AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

RESEARCH ARTICLE
10.1029/2018EA000423

Analyzing El Niño–Southern Oscillation Predictability Using Long-Short-Term-Memory Models

Andrew Huang[1], **Ben Vega-Westhoff**[1], **and Ryan L. Sriver**[1]

[1]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Key Points:
- Linear regression and long-short-term-memory model capabilities were compared in their ability to predict monthly Niño-3.4 indices, from 1 to 11 months ahead, using monthly and daily inputs of sea surface temperatures, zonal winds, and warm water volumes in the central, western, and overall equatorial Pacific, respectively
- Results show that LSTM had little advantage over LR in terms of correlation when monthly data were used as predictors; both models showed high skill at early leads, dropping off rapidly with longer leads; however, advantages became noticeable when daily data were used, particularly with predictions involving SST and predictions at longer leads
- Results indicate that some nonlinear information is encoded in the daily data of SST that LSTM can detect and utilize in predictions at longer leads

Supporting Information:
- Supporting Information S1

Correspondence to:
A. Huang,
ahuang11@illinois.edu

Citation:
Huang, A., Vega-Westhoff, B., & Sriver, R. L. (2019). Analyzing El Niño-Southern Oscillation predictability using long-short-term-memory models. *Earth and Space Science*, *6*, 212–221. https://doi.org/10.1029/2018EA000423

Received 14 JUN 2018
Accepted 19 DEC 2018
Accepted article online 13 JAN 2019
Published online 18 FEB 2019

Abstract El Niño–Southern Oscillation (ENSO) can have global impacts, affecting daily temperature and precipitation, and extreme weather, such as hurricanes and tornadoes. Because of its importance, scientists strive to understand the processes that govern ENSO and develop models to predict its evolution and changes in variability. Here long-short-term-memory models (LSTMs) were compared to linear regression models (LR) to explore the benefits of simple, deep neural networks in predicting ENSO, in addition to quantifying the relative importance of the sources of ENSO's predictability. The models use central Pacific sea surface temperatures (SST), equatorial Pacific warm water volumes, and western Pacific zonal winds as predictors, individually and in combinations, on monthly and daily resolutions, from 1- to 11-month leads. By using these predictors, many characteristic time scales are encompassed—from days-to-weeks in the atmosphere, to months-to-seasons in the coupled system, and interseasonal-to-interannual in the subsurface ocean. Results show, with monthly input, predictions from LSTM were like predictions from LR. However, with daily SST at longer leads, LSTM exhibited some advantage over LR in terms of the correlation coefficient. This suggests that daily SST may contain some nonlinear element that improves LSTM predictability compared to LR. In addition, this suggests that more information, such as gridded data and additional variables, would likely improve predictability using LSTM, but results would be more difficult to interpret. Overall, LSTM may be appealing because once the computationally expensive training of LSTM is complete, the predictions employing the trained model can be relatively cheap to perform thereafter.

1. Introduction

A rock in a stream can affect the flow of the stream. Similarly, the El Niño–Southern Oscillation (ENSO) can influence the atmospheric flow with its warmer than normal sea surface temperatures (SST) in the eastern central Pacific during its El Niño phase, or cooler than normal SST during its La Niña phase. These different phases can influence large-scale atmospheric dynamics and jet stream variability, which have been shown to lead to anomalous temperature and precipitation patterns across the United States and the world (Glantz, 2001; Mason & Goddard, 2001; Ropelewski & Halpert, 1987, 1989). ENSO can also influence severe and hazardous weather activity, such as tropical cyclones and tornadoes (Allen et al., 2015; Pielke & Landsea, 1999). These impacts will likely persist into the future, thus analysis of ENSO variability in climate models is a major area of current research (Bellenger et al., 2014), including estimating ENSO changes under parametric model uncertainties (Sriver et al., 2014) and internal variability (Vega-Westhoff & Sriver, 2017). Overall, knowledge acquired from these types of studies can then be applied to improve the operational models that forecast ENSO phases from the next month to the next year, ultimately helping countries prepare for the impending impacts (Chapman et al., 2015).

Statistical models, simple and complex, have been used to predict ENSO evolution. These types of models extrapolate empirical relationships and are usually computationally inexpensive. One example used is the canonical correlation analysis model, a multidimensional analysis between a set of predictands and predictors that maximizes the relationship between the two sets. The sets can be gridded time series, so that the evolution is captured and can be projected outward for a prediction (Barnston & Ropelewski, 1992). Another example is the constructed analog model, which retrieves several analogs (comparable months of the predictor in terms of least error) and applies weights to match the corresponding predictand at those times of the analogs to make a prediction (Van den Dool, 1994). These two statistical models have previously been found to be competitive with older dynamical models in forecasting ENSO, with all

correlation coefficients averaging to about 0.6 at 6-month lead in the historical period, 1956–1993 (Barnston et al., 1994).

A more recent study found that dynamical models generally outperform statistical models in forecasting ENSO (Barnston et al., 2011). At a lead time of 6 months in the historical period, 2002–2011, the best dynamical model boasted an average correlation value of 0.7 while the best statistical model at that time only had a correlation value at 0.5, lower than the earlier study, attributed to an interdecadal shift in variability and the mean state of the tropical Pacific Ocean (Hu et al., 2012. 2017). Now, operational forecasters often use the North American Multi-Model Ensemble (NMME), which combines many individual dynamic model ensembles (Kirtman et al., 2013). This approach, on average, yields relatively better prediction skill than any individual model ensembles (Becker et al., 2014; Smith et al., 2013; Tippett et al., 2011).

However, there are also other models that have skill comparable to, or in some regions, even better than, the NMME. One such model is the linear inverse model, with a premise that the tropical atmosphere-ocean system is a stable linear system driven by spatially coherent Gaussian white noise (Penland & Sardeshmukh, 1995). The linear inverse model is found to be much more skillful than the NMME in much of the equatorial strip (Newman & Sardeshmukh, 2017). There are also hybrid models—statistical techniques used to postprocess dynamical models with the goal of further improving dynamical models' performance like the Bayesian updating technique (BU) utilized on the NMME where BU essentially weights the individual model forecasts based on the strength of their regressed fit to observations (Zhang et al., 2017). They find that for almost all leads, especially short leads, BU model outperforms NMME.

