

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelorarbeit Kognitionswissenschaft

Spatial and frequency dependence of westerly wind events causing ENSO diversity

Jakob Unterholzner

28.06.2022

Gutachter

Dr. Bedartha Goswami
Machine Learning in Climate Science
Cluster of Excellence "Machine Learning"
University of Tübingen

Betreuer

Jakob Schlör
Machine Learning in Climate Science
Cluster of Excellence "Machine Learning"
University of Tübingen

Unterholzner, Jakob:

*Spatial and frequency dependence of westerly wind events
causing ENSO diversity*

Bachelorarbeit Kognitionswissenschaft

Eberhard Karls Universität Tübingen

Bearbeitungszeitraum: 01.03.2022 - 01.07.2022

Acknowledgements

I would like to thank Dr. Bedartha Goswami, the group leader of the Machine Learning in Climate Science (MLCS) junior research group, for supervising me. I want to thank the whole MLCS group for their helpful tips and discussions. I am grateful for the insight in your work. Special thanks go to Jakob Schlör for the support and help I have received. I have very much enjoyed working with you and this project would not have been possible without your efforts.

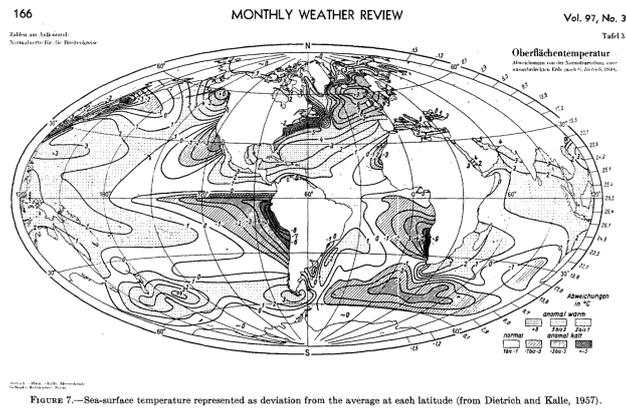


Figure 0. Sea surface temperature anomalies from Bjerknes, 1969

Spatial and frequency dependence of westerly wind events causing ENSO diversity

Jakob Unterholzner, Jakob Schlör, and Bedartha Goswami

Cluster of Excellence Machine Learning: New Perspectives for Science

Universität Tübingen

72072 Tübingen, Germany

Correspondence: Jakob Unterholzner (jakob.unterholzner@student.uni-tuebingen.de)

Abstract. The El Niño Southern Oscillation (ENSO) shows different characteristics regarding the location of its warmest and coldest Sea Surface Temperature (SST) anomalies, ranging from Eastern Pacific (EP) to Central Pacific (CP) events. Both types of ENSO events inhabit different teleconnections and impacts around the world. While stochastic wind forcing is one of the known drivers of the ENSO, the influence of westerly wind anomalies over the equatorial Pacific on ENSO development, regarding its diversity, remain unclear. A spectral analysis of daily wind data, yields high and low frequency components of equatorial wind anomalies for different Pacific regions. We investigate their respective influence on ENSO diversity by means of multiple linear regression models and use the causal network discovery PCMCI⁺ algorithm to infer the important westerly wind events (WWE) that drive ENSO diversity. The obtained results are tested in a causally informed ENSO prediction experiment. We provide evidence for the key role of low frequency wind variations in ENSO dynamics and reveal a regional correspondence of wind and SST anomalies. This study might help climate models in their prediction ability of ENSO diversity by including relevant wind information and thereby improve forecasting of ENSO and its implications.

Keywords. Westerly Wind Events · El Niño diversity · causal discovery

1 Introduction

The El Niño Southern Oscillation is the main source of interannual variability in Earth's climate system (e.g., Wallace et al., 1998). During neutral ENSO conditions an equatorial atmospheric circulation, called Walker circulation, with east-to-west surface winds, accumulates warm water in the western Pacific, while causing upwelling of colder water in the eastern Pacific. During cold ENSO phases this circulation intensifies, whereas in warm ENSO conditions the circulation weakens, which reduces upwelling and allows the western warm water pool to expand eastward. ENSO warm events exhibit different patterns regarding the spatial location of warmest sea surface temperature anomalies: In some years SST anomalies are highest in the Central Pacific, while during others they expand all the way to the Eastern Pacific. As the different types of warm ENSO events display different impacts around the globe (e.g., Strnad et al., 2022), predicting the location of highest SST anomalies is valuable but yet still difficult (Kirtman et al., 2014; Ren et al., 2019).

Equatorial winds play a central role in the ocean-atmospheric interaction of ENSO (Bjerknes, 1969). Wind anomalies arise in response to SST anomalies but on the other hand can also enhance or decrease them (Eisenman et al., 2005; Lian et al., 2014). Especially, westerly wind events in the western and central Pacific are known precursors of ENSO warm events (Chen et al., 2015), as they extend the warm pool eastward which weakens the Walker Circulation.

However, it is not entirely clear how WWE are influencing the spatial patterns of highest SST anomalies. Different definitions of WWE have been proposed and caused confusion about stochastic and deterministic wind forcing as illustratively pointed out by Capotondi et al. (2018): Most definitions of WWE set duration and amplitude thresholds for wind anomalies from the annual cycle. These anomalies consist of low and high frequencies, yet WWE are often perceived as strictly high frequency

wind variations. In contrast, there is evidence that only the low frequency spectrum of the wind forcing can effectively modulate the slowly evolving SST irregularities (Capotondi et al., 2018; Roulston and Neelin, 2000).

Motivated by the misconception of WVE and their unknown influence onto ENSO diversity, we address the question how the different ENSO types are influenced by low and high frequency zonal wind.

In this study, we first resolve the equatorial wind signal, obtained from the ERA5 reanalysis data, for different regions and frequencies, following the approach of Capotondi et al. (2018). We use two different ENSO indices capturing the EP and CP characteristics of ENSO diversity (Sullivan et al., 2016). As a baseline we investigate the influence of high and low frequency westerly wind anomalies for the EP and CP index by means of multiple linear regression models. In addition, we apply the recently developed PCMCI⁺ causal discovery algorithm (Runge, 2020), to reveal causal relationships between wind anomalies and ENSO indices. Applications of the algorithm have been used to identify teleconnection patterns (e.g., Di Capua et al., 2020) and precursor dimensionality reduction for a drought forecasting model (e.g., Lehmann et al., 2020). We use the PCMCI⁺ results to conduct a causally informed neural network experiment, testing the ability to predict ENSO diversity only with wind variations as input features. By removing spurious relations between wind anomalies and ENSO indices with the PCMCI⁺ algorithm, we reduce the input dimensionality for prediction models to improve training consistency and stability (Beucler et al., 2022). Predicting with causally informed linear and nonlinear regression models, we underline the importance of PCMCI⁺ revealed relevant spatial and frequency dependent wind forcing on the different ENSO types.

2 Data and Methods

2.1 Data: ERA5 and ERSSTV5

We use two ERA5 reanalysis datasets, covering 1950 to 1978 (Bell et al., 2020) and 1979 to 2020 (Hersbach et al., 2018) spanning a total of seventy years of data. From the ERA5 datasets we obtain the six hourly eastward wind component at a height of 10 meters above the surface, denoted as u10 wind component. Monthly Nino3, Nino4 and Nino3.4 indices are taken from the ERSSTV5 data (NWS, 2020).

2.2 Data preparation and approach

In order to spatially resolve the eastward wind component, we divide the equatorial Pacific into six regions from West to East (see Fig. 1), following the approach of Capotondi et al. (2018) and Harrison and Vecchi (1997). Starting at 130°E each region is spanning 25° in longitude and 10° in latitude (from 5°S to 5°N). We average the linearly detrended data within each region, yielding one time series of eastward wind speeds for each of the six rectangular regions.

