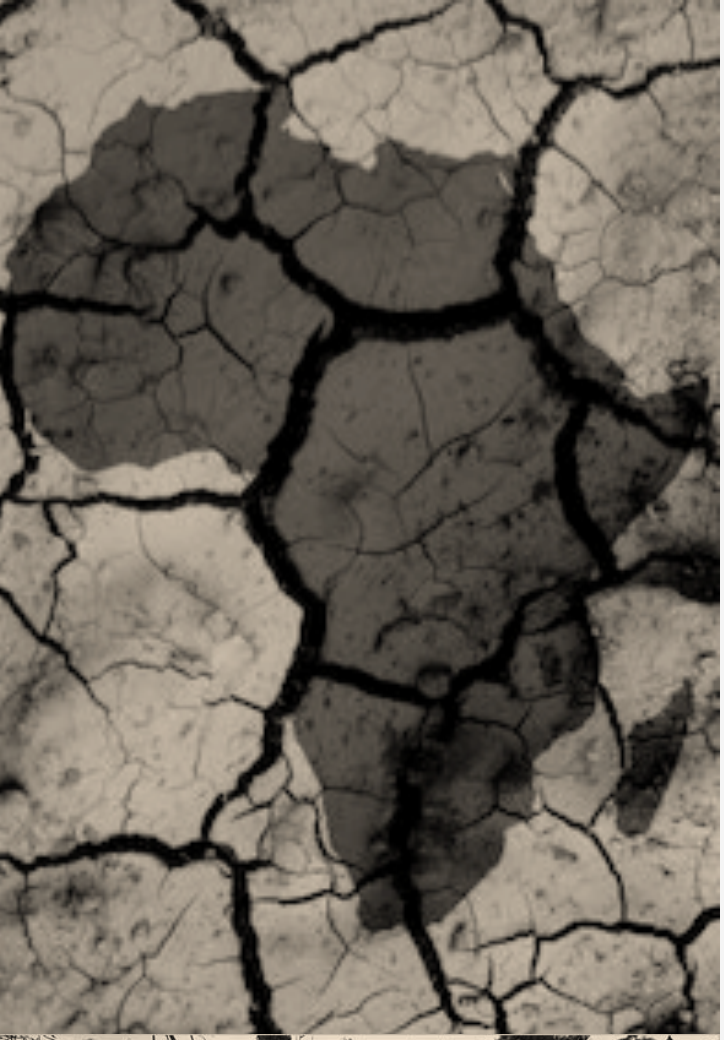


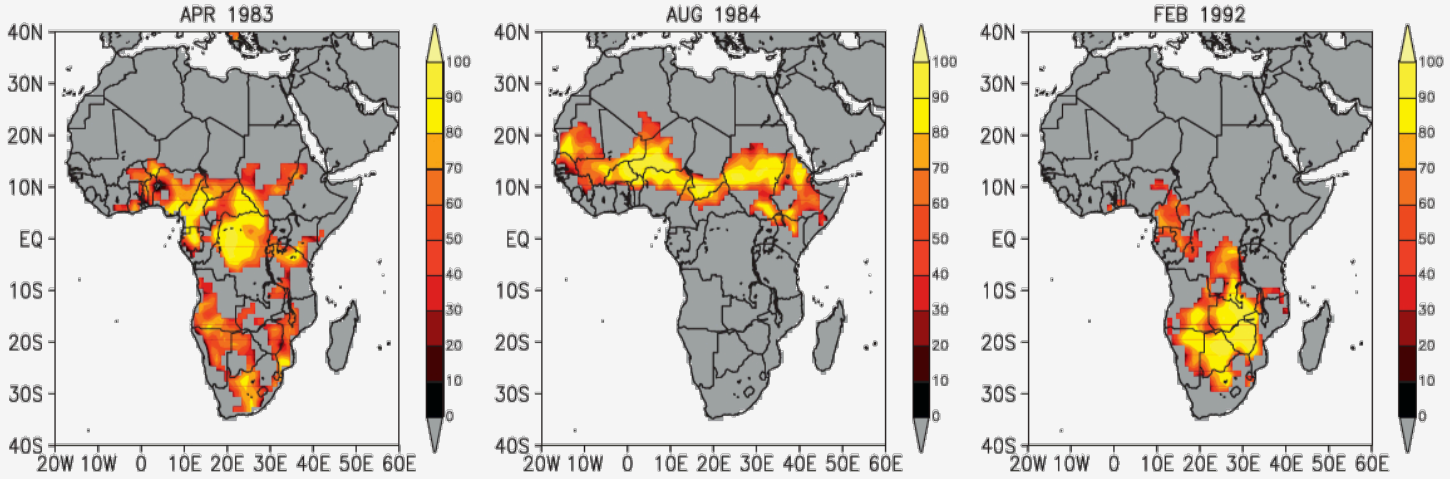
STATISTICAL PREDICTIVE MODELS FOR SEASONAL RAINFALL ANOMALIES OVER SAHEL

H. S. Badr, B. F. Zaitchik, S. D. Guikema

Johns Hopkins University

Motivation

- Africa is characterized by considerable variability of precipitation.
- Africa's average annual rainfall has decreased since 1968,
- There is also some evidence that natural disasters have increased in frequency and severity, particularly drought in the Sahel.



Sheffield et al. 2009

Motivation

- Advanced data analytic techniques can be applied to generate improved predictions while still allowing for conceptually meaningful results.
  - using nonlinear but robust methods informed by expert variable selection and variable importance analysis.

