Analysis of El Niño Southern Oscillation using Koopman Operator Framework using Reproducing Kernel Hilbert Space

Peilin Zhen
Advisor: Dimitris Giannakis

Abstract

The climate system consists of coupled interactions among different components of the Earth's atmospheres and other layers. The asymmetry and quasiperiodicity of the El Niño Southern Oscillation (ENSO), in particularly, hinder the ability to forecast future climate trends. There is a need to understand the predictable signal of this ocean-atmosphere exchange. This study aims to investigate the dominant frequencies and reduce the unpredictable noise using a newly introduced data-driven method called the Koopman Operator. This approach can extract favorable predictability properties from the time-ordered data without any prior knowledge of the underlying dynamical evolution equations. The results are then compared with other classical methods. We hope to obtain critical information that permits longer lead ENSO forecasts.

1. Introduction

The climate system is a highly complex system consisting of the interrelations among various components on a whole range of space and time series. Among these interactions, El Niño Southern Oscillation (ENSO) is the largest internal climate variation on the interannual timescales. ENSO is induced by an ocean-atmosphere exchange pattern called the Walker Circulation in the tropical Pacific Ocean [1]. When the Walker Circulation is altered due to intrinsic forces, two extrema phases of ENSO, warm El Niño and cool La Niña, occur alternately in a frequency of about every two to eight years. However, there has been a clear asymmetry between El Niño and La Niña, with positively skewed sea surface temperature (SST) anomalies brought by strong El Niño events [2, 3].

The correct simulation of the asymmetry in ENSO has been challenging in general circulation models (GCMs). The observed ENSOs demonstrate irregularity and quasiperiodicity, with its amplitude locked to the annual cycle [4]. Many models either fail to reproduce this asymmetry or demonstrate different observations than the expected results. The intrinsic physical mechanisms of the asymmetry have been not been fully understood [2]. Some suggest that nonlinearlies of the ocean dynamics and temperature advection [5], instability of the tropical ocean waves [6, 7], and many others are responsible for the complexity. Some studies also discovered the existence of nonlinear interdecadal changes in ENSO behavior. The frequent occurrence of strong El Niño events in the recent half century mirrored the changes in the characteristics of ENSO [4]. The variations in different time scales have limited climate modeling and forecasting with lead times of only several months to a year.

Such limitations were shown in recent strong El Niño predictions. For instance, none of the dynamical and statistical models were able to predict the intensity of the 1997/1998 El Niño events [8]. Even when several signs of onset were detected and expected a development of a major El Niño in 2014, the event was delayed until 2015/2016 [9]. Thus, uncertainty still remains in climate predictions.

Climate variability is a function of predictable sign and unpredictable, chaotic noise. In other words, climate forecasting can be potentially improved by isolating the noise and enhancing the predictable aspect of ENSO variability. Many classical methods such as empirical orthogonal function (EOF) and principal component analysis (PCA) have been widely applied to identify the key components and reduce the dimensionality of the predictors [10, 11, 12]. However, PCA assumes the linearity of training data and may not be the best representation of data sampled from the nonlinear dynamical system. Kernel PCA (KPCA) is an

alternate method that serves as an extension of PCA for data on the nonlinear space. Some studies used this method of dimensionality reduction on the time series of ENSO SST and verified a comparable predictability for a short lead time with SST [11].

Other algorithms that characterize the evolution of complex spatio-temporal phenomena also exist. In recent years, the Koopman operator have become popular in fluid applications and many other dynamical systems [13]. Koopman operator is a linear time evolution operator that can approximate the unknown, nonlinear, stochastic rules of evolution purely from data [14]. More specifically, some advantages of the Koopman operator include its ability to model nonlinear systems using linear techniques relying on data alone without any access to the underlying functions. It can also be decomposed linearly to eigenfunctions and truncate out some non-important terms in the computation. These eigenfunctions characterize the underlying system dynamics collectively, and the eigenvalues that are associated with the Koopman eigenfunctions can be used to predict the system state at any time later [13, 14, 15].

