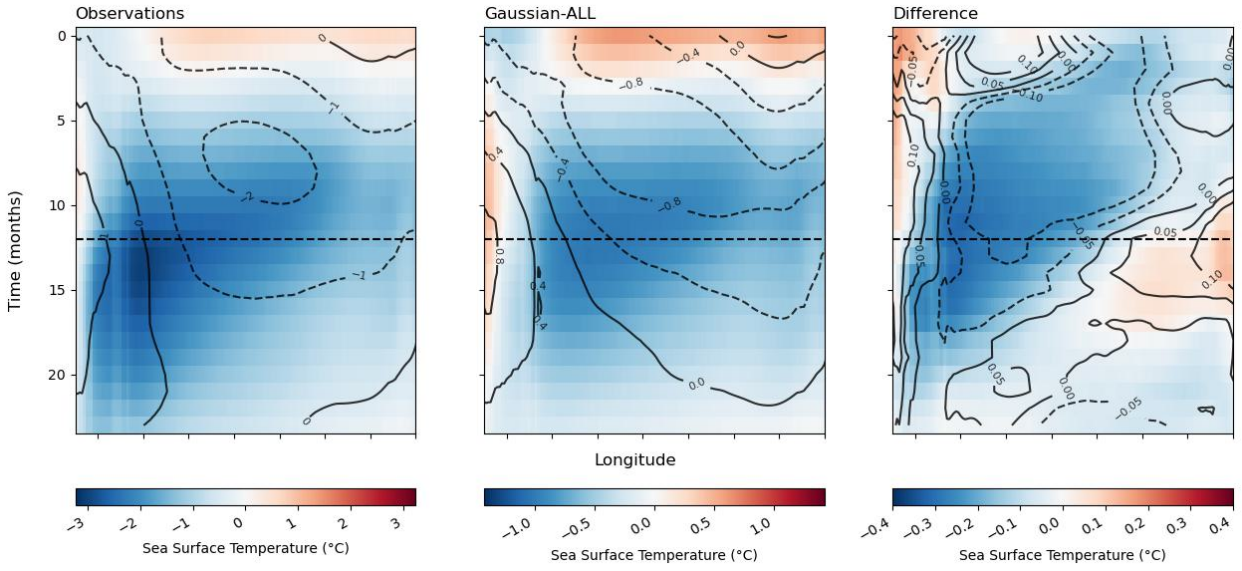

FIG. 8. Hovmoller diagram of the SST and SSH composites of the development of the 1% strongest La Niña events (peak taken as the 240 lowest monthly Niño 3.4 index cases). The color shadings represent SST, units in Kelvin degrees. The solid and dashed contours, the positive and negative SSH values, units in meters. The latitude averages have been taken over the -5 to 5 degrees north and south. The left panel shows the model observations starting from 12 months before the peak of the event. The middle one shows the forecasts made with the $\mathbf{K}_{200}$ Gaussian-ALL kernels, and the right one the different between the Gaussial-ALL and the linear one. The forecasts are initialized 12 months before the peak of the event (month zero), and the peak is indicated by the horizontal dashed line.