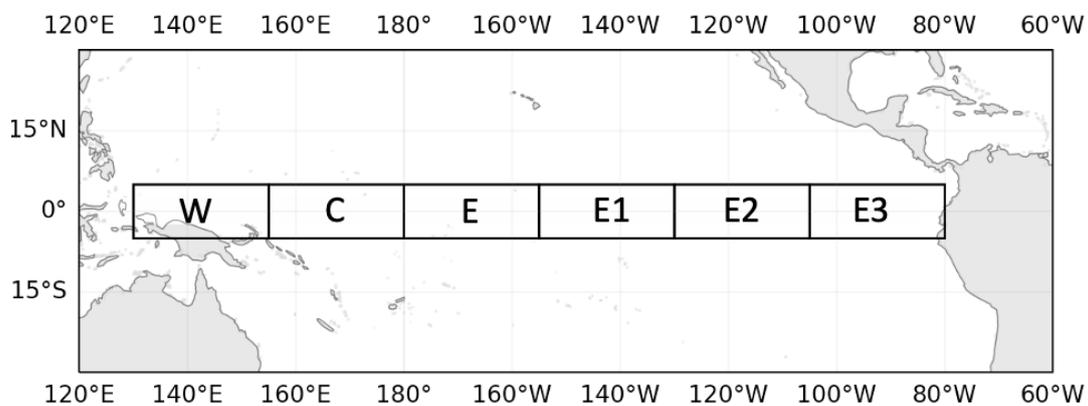


Figure 1. Six regions to spatially resolve equatorial winds as introduced by Capotondi et al. (2018). The longitudes from 130°E to 80°W are divided into six regions each spanning 25° in longitude and 10° in latitude (from 5°S to 5°N).

Replicating Capotondi et al. (2018), we conduct a spectral analysis of the six-hourly eastward wind component for each of the six equatorial regions (see Fig. 2). Our results resemble the ones from Capotondi et al. (2018), who used the cross-calibrated multiplatform vector wind analysis (Atlas et al., 2011), and we therefore argue that the ERA5 reanalyses datasets are well suited for our study.

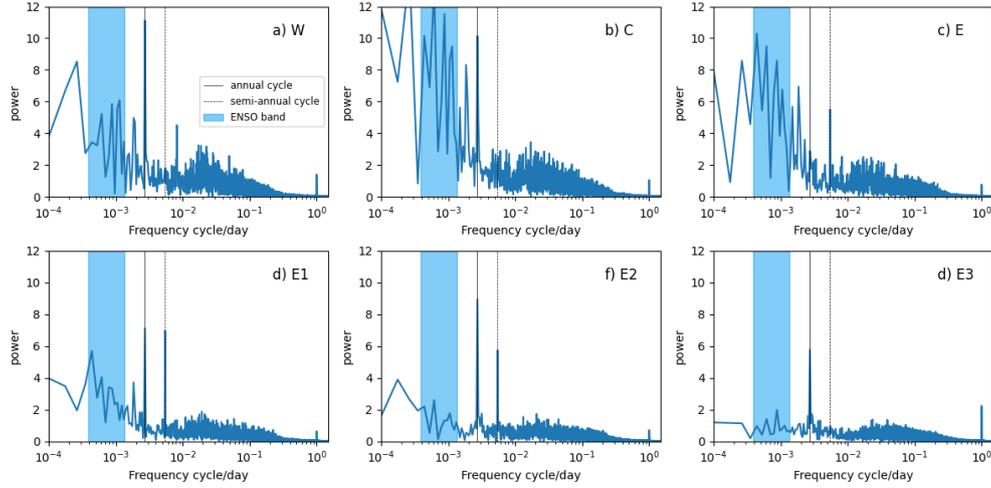


Figure 2. Fourier Spectra of the eastward wind component for the six equatorial regions (as defined in Fig. 1). The blue shaded area shows the ENSO (2-7 Years) frequency band. The annual and semiannual cycles are marked with a solid and a dashed line respectively.

As observed by Capotondi et al. (2018), there is a "spectral gap between interannual and intraseasonal bands", suggesting "a natural way to separate the variability in two bands". Again following their approach, we remove the annual cycle of the eastward wind and decompose the resulting anomalies into their low frequency (LF) and high frequency (HF) component by a frequency limit of 250 days. We show the zonal wind anomalies for region W, E and C as a superposition of their HF and LF component in Fig. 3, reproducing Capotondi et al. (2018).

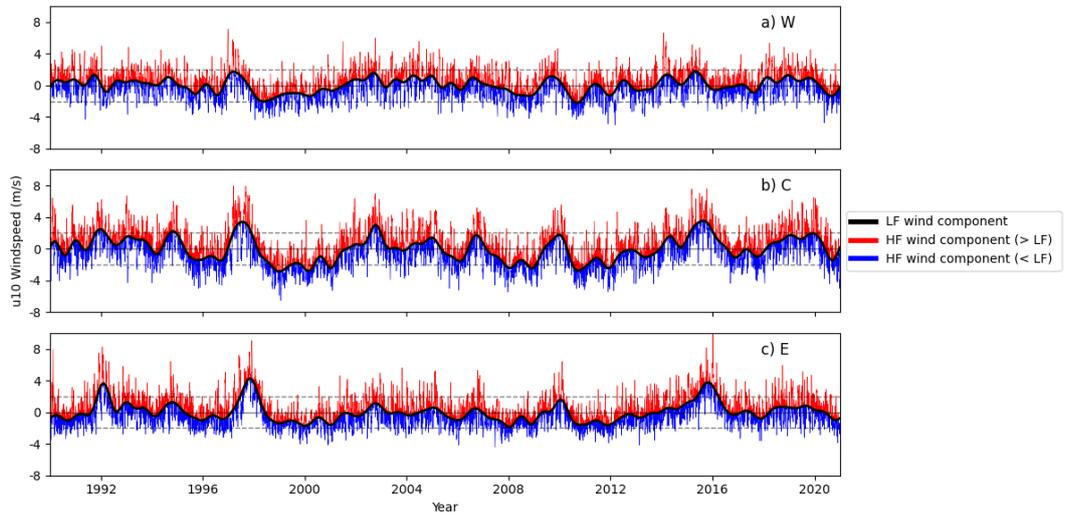


Figure 3. Zonal wind anomalies in regions W, C, E, replicating Capotondi et al. (2018). The anomalies are obtained by removing the annual cycle from the original wind signal. Fourier filtering and separating the anomalies by a frequency limit of 250 days yields a high frequency (HF, periods of 1 to 250 days) and low frequency (LF, periods longer than 250 days) component. The LF component is shown as the solid black line. The HF component is superposed on the LF component, with values larger and smaller than the LF component marked red and blue respectively. An amplitude threshold of $\pm 2\text{m s}^{-1}$ is shown as thin dashed lines, and exceeded by the LF component alone for instance in the 1998 ENSO warm event in regions C & E.

The importance of the LF variation on "commonly used definitions of westerly wind events and easterly wind events based on amplitude and duration thresholds" (Capotondi et al., 2018) is clearly visible, as LF variation alone exceeds a commonly used amplitude threshold of $\pm 2m s^{-1}$ at various instances in time (see Fig. 3). Nonetheless, wind events "are often perceived as high-frequency wind variations at the high-frequency tail of the subseasonal band" (Capotondi et al., 2018). Frequency decomposing the zonal wind anomalies avoids this confusion of deterministic and stochastic wind forcing.

2.3 Frequency dependent zonal wind activity and ENSO indices

Motivated by the spectral gap of the eastward winds (Fig. 2) and the importance of LF variations on the regional anomalies (Fig. 3), we construct an index of HF (Capotondi et al., 2018) and LF wind activity for each of the six equatorial regions (Fig. 1). Fourier decomposing the average wind anomaly of each region yields a HF and LF component. These frequency dependent anomaly time series are used to define the the HF and LF indices,

$$HF - Idx_{region}(t) := \sqrt{g(HF_{region}(t)^2)} \quad LF - Idx_{region}(t) := g(LF_{region}(t)) \quad (1)$$

where $g(x)$ is a 360 days rolling mean calculation performed on each day. Note that we do not square the original LF time series, as it yields a better fit to the Nino indices (see Appendix A). Depending on the centering, $g(x)$ uses either values from the past and the future or just from the past. We set $g(x)$ to include 360 days of past values and assume that Capotondi et al. (2018) used past and future values, as our results differ: Our HF index shows a shorter lead time to the Nino3.4 index (see Fig. 4a). The HF and LF index for the different regions correspond well to the Nino3.4 Index as shown in Fig. 4. Since ENSO displays its dynamic not on daily timescales, we resample the daily wind indices to monthly time series and standardize them to allow interpretation of the regression coefficients.