In this work, we attempt to improve climate forecasting through analyzing the ENSO sea surface temperature indices over a range of different regions using the Koopman Operator. Since the Koopman Operator is an abstract concept, we can approximate the Koopman Operator using a set of scalar observables, which are functions that map the states into scalars, that is determined by a kernel function [13]. Then we decompose the approximation of the linear operator into its corresponding eigenfunctions and eigenvalues and extract the intrinsic features of the system dynamics. This procedure can identify the dominant frequencies of the climate data which can potentially predict the future trends.

2. Mathematical Model

In this section, we summarize the Koopman Spectral Theory and its algorithms developed in [14, 15, 16, 17, 13, 19] to introduce notations for later use. The subsections are divided into (1) theories such as (1.1) the brief overview of the Koopman Operator Theory and (1.2) spectral analysis of the Koopman Operator with Reproducing Kernel Hilbert Space, and (2) methodology to compute the Koopman spectra.

2.1. Theory

2.1.1. The Koopman Operator

Consider a data space Y which is governed by a high-dimensional time series or state space X with an observable or function $f : x \to y$ of the function space F, where $y \in \mathbb{C}$, such that the n-th point of the data set is computed by $y_n = f(x_n)$, where $n \in \mathbb{Z}$ is discrete time. The dynamical evolution map of X is given by $\Phi^t : X \to X$, where $t \in \mathbb{R}$ is continuous time. Specifically, if a system is consecutive of the previous system by Δt, they are related with $\Phi^{\Delta t}(x_t) = x_{t+\Delta t}$. Then, the Koopman Operator $U^t : F \to F$, where F consists of scalar observables or functions of state space, hypothetically computes the value of y_{n+t} at t step in the future by introducing a new dynamical system of the same evolution

$$(U^t f)(x) = f(\Phi^t(x)). \tag{1}$$

Note that U^t is infinite dimensional (for acting on the function space F) and linear even if the evolution map Φ^t is nonlinear. Thus, it enables the ability to perform decomposition to obtain its spectral properties, such as the Koopman eigenfunctions $\{z_k\}$ and the corresponding eigenvalues $\{\lambda_k\}$

$$U^t z(x) = \lambda^t z(x), \tag{2}$$

where the Koopman eigenfunction z is also an observable $z \in F$ [14].

In addition to that, since the Koopman operator has a unitary property, the Koopman eigenfunctions form a set of observables that evolve by a periodic factor $e^{i\omega t}$, with ω as the true eigenfrequency, even if the dynamics is aperiodic [15]. Then the time series $\tilde{z}(x) = U^t z(x)$ behaves like a Fourier function on orbits of the dynamics

$$\tilde{z}(t) = e^{i\omega t}\tilde{z}(0). \tag{3}$$

2.1.2. Koopman Spectral Analysis in Reproducing Kernel Hilbert Space

Spectral analysis (or decomposition) of the dynamical systems is able to extract low-dimensional dynamics from the data. The linear decomposition of a Koopman Operator only produces a local approximation of the Koopman eigenfunctions [18]. This is certainly not applicable to all systems. We can construct a spectral analysis of this operator using reproducing kernel Hilbert space (RKHS) theory. Some benefits of this kernel-based approach are that the basis functions need not to be defined explicitly, which allows us to apply the technique to the infinite dimensional feature spaces [17].

An RKHS H is a Hilbert space of functions whose abstract vector space is equipped with the structure of an inner product that allows operations such as orthogonal projections, and completeness which every Cauchy sequence points in the space has a limit. These properties enable the ability to evaluate functions in RKHS with an inner product. An RKHS is associated with a a bivariate function kernel $k : X \times X \to \mathbb{C}$ that reproduces every function, also known as the reproducing property,

1. $\langle f, k(x, \cdot)\rangle_H = f(x), \forall f \in H$,

such that the function evaluation of f at a given point x can be regarded as an inner product evaluation in H between the representor $k(x, \cdot)$ and the function itself. As stated in Moore-Aronsajn theorem, every RKHS is uniquely defined with the reproducing kernel with these additional properties[15, 13]:

2. k is symmetric such that $k(x, x') = k(x', x)*$ for $x, x' \in X$;
3. For every $x_0, ..., x_{N-1} \in X$ and $c_0, ..., c_{N-1} \in \mathbb{C}$, k is positive definite such that

$$\sum_{m,n=0}^{N-1} k(x_m, x_n) \geq 0.$$

This theorem gives a construction of an RKHS without any presumption on X or the kernel k. Conversely, given any kernel satisfying properties 2 & 3, there exists a unique RKHS on X for which k is is the reproducing kernel. With these properties, we can extend a candidate eigenfunction from its values on the sample trajectory to the entire space, in a RKHS of functions. Note that the functions on the RKHS have a notion of regularity such that the functions are smooth. This guarantees the identification of Koopman eigenfunctions on RKHS would have uniform norms on the space of continuous functions, and a Hilbert space structure that allows the construction of data-driven algorithms using standard linear algebra tools [15]. We can construct orthonormal sets in H using kernel integral operators $G : F \to F, F = \{f : X \to \mathbb{C}\}$ defined by

$$G := f \to \int_X k(\cdot, x) f(x) du(x)$$

If $f \in F$ with $Gf = g$, we can discretize the linear integral operator as

$$g(x_m) = Gf(x_m) = \int_X k(x_m, x_n) f(x_n) du(x_n)$$

$$= \frac{1}{N} \sum_{n=0}^{N-1} k(x_m, x_n) f(x_n)$$

Since this is a symmetric kernel, there exists an orthonormal basis $\{\phi_0, ..., \phi_{N-1}\}$ of F coexisting as the eigenvectors of G. Then, let corresponding eigenvalues of the orthonormal eigenfunctions be $\{\lambda_0, ..., \lambda_{N-1}\}$ and let $J \geq N - 1$ be numbers of eigenvalues with $\{\lambda_0, ..., \lambda_{J-1}\} > 0$. For $j \leq J$,

$$\psi_j(x) = \frac{1}{\sqrt{\lambda_j}} \int_X k(x, x_n) \phi_j(x_n) du(x_n)$$

such that $\{\psi_0, ..., \psi_{J-1}\}$ is an orthonormal set in H.

2.2. Numerical Algorithms of the RKHS Norm

Recall the sampled dynamical states $\{x_0, ..., x_{N-1}\} \in X$ are unknown, and the only information that is accessible is the values $\{F(x_0), ..., F(x_{N-1})\}$ of an observation map F on the data space $Y = \{y_0, ..., y_{N-1}\}$ with $y_n = y(t_n) \in R^m$. One can analyze the evolution of the underlying dynamical states by approximating the spectral analysis of the Koopman Operator with the following algorithms derived from ref [16, 15].

1. Construct a Gaussian kernel matrix K to find the correlation of the time-ordered scalar observations from the data space Y

$$K_{ij} = K_\epsilon(y_i, y_j) = \exp\left(\frac{-d_Q^2(y_i, y_j)}{\epsilon}\right), \tag{4}$$

where d_Q^2 is the pairwise distance between $y_i, y_j \in Y$ calculated using the method of delay-coordinate map with Q delays.

$$d_Q^2(y_i, y_j) = \frac{1}{Q}\sum_{q=0}^{Q-1} ||y_{i+q} - y_{j-q}||^2. \tag{5}$$

2. Perform a Markov normalization of the kernel matrix such that the square matrix has the property of each row and column summing to 1.
3. Calculate the eigenfunctions $\{\phi_j\}$ and eigenvalues $\{\lambda_j\}$ of the kernel matrix using singular value decomposition on the normalized kernel such that

$$K\phi_j = \lambda_j \phi_j, \tag{6}$$

which gives sufficient conditions of the Koopman eigenfunctions.
4. Find the spectrum of the kernel eigenfunctions by taking the Fourier transform of each eigenfunction, then computing the row norm of the Fourier space as described by the Nyström extension operator $T : S \to H, S \subset F$ s.t. $S = \{\phi_0, ...\phi_N\}$ in (7),

$$T\left(\sum_{n=0}^{N-1} c_n\phi_n\right) = \sum_{n=0}^{N-1} c_n/\sqrt{\lambda_j}\psi_n \in H, \tag{7}$$

where $\{c_n\} \in \mathbb{C}$ and $\{\psi_n\}$ is an orthonormal set in H. Compare this RKHS norm with respect to the corresponding frequencies.
5. Select the Fourier frequencies with the largest norms if using a relatively smaller number of eigenfunctions, or smallest if more eigenfunctions, as the candidates of the Koopman eigenfrequencies that behaves with a periodic factor $e^{i\omega t}$.