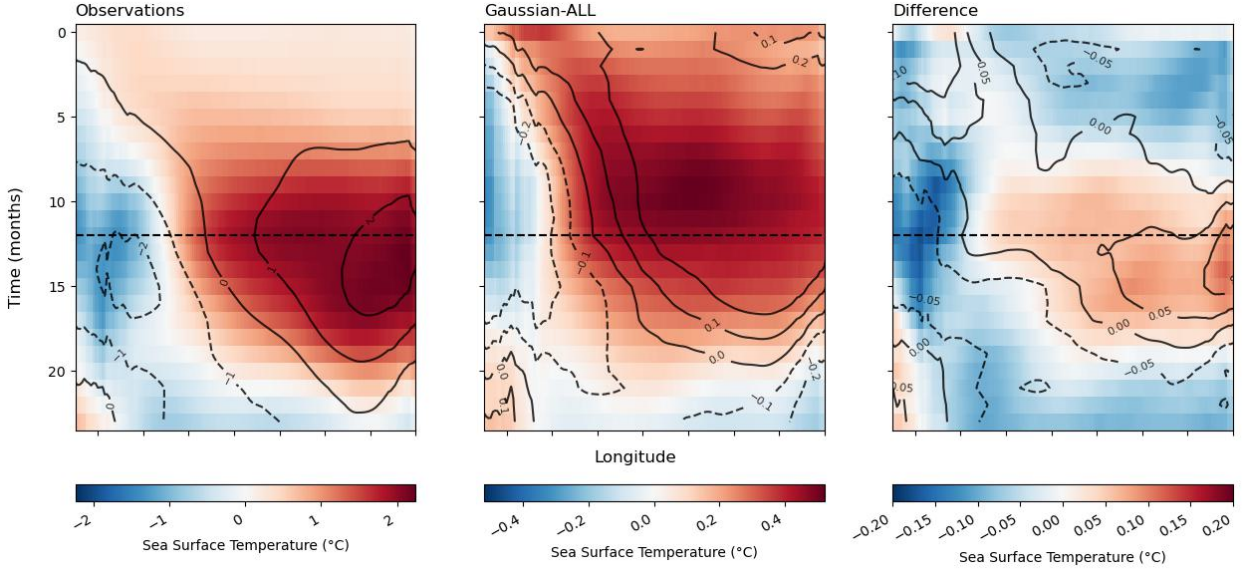

FIG. 9. As Figure 8, but for the 1% strongest El Niño events (peak taken as the 240 highest monthly Niño 3.4 index cases).

more robust, leading to sharper probabilistic distributions but less reliable forecasts. On the other hand, the Gaussian forecasts are less sharp, but present more reliable forecasts. This difference is particularly meaningful at the 0 lead time. The linear kernel does not present any case outside the 0 and 1 probabilities, meaning that all the forecast agree in their ENSO outcome. The Gaussian kernel, though, does not yield a perfect reconstruction of the initial conditions, leading to a small fraction of events presenting mixed forecasts (probabilities different from 0 or 1). As sample size in these bins is very small, the MoE bars are very large, indicating high uncertainty in the reliability.

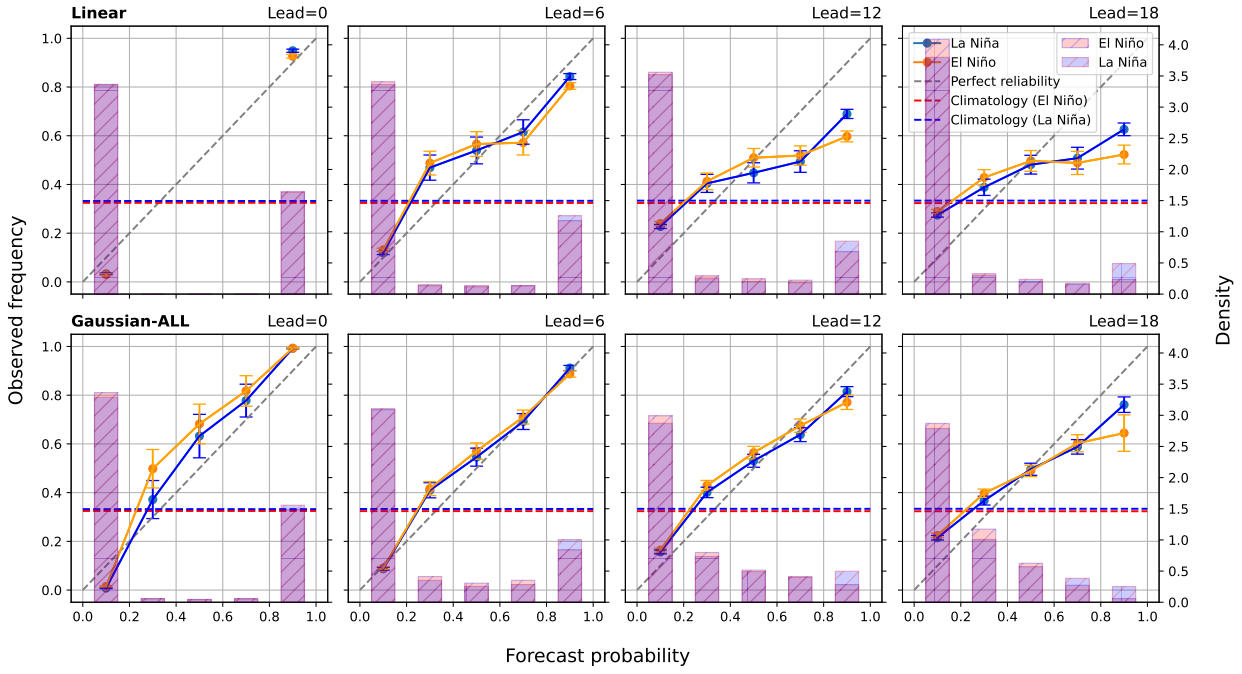

FIG. 10. Reliability diagram of the Niño 3.4 forecasts made with the $\mathbf{K}_{200}$ using the linear (top row) and Gaussian-ALL (bottom row) kernel configurations for the 0, 6, 12 and 18 months lead time. The orange and blue dots represent, respectively, the observed frequency of El Niño and La Niña events depending on the event's forecasted probability. The threshold used for the classification of both the observed frequency and the forecast's probability has been chosen as the top and bottom terciles of the model observations data. The error bars have been computed using the margin of error (MoE). The horizontal red and blue dashed lines represent the climatological frequency of El Niño and La Niña events, and the histograms the distribution of cases belonging to each probability bin.

