<u>Breaking the Spring Barrier: Forecasting El Nino with an Improved ENSO Index</u>

By: Allison Du and Qi Jia (Jerry) Sun

Introduction:

The El Nino Southern Oscillation (ENSO) is an interannual phenomenon that is caused by a combination of weakened trade winds and oceanic processes. When ENSO is not in effect, the waters off the West Coast of South America are cooler than the rest of the Pacific. Thus due to the weakened trade winds, a convection cell causes air to descend near South America and rise near Australia. The Warm ocean currents then move East and the atmospheric convection follows, bringing heavy rains to South America and drought to Australia, drastically altering the weather patterns in the Pacific.

ENSO phases are unpredictable; records demonstrate that cycles have lasted between two and seven years. Generally, events reach their peak during the Northern Hemisphere winter and decay during the proceeding spring or summer. The accuracy of prediction models reaches a valley during this spring period in which ENSO is transitioning when signals are low and noise is high. The difficulty to predict the status of the upcoming winter (El Nino, La Nina or neutral conditions) in the springtime has been coined as the Spring Predictability Barrier. Dynamical models outperform statistical models in the spring, due to their ingestion of recently observed data in comparison to statistical models' adherence to monthly or seasonal average data, which hints at the existence of this barrier as these models return similar results during other seasons.

The ENSO cycle causes global changes in temperature and rainfall reflected anywhere from the Gulf of Mexico to the East Coast of Kenya. While more economically developed countries like the United States have accumulated an estimated $4 billion in costs from the 1997-1998 ENSO, it is offset by $19 billion in benefits. The World Health Organization (WHO) predicts that over 60 million people were impacted by the 2015-2016 ENSO, many of which are rendered helpless in smaller and less economically developed countries. Patching up ENSO's natural consequences such as drought and flooding is extremely costly with $3.6 billion U.S. dollars required for the humanitarian response by WHO.

The unpredictability of ENSO drastically affects farmers on the West Coast of South America. This area of South America, Columbia, Ecuador, and Peru, consists mainly of traditional subsistence farming and fishing heavily dependent on rainfall and sea surface temperatures. The farmers of these developing countries, due to limitations on ENSO forecastability, are unable to match the seasons with their crops, leading to crop failures and famines.

ENSO increases flooding risks, with the most prominent rise in the Southwest U.S, parts of southern South America and the Horn of Africa. Lima, Peru suffered greatly when hit by ENSO-fuelled flash floods and landslides due to their lack of preparation. While heavy rainfall may not always induce floods, it can allow for cholera and other water-borne disease outbreaks. Communities geographically isolated from aid or with insufficient health services such as those in Mozambique.

An ENSO prediction model is beneficial to the entire globe as forewarnings of drought or flooding will help communities prepare to face the adversities or reap the benefits. Knowing when ENSO may bring heavy rainfall to your region allows you to determine when to start your crop rotation for better irrigation and bountiful harvests. Furthermore, knowledge on the ENSO phenomena can elucidate the intricacies of the unpredictable atmospheric and oceanic forces at work on the planet.

Related Work:
ENSO has been a defining weather phenomenon in history and there have been a plethora of attempts to quantify its features. For example, in the network theory paper, researchers incorporated variables such as the integrated warm water volume or WWV, to predict ENSO on a 12-month scale. However, an underlying bias is identified in all of the existing papers. They each heavily incorporate sea-surface temperature as a critical variable. Evidently it is the most crucial variable in classification as it is a direct indication of ENSO, but models may recognize an upcoming ENSO event solely due to sea surface temperature increases rather than a combination of other subtle but significant indicators. Furthermore, sea-surface temperatures are hard to predict in advance due to the multitude of factors at play. Other work involves using a grid of 15 coordinate points within the ENSO basin and 193 outside. These points are connected as nodes and their link strength determined by their cross-correlation

Objective:
The purpose of this project is to statistically determine the key indicators of ENSO and to devise a novel index of the importance of the aforementioned features that correctly encapsulates the essence of ENSO without the biased SST factor.