3. Description of data

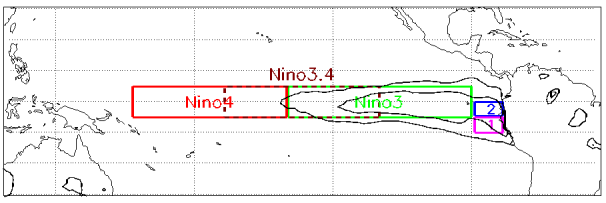


Figure 1: ENSO Index Map[1]

We use the Niño 1.2, Niño 3, and Niño 4 indices provided by NOAA Earth System Research Laboratory. These index regions are used to measure the strength of an ENSO (Figure 1). They capture different

[1] Connolley, W.M..Retrieved from *https://commons.wikimedia.org/wiki/File:Enso-index-map.png.*

properties of that particular region. These data values are monthly average time series of standardized sea surface temperature (SST) mean in the corresponding regions of the Pacific Ocean. In that sense, the index values are *an average* of the ocean gridded values that contains monthly temperature across that region on a 5 deg × 5 deg grid. The time series cover the January 1848 - May 2018 period. The indices were recorded monthly and calculated from the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST). Niño 1.2 measures the SST mean averaged over 0-10°South, 90°West-80°West. Likewise, Niño 3 and Niño 4 cover the range over 5°North-5°South, 150°West-90°West; and 5°North-5°South, 160°East-150°West respectively [20].

The SST indices of the three regions are gathered in a $m \times n$ matrix, where $m = 3$, the number of Niño sets and $n = 1781$, the number of monthly averaged sea surface temperature data recorded from 1870 to 2018. A symmetric Gaussian Kernel matrix was constructed for these time-ordered observations using the formula (4) with a rational delay-coordinate embeddings, and the spectral analysis was then performed using the methodology described in previous section.

4. Results

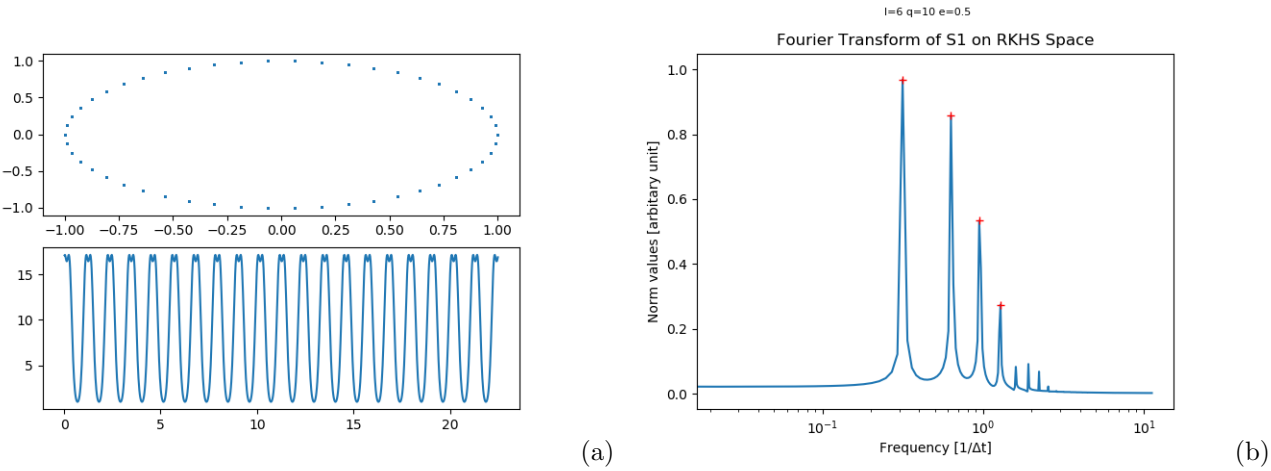
In this section, we applied the methods described in Sections 2-3 to a sample data with known state space and function space (1) and ENSO averaged SST mean data (2). The goal was to verify the applications of the spectral analysis of Koopman Operator technique in RKHS and to identify dominant frequencies in the underlying dynamical systems.