However, and even when taking the MoE into account, the Gaussian kernel presents underconfident representation of both the El Niño and La Niña events when reconstructing the initial conditions. The initial bias gradually dissipates over time, demonstrating enhanced reliability as early as six months into the forecast horizon. It is important to note, though, that these reliability diagrams have been computed using the same threshold for the observed frequency and the forecast probability. If different terciles had been considered (one coming from the observed data and one from the forecasts, as usually done when computing reliability diagrams) this bias at zero lead time would not appear.

8. Application to shorter data records

To better match the length of the available observed data records, we have repeated the analyses of Section 5 restricting the data to 150 years in length. The Koopman operators are now computed from subsamples with new lengths of 15, 25, 30, 50 and 75 years. This process was carried out for all the possible non-overlapping samples of data (leaving outside the last 50 years), and the results averaged in a cross-validated way. Following previous results, the Gaussian KEF means outperform the linear ones, especially for the lead times ranging between 9 and 18 months. The results for the Niño 3.4 ACC at the 12 month lead time are presented in the right panel of Figure 11 (solid lines). The results reflect similar properties to Figure 5, with an optimal distribution of the data that presents a pseudo-plateau. However, the best performances are located between the $\mathbf{K}_{25}$ and $\mathbf{K}_{50}$ KEFs, in contrast with the $\mathbf{K}_{100}$/$\mathbf{K}_{400}$ range of the 2000 years long total series. As expected, the position of the peak has been relocated depending on the length of the total available data, as well as the magnitude of the skill gain.

The dependence of the skill gain on the total available data is featured in the left panel of Figure 11, where the Niño 3.4 ACCs of the best Gaussian KEF means are presented for 4 different lengths of total available data (150, 200, 400 and 2000 yrs, solid lines). The coloured dots indicate the intersection points of the forecasts skill with the 0.6 ACC line. The best performing gaussian KEF for the total 2000 years of available data is $\mathbf{K}_{200}$, while it is $\mathbf{K}_{30}$ for 150 years of total available data, $\mathbf{K}_{50}$ for 200 and $\mathbf{K}_{100}$ for the 400 years one. The skill gain of the Gaussian KEF over the linear one is proportional to the total data record too. For the total length of the data (2000 years), the best Gaussian KEF hits the 0.6 ACC mark more than 4 months later than the linear one. This is reduced to 3 months for a data record of 400 years, slightly less than 2

