# SEASONAL PREDICTION USING UNSUPERVISED FEATURE LEARNING AND REGRESSION

Mahesh Mohan*[1], Cheng Tang*[1], Claire Monteleoni[1], Timothy DelSole[2,3], Benjamin Cash[2,3]

***Abstract*—We propose to use machine learning to discover indices from the SST field data, and to compare their prediction performance to that of the Niño3.4 index, on tasks related to ENSO. As a first step in this direction, this work focuses on predicting the time-series of monthly temperature anomalies in Texas, from time series for the whole ocean SST field, ending 6 months prior.**

## I. INTRODUCTION

Understanding the El Niño–Southern Oscillation (ENSO) is crucial for predictions of regional and seasonal conditions, as it tends to have strong (positive or negative) correlation with a variety of weather conditions and extreme events throughout the globe. However there is currently significant room for improvement in predicting even this extremely well-studied oscillation with such high global impact. For example, most statistical and climatological models yielded questionable predictions or else erred significantly in predicting an El Niño event for 2014; their predictions were off by several months. The 2015 event is still ongoing, with some predictions that it could be one of the biggest of the century. Better tools to predict such phenomena are critical for seasonal and regional climate prediction, and would thus address grand challenges in the study of climate change [**?**].

The index currently most widely-used to predict ENSO, known as Niño3.4, is the mean sea surface temperature (SST) anomaly in a fixed region of the ocean. Many of the currently used climate indices, including Niño3.4, were originally defined by human experts, as opposed to having been learned from the data in an automated fashion. Data mining techniques have shown recent success in discovering climate patterns [**?**]. We propose to use machine learning to discover indices from the SST field data, and to compare their seasonal prediction performance to that of the Niño3.4 index, on tasks related to ENSO. In this work we focus on predicting the time-series of monthly temperature anomalies in Texas, from time series for the whole ocean SST field, ending several months prior. We focus on predicting seasonal-mean temperature in Texas because the drought/heat wave there in 2011 raised critical questions about the role of ocean temperatures and the extent to which such events can be predicted in the future [**?**].

To understand which kind of patterns ("features" in the language of machine learning) in the global ocean SST field are important for capturing its relation to Texas temperature, we explore the effect of various unsupervised feature learning techniques—linear and nonlinear—applied to the ocean SST field followed by a unified regression step, on the prediction performance of Texas mean temperature.

## II. FRAMEWORK

Specifically, our approach has two stages:

1) Unsupervised feature learning: Given temperature anomalies over the ocean [**?**] we use unsupervised learning to generate a set of task-independent features. The methods we consider are k-Means clustering, Singular Value Decomposition (SVD/EOF), and Nonlinear Laplacian Spectral Analysis (NLSA) [**?**]. [1] The former two are commonly used for feature extraction that essentially factorize the data matrix into low-rank parts and thus both linear. The latter is a nonlinear method, which exploits the prior knowledge that spatial climate patterns at two consecutive time stamps should be "close" to each other. Based on this insight, it constructs a nonlinear embedding of the original data and performs SVD in the embedded spatial-temporal space. It has been empirically demonstrated to find low-frequency (rare) climate patterns, which are typically not captured by traditional methods such as SVD [**?**], [**?**].
2) Regression: With the learned features, we apply regularized least squares regression to learn a predictive model for the Texas temperature anomalies.

*Equal first authorship: Mohan (mahesh_mohan@gwu.edu) and Tang (tangch@gwu.edu); [1]George Washington University, Washington DC. [2]George Mason University, Fairfax, VA. [3]Center for Ocean-Land Atmosphere Studies, George Mason University, Fairfax, VA.

[1]These all outperformed hierarchical agglomerative clustering.

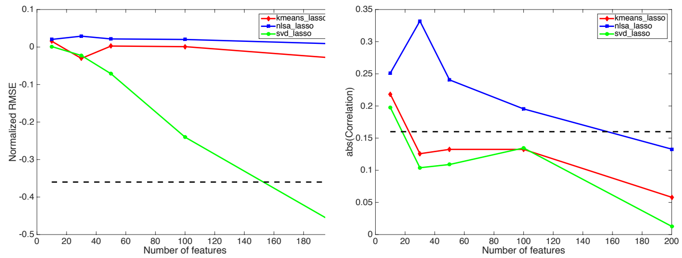


Fig. 1: Normalized RMSE (left) and absolute value of the correlation (right) for a 6 month lookahead task (with black lines for the Nino3.4 index). NLSA+LASSO was observed to have the best prediction error among all the methods tested.

