**On the Predictability of Rainfall Anomalies over the Southern Amazonia:**

**A Comparison between NMME and Statistical Models**

1. Summary:

Prediction of rainfall over the Amazonian rainforest during wet season is fundamental to assess the regional water and energy balance and global carbon-climate feedbacks. Previous observational analysis has identified some large-scale atmospheric dynamic and thermodynamics conditions that can influence the rainfall anomalies during the wet season. Based on these observed persistent conditions that started between June and August(JJA, dry season), we have developed and evaluated several statistical models to predict rainfall conditions during September to November (SON, early wet season) for the Southern Amazonia (5-15oS, 50-70oW). Multivariate Empirical Orthogonal Function (EOF) Analysis is applied to the following four fields during JJA from the ECMWF Reanalysis (ERA-Interim) spanning from year 1979 to 2015: geopotential height at 200 hPa, surface relative humidity, convective inhibition energy (CIN) index and convective available potential energy (CAPE), to filter out noise and highlight the most coherent spatial and temporal variations. The first 10 EOF modes are retained for inputs to the statistical models, accounting for at least 70% of the total variance in the predictor fields. Then the 12-fold cross-validation method is used to estimate the tuning parameters used in the regression algorithms.

Ridge Regression and Lasso Regression are able to capture the spatial pattern and magnitude of rainfall anomalies. Compared with the seasonal prediction based on dynamical models, this statistical prediction system has better predictions than the seasonal predictions of the dynamic climate model. The statistical models show longer and more accurate predictive persistence of the rainfall anomalies. In addition, we use Logistic regression and Neural Networks to predict the categorical states of rainfall over the Southern Amazon by classifying the rainfall states into two categories, i.e., dry and wet. Our statistical models show overall better predictions of categorical rainfall states than the magnitudes of rainfall in our study region. The accuracy of the statistical prediction based on Neural Networks method can reach greater than 90%, which is much higher than the simple logistic regression method, indicating the non-linearity of the atmospheric processes. Both predictions of the magnitudes and states of rainfall anomalies can be combined to provide more accurate information. The models we have developed have broad implications on the future development of seasonal climate predictions and can be used for real-time forecasts in the future.

1. Dataset:

In this study, I use the monthly fields of 200 hPa geopotential height, temperature at 700 hPa, surface dewpoint temperature, surface relative humidity and CAPE for June through August from the ERA Interim reanalyses spanning from 1979-2015. The spatial resolution is 2.5o latitude by 2.5o longitude. The dataset for rainfall is from the Global Precipitation Climatology Centre (GPCC) with the spatial resolution of 1o latitude by 1o longitude.

All the input data fields cover the domain 0-30oS, 40-80oW. We use these fields to predict rainfall conditions over the Southern Amazonia.

National Center for Environmental Prediction (NCEP) CFS version 2 (CFSv2) coupled atmosphere–ocean seasonal climate prediction systems is used as a representative of dynamic models for seasonal prediction. The CFS is widely used as an operational coupled seasonal forecast system. NCEP CFSv2 is an upgraded version of CFSv1. It produces a set of 9-month reforecast initiated from every 5th day with four ensemble members for the period 1982–2010. Initial conditions for the atmosphere and ocean come from NCEP Climate Forecast System Reanalysis (CFSR, Saha et al. 2010). 6-hourly Climate Forecast System version 2 (CFSv2) realtime data for the period 2011-2015 are also provided. CFSv2 data can be accessed through the Data Library of the International Research Institute for Climate and Society (http://iridl.ldeo.columbia.edu). The Climate Forecast System Reforecasts for hindcasts are initialized at June, July and August to forecast rainfall at September, October and November. For each year, 4 predictions were produced every 5 days starting January 1st with ocean and atmosphere initial conditions (ICs) from the NCEP Climate Forecast System Reanalysis.

1. Methodology:

* 1. Multivariate EOF:

Prior to analysis, all the four fields are averaged during the dry season and then converted to standardized anomalies during the period of 1979-2015. Original predictor inputs could have multicollinearity, noise and variance irrelevant to rainfall prediction. Therefore, we apply Multivariate Empirical Orthogonal Function analysis (EOF) to the four predictor fields to filter out noise and highlight the most coherent spatial and temporal variances. EOF analysis uses principal component analysis (PCA) to compress datasets, such that a dataset containing a large number of samples is reduced to a dataset that captures the dominant modes. These modes explain a large fraction of the squared total co-variance among those fields. The new variables are linear combinations of the original variables and represent the highest possible proportion of co-variability found in the original datasets. We retain the first ten EOF modes, accounting for about 77 percent of the variance in the predictor fields, to minimize the potential multicollinearity and efficiently use the original predictor fields at the same time.