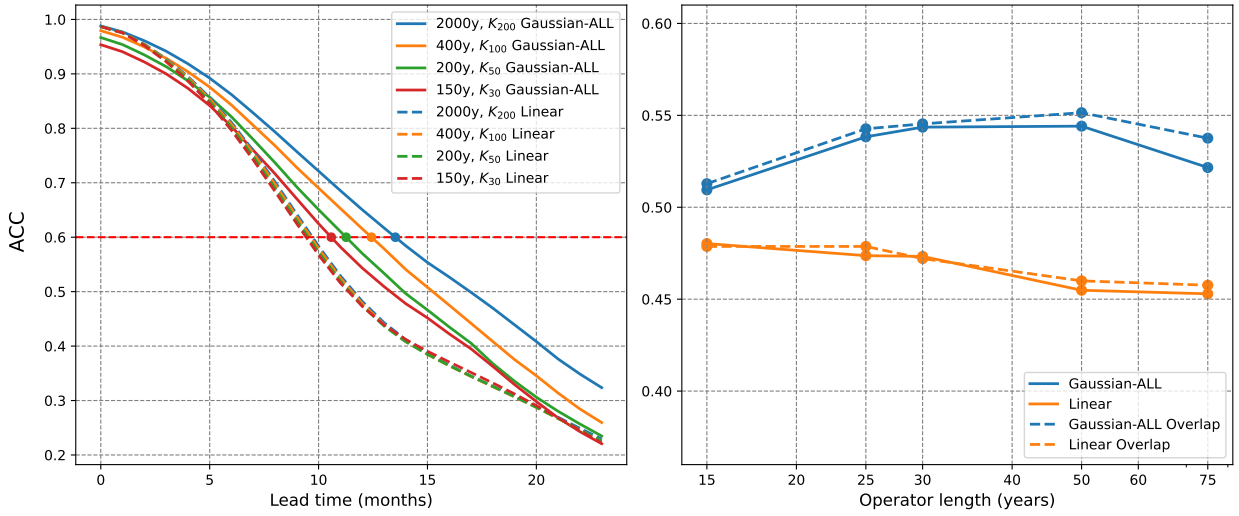

FIG. 11. Left panel: Niño 3.4 index anomaly correlation of the best forecast ensemble mean using the Gaussian-ALL configuration for 4 different lengths of total available data (150, 200, 400 and 2000 yrs). The forecasts have been computed using total sub-samples of those lengths. The ACC of each period has been subsequently averaged to be represented in this figure. Solid lines and dashed lines represent, respectively, the Niño 3.4 ACC computed with the Gaussian-ALL and linear kernels. The coloured dots indicate the point where each forecast intersects the 0.6 ACC line. Right panel: ACC of the KEF mean at the 12 month lead time for the different lengths of the operators (15, 25, 30, 50 and 75 yrs) computed over a total available period of 150 years. The solid and dashed lines represent, respectively, the linear and Gaussian-all KEF mean using non-overlapping operators. The dashed lines represent the KEF mean computed adding (n/2)-1 overlapping operators shifted half the length of the original ones, where n is the number of original non-overlapping operators.

months for 200 years and around just one month for the 150 years time series, reaching the 0.6 ACC mark around the 11 months lead mark.

In addition, there is a trade-off between the number of members and the length of the operators used for the forecast. We are interested into exploring if it could be possible to improve the forecasts skill by increasing the number of the ensemble members while maintaining their length. This can be done adding overlapping operators, specially for the $\mathbf{K}_{50}$ and $\mathbf{K}_{75}$ forecasts, where the number of members is very low. Therefore, (n/2)-1 overlapping operators have been computed by shifting the starting time-step half the length of the original ones, where n is the number of original non-overlapping operators. For example, the $\mathbf{K}_{75}$ KEF will have 3 members instead of 2; the $\mathbf{K}_{50}$ KEF will have 5 instead of 3, and so on. The overlapping KEF ACC at lead time 12 is represented by the dashed lines in the left panel of Figure 11. Although the ACC gains are limited, there are small improvements, particularly for the smaller KEFs, namely those with longer operators: $\mathbf{K}_{50}$ and $\mathbf{K}_{75}$. In fact, after adding the overlapping operators, the Gaussian $\mathbf{K}_{50}$ slightly outperforms the $\mathbf{K}_{30}$ and $\mathbf{K}_{25}$ forecasts. Contrarily, little or no gain is added by the overlapping operators to the shorter Gaussian KEFs, neither to the linear ones. Overall, and even if the the skill gain is reduced, adding overlapping operators can be a tool for improving the KEFs skill when we face reduced data availability.

9. Conclusions

A key finding of this study is the strong sensitivity of estimated Koopman operators to data availability and record length. With Gaussian kernels, even 1000-year records exhibit robustness issues, and still shorter ones lead to increased uncertainty and reduced forecast robustness. This poses a challenge to the use of the Koopman framework for climate forecasting, particularly given the limited length of the observational records. To address this problem, we introduce Koopman Ensemble Forecasts, which employ ensemble techniques to mitigate sampling uncertainty and improve forecast robustness. Results indicate that KEF means consistently outperform individual Koopman forecasts, particularly for longer lead times. In addition, by methodically sub-sampling the data record, we gained insight into the optimal approach for maximizing forecast skill depending on the total length of available data, demonstrating the existence of an optimal point with a pseudo-plateau around it.

Furthermore, the spatial analysis of Pacific SST forecasts reveals significant skill gains when using the Gaussian kernel, especially in regions known for their challenging predictability, such as the WPWP. Although these results have been empirically retrieved from a perfect model, and might not perfectly reflect the dynamics of the real-world ENSO system, they hint the ability of the Koopman nonlinear methods in capturing complex SST patterns.

In addition, the evaluation of El Niño and La Niña event forecasts through reliability diagrams highlights the higher performance of the Gaussian kernel, particularly from the six-month lead time onward. Also the study of the development of extreme events yields notable results. In particular, asymmetries between positive and negative phases of ENSO are successfully represented within the Gaussian kernel, facing one of the major drawbacks that linear models present for ENSO modelling. This leads to better dynamical representation, particularly in the case of La Niña events, resembling more accurately both the spatial distribution of SSTs and the timeline of the phenomena. In addition, the forecasts keep their skill up to long lead times and seem to overcome the spring predictability barrier, which are two of the main challenges that most-commonly used non-linear data-driven methods present for ENSO forecasting (Wang et al. 2023; Fang et al. 2022). However, both kernels exhibit struggles for the extreme El Niño events representation, especially for longer lead times (from 6 months ahead the peak). The reasons for the notably different behaviour between positive and negative extreme phases would need further investigation.