Because of advances in data, hardware, and techniques leading to the advent of deep learning, there are now new opportunities to explore statistical ENSO forecasting. Deep learning is a form of machine learning that uses layered, hierarchical representations of data, or neural networks, to learn iteratively which of the interconnected, abstract features of the data help advance its performance. Neural networks are inspired by biological brains—the input neurons are activated through stimuli, which may then lead to reactions, often nonlinear, from other neurons, based on assigned thresholds. The thresholds, or weights, between each neuron are designated by how effective the connection is in leading to the desired output during its training period (Schmidhuber, 2015). Training a neural network can be computationally expensive, but upon completion of its training, the neural network can be reused inexpensively to conduct predictions.

Different forms of neural networks have been applied previously in atmospheric sciences. For instance, convolutional neural networks have been used to estimate tropical cyclone intensity using a deep convolutional neural network, with the goal of superseding the Dvorak method (Pradhan et al., 2018). Their classification accuracies were found to be as high as 95.47% for the top two probable intensities. In addition, feed-forward neural networks have been used to forecast SST (Tangang et al., 1997) with correlations comparable to the older, aforementioned dynamical models. Here the capabilities of deep neural networks, specifically long-short-term-memory models (LSTM; Hochreiter & Schmidhuber, 1997), are investigated from a variety of aspects, focusing on the Niño-3.4 (N34) region and with a longer time period, extending past the 2000s. LSTM is an advanced recurrent neural network (RNN), and its specialty is that it can retain short segments of memory for long term, unlike a vanilla RNN which is only able to remember for short periods of time before suffering from vanishing and exploding gradient problems (Gers & Schmidhuber, 2001). This same logic can be applied to ENSO predictions. For example, if the observations show an elongated La Niña phase consisting of many time steps, then LSTM may realize this and predict an El Niño next.

For reference we compare results from LSTM with linear regression models (LR) from 1- to 11-month leads, where a 1-month-lead prediction is simply the past month's value used to predict the following month's value, for example, January SST predicting February SST. Daily and monthly inputs of SST, warm water volume (WWV), and 925-mb zonal winds (WND) were used individually and in combination as predictors. Each predictor captures a distinct, key process and characteristic time scale that helps modulate the evolution of ENSO from either the atmosphere, ocean, or coupled system, across daily to interannual time scales. The data and methods are discussed in depth in section 2, the results and discussion are presented in section 3, and a conclusion is offered in section 4.

2. Data and Methodology

The data sets used were all sliced from 1982 to 2017. N34 predictors were computed from the National Oceanic and Atmospheric Administration (NOAA) Optimum Interpolation (OI) SST v2 data set by taking the daily SST values and subtracting the monthly climatological (1981–2010 or the base period) mean, averaged over 5°S to 5°N and 170–120°W (Trenberth, 1997), reflecting the methods at the Climate Prediction Center. WND predictors were computed from the National Centers for Environmental Prediction (NCEP) Reanalysis data set by taking the daily WND values and subtracting the base period mean, averaged over 5°S to 5°N and 120–160°E (Wang et al., 2011). WWV predictors, defined as the volume above the 20 °C isotherm over the region 5°S to 5°N and 120°E to 80 W (Meinen & McPhaden, 2000), were retrieved from the Tropical Atmosphere Ocean (TAO) Array data, and its anomalies were taken in a similar fashion by subtracting the base period mean. (supporting information Figure S1).

WND was chosen as a predictor to try capturing the high frequency atmospheric variability, which includes westerly wind bursts and Madden-Julian Oscillation (MJO). WWV, another predictor, captures the low-frequency variations in ocean heat as it is transported to higher latitudes during the El Niño and must build up again before another El Niño is able to occur (Wyrtki, 1985). Lastly, SST acts as the interface between the ocean and atmosphere, capturing the coupled ocean-atmosphere variability, and SST anomalies also define many ENSO indices, such as the N34 index. These predictors' effectiveness was then tested on a monthly and daily resolution, individually and in combinations (SST-WWV, SST-WND, SST-WWV-WND, etc.). Similar indices were used in the Florida State University multivariate regression model (Clarke & Van Gorder, 2003).

ENSO predictions were performed with LR and LSTM. LSTM was constructed using Python's keras, and has three layers: two LSTM layers, the first with 50 nodes and the second with 150 nodes, connected to a dense layer, compiled with a mean-square error (MSE) loss function and the Adam optimizer (Kingma & Ba, 2014), and finally trained on eight epochs with a batch size of eight. The epochs and batch sizes were tested not only to attain appreciable skill but also to minimize runtime and maximize the number of forecasts made.

Using monthly data, predictors were shifted backward in time by the number of months equaling the number of leads (supporting information Figure S2). An example of a one-lead prediction is using averaged February predictors to predict the averaged March's N34 value. For four leads, averaged February predictors were used to predict the averaged June's N34 value, and so on. To utilize daily data, predictors were transposed next to monthly N34 values and shifted backward in time by the number of months equaling the number of leads. If a month does not contain 31 days, the last daily value was forward filled, or repeatedly copied over until the total length of the row equaled 31 (Figure S3). Since WWV was only available monthly, its monthly average had to be forward filled 31 days for predictions on the daily time scale.