* 1. Linear Regression:

In order to predict the magnitude of the rainfall anomalies, we use linear regression model. For each location, the rainfall anomaly (y) is modeled as a linear combination of ten predictor variables derived from the multivariate EOF analysis (X) with associated regression parameter matrix (a) and residuals (b).

* 1. Logistic Regression:

Based on rainfall anomalies, we create a dichotomous outcome, which takes values 0 (negative) or 1 (positive), which indicate below (dry) or above (wet) normal. The transformed mean of the binary response can be modeled as a linear combination of the ten predictor variables derived from the multivariate EOF analysis (X) with parameter matrix r.

* 1. Neural Networks:

The Neural Networks model is made up of neurons that have learnable weights and biases. Each neuron receives equivalent inputs, performs a dot product and transition with a non-linearity. In this study, Neural Networks receive an input (signal vector with 10 units), and transform it through a series of hidden layers with 5 units. In the hidden layer, each neuron is fully connected to all neurons in the previous layer, and where neurons are from the first 10 EOF modes. The last fully-connected layer, called the “output layer”, represents the class scores of the rainfall classifications (i.e., 0 or 1 in logistic regression) in this study. The Neural Networks model predict rainfall states similarly to the logistic regression. Since the climate is full of non-linear processes, the linearity in the logistic regression cannot fully represent the processes in the atmosphere. We use Neural Networks to better represent the nonlinear links between the input variables and rainfall states.

1. Conclusion and Discussion:

In this report, we use both dynamic models and statistical models to assess the predictability of rainfall anomalies during wet season (Sep-Nov) over the Southern Amazonia. The dynamic model offers insights into the potential sources of predictability from the slowly varying conditions from the ocean and land memories. However, there are large uncertainties in the application of seasonal predictions due to the limitation of the dynamic climate models in representing convection and land-surface and expansion of the uncertainties as predicted time forward grows. Therefore, we also use statistical models to compute seasonal predictions and compare the results with those from dynamic models.

The statistical models are based on these observed persistent conditions during June to August (JJA, dry season). We input the precondition fields into several statistical models to predict rainfall conditions during September to November (SON, early wet season) for the Southern Amazonia (5-15oS, 50-70oW). To decrease the noise and multicollinearity, we apply Multivariate Empirical Orthogonal Function (EOF) Analysis to the following four fields during JJA from the ECMWF Reanalysis (ERA-Interim) spanning 1979 to 2015: geopotential height at 200 hPa, surface relative humidity, convective inhibition energy (CIN) index and convective available potential energy (CAPE), so that we can hightlight the most coherent spatial and temporal variances. The first 10 EOF modes are retained for inputs to the statistical models, accounting for at least 70% of the total variance in the predictor fields. To choose the penalty parameters in the regression algorithms, we use 12-fold cross-validation.

This study shows that Ridge Regression and Lasso Regression are able to capture the spatial pattern and magnitude of rainfall anomalies. The statistical models generate better predictions than the seasonal predictions of the dynamic climate model. In addition, the statistical models generate more accurate predictions for the rainfall anomalies. To predict the two rainfall states, i.e., dry and wet, over the Southern Amazon, we use Logistic regression and Neural Networks. The predictions are better than predictions of the magnitudes of rainfall in our study region. The accuracy of the statistical prediction based on Neural Networks method can reach greater than 90%, which is much higher than the simple logistic regression method, indicating the non-linearity of the atmospheric processes. The statistical models have broad implications on seasonal climate prodictions and can be used for real-time forecasts in the future.

Despite the advantages of the statistical models, there are two major limitations to be highlighted:

1. The accuracy of the prediction results relies on the sample size of the observations and accurate input datasets. Since the satellite era (1970s), observations are very limited. In this study, we only have 37-year observations for each location. In the future, more datasets are needed to train the statistical models.

2) The statistical models require predictor fields. That is to say, we need to first understand the factors that influence the rainfall processes over the Southern Amazonia. The factors we find should be sufficient to provide enough information and should not be redundant. Understanding the physical mechanisms behind the processes is an area that still needs more effort from scientists in climate science. In general, the seasonal prediction of rainfall is a subject that is both important and challenging. Future work involves combining the advantages of dynamical models and statistical models to provide more accurate predictions in the globe.