Some additional challenges still need to be faced, in particular regarding the method's applicability to operational forecasting settings. The results from this study come from the analysis of CESM2 data. Although some of the main behaviors might hold for observations, we cannot assume that the optimal points and skill levels found will hold for observational records, where the dynamics could be more complex and both forcing and observational biases need to be taken into account. Moreover, in order to face our method with the state-of-the-art LIM frameworks, we would need to compare it to models such as the cyclo-stationary LIMs, allowing us to assess to which extent do our skill improvements come from purely nonlinear dynamics and which can be attributed to the explicit inclusion of the seasonal cycle instead. In addition, the comparison to other nonlinear kernels might be an interesting topic of study, exploring if they could help overcoming some of the limitations found here. Finally, but not less important, there are intrinsic challenges in the determination of robust estimates of the Koopman operator, particularly in shorter data spans, signaling avenues for further refinement of both the algorithms and forecasting frameworks.

Overall, this study helps to frame some of the potential features of Koopman theory for advancing ENSO forecasting. By addressing the limitations of traditional linear models and introducing nonlinear dynamics, Koopman-based approaches offer a pathway towards more accurate and reliable climate predictions, with significant implications for several societal sectors dependent on climate information. A key focus of our work is the fundamental question of data sensitivity—how much data is actually needed to obtain robust Koopman operator estimates? We show that employing ensemble techniques can effectively help to mitigate the uncertainty associated with data limitations when using nonlinear Koopman estimations, yielding enhanced forecast performances compared to the linear benchmark. Moreover, we present an approach to determine the most efficient way to employ the available data in order to optimize the ensemble's skill. Yet, multiple challenges remain, particularly regarding the impact of sampling uncertainty in shorter data spans, warranting continued research in this topic.

Acknowledgments. We thank our internal NOAA reviewer, Cécile Pendland, as well as Dimitrios Giannakis, as their feedback and insights have helped to notably improve the quality of this manuscript. In addition, ACS would like to acknowledge support from NOAA grants (NA22OAR4310599 and NA21OAR4310253) and NASA grant (21-OSST21-0026).

Data availability statement. Data is available from the authors. Code has been uploaded to https://github.com/Paulalo95/Koopman-Ensemble-Forecast

APPENDIX A

k-EDMD algorithm for Koopman Spectrum Estimation

It was not until the last decade when Williams et al. (2015) and Klus et al. (2018, 2019) demonstrated that the eigenvalues and eigenmodes of transfer operators can be estimated from data using EDMD-type algorithms. Assuming we have a dataset that describes the temporal evolution of a dynamical system, we can organize the data as follows

$$\mathbf{X} = \left[x_1, x_2, \cdots, x_m\right] \tag{A1}$$

where the vectors x_i of length n describe grid-point values of representative physical fields of the system at m different time-steps, which are lagged by a discrete time interval of τ. In this way, the discrete data represents the sampled evolution of the system according to the unknown dynamical operator $\mathbf{F}$. This organization of the data is similar to the one used in some traditional methods, such as LIMs or singular spectra analysis, where various forms of regression are used involving time covariance matrices ($\mathbf{C}_{xx} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$) of the data at different lag times Penland and Sardeshmukh (1995).

For the Koopman framework we adopt a slightly different approach. Instead of using covariance matrices we use Gramian matrices

$$\mathbf{G}_{xx} = [k(\mathbf{x}_i, \mathbf{x}_j)]$$

$$\mathbf{G}_{yx} = [k(\mathbf{y}_i, \mathbf{x}_j)]$$

where k is the chosen feature kernel and y_i are the components of the lagged vector:

$$\mathbf{Y} = \begin{bmatrix} y_1, y_2, \dots & y_m \end{bmatrix} = \begin{bmatrix} x_2, x_3, \dots & x_{m+1} \end{bmatrix}$$

The Gramian matrices can be used to set up the following eigenvalue problem

$$(\mathbf{G}_{xx} + n\epsilon I)^{-1}\mathbf{G}_{yx}\mathbf{w} = \mu\mathbf{w} \tag{A2}$$

where $(\mathbf{G}_{xx} + n\epsilon I)^{-1}\mathbf{G}_{yx}$ is known as the auxiliary matrix A and ϵ is a regularization parameter added to ensure that the inverse exist, particularly when dealing with finite datasets or highly correlated data points. The choice of ϵ is crucial for balancing numerical stability and spectral accuracy. A too-small ϵ can lead to near-singular matrices, causing numerical instability in eigenvalue computations. Conversely, a too-large ϵ can excessively dampen small eigenvalues, distorting the spectral properties of the Koopman operator. For this work, a ϵ of 10^8 has been chosen, with no relevant changes when ranging it from 10^6 to 10^{10}.