Methodology:

Data Collection:
30 reputable datasets from various governmental and non-governmental organizations — National Oceanic and Atmospheric Administration (NOAA), European Centre for Medium-Range Weather Forecasts (ECMWF), Earth Science Research Laboratory (ESRL), and Coral Reef Temperature Anomaly Database (CoRTAD) — were examined in detail for data-extractability, coverage, and completeness. These datasets each contained variables pertaining to specific environmental and chemical processes. After a thorough evaluation of the relevance of individual oceanic and atmospheric factors to ENSO, the NOAA v5, specifically, the NCEP/NCAR Reanalysis Monthly Means and Other Derived Variables was the clear optimal choice because it had the most variability in their datasets. All the latitude-longitude intervals were standardized for NCEP/NCAR Reanalysis datasets, which allowed for all our chosen variables to lie on the exact same coordinate points. Furthermore, the labels were taken from the Nino 3.4 Index, a dataset that labeled the change in temperatures of the sea surface in a specific region in the pacific ocean.

Data Processing:
Due to the multidimensional nature of the files provided by the NOAA, a program in R was

written to convert the files into readable ones (.csv) for machine learning. The files were concatenated into a single HDF5 file which was then used to train the Keras model in python. Only specific values that corresponded to regions inside the Nino 3.4 region were chosen to form a 6x27 point grid. This included values that are 5 degrees latitude North and South of the equator and 170-120 degrees West of the Prime Meridian. The Nino 3.4 region was selected as it sees the most effects of ENSO, and thus, ENSO indicators are at peak prominence.

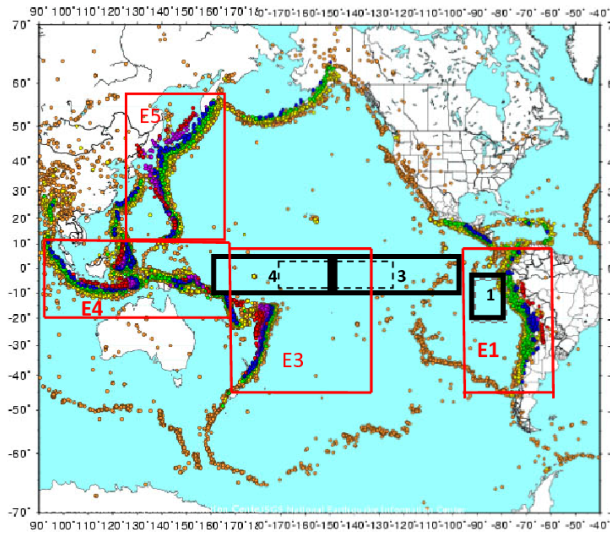


Figure 1: The dotted box represents the region covered by the Nino 3.4 index as well as our selected region.

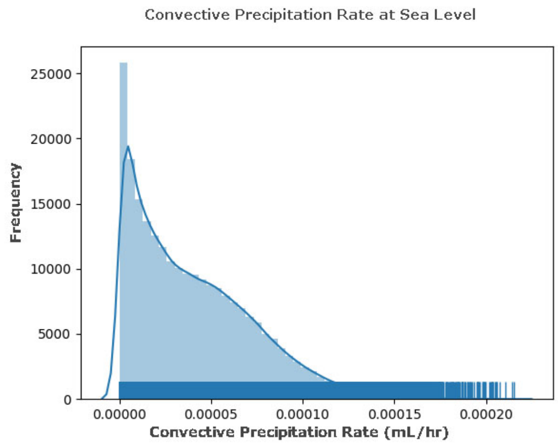
Model:

After careful selection, it was decided that a mean squared error regression model would be most appropriate because it is considered optimal at minimizing the loss function. Instead of a standard classification problem, a regression problem would be more appropriate because of its flexibility to not only classify the existence of ENSO, but also the magnitude. The regression was done by forecasting the future months based on the past 12 months of data. In this model, we processed the data of each variable and generated images corresponding to the Nino 3.4 region for the CNN as training data. The dataset was split 7:3 into training and validation data respectively.

<u>Results:</u>

The regression model reached a validation accuracy of 1.28%. However, if we take the inverse, this model has a validation accuracy of 98.72% due to the binary nature of the categories. This may be a result of a reversal of labels. The model is useful in the sense that it limits the outcomes that are possible due to its low accuracy rate. It also elucidates the difficulties of classifying without an SST model and highlights the unpredictability of the ENSO phenomenon. However, a prime criterion of the mean square loss regression model is that the variables must be independent of each other, and it may appear that this factor was overlooked. Although the variables are inherently independent, there exist many climate and atmospheric factors that act as lurking variables influencing each individual variable and not rendering them truly independent.