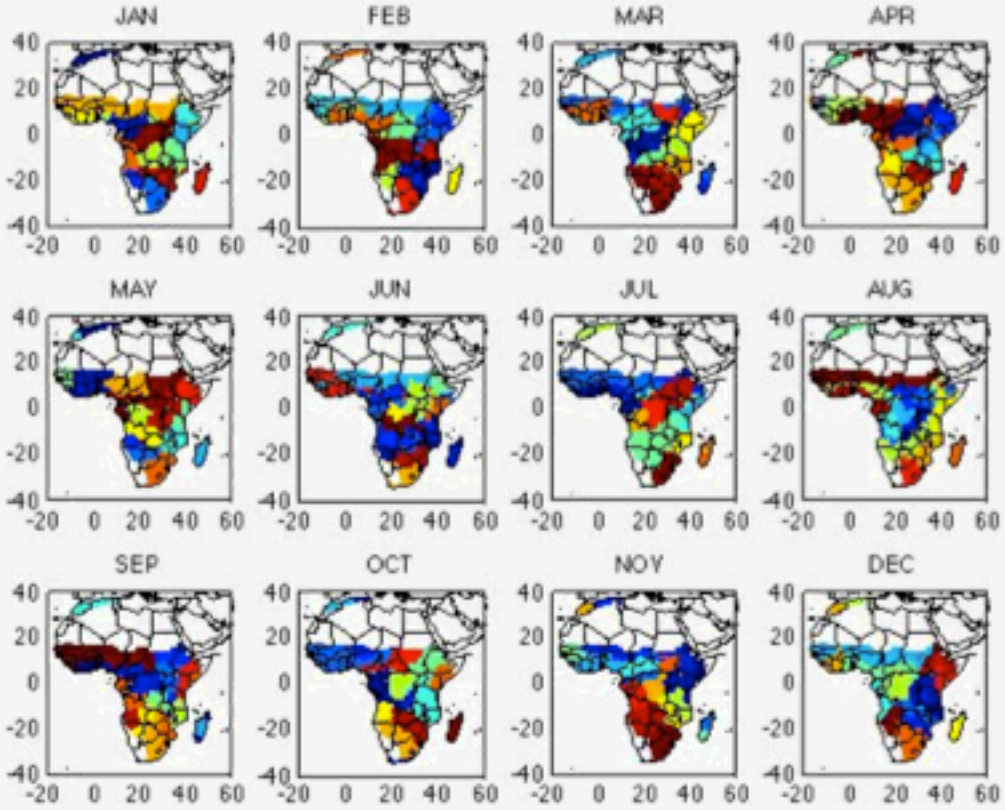
- Assessment of the developed statistical models based on both goodness-of-fit and predictive skill.
  - It is critical to avoid overfitting the date and make the distinction between goodness-of-fit and predictive skill.

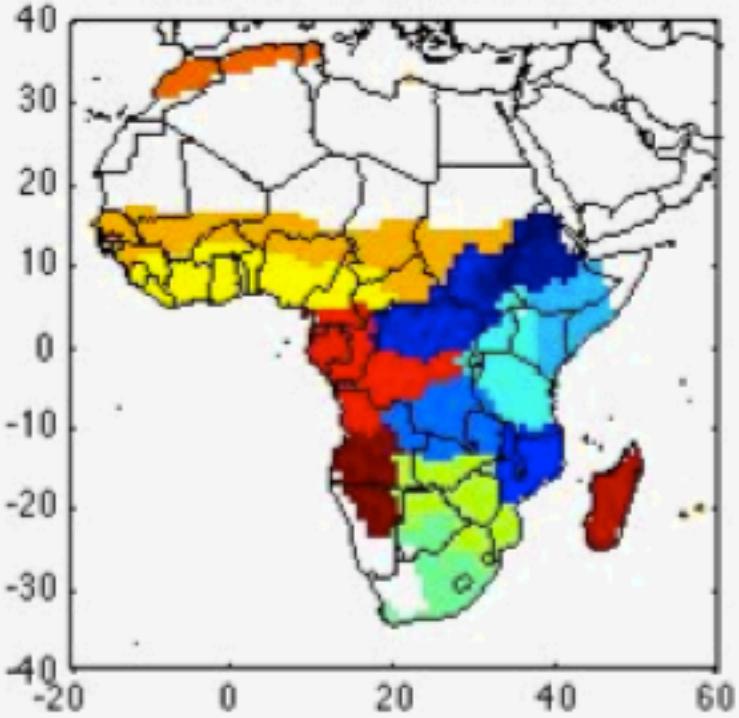
Objectives

- Regionalization of Africa into homogeneous precipitation climates using Hierarchical Clustering Analysis (HCA).
  - Definition of seasons, as sets of months, for each region…
- Prediction of precipitation variability for each region
  - Developing statistical models using advanced data analytic techniques.
  - Variable selection, model comparison and assessment with respect to both goodness-of-fit and predictive skill.
  - Variable importance analysis and interpretation of results according to fundamental climate processes.
- Association of precipitation variability with global patterns
  - Understanding mechanisms and representation in climate models…