The eigenvalues of the auxiliary matrix (μ) are the estimated eigenvalues of the Koopman operator, while its eigenvectors $\mathbf{w}$ can be used to determine the values taken by the eigenfunctions at the data points $\Phi(X)$ Klus et al. (2019):

$$\Phi(X) = \mathbf{G}_{xx}\mathbf{W}, \tag{A3}$$

where the columns of $\mathbf{W}$ are the eigenvectors $\mathbf{w}$.

The outputs of the algorithm are the estimated Koopman eigenvalues μ, the matrix $\Phi(X)$ of the values taken by the eigenfunctions $\phi_i(\mathbf{x}_j)$ at the sample points

$$\Phi(X) = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \phi_3(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots \\ \phi_1(\mathbf{x}_3) & \dots & \dots \end{bmatrix} \tag{A4}$$

and the coefficients w_{ik}, which are used for the expansion of the eigenfunctions in the feature functions derived from the kernel. With these outputs we can compute the Koopman modes. Having a vector-valued observable $\mathbf{g} = [g_1, g_2, g_3, \dots]^T$ that takes values on the sample points:

$$G(X) = \begin{bmatrix} g_1(\mathbf{x}_1) & g_1(\mathbf{x}_2) & g_1(\mathbf{x}_3) \\ g_2(\mathbf{x}_1) & g_2(\mathbf{x}_2) & \dots \\ g_3(\mathbf{x}_1) & \dots & \dots \end{bmatrix} \tag{A5}$$

the Koopman modes $\boldsymbol{V}$ of the state vectors are obtained by projecting the matrix of the observable's values on the eigenfunctions Mezic and Surana (2016):

$$V = \Phi(X)^{+}\mathbf{X}, \tag{A6}$$

where + denotes the pseudoinverse of a matrix: a generalization of the inverse matrix that can be applied to matrices that are not necessarily square or invertible, computing a "best fit" (least squares) approximate solution to a system of linear equations that lacks an exact solution.

The estimated evolution of the dynamical system described by the data set is then given by (cf. Eq. 5):

$$\mathbf{x}(t) = \sum_k \mathbf{v}_k e^{(\sigma_k + i\omega_k)t}\phi_k(\mathbf{x}) \tag{A7}$$

$$= \sum_{ik} \mathbf{v}_k e^{(\sigma_k + i\omega_k)t} w_{ki} p_i(\mathbf{x}) \tag{A8}$$

It is important to notice that the choice of the kernel determines the class of functions used to approximate the Koopman operator. If the standard inner product is used, we restrict the function space to linear functions. In this case, the Gram matrices become $\mathbf{G}_{xx} = \langle \mathbf{x}_j, \mathbf{x}_i \rangle$ and $\mathbf{G}_{yx} = \langle \mathbf{y}_j, \mathbf{x}_i \rangle$, containing the same information as the covariance matrices used in LIMs. The eigenvalues and eigenfuctions found in this setting are equivalent to those retrieved by LIM algorithms Tu (2013); Tu et al. (2014); Navarra et al. (2021). However, the Gram and covariance matrices have different dimensions. The dimension of the covariance matrix is given by the spatial length of the data, while for the Gramian by the length of the time series. As a consequence, the Koopman operator framework produces a much higher number of eigenmodes. When using a linear kernel (and a truncation of the state space, which is usually needed), most part of them equal zero, yielding a final number of non-trivial modes equal to the spatial dimensionality of the system and collapsing to the LIM results.

References

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: Enso representation in climate models: From cmip3 to cmip5. *Clim Dyn*, **42(7-8)**, 1999– 2018.

Berry, T., D. Giannakis, and J. Harlim, 2015: Nonparametric forecasting of low-dimensional dynamical systems. *Physical Review E*, **91 (3)**, https://doi.org/10.1103/physreve.91.032915, URL http://dx.doi.org/10.1103/PhysRevE.91.032915.

Berry, T., and J. Harlim, 2016: Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, **40 (1)**, 68–96, https://doi.org/https://doi.org/10.1016/j.acha.2015.01.001, URL https://www.sciencedirect.com/science/article/pii/S1063520315000020.

Capotondi, A., C. Deser, A. Phillips, Y. Okumura, and S. M. Larson, 2020: Enso and pacific decadal variability in the community earth system model version 2. *Physical Review E*, **12 (12)**, URL https://doi.org/10.1029/2019ms002022.

Danabasoglu, G., and Coauthors, 2020: The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, **12 (2)**.

Fang, W., Y. Sha, and V. S. Sheng, 2022: Survey on the application of artificial intelligence in enso forecasting. *Mathematics*, **10 (20)**, https://doi.org/10.3390/math10203793, URL https://www.mdpi.com/2227-7390/10/20/3793.