To evaluate skill, the correlation coefficient was computed between the observed and the predicted N34 values. To explore the variability across years, a jackknife resampling technique was employed (Efron, 1982). This involves reserving some years in the time series as the validation period and the rest as the training period (Figure S4). For instance, in one sample, the models were trained using 1986–2017 to predict 1983–1985; in another sample the model was trained using 1983–1985 and 1989–2017 to predict 1986–1988. This method produces a total of 11 samples, each with a validation sample period of 3 years. The 3-year validation sample period was selected based on the following considerations. If the validation sample period had been too small, autocorrelation may have biased the correlations. Alternatively, if the period had been too large, the variability across years would have been difficult to explore, and the number of samples would have been too small to test for statistical significance. The 3-year validation sample period minimized this tradeoff based on the length of the time series used in this study.

To test statistical significance, on each predictor combination, an overall average for each jackknife window was found by transforming all correlation coefficients to Fisher Z-space then averaging across the eleven leads, leaving 11 independent Fisher Z-values, one for each jackknife sample. These averaged Fisher Z-values were then used to perform two-sided t tests to measure whether the two types of overarching models' (LSTM and LR) correlations of the 11 jackknife samples differ significantly for each predictor combination. If the resulting p values are observed to be less than 0.05, then the null hypothesis, which the two models' average correlations are identical, can be rejected. Otherwise, the differences between correlations in the two models cannot be explicitly stated to be significantly different. These hypotheses tests were also repeated for the two input temporal resolutions: monthly and daily.

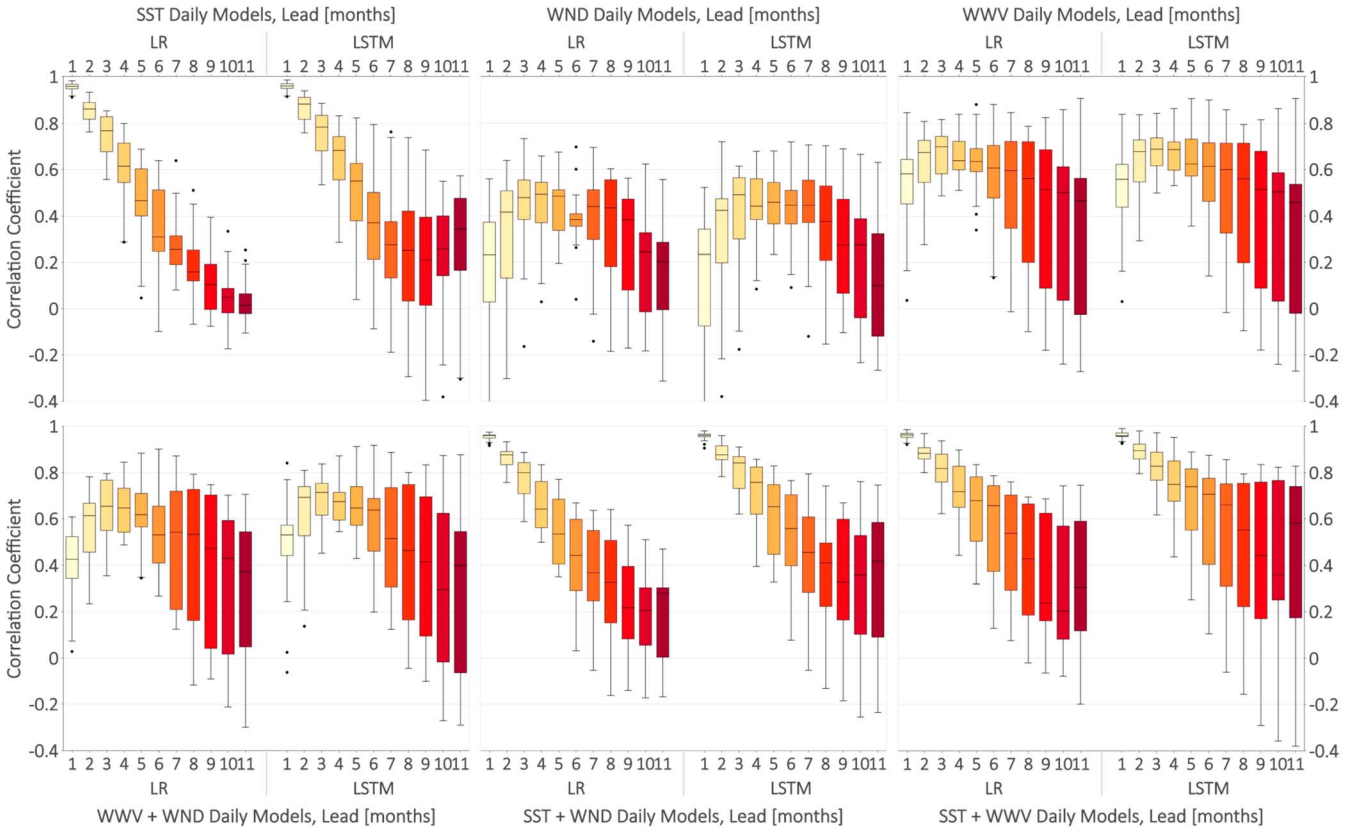

Figure 1. Box whisker plots of correlations for each daily model and lead month with the top row containing correlations from single predictors (left to right: SST, WND, and WWV). Results from LR and LSTM are shown for each variable. The bottom row shows correlations from two predictors (left to right: WWV+WND, SST+WND, SST+WWV). Results from LR and LSTM are shown for each pair of variables. SST has strong correlations at short leads, but steadily decreases to zero at longer leads; WWV and WND, however, maximize out at around five leads (top row). Models that combine SST with WND and/or WWV combine the strengths of the individual models to produce relatively high correlations at all leads (bottom row). Comparing LR and LSTM, with monthly data, LSTM has little advantage over LR (not shown). However, for predictions done with daily time scale data, LSTM generally has higher average correlation across all leads, but at the same time, it shows larger spread. SST = sea surface temperature; WWV = warm water volume; WND = zonal winds; LR = linear regression models; LSTM = long-short-term-memory models.