On Thursday Now (for Sahel) In progress

Regionalization of African Precipitation





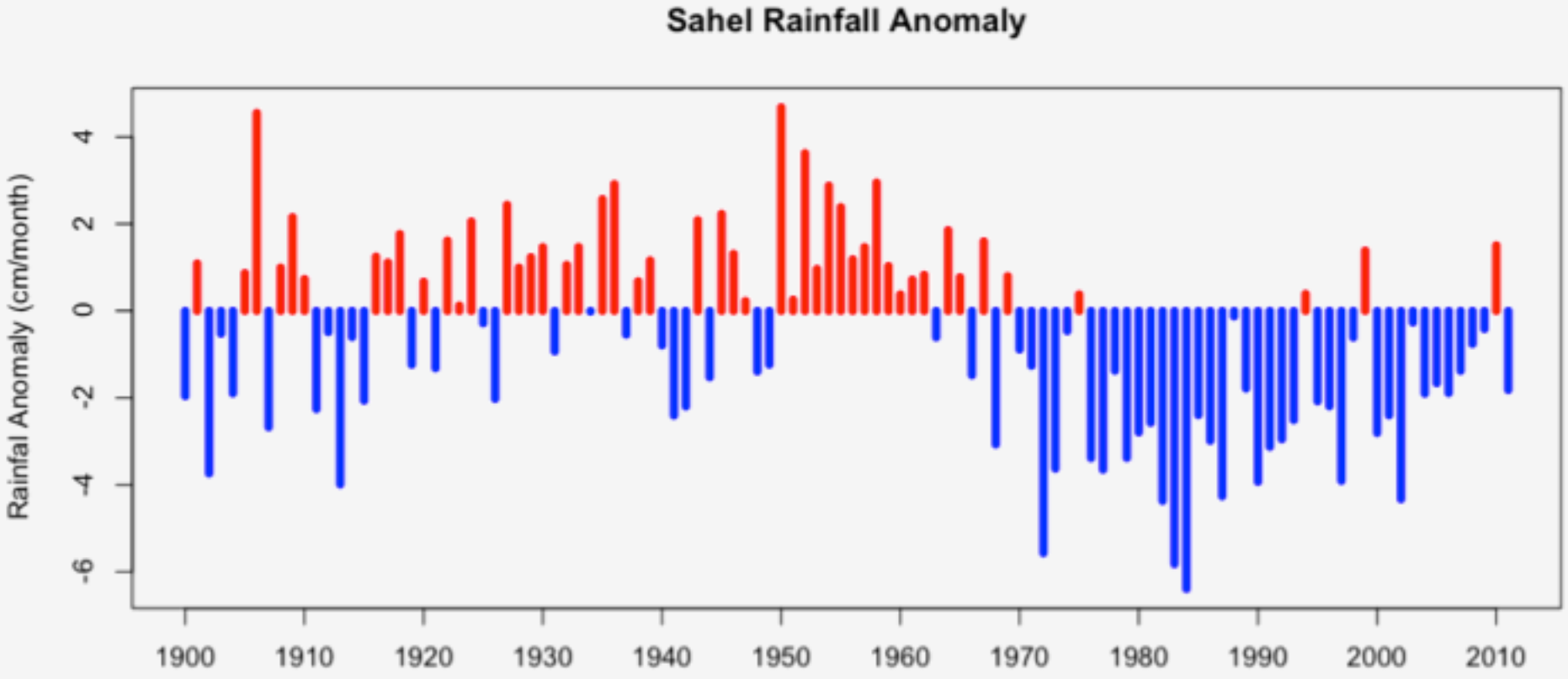
Thursday, 10 January 2013: 3:30 PM

Spatiotemporal Variability of Precipitation over Africa
Ballroom C (Austin Convention Center)

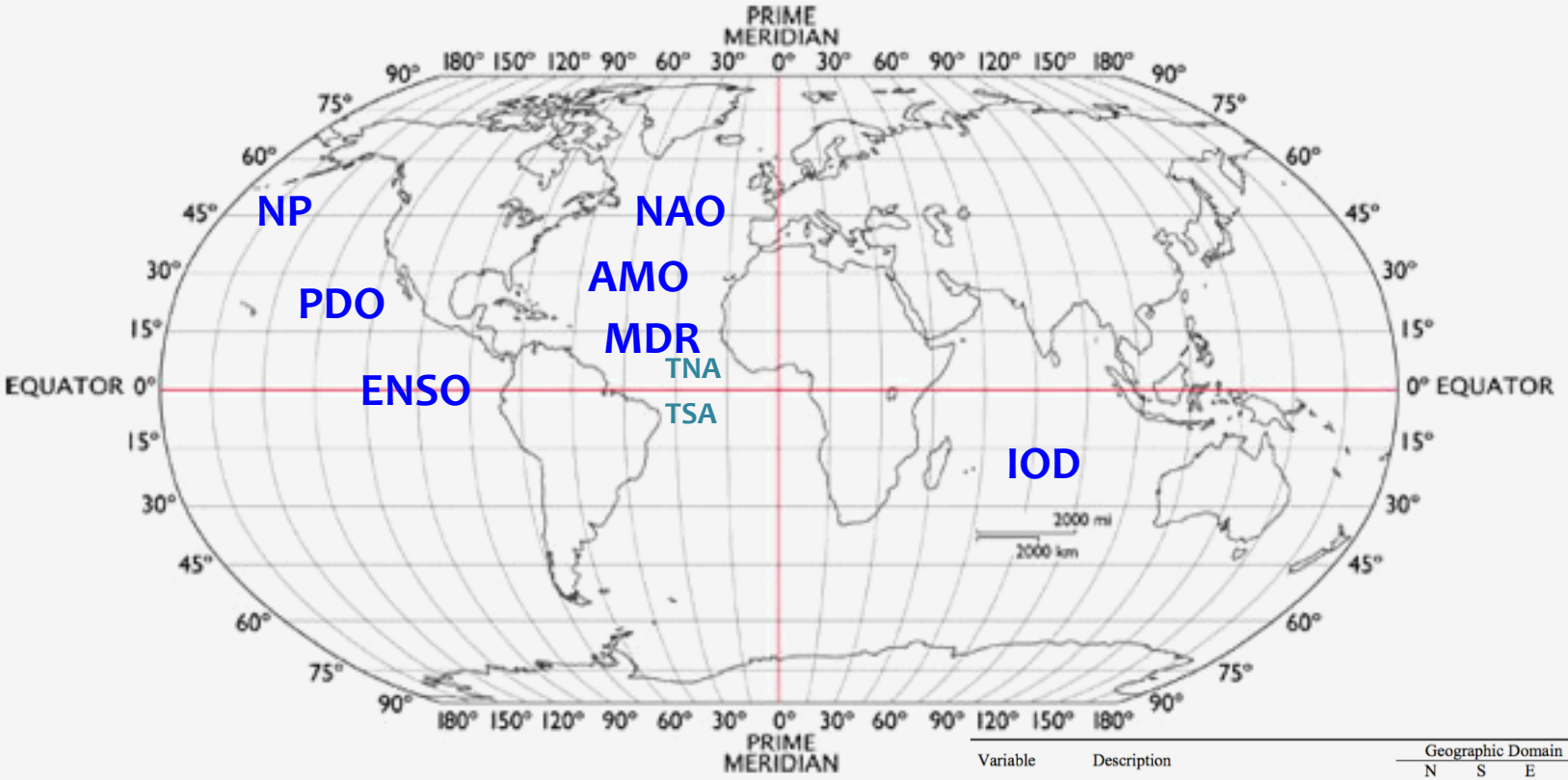
***Hamada S. Badr**, Johns Hopkins University, Baltimore, MD; and B. F. Zaitchik and A. K. Dezfuli*