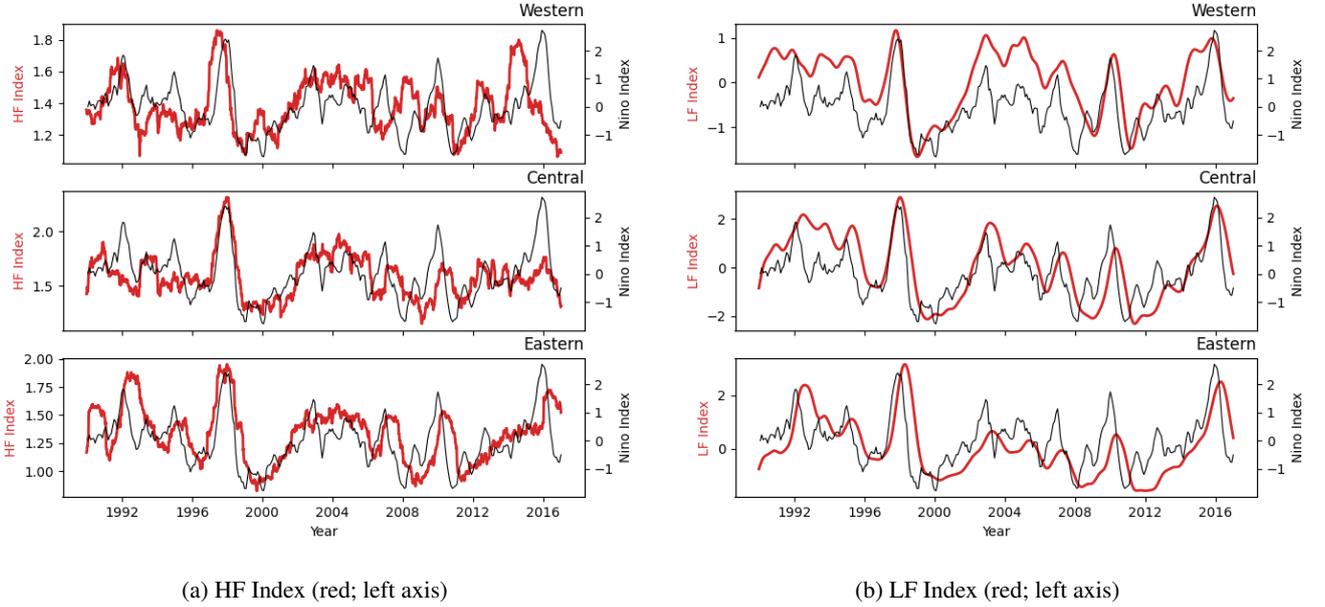


Figure 4. Indices of HF and LF zonal wind activity (red) for three equatorial regions, as defined by Equation 2, related to the Nino3.4 Index (black; right axis).

In order to capture the main modes of ENSO's diversity, we use the EP_{new} and CP_{new} index (Sullivan et al., 2016). We will refer to them as EP and CP index. The CP index corresponds to SST anomalies in the central pacific, while the EP index corresponds to SSTa in the eastern pacific. They are defined by the Nino3 and Nino4 index as follows:

$$EP(t) := Nino3_{normalized}(t) - 0.5 * Nino4_{normalized}(t) \quad CP(t) := Nino4_{normalized}(t) - 0.5 * Nino3_{normalized}(t) \quad (2)$$

90 The goal of our study is to investigate on the spatial and frequency dependent influence of irregular wind activity on SST anomalies spatial patterns captured by the EP and CP index. We want to know which regions at what lead times and frequencies drive the SST anomalies. In order to reveal relevant interactions, we conduct multiple linear regression analyses and a recently developed causal discovery method.

2.4 Linear Regression

95 Due to the simplicity and high degree of interpretability, we choose multiple linear regression models as a baseline of the present study. We introduce lead times by shifting the wind indices for up to twelve months. Shifted time series data intrinsically exhibit strong multicollinearity, which hinders regression coefficient interpretation. To avoid this pitfall, we set up a linear regression for each lead time and region separately. More precisely at each equatorial region, we fit two models (for CP and EP) for twelve different lead times. Each model regresses the ENSO index at time t_n with the HF and LF wind activity at a specific lead time
100 i and region $l \in \{W, C, E, E1, E2, E3\}$. The resulting models are interpretable and enable comparison of regional HF and LF importance via the $w_{HF_l}(t_{n-i})$ and $w_{LF_l}(t_{n-i})$ regression coefficients. For one region, the models can be written like this:

$$\begin{aligned} EP(t_n) &= w_l(t_{n-i}) + w_{HF_l}(t_{n-i}) * HF_l(t_{n-i}) + w_{LF_l}(t_{n-i}) * LF_l(t_{n-i}) \\ CP(t_n) &= w_l(t_{n-i}) + w_{HF_l}(t_{n-i}) * HF_l(t_{n-i}) + w_{LF_l}(t_{n-i}) * LF_l(t_{n-i}) \end{aligned} \quad (3)$$

2.5 PCMCI⁺ Causal Discovery Algorithm

The PCMCI⁺ algorithm (Runge, 2020) is a recently developed causal discovery algorithm for time series which is based-on
105 conditional independence (CI) tests. We apply the PCMCI⁺ method to reveal relationships between zonal HF winds, LF winds, the CP, and EP index. PCMCI⁺ allows us to identify conditional dependencies between time series (in the following referred as links), incorporating lead times and extending PCMCI (Runge et al., 2019) to include discovery of contemporaneous links, which we assume relevant in this ocean-atmospheric coupling investigation.

The PCMCI⁺ method consists of two main ideas: First, the skeleton edge removal phase of the PC Algorithm (Spirtes
110 et al., 2000) which tests for conditional dependencies between variables. For time series, this is separated into lagged and contemporaneous conditioning phase and heuristically only the strongest CI are tested. Secondly, the momentary conditional independence test from Runge et al. (2019) is used to increase effect size (Runge, 2020) by an improved choice of conditioning sets.

In other words, PCMCI⁺ iteratively deletes insignificant links between time series at different locations and lead times,
115 starting from a fully connected graph, resulting in the causal process graph with only relevant links left. These links are causal within the PCMCI⁺ setting and are therefore also restricted by several assumption like the Causal Markov Condition, Faithfulness, and Causal sufficiency. We refer to Runge et al. (2019) for a detailed explanation of the algorithm and discussion of its strengths and limitations.

The PCMCI⁺ method has free parameters that shape its outcome. The most important parameters are the type of CI test and
120 the chosen p-value for which the test is considered significant. We use partial correlation as CI test and set the significance threshold to $pc\text{-}\alpha=0.0005$. The maximum lead time in our analysis is set to twelve months.

2.6 Causally informed linear and non-linear prediction experiment

The main result of the PCMCI⁺ algorithm is the causal process graph displaying links between time series. In our case, this graph contains the information which regions respective to time and frequency are influencing the EP and CP index. Based on
125 this process graph, we determine two sets of input variables for our regression models: causal parents, with direct links to EP and CP (0-order, see Fig. 7a, 7b) and the causal parents plus grand parents, with direct and indirect links (1-order, see Fig. 7c, 7d). We compare the predictability of EP and CP index by fitting models for different inputs conditions using the PCMCI⁺

results. The three conditions are the 0-order sets, the 1-order sets, and an uninformed set containing all regions, frequencies and lead times.

130 We use multiple linear regression models (linear) and multi layer perceptrons (MLPs, non linear) for the three input conditions and compare the coefficient of determination (R^2) of their prediction on unseen test data.