4.1. Periodic flow

A periodic flow $\mathbb{R}^2$ on the unit circle S^1 is represented as $X \to S^1$ with $\Phi^{\Delta t} : X \to X$ and the observation map $F : S^1 \to \mathbb{R}^2$ is defined as

$$F(\theta) = \begin{bmatrix} e^{\cos(\theta)} \\ e^{\sin(\theta)} \end{bmatrix},$$

where $\Phi^{\Delta t} = (\theta + \omega t) \mod 2\pi$, ω is defined as frequency. Then, we used a sampling interval $\Delta t \approx 0.044$ for a sample number $N = 1000$, the initial point in the state space was (1,1) and the delay-coordinate embedding was $q = 10$ and $\epsilon = 0.5$. The quantity of RKHS norm was computed with $l = 6$ eigenfunctions.



(a) (b)

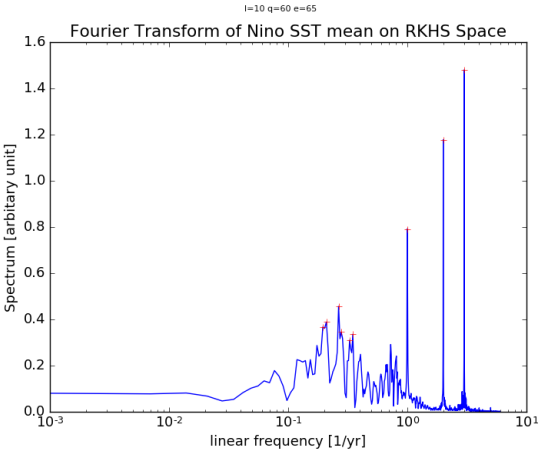
Figure (a) demonstrate the dynamical system X and time series of the dynamical system respectively. Some dominant frequencies f detected in (b) are $f = 0.315$, 0.630, 0.945, and 1.283. They are equivalent to $f \approx 1/\pi, 2/\pi, 3/\pi, 4/\pi$, or angular frequency of $\omega = 2$ as expected from the dynamical system, and the consecutive frequencies are multiple of the first frequency. The results extracted using the procedure described above was able to detect the correct information about the underlying systems.

4.2. ENSO Indices Time Series

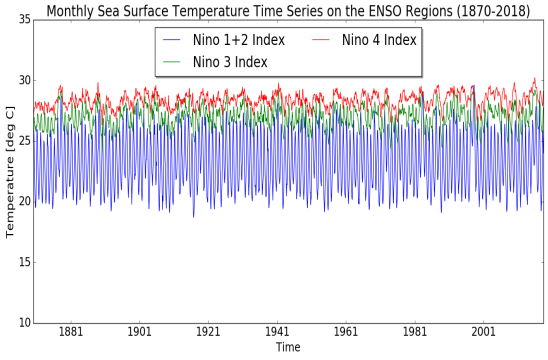
The time series of ENSO Indices are monthly averaged SST mean in different regions of the Pacific Ocean. The scalar observations of different regions with an unknown dynamical system X were consolidated as the observation map F

$$F(X) = \begin{bmatrix} Nino\ 1.2 \\ Nino\ 3 \\ Nino\ 4 \end{bmatrix},$$

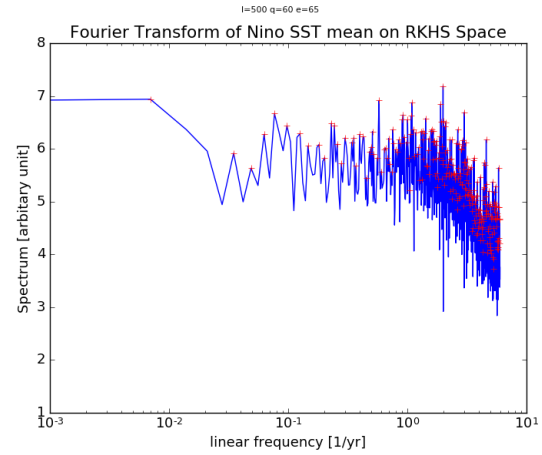
with a time interval of $\Delta t = 1/12\ year$. The delay-coordinate embedding was $q = 60$ and $\epsilon = 65$. The quantity of RKHS norm was computed with $l_0 = 10$ and $l_1 = 500$ eigenfunctions. The ratio of the RKHS norms in l_1, l_0 was then calculated to extract frequencies those with the least change of norms at the higher eigenfunction feature.