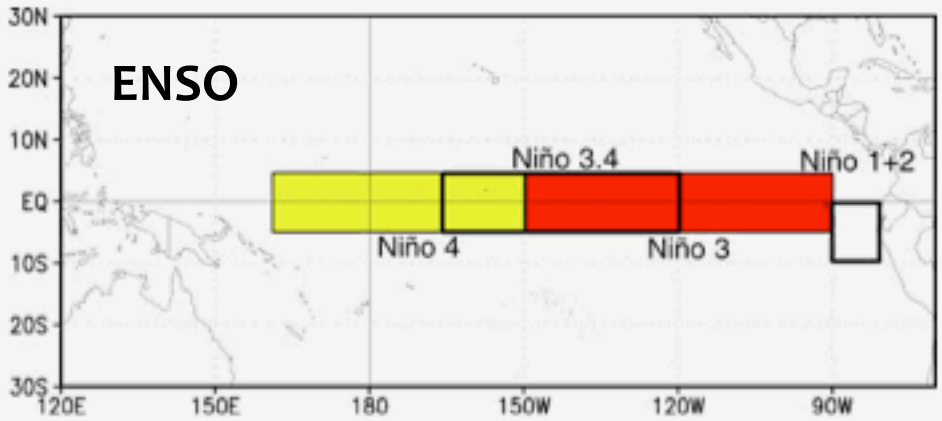
Data and Methods

- Sahel rainfall index is mean summertime (JAS) rainfall, while the oceanic indices are computed as mean springtime (AMJ).
  - The region's averages is taken for the region (10N-20N, 10W-20E) of gridded rain gauge data anomalies in the NOAA GHCN.