The MLPs consist of two hidden layers each with 16 units. A batchsize of 10 and a learningrate of 0.0001 is used and we stop training once the validation error increases for two epochs in a row. An ensemble of ten MLPs is trained for each input condition, from where we compute the mean and standard deviation of prediction skill.

135 3 Results

We apply two methods for discovering the spatial and frequency dependence of equatorial wind anomalies causing ENSO diversity, as captured by the EP and CP index. First, we fit multiple linear regression models that consist of regional HF and LF wind indices as regressors, with the EP and CP indices being the dependent variable. The second approach employs the PCMC I^+ method, retrieving a causal process graph which reveals connections of wind indices with the EP and CP indices. Finally, we test our results in a prediction experiment, comparing uninformed and PCMC I^+ informed ENSO index predictability.

3.1 Multiple Linear Regression

We fit multiple linear regression models (as defined in Eq. 3) to investigate on the relevance of regional HF and LF wind activity onto the EP and CP index respectively. The coefficient of determination (R^2) for each spatial, frequency and lead-time dependent model can be seen in the top row of Fig. 5. The models regression coefficients for the regional HF and LF wind activity are shown in the bottom row. Between model comparison in R^2 score reveals relevant regions and lead times, while within each model the regression coefficients display frequency dependence. We consider regression coefficients to be significant if their p-value is smaller than 0.05.

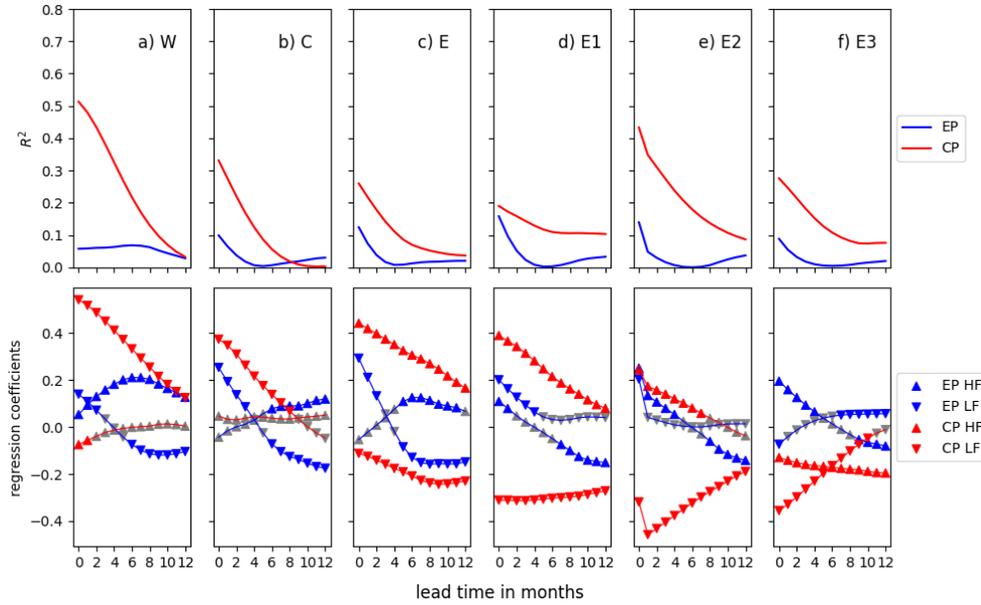


Figure 5. R^2 and regression coefficients of the linear models for the six equatorial regions (Fig. 1) are shown as the columns from a) to f). The regression models are defined by Eq. 3. Lead times are between zero to twelve months. For each region the top row shows the coefficient of determination (R^2) for EP (blue) and CP (red) regression models at different lead times. In the bottom row, one can see the regression coefficients for the EP (blue) and CP (red) models at different lead times. Significant regression coefficients for the HF wind index and LF wind index are shown as upside (\triangle , HF) and downside (∇ , LF) triangles. Insignificant regression coefficients ($p > 0.05$) are shown in gray.

The CP models reach their highest coefficient of determination in the W, E2 and C region (see Fig. 5 a), e), C)). For all six equatorial regions, the coefficient of determination of CP models decreases with an increase in lead time. The high R^2 values in region W and C (see Fig. 5 a) and b)) can be explained by the LF wind activity, as the HF regression coefficients are small and insignificant ($p > 0.05$). The CP models R^2 values are lower in regions E and E1 (see Fig. 5 c) and d)) and LF wind activity is negatively influencing the CP index, while HF wind activity positively correlates to the CP index. The wind activity in the two easternmost regions E2 and E3 (see Fig. 5 e) and f)) fits the CP variability better again, yielding high R^2 scores especially in the E2 region. The negative LF coefficients show the biggest relative influence in that regions. This indicates that a negative LF wind anomaly (stronger than average eastward wind) in regions E2 and E3, is correlated to high CP index values, while a positive LF wind anomaly (weaker than average eastward, or westward wind) corresponds to low CP index values. Generally, CP models show higher R^2 values compared to EP models, suggesting that regional wind activity explains CP variability better than EP variability.

The EP models reach their highest coefficient of determination in the E1, E2 and E region (see Fig. 5 c), d), e)). The R^2 decreases with increasing lead times up to five months and then increases again after seven months for all regions, except for region W. Wind activity in region W explains the EP index best at around six months lead time, at which the HF regression coefficients also reach their maximum. In regions C and E the LF wind component is explaining the peak in R^2 at low lead times. The regions E1 and E2 both display comparable patterns in R^2 values and within model parameters, where LF and HF winds are relevant at low lead times, but only HF winds contribute significantly at lead times over 7 months. This result implies that the increase in R^2 at high lead times can be attributed to the HF wind component. Nevertheless, as the R^2 is quite low for these high lead times, further analyses is needed to explain this observation and its significance, which is beyond the scope of this work.

In a nutshell, multiple regression analyses results provide evidence for the importance of LF wind activity at low lead times. The CP index is best explained by winds in the W, E2 and C region, while three highest R^2 scores for the EP index were found in E1, E2 and E regions. In four of these six regions the LF wind component is the important regressor.

3.2 PCMCI⁺ causal discovery algorithm

To further analyze the relationship of frequency dependent zonal wind forcing on ENSO diversity, we apply the PCMCI⁺ causal discovery method. Like in the multiple regression analyses, we use HF and LF wind indices for each of the six equatorial regions (see Fig. 1) and the EP and CP index as input. The resulting causal process graph for this network of wind activity and ENSO indices can be seen in Fig. 6. Each node in the graph represents one of the wind or ENSO indices. The wind indices are shown relative to their spatial location, with HF winds in the top row, LF winds in the middle and ENSO indices in the bottom row. The node color reflects the auto-MCI statistic, which is the partial correlation value of the node to a shifted version of itself, conditioned on a set of its causal parents (see Runge (2020) for details). The link color denotes the cross-MCI statistic, which is the partial correlation of the connected nodes at a certain lead time, conditioned on a superset of the causal parents of both (see Runge (2020)). The small numbers next to the arrows display lead times at which the cross-MCI statistic is significant. In case of significant links at various lead times the relevant months are sorted after their CI test statistic values from high to low. The darker tone in node colors compared to the link colors reflects the stronger partial auto correlation compared to the partial cross correlation. LF wind indices display higher partial auto correlation than HF and ENSO indices. The highest absolute cross-MCI value is found at the contemporaneous, undirected link between EP and CP index. As the EP and CP index are defined by Eq. 2, PCMCI⁺ correctly revealed their negative relationship.

Further, the process graph displays causal links from the LF to HF domain in the regions W, C, and E. Another LF to HF connection is displayed from regions E2 to E3. Vice versa the HF wind indices are not influencing LF nor ENSO indices and restricted to their HF domain. Additionally, if the PCMCI⁺ returns the absence of a link, one can assume that there really is no

causal relation between the time series (Runge et al., 2019). This observation again corroborates the lack of influence from the high frequency wind variations onto ENSO dynamics.