3. Results and Discussion

To achieve an overview of all 11 jackknife samples, 11 leads, and 7 combinations of predictors, the samples were consolidated into box and whisker plots while the leads and predictors are nested within separate sectors (Figure 1). Out of the three predictors, SST, WWV, and WND, individually, SST was found to have high correlations at early leads for both LR and LSTM, dropping off rapidly with longer leads. In contrast, WWV and WND both started with low correlations in the early leads, peaking at mid leads, and dropping off again at longer leads. When predictions were made with both SST and WWV together or SST and WND together, the drawbacks were offset by the partner variable, leading to comparatively higher correlations across all leads. Also, WWV alone had higher overall correlations than WND alone, and thus, the correlation of SST and WWV together exceeded SST and WND together. There was not much difference when using WWV and WND together. When all three predictors were used, SST, WWV, and WND, the correlations were only slightly improved relative to SST and WWV (Figure 2).

When comparing LSTM and LR using monthly resolution data, little to no difference in correlation was found across the samples for all parameters and leads. However, when using daily resolution data, predictions performed with LSTM at longer leads and involving SST resulted in relatively higher correlations compared to LR (Figure 3). This could be because daily data exhibits more nonlinear evolution of SST, which LSTM is able to extrapolate and utilize in its memory (Alahi et al., 2016), which is not captured by the monthly data.

10.1029/2018EA000423

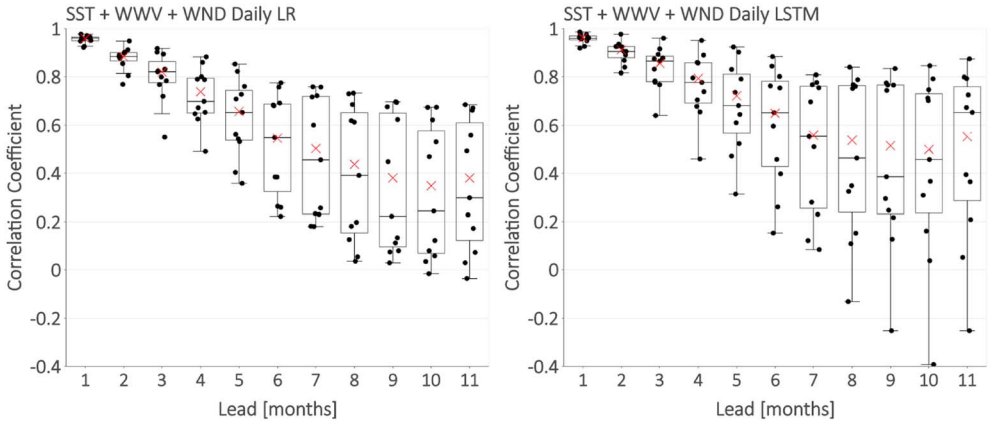

Figure 2. Box whisker plots of correlations for each daily model and lead month using all three of the predictors together: SST, WWV, and WND. The red X signifies the fisher Z-transformed average, while the black line within the boxes signifies the median. Left panel shows results from LR and right shows results from LSTM. Compared to LR, LSTM generally exhibits higher averages and medians, but also broader tails, at longer leads. SST = sea surface temperature; WWV = warm water volume; WND = zonal winds; LR = linear regression models; LSTM = long-short-term-memory models.

Predictions performed using daily, forward filled WWV data did not indicate any improvements in LSTM over LR, which is not surprising because the monthly resolution was artificially transformed into daily resolution. WND did have actual daily data, but unlike SST, the LSTM predictions were not generally better than the LR predictions. By subtracting the 30-day running mean of both SST and WND from its respective daily

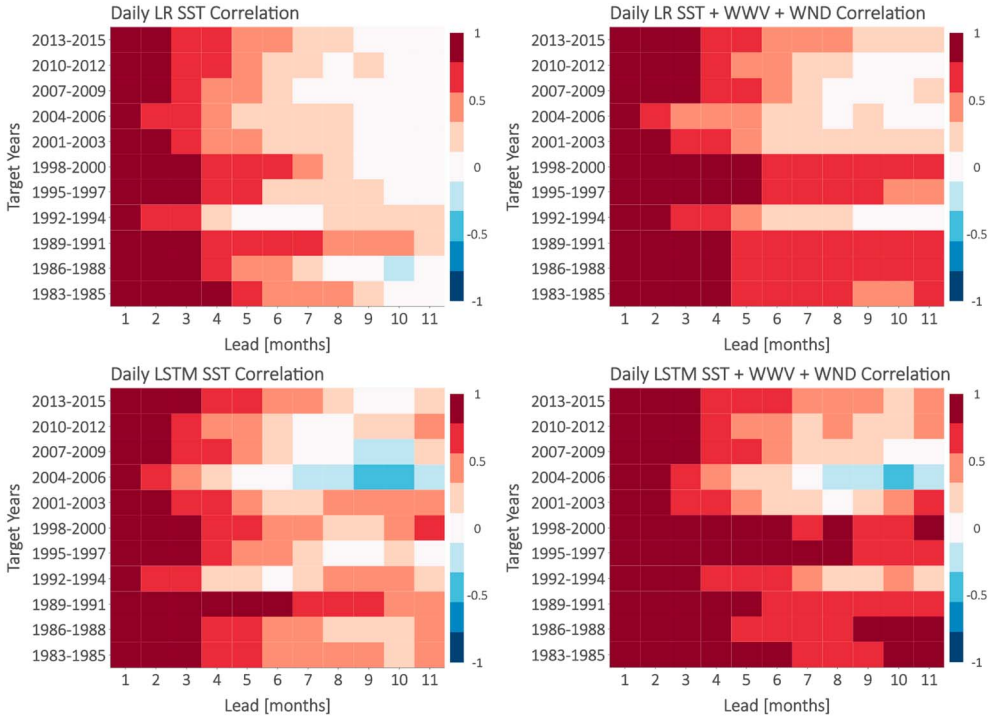

Figure 3. Correlation heatmaps of LR (top two plots) and LSTM (bottom two plots) using daily SST (left two plots) and a combination of SST, WWV, and WND (right two plots) as predictors. Red indicates positive correlations, and blue indicates negative correlations; the darker the color, the higher magnitude. There is high variability among the 3-year target sections, but also a noticeable trend toward lower correlations after the 2000s (lighter red shade near the top of each plot). LSTM at longer leads (near the right of each plot) generally shows higher correlations (more and darker red shade) when compared to LR, but there is a distinct period from 2004 to 2006 when there seems to be a forecast bust with negative correlations (blue). SST = sea surface temperature; WWV = warm water volume; WND = zonal winds; LR = linear regression models; LSTM = long-short-term-memory models.