Froyland, G., D. Giannakis, B. R. Lintner, M. Pike, and J. Slawinska, 2021: Spectral analysis of climate dynamics with operator-theoretic approaches. *Nat. Commun.*, **12 (1)**, 6570.

Klus, S., A. Bittracher, I. Schuster, and C. Schütte, 2018: A kernel-based approach to molecular conformation analysis. *The Journal of Chemical Physics*, **149 (24)**, 244 109, https://doi.org/10.1063/1.5063533, URL https://doi.org/10.1063/1.5063533, https://doi.org/10.1063/1.5063533.

Klus, S., I. Schuster, and K. Muandet, 2019: Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *J Nonlinear Sci*, https://doi.org/https://doi.org/10.1007/s00332-019-09574-z.

Koopman, B., 1931: Hamiltonian systems and transformation in hilbert space. *Proc. Natl. Acad. Sci. USA*, **10**, 315–318.

Koopman, B., and J. Neumann, 1932: Dynamical systems of continuous spectra. *Proc. Natl. Acad. Sci. USA*, **18**, 255–263.

Lenssen, N., P. DiNezio, L. Goddard, C. Deser, Y. Kushnir, S. J. Mason, M. Newman, and Y. Okumura, 2024: Strong El Niño events lead to robust Multi-Year ENSO predictability. *Geophysical Research Letters*, **51 (12)**, https://doi.org/10.1029/2023gl106988, URL https://doi.org/10.1029/2023gl106988.

L'Heureux, M., A. Levine, M. Newman, C. Ganter, J.-J. Luo, M. Tippett, , and T. Stockdale, 2020: *"Chapter 10: ENSO Prediction". AGU Monograph: El Niño-Southern Oscillation (ENSO) in a Changing Climate*. M. McPhaden, A. Santoso, W. Cai (Eds.), Wiley, 506 pp.

Lin, Y., L. Wang, and J. F. Li, 2023: Effects of equatorial ocean current bias on simulated el niño pattern in cmip6 models. *Geophysical Research Letters*, **50 (8)**, https://doi.org/doi:10.1029/2023gl102890.

Liu, X., A. Liu, J. L. Chen, and G. Li, 2023: Impact of decomposition on time series bagging forecasting performance. *Tourism Management*, **97**, 104 725, https://doi.org/https://doi.org/10.1016/j.tourman.2023.104725, URL https://www.sciencedirect.com/science/article/pii/S0261517723000079.

Mezić, I., 2013: Analysis of fluid flows via spectral properties of the koopman operator. *Annual Review of Fluid Mechanics*, **45 (1)**, 357–378, https://doi.org/10.1146/annurev-fluid-011212-140652.

Mezic, I., and A. Surana, 2016: Koopman mode decomposition for periodic/quasi-periodic time dependence. *IFAC-PapersOnLine*, **49 (18)**, 690–697, https://doi.org/https://doi.org/10.1016/j.ifacol.2016.10.246, URL https://www.sciencedirect.com/science/article/pii/S2405896316318262, 10th IFAC Symposium on Nonlinear Control Systems NOLCOS 2016.

Mezić, I., 2005: Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, **41**, https://doi.org/https://doi.org/10.1007/s11071-005-2824-x.

Mu, B., C. Peng, S. Yuan, and L. Chen, 2019: Enso forecasting over multiple time horizons using convlstm network and rolling mechanism. *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8, https://doi.org/10.1109/IJCNN.2019.8851967.

Muandet, K., K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, 2017: Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, **10 (1-2)**, 1–141, https://doi.org/10.1561/2200000060, URL http://dx.doi.org/10.1561/2200000060.

Navarra, A., J. Tribbia, and S. Klus, 2021: Estimation of koopman transfer operators for the equatorial pacific sst. *Journal of the Atmospheric Sciences*, **78(4)**, 1227–1244.

Navarra, A., J. Tribbia, S. Klus, and P. Lorenzo-Sánchez, 2024: Variability of sst through koopman modes. *Journal of Climate*, https://doi.org/10.1175/JCLI-D-23-0335.1, URL https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-23-0335.1/JCLI-D-23-0335.1.xml.

Newman, M., M. Alexander, and J. Scott, 2011: An empirical model of tropical ocean dynamics. *Clim Dyn*, **37**, 1823–1841.

Newman, M., and P. D. Sardeshmukh, 2017: Are we near the predictability limit of tropical indo-pacific sea surface temperatures? *Geophysical Research Letters*, **44**, 8520– 8529.

Penland, C., 1996: A stochastic model of indopacific sea surface temperature anomalies. *Physica D: Nonlinear Phenomena*, **98**, 534–558, https://doi.org/https://doi.org/10.1016/0167-2789(96)00124-8.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J Clim*, **8**, 1999–2024.