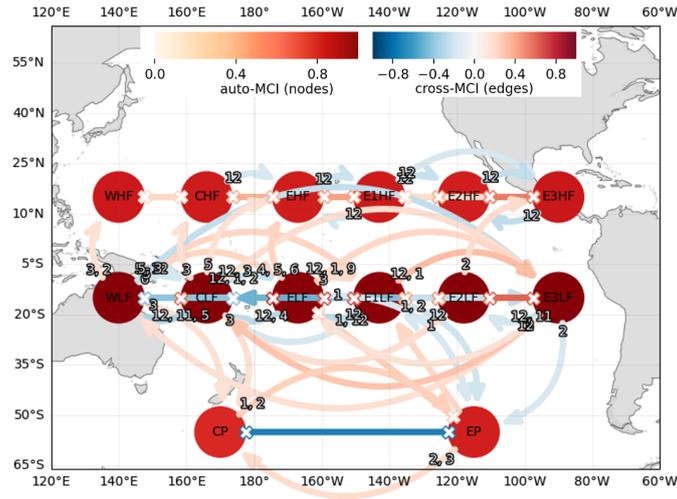


Figure 6. The causal network graph resulting from the PCMCIC⁺ algorithm. For six equatorial regions the top nodes displays the HF indices, while the middle row displays their LF indices. The ENSO EP and CP index are shown at the bottom. Curved arrows denote lagged causal links, the number next to them denote the lag time at which this causal link is strongest. Straight links denote contemporaneous links. Straight arrows with cross symbols at the end indicate that the contemporaneous link directionality is undecided, due to conflicting orientation rules.

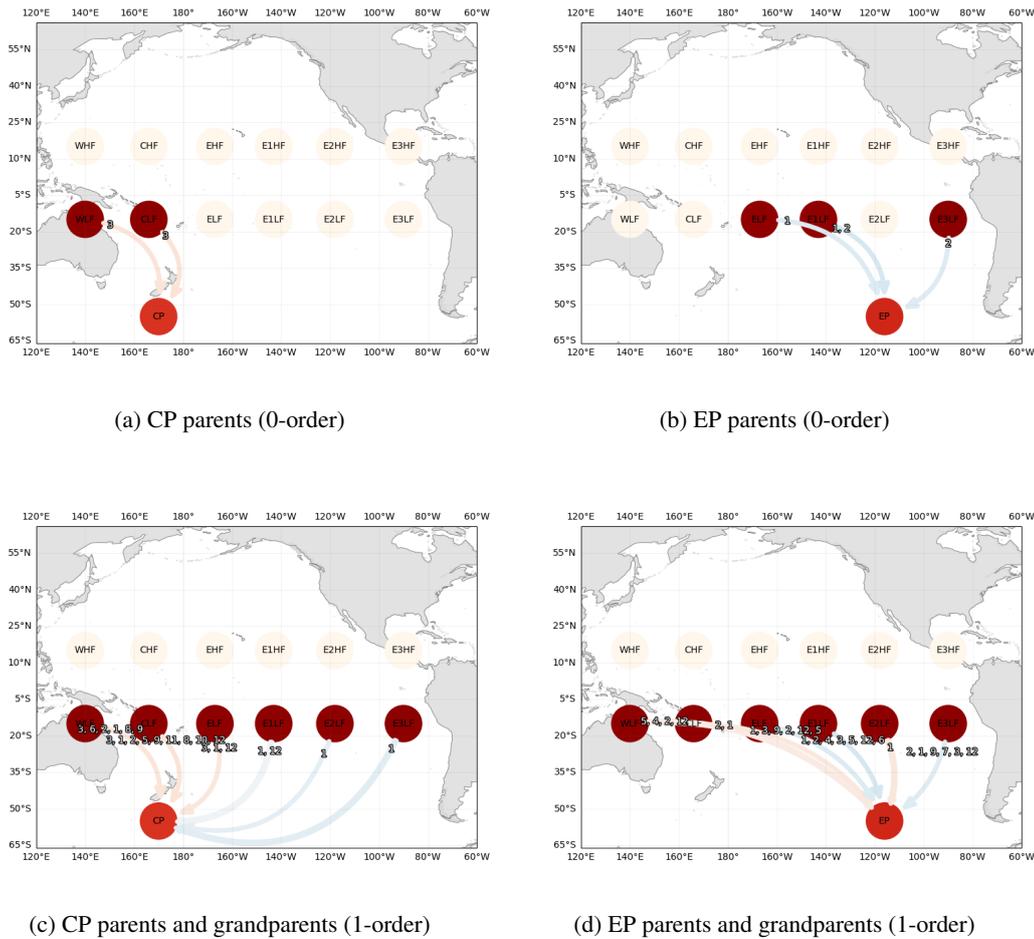


Figure 7. Causal parents (0-order links) for CP (left) and EP (right) can be seen in the top row. Below causal parents and grandparents (1-order) for for CP (left) and EP (right) are shown. These are the input sets for the informed ENSO prediction experiment.

In our study, we are especially interested in the links causing EP and CP indices. The CP index has two direct causal parents (0-order), the LF wind activity in regions W and C (see Fig. 7a) at a lead time of three months. The EP index has three direct causal parents, namely the LF winds in the regions E, E1 and E3 at shorter lead times between one and two months (see Fig. 7b). PCMCI⁺ established causal relevant links to the EP and CP index that stem from regions which resemble the respective spatial pattern of eastern Pacific and central Pacific ENSO events. The sign of the partial correlations causing the CP index is positive, while the EP index is caused by negative partial correlations. Rather than a direct relation these differences in sign denote the different relative influence, keeping causal relevant parents of both connected time series constant. Therefore, the partial correlation values are inconvenient for interpretation. In contrast, the regression coefficients in Fig. 5, at the regions and lead time revealed by PCMCI⁺, offer a direct measure of proportional influence and allow for physical interpretation. For example, the causal parents of the CP index, namely LF wind variations in region W and C at a lead of 3 months, are contributing positively to CP index, since LF regression coefficients of the zonal regression models are positive at these lead times (see Fig. 5). Westerly winds events (or a weakening of eastward trade winds) in regions W and C correspond to higher CP values.

We also consider the causal grandparents (1-Order) of the ENSO indices, as they might have an indirect influence. Taking all regions that are linked to a causal parent of EP and CP into account makes all LF regions relevant at various lead times for EP and CP (see Fig. 7c, 7d). Again the absence of any HF contribution to the ENSO dynamics is striking.

3.3 Causally informed prediction experiment

In order to test the PCMCI⁺ results, we conduct a prediction experiment, where we compare EP and CP index predictability by three different sets of wind activity inputs: First, in the uninformed condition, we use HF and LF indices of all regions at twelve different lead times to predict EP and CP. Secondly, we use the 0-order set of causal parents of EP and CP only, as shown in Fig. 7a and 7b, at the relevant lead times. Third, we use causal parents and grandparents (1-order), as shown in Fig. 7c and 7d, again at the PCMCI⁺ revealed lead times. The prediction skill of uninformed, 0-order and 1-order conditions are compared using a linear regression model and an ensemble of multi layer perceptrons (see Fig. 8). We hypothesize that the causally relevant, zonal LF winds revealed by PCMCI⁺ should exhibit comparable prediction skill as the uninformed condition. In other words, if the wind activities in 0-order and 1-order regions are causing the EP and CP index, information that is additionally given should not substantially increase prediction skill.