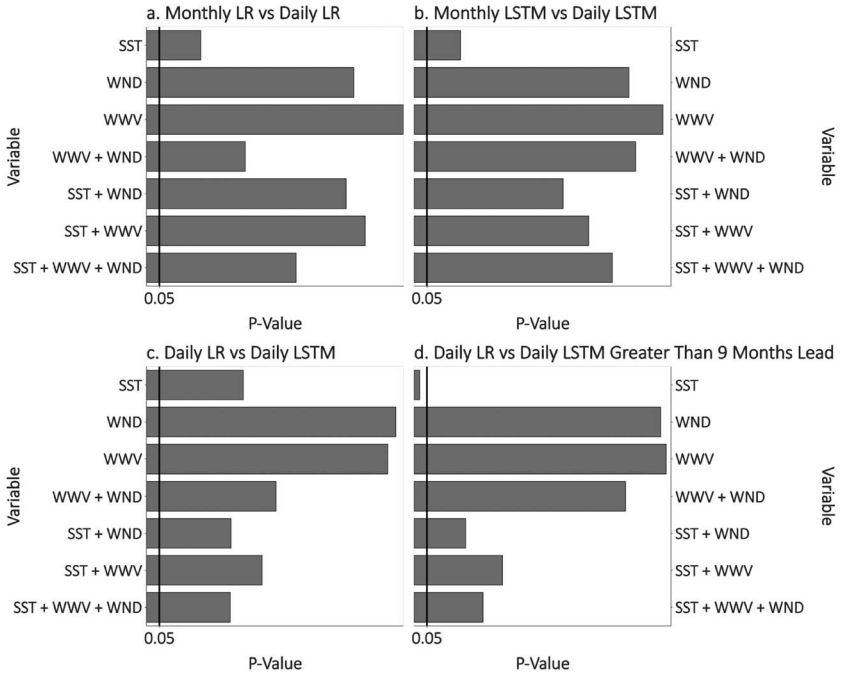

Figure 4. Bar plots of *p* value versus predictors. (top row) *P* values for the null hypothesis that there is no significant difference between monthly and daily correlations using LR (a**)** and LSTM (b). (bottom row) *P* values for the null hypothesis that there is no significant difference between daily LR and daily LSTM Fisher Z-values at all leads (c), and only at leads greater than 9 months (d) for SST. LR = linear regression models; LSTM = long-short-term-memory models; SST = sea surface temperature.

data and examining the autocorrelation, we find that there is some lower-lag correlation (2–3 days) in SST not in WND that might have contributed to higher skill for SST in LSTM predictions. In most cases, the correlations do not differ significantly across samples, having *p* values greater than 0.05, signifying that the null hypothesis cannot be rejected at a 95% confidence level (Figure 4). However, there is an exception to this: LSTM having daily SST as input, especially at leads greater than nine which holds a *p* value near 0, relative to other predictor combinations.

Recreating the lagged WWV versus SST Figure 7 in Meinen and McPhaden (2000), the asymmetry between WWV and SST positive and negative anomalies is found to be maximized at around 4 months back rather than 7 months back. The difference can be primarily attributed to the separate time periods used and is overall consistent with what McPhaden (2012) found in his later study: WWV led SST anomalies by two to three seasons during the 1980s and 1990s, but in the 2000s, this phenomenon was reduced to just one season, or approximately 4 months. This same idea was also applied to SST, and interestingly, at leads of 7 months and up, correlations of lead SST versus observed SST diverged for positive and negative anomalies. That is, positive anomalies switched to negative anomalies and negative anomalies preserved a positive correlation (Figure 5). This could possibly signify that El Niño tends to lead to a La Niña while La Niña tends to stay as a La Niña.

Inspired by this result, one final test was performed by splitting the time series into a training and validation period, and then further separating the monthly SST predictors into two groups: positive and negative anomalies. Within the training time series, two submodels were developed for both LSTM and LR, where the first submodel trained on positive anomalies while the second submodel trained on negative anomalies. The first submodel was used to predict during positive anomaly conditions in the validation time series and the second submodel was used to predict during negative anomaly conditions. Then, the predictions from both submodels were arranged back into a single chronological time series and the correlation was computed between the N34 observations and the merged predictions.

Most of these split submodels suffered lower correlations compared to the nonsplit models at early to middle leads, but exhibited higher correlations beginning at lead eight (Figure 6). This is true on all