Data and Methods

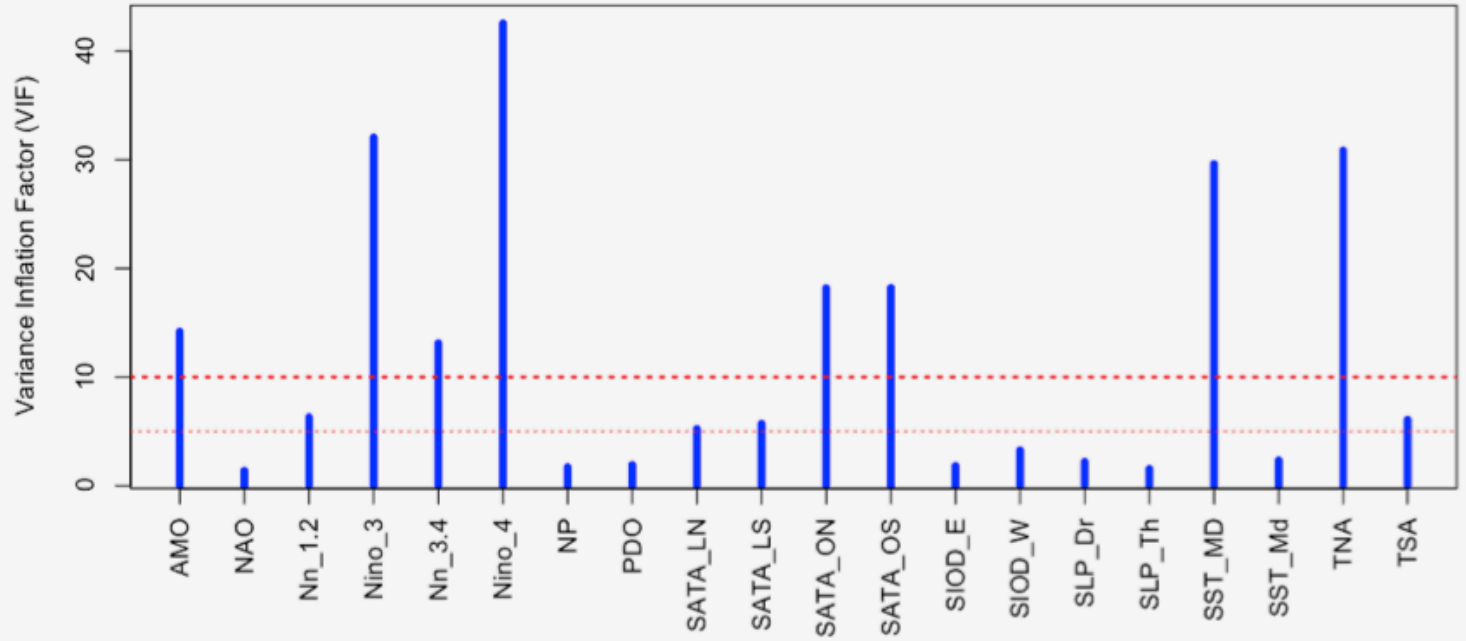




Variable	Description	Geographic Domain			
		N	S	E	W
Sahel_Precip	Sahel Rainfall Anomaly	20	10	20	10
AMO	Atlantic Multidecadal Oscillation	70	0	-10	-75
NAO	North Atlantic Oscillation Index (PC)	80	20	40	-90
Nino_1.2	Niño Region 1+2 SST	0	-10	-80	-90
Nino_3	Niño Region 3 SST	5	-5	-90	-150
Nino_3.4	Niño Region 3.4 SST	5	-5	-120	-170
Nino_4	Niño Region 4 SST	5	-5	160	-150
NP	Northern Pacific Pattern	65	30	160	-140
PDO	Pacific Decadal Oscillation (PC)	70	-60	-60	100
SATA_LNH	Global Mean SATA over Land (NH)	90	0	-- Land --	
SATA_LSH	Global Mean SATA over Land (SH)	0	-90	-- Land --	
SATA_ONH	Global Mean SATA over Ocean (NH)	90	0	-- Ocean --	
SATA_OSH	Global Mean SATA over Ocean (SH)	0	-90	-- Ocean --	
SIOD_E	Eastern Subtropical Indian Ocean SST	-18	-28	100	90
SIOD_W	Western Subtropical Indian Ocean SST	-27	-37	65	55
SLP_Darwin	Sea Level Pressure at Darwin	-13	-13	131	131
SLP_Tahiti	Sea Level Pressure at Tahiti	-18	-18	-150	-150
SST_MDR	Hurricane Main Development Region SST	20	10	-20	-85
SST_Med	Mediterranean Sea SST	45	30	25	0
TNA	Tropical Northern Atlantic SST	25	5	-15	-55
TSA	Tropical Northern Atlantic SST	0	-20	10	-30

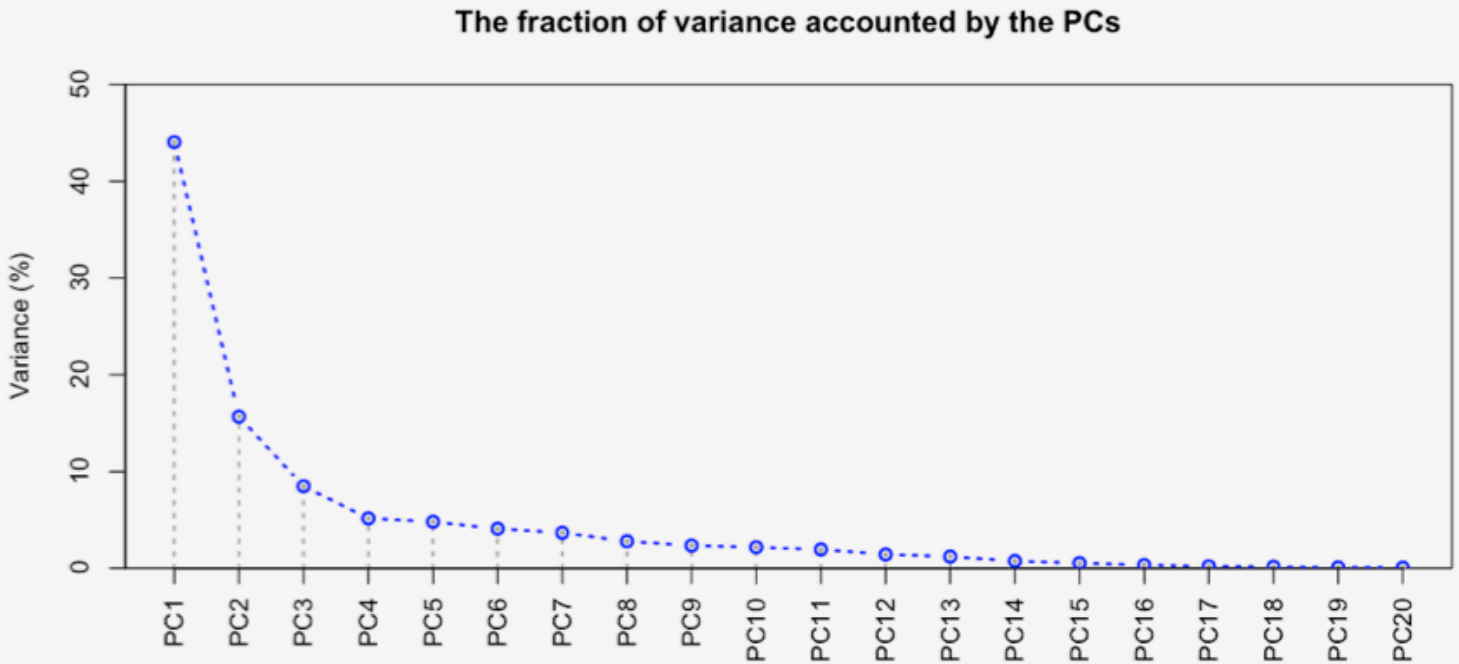
Multicollinearity Problem

- Multicollinearity increases the standard errors of the coefficients.
  - The coefficients for some predictor variables may be found not statistically significant, while when solving the multicollinearity problem these same coefficients might have been found to be statistically significant.
- Variance Inflation Factor (VIF) was computed.
  - A VIF of 5 or 10 and above indicates a multicollinearity problem.