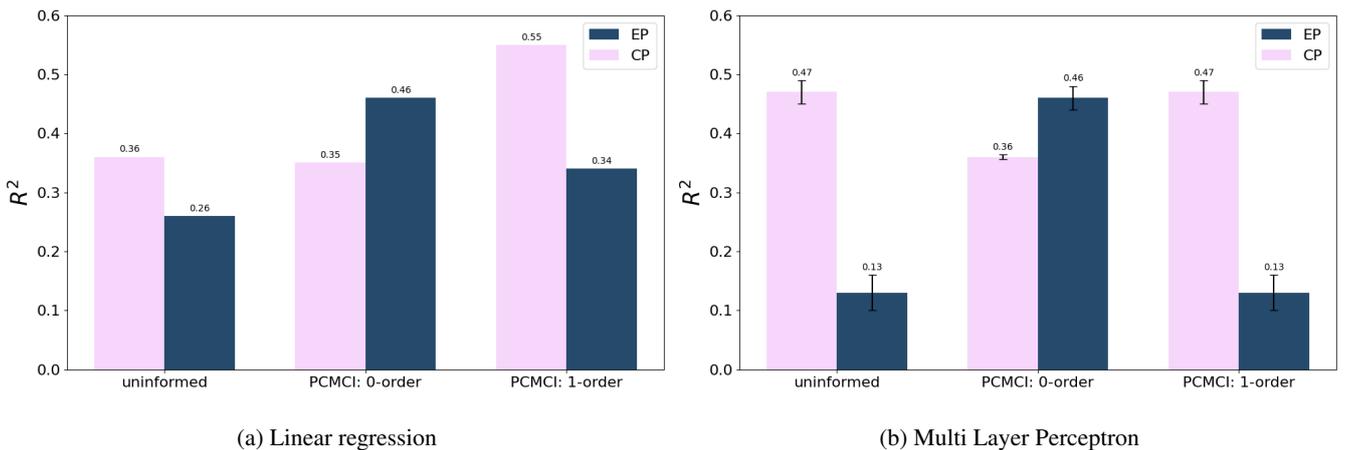


Figure 8. The coefficient of determination (R^2) from predicted and real ENSO indices on unseen testdata for linear regression models and ensemble multi layer perceptrons. For each architecture we test uninformed (all inputs), and PCMCI⁺ informed models, which only use 0-order and 1-order causal ancestors (see Fig. 7) of EP and CP respectively. The numbers above the bars display the (mean) R^2 score and the error bars show the standard deviation of an ensemble of 10 trained MLPs.

The pre-selected input features improve the prediction skill of linear regression models, except for the 0-order CP model, where R^2 decreases marginally by 0.01 (see Fig. 8a). Linear regression predicts the EP index best by using the 0-order PCMCI⁺ inputs ($R^2 = 0.46$), while CP prediction skill is best for the 1-order PCMCI⁺ inputs ($R^2 = 0.55$). As our dataset exhibits strong
220 multicollinearity (since the shifted timeseries), we expected the multi layer perceptrons to perform better compared to the linear regression, especially in the uninform condition. While this is the case for the CP index, the EP index prediction skill drops.

The 0-order results are very similar between the linear and non-linear model, which indicates a linear relationship between the wind and ENSO indices. The MLP 1-order prediction skill equals the uninform prediction skill. While the MLPs did not profit from additional HF information in the uninform condition, it does also seem clear that especially the EP model
225 overfits on some structure in the data. In comparison to the linear prediction, the EP 1-order MLPs underperform. We tested regularization methods like drop-out layers to avoid the overfitting problem, but did not yield better results due to the sparsity of datapoints. The CP prediction in contrast benefits from including more eastern regions LF wind activity, as 0-order skill is lowest and improves with additional inputs in the 1-order condition for both model types. One can see in Fig. 5 e) that the E2 region does have the second highest R^2 values for the CP index, which could explain why the CP prediction skill improves if
230 this region is given as input.

This experiment demonstrates the ability of PCMCI⁺ method to considerably reduce input dimensionality, while keeping the relevant information when working with shifted time series data. The results further underline the importance of the regional LF wind activity revealed as causally relevant by PCMCI⁺.

4 Discussion

235 In this article, we present a study of spatial and frequency dependent wind forcing ENSO diversity and test ENSO predictability by winds alone. In summary, our results demonstrate the crucial role of LF wind variations on ENSO diversity. Their importance is evident in the regression coefficients of spatio-temporal relevant linear regression models, as well as in the PCMCI⁺ revealed 0-order and 1-order causal relevant sets, which contain solely LF wind indices. In addition, including HF wind activity as input features to the linear and non-linear regression models did not yield better ENSO predictability as excluding them.

240 We obtained these results using observational data from the ERA5 reanalyses dataset, and therefore empirically validate a numerical consideration proposed by Capotondi et al. (2018). They elegantly exploited a two-component linear model containing a slow variable, which is forced by a fast, noise dependent variable. Running the model with a high pass filtered version of the fast variable, does not excite the slow variable. In contrast, the low frequency band of the fast variable is in fact forcing the slow variable (Capotondi et al., 2018). If ENSO dynamics are linear and stochastically forced, then the interannual part
245 of the stochastic wind spectrum is the only relevant frequency range that can effectively trigger SST anomalies (Newman and Sardeshmukh, 2017; Capotondi et al., 2018).

In addition, our results display a spatial coherence between zonal wind and patterns of SST anomalies. Relevant LF wind forcing for the CP index stems from the western Pacific while eastern LF wind activity is relevant for EP. This pattern is evident in the multiple regression analyses and PCMCI⁺ results. The LF wind component in regions E, E1, E2 and E3 seems relevant
250 for EP, while regions W, C and E2 are relevant for CP. Multiple linear regression revealed that the strongest correlation of zonal winds and SST happens concurrently, while PCMCI⁺ shows the strongest partial cross correlation if the wind activity leads the SST by one to three months. We hypothesize, that this discrepancy between the two methods occurs due to the definition of the link strength via the partial correlation. Partial correlation tests the correlations between two time series while the effect of the conditioning set is removed. Therefore, in the PCMCI⁺ algorithm wind activity at a lead time of one month could be
255 insignificant, if the wind activity at a lead time of two months is included in the conditioning set, while linear regression models profit from lower lead times as information of previous months is not explicitly removed.

Our linear regression analyses does not allow to discriminate the cause and effect relationship between the wind and ENSO indices. Models with highest coefficient of determination are found at zero months lead time. Wind variations could either be

a deterministic response to SST anomalies, or they could act as direct triggers of them. In both cases, we would expect the
260 observed linear regression results. In contrast, the PCMCI⁺ algorithm attributes a directionality to the relationship of the wind
indices to EP and CP index. The CP index is caused by LF wind variations at a lead time of three months. The EP index is
caused by LF wind variations at one to two months lead time and contemporaneously causing LF wind variations in region
E and E1, indicating a feedback circle of wind and SST variations over the eastern Pacific. Notably, PCMCI⁺ did not reveal
265 a causal connection of SST variations to the HF wind indices. This absence is somewhat surprising, as positive regional SST
anomalies are causing convection and low air pressure which causes wind.

Another point we want to discuss, is the well studied importance of westerly wind events (WWE) in the western Pacific for
EP events (Chen et al., 2015; Fedorov et al., 2015; Harrison and Chiodi, 2009). Looking at the time series in Fig. 2 in region W
suggests, that especially HF but also LF wind activity acted rather as a forcing than a response for the 2016 EP ENSO event,
as they lead the Nino3.4 Index. Further, we observe a peak in R^2 for EP regression models in the western Pacific region at
270 a lead time of around six months. Interpreting the regression coefficients highlights the importance of the HF wind variation,
while also LF coefficients are significant at higher lead times. We argue that this result could reflect the importance of WWE in
the western Pacific before strong EP events. However, absolute R^2 is still quite low and we can not find any evidence for this
phenomenon in the PCMCI⁺ results. This could be due to the fact that we examined the importance of zonal wind variations
for the whole time period from 1950 to 2020 and not specifically for ENSO events. WWE can cause strong EP events, but
275 they might only correspond well to the EP index at EP ENSO warm events and not during neutral or cold phases. It has been
shown, that WWE excite an EP event if the ocean is in a charged condition, while an CP event occurs if the oceans state is
uncharged (Fedorov et al., 2015; Lian et al., 2014). Since we did not include the ocean state in our analyses, our results are
somewhat limited. Moreover, the real effect of atmospheric forcing can not be seen separately from other physical states like
the upper ocean heat content and known ENSO precursors states, since there are significant interactions (Fedorov et al., 2015;
280 Chen et al., 2015; Lian et al., 2014). It could be therefore that PCMCI⁺ did not reveal the effect of wind activity in the western
Pacific for the EP index. It is the goal of future work to analyze how zonal and frequency depended forcing differs for neutral,
warm and cold ENSO phases.