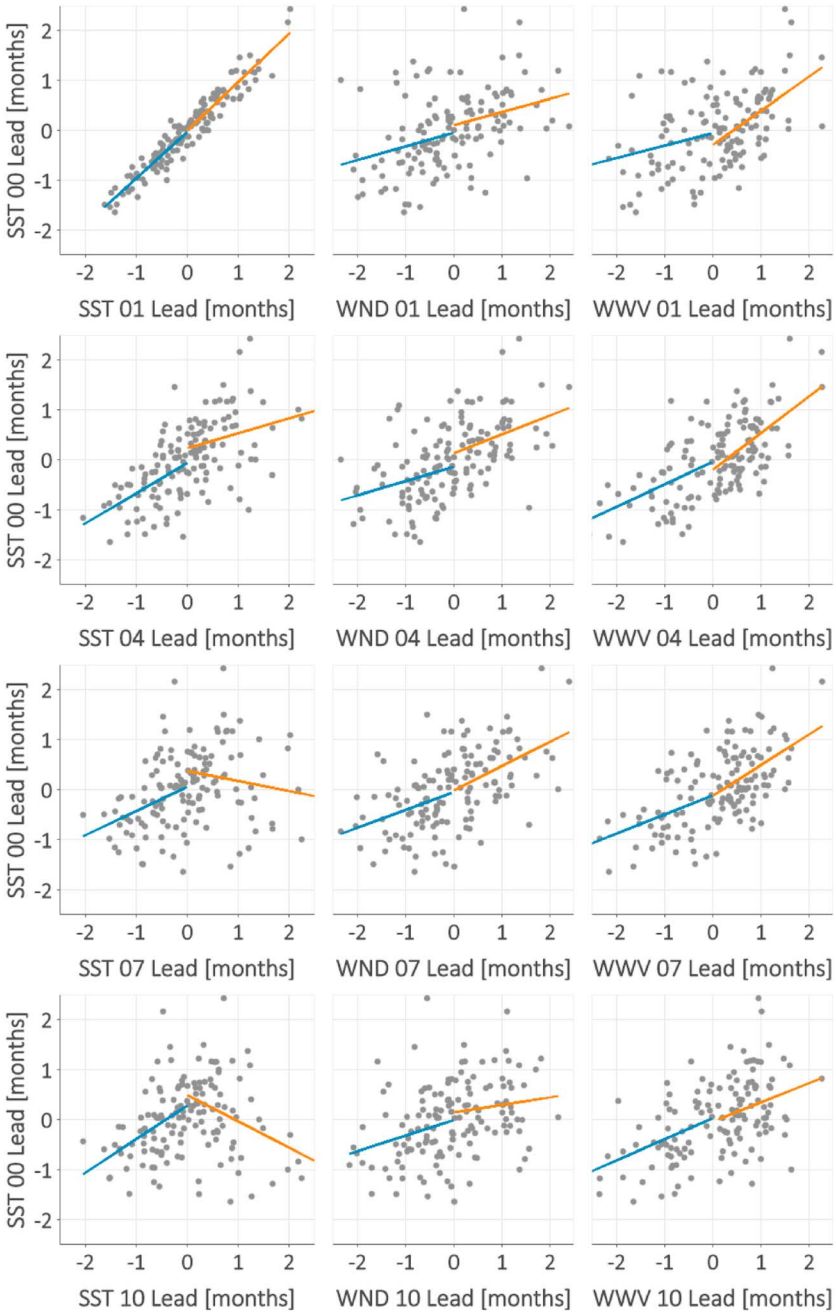

Figure 5. Scatter plots of seasonally averaged (December-January-February, March-April-May, etc.) SST versus WWV, WND, or SST anomalies. Each given variable has its time series shifted backward by 1, 4, 7, and 10 months. Lines represent least square fits, split by negative and positive anomalies of the given lead variable. At lead 10, negative SST anomalies tend to preserve a positive slope/correlation while positive anomalies tend to preserve a negative slope/correlation; in other words, out to 1 year later, negative anomalies tend to persist, whereas positive anomalies tend to have switched to negative anomalies. Both positive and negative anomalies of WWV and WND maintain a positive correlation at all leads. WWV = warm water volume; WND = zonal winds; SST = sea surface temperature.

time scales for both LSTM and LR except one: LSTM using daily data. The nonsplit LSTM had higher correlations than the split submodels across all leads. This suggests that LSTM possesses the ability to recognize these nonlinear relationships between predictors and predictands on its own through training and could partly explain why LSTM predictions involving SST had higher correlations than their LR counterparts at longer leads.

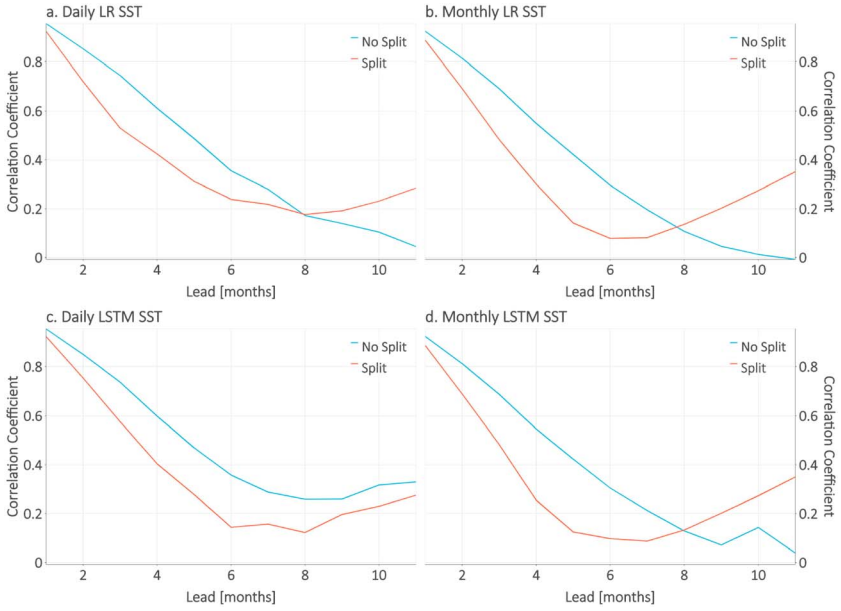

Figure 6. Line plots of lead versus correlation using just SST as a predictor for (a) daily LR SST, (b) monthly LR SST, (c) daily LSTM SST, and (d) monthly LSTM SST. Split models (red lines) have two separate submodels: One submodel is trained with positive predictors and the other with negative predictors. Forecasts are then assigned accordingly: When the predictor SST anomalies are positive, utilize the positive model, and when the predictor SST anomalies are negative, utilize the negative model. At longer leads, these split models demonstrate higher correlations than the no-split models, except LSTM using daily data (c). This suggests that LSTM can recognize the nonlinearity of the lagged SST autocorrelation and learn it automatically.