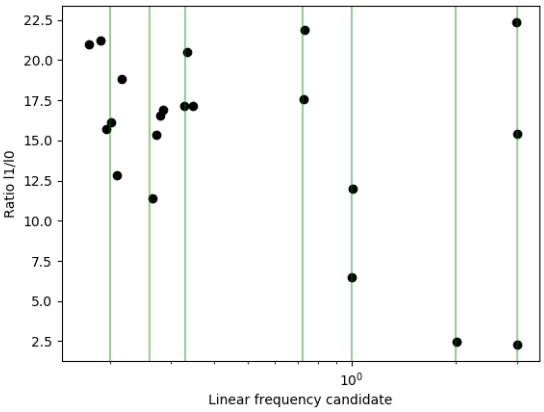
(c)



(e)



(d)



(f)

Figure (e) demonstrates the monthly sea surface temperature mean in different ENSO regions. After applying the spectral analysis procedure, (c) shows significant peaks on $f = 1, 2, 3(1/yr)$, which corresponds to the annual, biennial, and triennial cycles of the ENSO occurrence. There are some peaks in between the ranges of [0.18, 0.70], which was further analyzed by computing the RKHS norm with more eigenfunctions. Figure (d) indicates the increasing quantity of RKHS norm as the number of eigenfunction increases. Troughs in RKHS norm with dense eigenfunctions like this would represent the significance of the corresponding frequency. Again, the annual, biennial, and triennial cycles are obvious from observation. Then, figure (f) shows the ratio of RKHS norm between those of more eigenfunctions l_1 versus less eigenfunctions l_0, and 20 frequencies with the smallest ratios are shown in the graph. The vertical lines represent the potential frequencies $f = 0.2, 0.26, 0.33, 0.71, 1, 2, 3$. These frequencies were selected based on observations at l_0 and

other graphs with lower number of eigenfunctions.

5. Discussions

The approach of using reproducing kernels to detect eigenfrquencies as the candidate of the Koopman eigenfrequencies demonstrates its ability to recover the behaviors of the unknown underlying dynamical systems. This process allows us to reduce dimensionality by truncating noise and enhancing the significant signals.

Furthermore, in section 4.2, the dominant frequencies of the Southern Oscillation, such as annual, biennial, and triennial cycles, are easily detected and verify by the interannual climate variation of ENSO. Some other frequencies such as $f \approx 0.2, 0.26, 0.33, 071(1/yr)$ are also detected with significant RKHS norms with some small variances. These frequencies correspond to 5 years, 4 years, 3 years, and 16 months of the ENSO cycle. The first three frequencies agree with the observed ENSOs, that the El Niño and La Niña occur alternately in a frequency of approximately 2-8 years. The last frequency, $t \approx 16 months$, is worth an in-depth investigation. Some explanations of the detection of this frequency can be the Koopman Operator technique, that is, the frequency can be a combination/summation or multiple of other frequencies.

The scalar observation data of the ENSO, sampling over the average temperature from each grid of the corresponding regions, reveals some potential behaviors that can characterize the temporal patterns of the Southern Oscillation. The frequencies mention above still need further analysis and verification. Future studies should include verifying and modeling those potential frequencies geographically.

Acknowledgement

I would like to express my sincere gratitude to my advisor, Professor Dimitris Giannakis, for all his help and guidance that he has given me throughout the summer program. His knowledge has helped me to stay on track. I would also like to thank Suddhasattwa Das for his patience and time to explain the theories underlying this research and debug my codes. I heartily thank our program coordinators, Jason Kaye and Pejman Sanaei for their constant encouragement and support. Last but not least, I would like to thank Professor Aleksandar Donev for organizing the summer research program and providing constructive feedback to my project work.