We are also aware that the ENSO predictability would benefit from inclusion of information about the state of the ocean in
our prediction experiment. However, we restricted ourselves in this study exclusively to the wind forcing on ENSO diversity.
285 This approach allowed us to empirically validate theoretical considerations of Capotondi et al. (2018) and reveal zonal wind
activity relevant for different ENSO spatial patterns and explore ENSO predictability by winds alone. We thereby might provide
information, that has the potential to improve ENSO predictability via detecting the zonal relevant wind activity. Furthermore,
our results support the idea of linear inverse modeling (LIM) for ENSO prediction by empirically validating the importance
of the low frequency spectra of the broadband wind forcing. Significant HF wind regression coefficients of spatio-temporal
290 regression models could depict regions, where the noise forcing on ENSO diversity is relevant. If this relevant noise forcing
could be integrated in linear inverse models, it has the potential to improve their ENSO forecasts (Thomas et al., 2018).

Methodologically, we found that PCMCI⁺ is a practicable regularization method for regression models and allowed for
effective input dimension reduction. Causally informed linear and nonlinear regression models benefited from PCMCI⁺ in-
formation. Climate data records are often limited in data points and neural networks with lot of parameters easily over fit.
295 Especially when data is sparse, the normally expensive PCMCI⁺ algorithm is feasible and can be a convenient regularization
method for neural networks before their training even started. On the other hand, one has to be aware of the limitations and as-
sumptions of PCMCI⁺ method (Runge, 2020). Violations against its assumptions (see Runge et al. (2019)) can result in biased
outcomes. We argue the Causal Markov Condition and Faithfulness to be true in our case, but the Causal Sufficiency Condition
is most likely violated, as we only considered winds as ENSO drivers. This implies that our results could significantly change
300 if we include other known drivers of ENSO. The term causality itself is a precarious word to use and in this context defined
via statistically significant partial correlations. Therefore, we point out that the causal parents set of EP and CP as revealed by

PCMCI⁺, are not assumed to actually cause these indices, but that they correspond strongest to these indices relative to the other wind indices.

In future work, we would like to use wind fields rather than the six equatorial regions, whereby we could improve our spatial resolving of equatorial winds and enable an hemispherical comparison. Instead of relating wind indices to the whole time window, with all ENSO states, we would like to separate our analyses into the different ENSO phases and examine wind forcing especially at times before ENSO warm events. Furthermore, we want to incorporate the ocean heat content into our investigation of spacial and frequency dependent wind forcing ENSO diversity, and investigate wind forcing onto a charge-discharge oscillating background state. Dividing time periods of data depending on a thermocline height limit into recharged or discharged state would allow us to discriminate and evaluate wind forcing onto the oceans state.

5 Conclusions

In a nutshell, we revealed that the low-frequencies of broad band wind forcing are relevant for long term ENSO dynamics. EP and CP are best explained by zonal wind activity in the eastern and western Pacific, respectively. We were able to successfully reduce variable dimension with the causal discovery PCMCI⁺ algorithm and demonstrated a practicable approach that considerably improved prediction skill for linear and non-linear analyses in case of sparse data.

Code and data availability. The code for replicating the presented results the shown graphics is made publicly available under the authors GitHub repository. While one can process the original ERA5 wind data (Hersbach et al., 2018; Bell et al., 2020) with the available code, one can also find a dataset of lagged HF and LF wind indices in the GitHub repository.

Appendix A

Following the approach of Capotondi et al. (2018), we set up the index for HF wind activity denoted in Equation 1. At first, we used the same formula for the LF index, but looking at a comparison of the resulting LF wind index and the Nino3.4 index (see Fig. A1) made us think, that without squaring the LF wind component might yield a better fit to the ENSO index. For example, in the cold ENSO year of 2011 the wind index is high while the ENSO index is low. Motivated by that observation, we construct the LF index without squaring as defined in Equation 1.

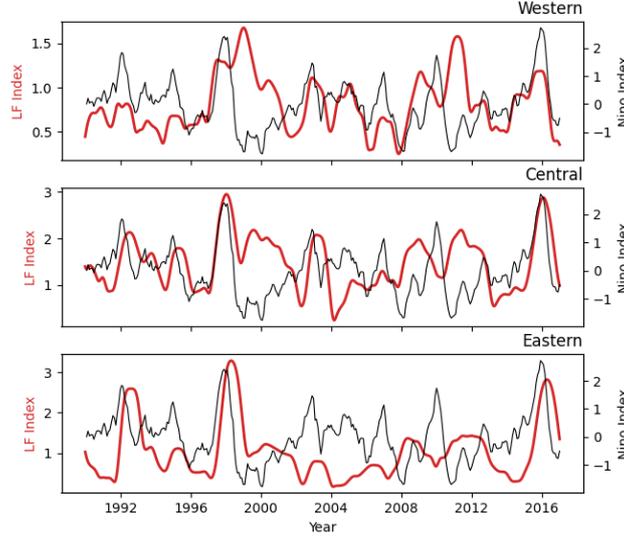


Figure A1. Indices LF zonal wind activity (red; left axis) for three equatorial regions, obtained via Eq. A1, related to the Nino3.4 Index (black; right axis).

Further, we compare the fit of different wind indices definitions and the EP and CP index (see Eq. 2) via multiple linear regression analyses. We test three different wind index definitions: The first index uses the HF formula from Capotondi et al. (2018) for LF and HF (see Eq. A1). The second index uses the formula without squaring for HF and LF index calculation (see Eq. A2). The third index combines both, squaring the HF while keeping the LF index unsquared (see Eq. A3).

$$C - Idx : \quad HF_{region}(t) := \sqrt{g(HF_{region}(t)^2)} \quad LF_{region}(t) := \sqrt{g(LF_{region}(t)^2)} \quad (A1)$$

$$D - Idx : \quad HF_{region}(t) := g(HF_{region}(t)) \quad LF_{region}(t) := g(LF_{region}(t)) \quad (A2)$$

$$M - Idx : \quad HF_{region}(t) := \sqrt{g(HF_{region}(t)^2)} \quad LF_{region}(t) := g(LF_{region}(t)) \quad (A3)$$

335 We fit multiple linear regression models, for each equatorial zone and lead time, using the different wind indices as regressors onto the EP and CP index. We compare the mean RMSE for EP and CP models prediction over the train period from 1952 to 2020 in Fig. A2. We argue that the M-Idx displays the best overall fit. All further analyses is conducted with the M-Idx, as defined in Equations A3 and 1.

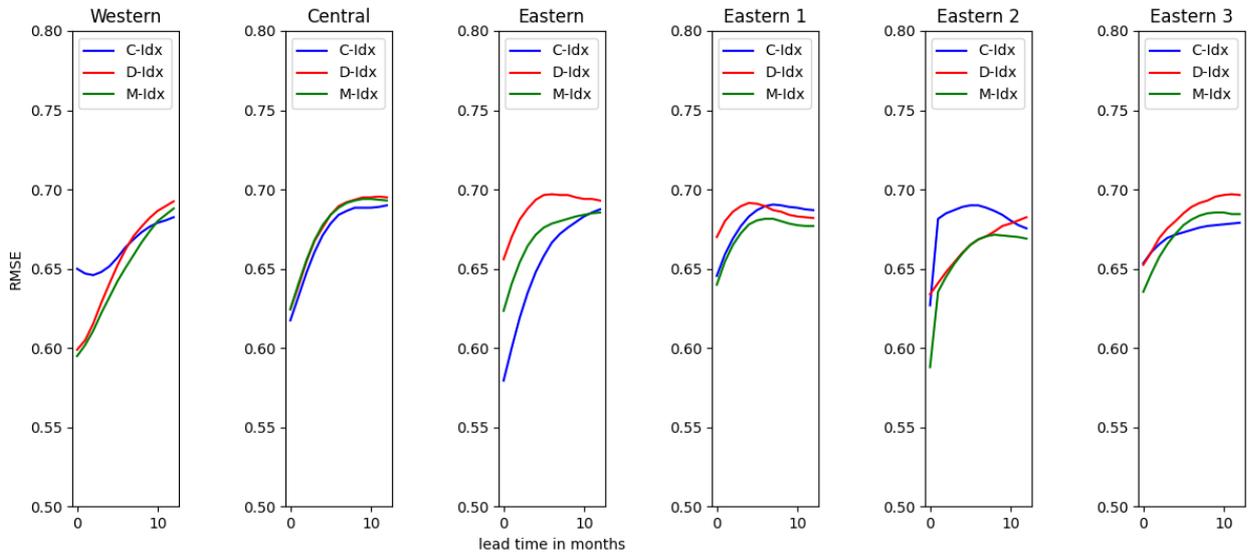


Figure A2. RMSE for the fit of three different wind index calculations for each of the six equatorial regions. See the text for definitions of the indices.