4. Conclusions

This study explored seven predictor combinations with LSTM against LR across 11-month leads using either monthly or daily resolution data, with 11, 3-year jackknife samples spanning from 1983 to 2015. This study demonstrates that LSTM has some advantage over LR with predictions involving SST and longer leads. However, beyond that, the differences between LSTM and LR are statistically insignificant based on the predictors used.

A key goal of this study is to explore how LSTM can build on LR techniques for ENSO prediction. While both methods exhibited similar results for the variables considered, there are several important differences. One notable finding relates to the nonlinear component in daily SST that improves LSTM predictability over LR. We also find that LSTM has a statistically significant advantage over LR at long leads, and this may be because of LSTM's memory—through training, it may have recognized the typical length of a given ENSO phase and was able to use this knowledge in its predictions. This capability could hypothetically add value to climate predictions and projections. The benefits of neural networks and LSTM can potentially be maximized by additional information, such as gridded data and more variables.

Due to the study's exploratory nature, we made certain strategic methodological choices to preserve computational tractability, such as limiting our analysis to a small number of key ENSO variables, restricting analysis to N34 region, and using limited training periods that only consider temporal evolution of N34. However, it is important to note that, although the N34 region is often used as a measure of skill in the models, ENSO is a dynamic process. Most of ENSO's teleconnections over North America are due to its effect on the warm pool in the western Pacific. While South America is also affected by the result of this, it is additionally affected by the warming in the far eastern Pacific as well (Trenberth et al., 1998).

Future studies can provide both the spatial and temporal dimensions for training so more nonlinear patterns are available for LSTM to take advantage of besides the nonlinear progression of SST. In addition to that, this work can be considered a proof of concept to isolate effects and interactions between different predictors of ENSO to gain some scientific insight using deep learning methods contrasted with linear models. Future

work could expand on this work using more variables to potentially improve predictability at the expense of straightforward physical interpretation.

Additionally, this work suggests that LSTM could be used to test the assumptions inherent in stochastic models that multiscale interaction between a rapidly varying chaotic system and a slower system can be resolved at some coarser time scale. Measurements of this slower process on time scales of months include the effects of the rapid nonlinearities that vary on time scales of days but cannot distinguish these nonlinearities from white noise forcing of the slower process (C. Penland, personal communication, 2018). The literature on how a slow system connected to a fast system may be treated as a stochastic differential equation is old and vast, starting with Wong and Zakai (1965) where they clarified the differences between ordinary and stochastic differential equations, then Khasminskii (1966), who showed that the solutions at suitable conditions weakly converge to zero in probability, and continuing with Papanicolaou and Kohler (1974), expanding on the previous work and demonstrating effectiveness in a variety of applicable problems. It is possible that LSTMs could build new research in this area by aiding in identifying the physical mechanisms seen by coarse-grained statistical models as dynamical stochastic noise.

Acknowledgments
NOAA High Resolution SST data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/. NCEP Reanalysis data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/. Warm water volume data were provided by TAO Project Office/NOAA/PMEL/Seattle, from their Web site at http://www.pmel.noaa.gov/tao/. In Python, visuals were made with HoloViews and matplotlib, LSTM models were built with Keras, LR models were built with scikit-learn, and data management and wrangling were done with xarray, pandas, and numpy. We thank Cécile Penland for her insightful feedback and comments that greatly improved an earlier version of this paper.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). "Social LSTM: Human trajectory prediction in crowded spaces." In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 961–71. https://doi.org/10.1109/CVPR.2016.110

Allen, J. T., Tippett, M. K., & Sobel, A. H. (2015). Influence of the El Niño/Southern Oscillation on tornado and hail frequency in the United States. *Nature Geoscience*, *8*(4), 278–283. https://doi.org/10.1038/ngeo2385

Barnston, A. G., & Ropelewski, C. F. (1992). Prediction of ENSO episodes using canonical correlation analysis. *Journal of Climate*, *5*(11), 1316–1345. https://doi.org/10.1175/1520-0442(1992)005<1316:POEEUC>2.0.CO;2

Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S., & DeWitt, D. G. (2011). Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society*, *93*(5), 631–651. https://doi.org/10.1175/BAMS-D-11-00111.1

Barnston, A. G., van den Dool, H. M., Zebiak, S. E., Barnett, T. P., Ji, M., Rodenhuis, D. R., et al. (1994). Long-Lead seasonal forecasts—Where do we stand? *Bulletin of the American Meteorological Society*, *75*(11), 2097–2114. https://doi.org/10.1175/1520-0477(1994)075<2097:LLSFDW>2.0.CO;2

Becker, E., Van den Dool, H., & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate*, *27*(15), 5891–5906. https://doi.org/10.1175/JCLI-D-13-00597.1

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., & Vialard, J. (2014). ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dynamics*, *42*(7–8), 1999–2018. https://doi.org/10.1007/s00382-013-1783-z

Chapman, D., Cane, M. A., Henderson, N., Lee, D. E., & Chen, C. (2015). A vector autoregressive ENSO prediction model. *Journal of Climate*, *28*(21), 8511–8520. https://doi.org/10.1175/JCLI-D-15-0306.1

Clarke, A., & Van Gorder, S. (2003). Improving El Niño prediction using a space-time integration of indo-Pacific winds and equatorial Pacific Upper Ocean heat content. *Geophysical Research Letters - GEOPHYS RES LETT*, *30*(April). https://doi.org/10.1029/2002GL016673

Efron, Bradley. (1982). "7. Cross validation, the jackknife and the bootstrap." In The jackknife, the bootstrap and other resampling plans, 49–59. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970319.ch7

Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, *12*(6), 1333–1340. https://doi.org/10.1109/72.963769