References

[1] Baede, A.P.M., Ahlonsou, E., Ding, Y., Schimel, D. The Climate System: an Overview. *IPCC Third Assessment Report: Climate Change*, 87-98 (2001).

[2] Dong, B. Asymmetry between El Niño and La Niña in a Global Coupled GCM with an Eddy-Permitting Ocean Resolution. *Journal of Climate* **18**, 3373-3387 (2005).

[3] Slawinska, J., Giannakis, D. Indo-Pacific Variability on Seasonal to Multidecadal Time Scales. Part I: Intrinsic SST Modes in Models and Observations. *Journal of Climate* **30**, 5265-5294 (2017).

[4] An, S. A review of interdecadal changes in the nonlinearity of the El Niño-Southern Oscillation. *Theoretical and Applied Climatology* **97**, 29-40 (2009).

[5] Zhang, T., Sun, D. ENSO Asymmetry in CMIP5 Models. *Journal of Climate* **27**, 4070-4093 (2014).

[6] Liang, J., Yang, X., Sun, D. Factors Determining the Asymmetry of ENSO. *Journal of Climate* **30**, 6097-6106 (2017).

[7] Hoerling, M., Kumar, A., Zhong, M. El Niño, La Niña, and the Nonlinearity of Their Teleconnections. *Journal of Climate* **10**, 1769-1786 (1997).

[8] Barnston, A.G., Glantz, M.H., He, Y. Predictive Skill of Statistical and Dynamical Climate Models in SST Forecasts during the 1997–98 El Niño Episode and the 1998 La Niña Onset. *Bulletin of the American Meteorological Society* **80**, 217-244 (1999).

[9] Masuda, S., Matthews, J.P., Ishikawa, Y., Mochizuki, T., Tanaka, Y., Awaji, T. A new Approach to El Niño Prediction beyond the Spring Season. *Scientific Reports* **5**, srep16782 (2015).

[10] Zheng, Z., Hu, Z., L'Heureux, M. Predictable Components of ENSO Evolution in Real-time Multi-Model Predictions. *Scientific Reports* **6**, srep35909 (2016).

[11] Lima, C., Lall, U., Jebara, T., Barnston, A. Statistical Prediction of ENSO from Subsurface Sea Temperature Using a Nonlinear Dimensionality Reduction. *Journal of Climate* **22**, 4501-4518 (2008).

[12] Lu, W., Atkinson, D., Newlands, N. ENSO climate risk: predicting crop yield variability and coherence using cluster-based PCA. *Modeling Earth Systems and Environment* **3**, 1343-1359 (2017).

[13] Williams, M.O., Rowley, C.W., Kevrekidis, I.G.A Kernel-Based Approach to Data-Driven Koopman Spectral Analysis. *arXiv* 1411.2260 (2014).

[14] Hua, J., Noorian, F., Moss, D., Leong, P.H.W., Gunaratne, G.H. High-dimensional time series prediction using kernel-based Koopman mode regression. *Nonlinear Dynamics* **90** 1785-1806 (2017).

[15] Das, S., Giannakis, D. Koopman spectra in reproducing kernel Hilbert spaces. *arXiv* 1801.07799 (2018).

[16] Das, S., Giannakis, D. Delay-coordinate maps and the spectra of Koopman operators. *arXiv* 1706.08544 (2017).

[17] Kawahara, Y. Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis. *Proc. of Advances in Neural Information Processing Systems* 911-919 (2016).

[18] Fujii, K., Inaba, Y., Kawahara, Y., Koopman Spectral Kernels for Comparing Complex Dynamics: Application to Multiagent Sport Plays. *Machine Learning and Knowledge Discovery in Databases* 127 - 139 (2017).

[19] Giannakis, D., Slawinska, J., Zhao, Z. Spatiotemporal Feature Extraction with Data-Driven Koopman Operators. *JMLR: Workshop and Conference Proceedings* **44** 103-155 (2015).

[20] Earth System Research Laboratory El Niño Southern Oscillation (ENSO) Index Dashboard. (2018) Retrieved from https://www.esrl.noaa.gov/psd/enso/dashboard.html