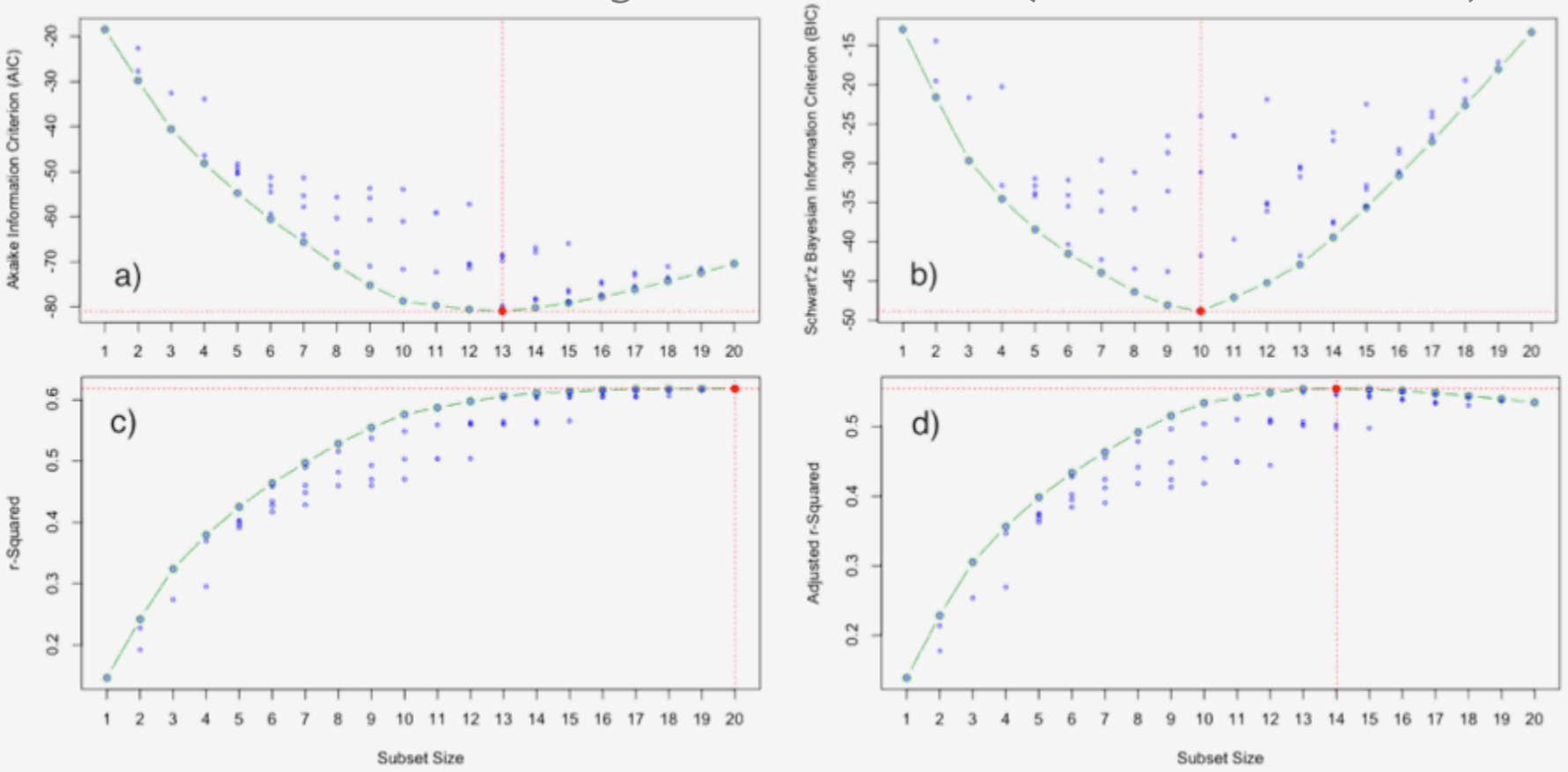
Multicollinearity Problem

- Multicollinearity problem is solved by PCA for the complete set of standardized predictor variables.
  - it is not convenient to remove a variable or more outside the cross-validation analysis.



Variable Selection

- The number of the covariates is relatively large and better predictions may be obtained by removing some covariates.
  - Models with many covariates have low bias/high variance, while models with few covariates have high bias/low variance (bias-variance tradeoff).



Statistical Models

- Multiple statistical models were developed to predict the seasonal rainfall anomalies using the large-scale SST and SAT predictors.
  - All models are compared with each other.
  - The regularization of the ANN was based on weight decay method.

ID	Name	Model Description
1	GLM	Full-Covariate Generalized Linear Model
2	SGLM	Selected Generalized Linear Model based on Stepwise Selection
3	GAM	Full-Covariate Generalized Additive Model
4	SGAM	Selected Generalized Additive Model based on Penalized Terms
5	MARS	Multivariate Adaptive Regression Spline
6	CART	Classification and Regression Trees Model
7	BCART	Bagged Classification and Regression Trees Model
8	BART	Bayesian Additive Regression Trees Model
9	RF	Random Forest Model
10	ANN	Artificial Neural Network
11	Average	The prediction is the mean of the predictions of all models (1-10)
12	Null	The prediction is the mean of the response variable in the training data