Furthermore, there exists some data that was not fully useful, due to the difficulty of collecting the data, and the errors in the data collected (months with no data entries or 'NA' entries).



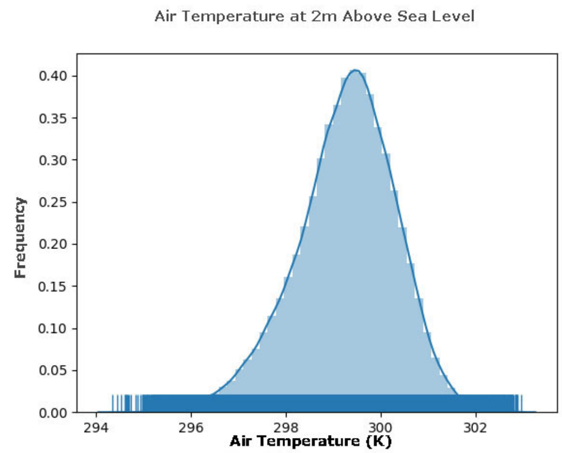


Figure 2 and Figure 3: While most data had decent variability like Air Temperature, Convective Precipitation Rate has a range of around 0.0012 (mL/hr) which the model may mistake for noise.

Conclusion:

This regression model reached a validation accuracy of 1.28%. However, this can be reversed and the model can be seen as a means to limit the possible outcomes. Once the accuracy is inverted, this model provides a foundation of significant importance. It reveals the importance of SST in classification and shows the difficulty of dependence on other variables. Although ENSO is a result of a multitude of factors, it appears that early stage detection still has its limitations with things that are not SST. However, SST extrapolation is limited and perhaps this is an indication to the existence of the Spring Barrier.

Future Work:

The main pitfall of our model was a lack of training data. As ENSO is a global phenomenon, our selected spatial area (based on the Nino 3.4 index) is small in comparison to the total regions affected. The changes in our selected data per month were nuanced and perhaps seen as mere noise by the model. Furthermore, some datasets such as pevpr.csv (potential evaporation rate) had columns missing from certain latitude-longitude coordinates, which takes away over 140,000 data points from the training dataset. To bolster the accuracy of this model, more datasets can be analyzed from different sources to select the most optimal in terms of completion and relevancy. Future work entails datasets with a higher concentration of data points per area and expansion of the selected area so that larger patterns at play can be detected. It was assumed that 162 data points per variable would be sufficient since the separation between each point was a mere longitude (1.875) and latitude (1.91). However, due to the data-dependent nature of computer vision models, more data will drastically improve the validation accuracy and overall improve the performance of the classifier.

<u>**References**</u>

Convolutional Layers. (n.d.). Home - Keras Documentation. https://keras.io/layers/convolutional/

El Niño Southern Oscillation as an early warning tool for dengue outbreak in India. (2020). https://doi.org/10.21203/rs.2.21293/v1

El Nino's trail of destruction in Africa. (n.d.). News and current affairs from Germany and around the world | DW. https://www.dw.com/en/el-ninos-trail-of-destruction-in-africa/

Fig. 1. Map of data collection. (2013, December 1). ResearchGate. https://www.researchgate.net/figure/Map-of-data-collection-Thick-rectangles-show-zones-of-SSTOI-indices-registration-Nino_fig1_41908083

Impacts. (n.d.). NOAA Pacific Marine Environmental Laboratory (PMEL) |. https://www.pmel.noaa.gov/elnino/impacts-of-el-nino

Ludescher, J. (2019, October). Very early warning signal for El Nino in 2020 with a 4 in 5 likelihood. https://arxiv.org/pdf/1910.14642.pdf

PSD Web Team. (n.d.). ESRL : PSD : PSD Data: NCEP/NCAR Reanalysis Monthly Means and Other Derived Variables. NOAA Earth System Research Laboratory. https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.surfaceflux.html

PSD Web Team. (n.d.). ESRL : PSD : NCEP-DOE AMIP-II Reanalysis (AKA Reanalysis 2). NOAA Earth System Research Laboratory. https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html

World Health Organization. (2016, April 26). El Niño affects more than 60 million people. WHO | World Health Organization.

Salman, A. G. (2015, October 1). Weather forecasting using deep learning techniques. ResearchGate. https://www.researchgate.net/publication/304256826_Weather_forecasting_using_deep_learning_techniques