References

- 340 Atlas, R., Hoffman, R. N., Ardizzone, J., Leidner, S. M., Jusem, J. C., Smith, D. K., and Gombos, D.: A Cross-calibrated, Multiplatform Ocean Surface Wind Velocity Product for Meteorological and Oceanographic Applications, *Bulletin of the American Meteorological Society*, 92, 157 – 174, <https://doi.org/10.1175/2010BAMS2946.1>, 2011.
- Bell, B., Hersbach, H., Berrisford, P., Dahlgren, P., Horányi, A., Muñoz Sabater, J., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1950 to 1978 (preliminary version), Copernicus Climate Change Service (C3S) Climate Data Store (CDS), <https://cds.climate.copernicus-climate.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-preliminary-back-extension?tab=overview>, accessed on 25-04-2022, 2020.
- 345 Beucler, T., Iglesias-Suarez, F., Eyring, V., Pritchard, M., Runge, J., and Gentine, P.: Climate-Invariant, Causally Consistent Neural Networks as Robust Emulators of Subgrid Processes across Climates, Tech. rep., EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-722, <https://doi.org/10.5194/egusphere-egu22-722>, 2022.
- 350 Bjerknes, J.: Atmospheric teleconnections from the equatorial Pacific, *Monthly Weather Review*, 97, 163 – 172, [https://doi.org/10.1175/1520-0493\(1969\)097<0163:ATFTEP>2.3.CO;2](https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2), 1969.
- Capotondi, A., Sardeshmukh, P. D., and Ricciardulli, L.: The Nature of the Stochastic Wind Forcing of ENSO, *Journal of Climate*, 31, 8081 – 8099, <https://doi.org/10.1175/JCLI-D-17-0842.1>, 2018.
- Chen, D., Lian, T., Fu, C., Cane, M., and Tang, Y.: Strong influence of westerly wind bursts on El Niño diversity, *Nature Geoscience*, 8, 339–345, <https://doi.org/10.1038/ngeo2399>, 2015.
- 355 Di Capua, G., Runge, J., Donner, R. V., van den Hurk, B., Turner, A. G., Vellore, R., Krishnan, R., and Coumou, D.: Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales, *Weather and Climate Dynamics*, 1, 519–539, <https://doi.org/10.5194/wcd-1-519-2020>, 2020.
- Eisenman, I., Yu, L., and Tziperman, E.: Westerly Wind Bursts: ENSO’s tail rather than the dog?, *Journal of Climate*, 18, <https://doi.org/10.1175/JCLI3588.1>, 2005.
- 360 Fedorov, A., Hu, S., Lengaigne, M., and Guilyardi, E.: The impact of westerly wind bursts and ocean initial state on the development, and diversity of El Niño events, *Climate Dynamics*, 44, 1381–1401, <https://doi.org/10.1007/s00382-014-2126-4>, 2015.
- Harrison, D. E. and Chiodi, A. M.: Pre- and Post-1997/98 Westerly Wind Events and Equatorial Pacific Cold Tongue Warming, *Journal of Climate*, 22, 568 – 581, <https://doi.org/10.1175/2008JCLI2270.1>, 2009.
- 365 Harrison, D. E. and Vecchi, G. A.: Westerly Wind Events in the Tropical Pacific, 1986–95, *Journal of Climate*, 10, 3131 – 3156, [https://doi.org/10.1175/1520-0442\(1997\)010<3131:WWEITT>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<3131:WWEITT>2.0.CO;2), 1997.
- Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1979 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), <https://doi.org/10.24381/cds.adbb2d47>, accessed on 25-04-2022, 2018.
- 370 Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippet, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F.: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction, *Bulletin of the American Meteorological Society*, 95, 585 – 601, <https://doi.org/10.1175/BAMS-D-12-00050.1>, 2014.
- 375 Lehmann, J., Kretschmer, M., Schauburger, B., and Wechsung, F.: Potential for Early Forecast of Moroccan Wheat Yields Based on Climatic Driver, *Geophysical Research Letter*, 47, <https://doi.org/10.1029/2020GL087516>, 2020.
- Lian, T., Chen, D., Tang, Y., and Wu, Q.: Effects of Westerly Wind Bursts on El Niño: A New Perspective, *Geophysical Research Letters*, <https://doi.org/10.1002/2014GL059989>, 2014.
- Newman, M. and Sardeshmukh, P. D.: Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures?, *Geophysical Research Letter*, 44, 8520–8529, <https://doi.org/10.1002/2017GL074088>, 2017.
- 380 NWS: Monthly ERSSTv5 (1950-2022) Nino Indices, NOAA Center for Weather and Climate Prediction, <https://www.cpc.ncep.noaa.gov/data/indices/ersst5.nino.mth.91-20.ascii>, accessed on 25-04-2022, 2020.
- Ren, H.-L., Scaife, A. A., Dunstone, N., Tian, B., Liu, Y., Ineson, S., Lee, J.-Y., Smith, D., Liu, C., Thompson, V., et al.: Seasonal predictability of winter ENSO types in operational dynamical model predictions, *Climate Dynamics*, 52, 3869–3890, <https://doi.org/10.1007/s00382-018-4366-1>, 2019.
- 385

- Roulston, M. and Neelin, D.: The response of an ENSO Model to climate noise, weather noise and intraseasonal forcing, *Geophysical Research Letters*, 27, 3723–3726, <https://doi.org/10.1029/2000GL011941>, 2000.
- Runge, J.: Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, http://auai.org/uai2020/proceedings/579_main_paper.pdf, 2020.
- 390 Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, *Science Advances*, 5, eaau4996, <https://doi.org/10.1126/sciadv.aau4996>, 2019.
- Spirtes, P., Glymour, C., and Scheines, R.: *Causation, Prediction, and Search*, 2nd Edition, MIT Press, Boston, 2000.
- Strnad, F. M., Schlör, J., Fröhlich, C., and Goswami, B.: Teleconnection patterns of different El Niño types revealed by climate network curvature, *arXiv preprint*, <https://doi.org/10.48550/ARXIV.2203.07035>, 2022.
- 395 Sullivan, A., Luo, J., Hirst, A. C., Bi, D., Cai, W., and He, J.: Robust contribution of decadal anomalies to the frequency of central-Pacific El Niño, *Scientific Reports*, 6, <https://doi.org/10.1038/srep38540>, 2016.
- Thomas, E. E., Vimont, D. J., Newman, M., Penland, C., and Martínez-Villalobos, C.: The Role of Stochastic Forcing in Generating ENSO Diversity, *Journal of Climate*, 31, 9125 – 9150, <https://doi.org/10.1175/JCLI-D-17-0582.1>, 2018.
- 400 Wallace, J., Rasmusson, E., Mitchell, T., Kousky, V., Sarachik, E., and Von Storch, H.: On the structure and evolution of ENSO-related climate variability in the tropical Pacific: Lessons from TOGA, *Journal of Geophysical Research: Oceans*, 103, 14 241–14 259, <https://doi.org/10.1029/97JC02905>, 1998.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum

Unterschrift