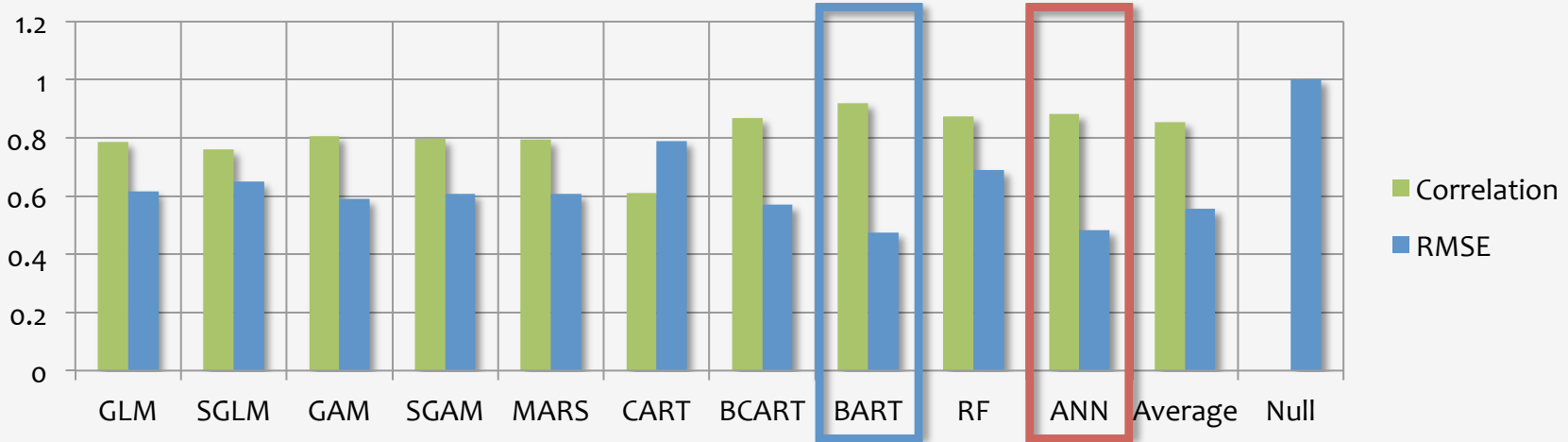
Model Assessment

- The goodness of fit was assessed through the temporal correlation between the predicted and observed rainfall anomalies.
- The predictive skill was assessed through the out-of-sample error measures obtained from different cross-validation methods.

	Repeated *k*-fold Cross-Validation	Leave-One-Out Cross-Validation
Method	A random holdout of 10% from the data used for validation, and the remaining observations as the training data.	A single observation from the sample used for validation, and the remaining observations as the training data.
Pros	Large number of performance estimates	Unbiased performance estimation
Cons	Overlapped training and test data between each round; Underestimated performance variance / overestimated degree of freedom for comparison	Very large variance

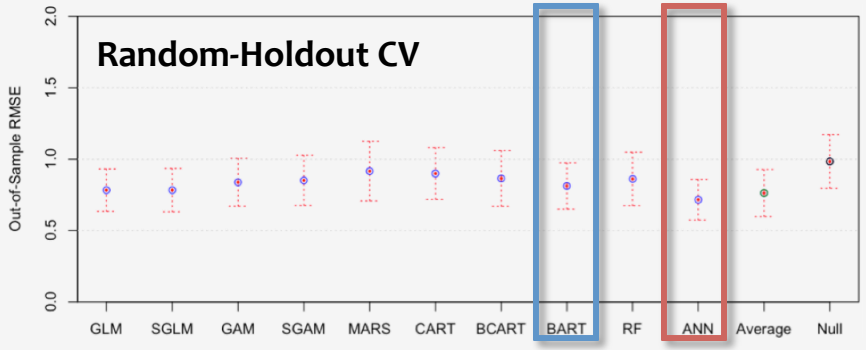
Goodness of Fit

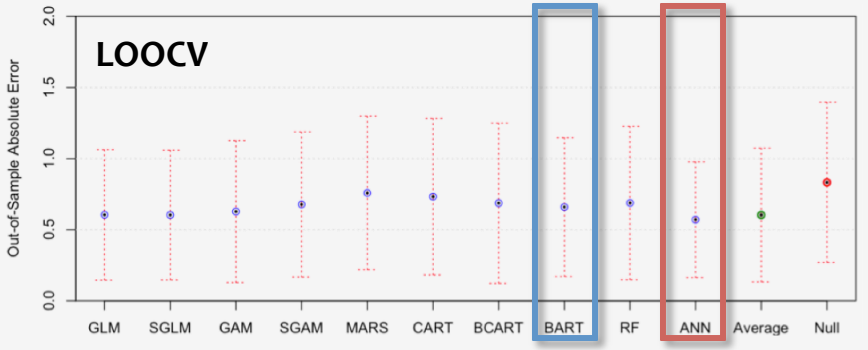
ID	Model	Correlation	MAE	RMSE	MSE	MAD
1	GLM	0.786	0.493	0.616	0.379	0.657
2	SGLM	0.759	0.528	0.649	0.421	0.650
3	GAM	0.806	0.462	0.589	0.347	0.633
4	SGAM	0.796	0.488	0.607	0.368	0.624
5	MARS	0.793	0.473	0.607	0.369	0.534
6	CART	0.611	0.612	0.788	0.621	0.755
7	BCART	0.867	0.431	0.569	0.324	0.489
8	BART	0.919	0.373	0.475	0.225	0.435
9	RF	0.873	0.548	0.688	0.473	0.650
10	ANN	0.881	0.393	0.482	0.232	0.495
11	Average	0.855	0.441	0.555	0.309	0.507
12	Null	---	0.826	1.000	1.000	1.095