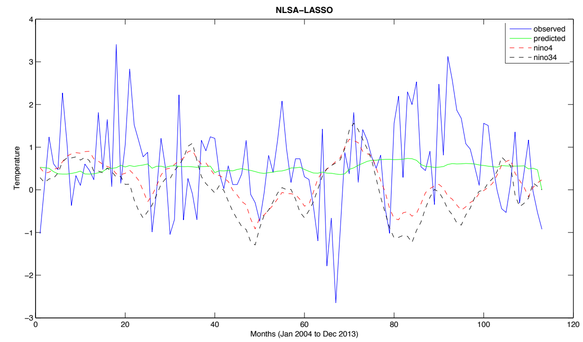


Fig. 2: The time series predicted by NLSA+LASSO, is close to the mean of the observed temperature anomalies. The Niño4 and Niño3.4 indices had significantly higher prediction errors.

The methods we consider are Ridge ($L_2$-norm regularization) and LASSO ($L_1$-norm regularization).

## III. Experimental Results

We used NOAA's Merged Land-Ocean Surface Temperature Analysis (MLOST) V3.5.4 dataset [**?**] for the monthly temperature anomalies. All methods were trained with data for the period 1964-2003. Prediction performance was then determined by computing prediction errors over 2004-2013. Performance was evaluated on two metrics, the Normalized RSME, and the correlation coefficient (between a method's lagged predictions and the Texas time-series). The Normalized RMSE can be computed from a prediction model's RMSE, and the standard deviation of the observation sequence, $\sigma_y$, as, $\mathtt{NRMSE} = 1 - \frac{\mathtt{RMSE}}{\sigma_y}$. We observed that Ridge regression performed worse than LASSO, when paired with each feature learning technique. We also tried linear regression without regularization, which fared worse. Both of these findings suggest the need for sparse solutions for this task.

The NRMSE error, and the correlation obtained by predicting with the Niño3.4 index for the test period (with a 6 month lag, corresponding to the test framework for the learning methods) were -0.36 and -0.16, respectively. Fig. 1 shows the results obtained using machine learning as the number of features varied. This shows that NLSA performs better than Niño3.4 for most feature sizes, which may be explained by its ability to capture nonlinear patterns that are not captured by linear methods due to their low frequency (rare) nature. k-means and SVD outperform Niño3.4 when the number of features is small (around 10).

Although dimension reduction plus LASSO perform better than Niño3.4 on these metrics, the predicted time series is close to the mean of the observed temperature anomalies (see Fig. 2). Thus the predictions do not match the large spikes in the observation time-series, corresponding to extreme weather. While the Niño3.4 and 4 indices capture some spikes, they also err much more significantly on others.

## IV. Future Directions

To further test our conjecture that prediction models learned via machine learning can outperform existing ENSO indices, we plan to study other supervised regression tasks associated with ENSO, both individually, and simultaneously as a multi-task problem. We also plan to explore different notions of error, in order to quantify prediction of extremes. Additionally, identifying regions in the SST field that have significant predictive power with respect to the Texas temperatures is a subject of ongoing research.

## References

[1] WCRP, "World Climate Research Programme: Grand Challenges." http://www.wcrp-climate.org/grandcha.shtml, 2013.

[2] S. Liess, A. Kumar, P. K. Snyder, J. Kawale, K. Steinhaeuser, F. H. M. Semazzi, A. R. Ganguly, N. F. Samatova, and V. Kumar, "Different modes of variability over the tasman sea: Implications for regional climate," *Journal of Climate*, vol. 27, pp. 8466–8486, 2015/05/19 2014.

[3] M. Hoerling, A. Kumar, R. Dole, J. W. Nielsen-Gammon, J. Eischeid, J. Perlwitz, X.-W. Quan, T. Zhang, P. Pegion, and M. Chen, "Anatomy of an extreme event," *Journal of Climate*, vol. 26, no. 9, pp. 2811–2832, 2013.

[4] T. M. Smith, R. W. Reynolds, T. C. Peterson, and J. Lawrimore, "Improvements to noaa's historical merged land-ocean surface temperature analysis (1880-2006)," *Journal of Climate*, vol. 21, no. 10, pp. 2283–2296, 2008.

[5] D. Giannakis and A. J. Majda, "Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability," *Proceedings of the National Academy of Sciences*, vol. 109, no. 7, pp. 2222–2227, 2012.

[6] E. Szekely, D. Giannakis, and A. J. Majda, "Extraction and predictability of intraseasonal signals in infrared brightness temperature data," *4th International Workshop on Climate Informatics*, Sept. 2014.