Glantz, M. H. (2001). *Currents of change: Impacts of El Niño and La Niña on climate and society*, (p. 252). Cambridge, UK; New York: Cambridge University Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, Z.-Z., Kumar, A., Ren, H.-L., Wang, H., L'Heureux, M., & Jin, F.-F. (2012). Weakened interannual variability in the tropical Pacific Ocean since 2000. *Journal of Climate*, *26*(8), 2601–2613. https://doi.org/10.1175/JCLI-D-12-00265.1

Hu, Z.-Z., Kumar, A., Zhu, J., Huang, B., Tseng, Y.-h., & Wang, X. (2017). On the shortening of the lead time of ocean warm water volume to ENSO SST since 2000. *Scientific Reports*, *7*, 4294. https://doi.org/10.1038/s41598-017-04566-z

Khasminskii, R. Z. (1966). A limit theorem for solutions of differential equations with random right-hand side. *Theory of Probability and its Applications*, *11*(3), 390–406. https://doi.org/10.1137/1111038

Kingma, D. P., & Ba, J. (2014). "Adam: A method for stochastic optimization." ArXiv:1412.6980 [Cs], December. http://arxiv.org/abs/1412.6980

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., et al. (2013). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing Intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Mason, S. J., & Goddard, L. (2001). Probabilistic precipitation anomalies associated with ENSO. *Bulletin of the American Meteorological Society*, *82*(4), 619–638. https://doi.org/10.1175/1520-0477(2001)082<0619:PPAAWE>2.3.CO;2

McPhaden, M. J. (2012). A 21st century shift in the relationship between ENSO SST and warm water volume anomalies. *Geophysical Research Letters*, *39*, L09706. https://doi.org/10.1029/2012GL051826

Meinen, C. S., & McPhaden, M. J. (2000). Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña. *Journal of Climate*, *13*(20), 3551–3559. https://doi.org/10.1175/1520-0442(2000)013<3551:OOWWVC>2.0.CO;2

Newman, M., & Sardeshmukh, P. D. (2017). Are we near the predictability limit of tropical sea surface temperatures? *Geophysical Research Letters*, *44*, 8520–8529. https://doi.org/10.1002/2017GL074088

Papanicolaou, G. C., & Kohler, W. (1974). Asymptotic theory of mixing stochastic differential equations. *Communications on Pure and Applied Mathematics*, *27*, 641–668.
Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, *8*(8), 1999–2024. https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2
Pielke, R. A., & Landsea, C. N. (1999). La Niña, El Niño, and Atlantic hurricane damages in the United States. *Bulletin of the American Meteorological Society*, *80*(10), 2027–2033. https://doi.org/10.1175/1520-0477(1999)080<2027:LNAENO>2.0.CO;2.\
Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2018). Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing*, *27*(2), 692–702. https://doi.org/10.1109/TIP.2017.2766358
Ropelewski, C. F., & Halpert, M. S. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, *115*(8), 1606–1626. https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2
Ropelewski, C. F., & Halpert, M. S. (1989). Precipitation patterns associated with the high index phase of the Southern Oscillation. *Journal of Climate*, *2*(3), 268–284. https://doi.org/10.1175/1520-0442(1989)002<0268:PPAWTH>2.0.CO;2
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*(January), 85–117. https://doi.org/10.1016/j.neunet.2014.09.003
Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J., Guemas, V., et al. (2013). Real-time multi-model decadal climate predictions. *Climate Dynamics*, *41*(11–12), 2875–2888. https://doi.org/10.1007/s00382-012-1600-0
Sriver, R. L., Timmermann, A., Mann, M. E., Keller, K., & Goosse, H. (2014). Improved representation of tropical Pacific Ocean–atmosphere dynamics in an intermediate complexity climate model. *Journal of Climate*, *27*(1), 168–185. https://doi.org/10.1175/JCLI-D-12-00849.1
Tangang, F. T., Hsieh, W. W., & Tang, B. (1997). Forecasting the equatorial Pacific sea surface temperatures by neural network models. *Climate Dynamics*, *13*(2), 135–147. https://doi.org/10.1007/s003820050156
Tippett, M. K., Barnston, A. G., & Li, S. (2011). Performance of recent multimodel ENSO forecasts. *Journal of Applied Meteorology and Climatology*, *51*(3), 637–654. https://doi.org/10.1175/JAMC-D-11-093.1
Trenberth, K. E. (1997). The definition of El Niño. *Bulletin of the American Meteorological Society*, *78*(12), 2771–2778. https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2
Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., & Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research*, *103*(C7), 14,291–14,324. https://doi.org/10.1029/97JC01444
Van den Dool, H. (1994). Searching for analogues, how long must we wait? In *Tellus A* (pp. 314–324). Wiley Online Library." n.d. Accessed March 2, 2018. http://onlinelibrary.wiley.com/doi/10.1034/j.1600-0870.1994.t01-2-00006.x/abstract
Vega-Westhoff, B., & Sriver, R. L. (2017). Analysis of ENSO's response to unforced variability and anthropogenic forcing using CESM. *Scientific Reports*, *7*(1), 18,047. https://doi.org/10.1038/s41598-017-18459-8
Wang, W., Chen, M., Kumar, A., & Xue, Y. (2011). How important is Intraseasonal surface wind variability to real-time ENSO prediction? *Geophysical Research Letters*, *38*, L13705. https://doi.org/10.1029/2011GL047684
Wong, E., & Zakai, M. (1965). On the convergence of ordinary integrals to stochastic integrals. *Annals of Mathematical Statistics*, *36*(5), 1560–1564. https://doi.org/10.1214/aoms/1177699916
Wyrtki, K. (1985). Water displacements in the Pacific and the genesis of El Niño cycle. *Journal of Geophysical Research, Oceans*, *90*(C4), 7129–7132. https://doi.org/10.1029/JC090iC04p07129
Zhang, W., Villarini, G., Slater, L., Vecchi, G. A., & Allen Bradley, A. (2017). Improved ENSO forecasting using Bayesian updating and the north American multimodel ensemble (NMME). *Journal of Climate*, *30*(22), 9007–9025. https://doi.org/10.1175/JCLI-D-17-0073.1