Petropoulos, F., R. J. Hyndman, and C. Bergmeir, 2018: Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, **268 (2)**, 545–554, https://doi.org/https://doi.org/10.1016/j.ejor.2018.01.045, URL https://www.sciencedirect.com/science/article/pii/S037722171830081X.

Planton, Y. Y., and Coauthors, 2021: Evaluating climate models with the clivar 2020 enso metrics package. *Bulletin of the American Meteorological Society*, **102 (2)**, E193 – E217, https://doi.org/10.1175/BAMS-D-19-0337.1, URL https://journals.ametsoc.org/view/journals/bams/102/2/BAMS-D-19-0337.1.xml.

Richter, I., P. Chang, and X. Liu, 2020: Impact of systematic gcm errors on prediction skill as estimated by linear inverse modeling. *J Climate*, **33 (1)**, 10 073–10 095, https://doi.org/https://doi.org/10.1175/jcli-d-20-0209.1.

Ropelewski, C. F., and M. S. Halpert, 1986: North american precipitation and temperature patterns associated with the el niño/southern oscillation (enso). *Monthly Weather Review*, **114 (12)**, 2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:napatp>2.0.co;2.

Rowley, C. W., I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, 2009: Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, **641**, 115–127, https://doi.org/10.1017/S0022112009992059.

Scholköpf, B., and A. J. Smola, 2002: *Learning with kernels*. MIT Press, 626pp pp.

Schütte, C., P. Koltai, and S. Klus, 2016: On the numerical approximation of the perron-frobenius and koopman operator. *Journal of Computational Dynamics*, **3 (1)**, 1–12, https://doi.org/10.3934/jcd.2016003.

Sharmila, S. e. a., 2023: *Contrasting el niño–la niña predictability and prediction skill in 2-year reforecasts of the Twentieth Century*, **36 (5)**, 1269–1285, https://doi.org/doi:10.1175/jcli-d-22-0028.1.

Shin, S., P. D. Sardeshmukh, M. Newman, C. Penland, and M. Alexander, 2021: Impact of annual cycle on enso variability and predictability. *Journal of Climate*, **34(1)**, 171–193.

Siegert, F., G. Ruecker, A. Hinrichs, and A. Hoffman, 2001: Prediction of sea surface temperature using long short-term memory. *IEEE Geoscience and Remote Sensing Letters*, **414**, https://doi.org/https://doi.org/10.1038/35106547.

Tu, J. H., 2013: Dynamic mode decomposition: theory and applications. Ph.D. thesis, Princeton University, Mechanical and Aereospace Enigineering Department.

Tu, J. H., C. W. Rowley, D. M. Luchtenburg, S. B. Brunton, and J. N. Kutz, 2014: On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, **1 (2)**, 391, https://doi.org/10.3934/jcd.2014.1.391.

Wang, X., J. Slawinska, and D. Giannakis, 2016: Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, **29: 2888–2939**, https://doi.org/doi:10.1088/0951-7715/29/9/2888.

Wang, X., J. Slawinska, and D. Giannakis, 2020: Extended-range statistical enso prediction through operator-theoretic techniques for nonlinear dynamics. *Scientific Reports*, **10(1)**, https://doi.org/doi.org/10.1038/s41598-020-59128-7.

Wang, Y., Y. Zhang, and G.-G. Wang, 2023: Forecasting enso using convolutional lstm network with improved attention mechanism and models recombined by genetic algorithm in cmip5/6. *Information Sciences*, **642**, 119 106, https://doi.org/https://doi.org/10.1016/j.ins.2023.119106, URL https://www.sciencedirect.com/science/article/pii/S0020025523006916.

Williams, M. O., I. G. Kevrekidis, and C. W. Rowley, 2015: A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, **25 (6)**, 1307–1346, https://doi.org/10.1007/s00332-015-9258-5.

Xu, T., M. Newman, M. A. Alexander, and A. Capotondi, 2024: A forecast test for reducing dynamical dimensionality of model emulators. *Journal of Advances in Modeling Earth Systems*, **16 (1)**, e2022MS003 599, https://doi.org/https://doi.org/10.1029/2022MS003599, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003599, e2022MS003599 2022MS003599, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003599.

Zhang, G. P., 2012: Neural networks for time-series forecasting. *Handbook of Natural Computing*, 461–477, https://doi.org/10.1007/978-3-540-92910-9_14.

Zhang, R.-H., C. Gao, and L. Feng, 2022: Recent ENSO evolution and its real-time prediction challenges. *National Science Review*, **9 (4)**, https://doi.org/10.1093/nsr/nwac052, URL https://doi.org/10.1093/nsr/nwac052, https://academic.oup.com/nsr/article-pdf/9/4/nwac052/44253661/nwac052.pdf.