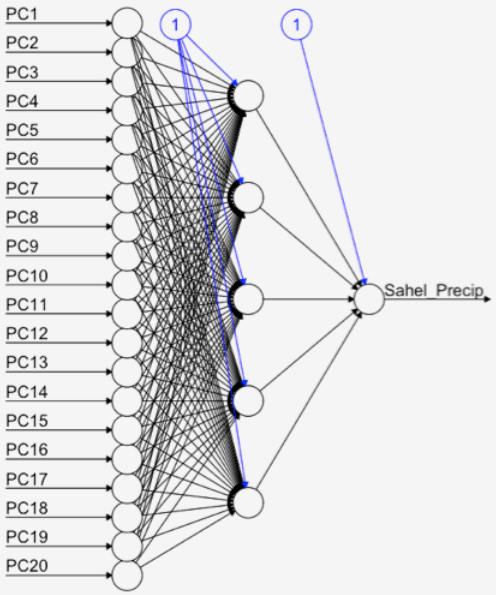
Predictive Skill

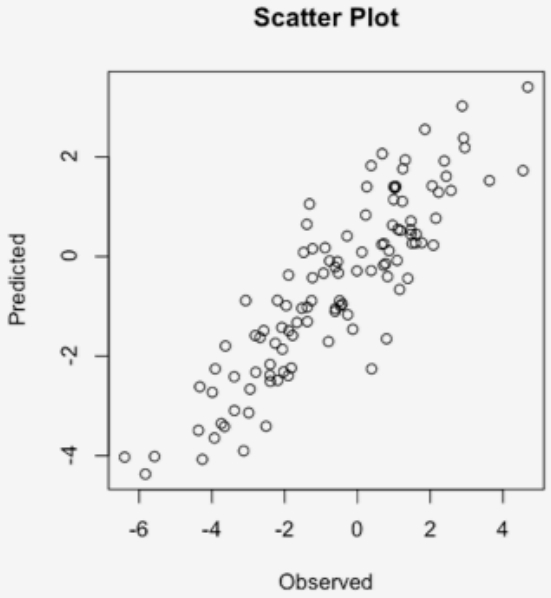
Model		Repeated *k*-Fold Random-Holdout Cross-Validation								LOOCV	
		MAE		RMSE		MSE		MAD		Abs. Error	
ID	Name	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	GLM	0.637	0.133	0.783	0.149	0.635	0.231	0.742	0.260	0.606	0.458
2	SGLM	0.643	0.135	0.783	0.152	0.636	0.235	0.776	0.289	0.605	0.456
3	GAM	0.675	0.150	0.838	0.168	0.730	0.283	0.782	0.268	0.628	0.498
4	SGAM	0.695	0.156	0.851	0.176	0.756	0.304	0.813	0.289	0.679	0.510
5	MARS	0.764	0.170	0.916	0.208	0.882	0.423	0.885	0.304	0.760	0.539
6	CART	0.739	0.155	0.900	0.181	0.842	0.349	0.850	0.302	0.734	0.549
7	BCART	0.693	0.169	0.865	0.195	0.786	0.347	0.779	0.289	0.688	0.563
8	BART	0.668	0.146	0.812	0.162	0.685	0.262	0.782	0.282	0.660	0.487
9	RF	0.703	0.166	0.861	0.188	0.777	0.336	0.811	0.283	0.689	0.539
10	ANN	0.588	0.125	0.716	0.142	0.532	0.200	0.685	0.240	0.572	0.407
11	Average	0.605	0.142	0.763	0.165	0.609	0.250	0.722	0.256	0.605	0.470
12	Null	0.823	0.167	0.984	0.188	1.003	0.386	0.972	0.287	0.834	0.563

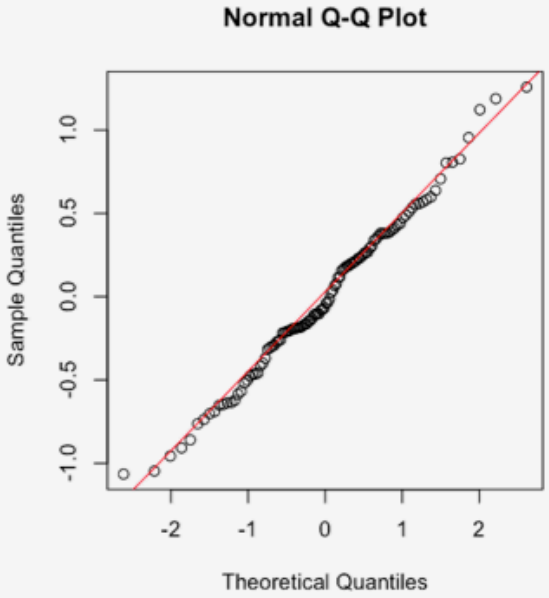


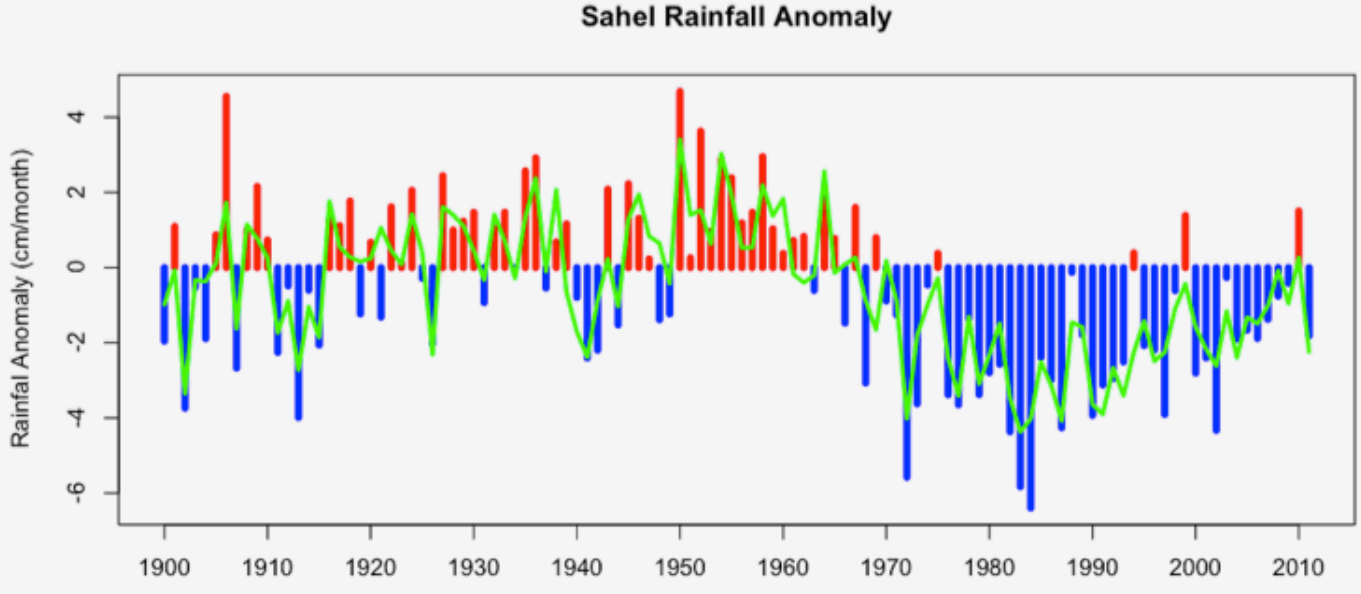


Predictions vs. Observations

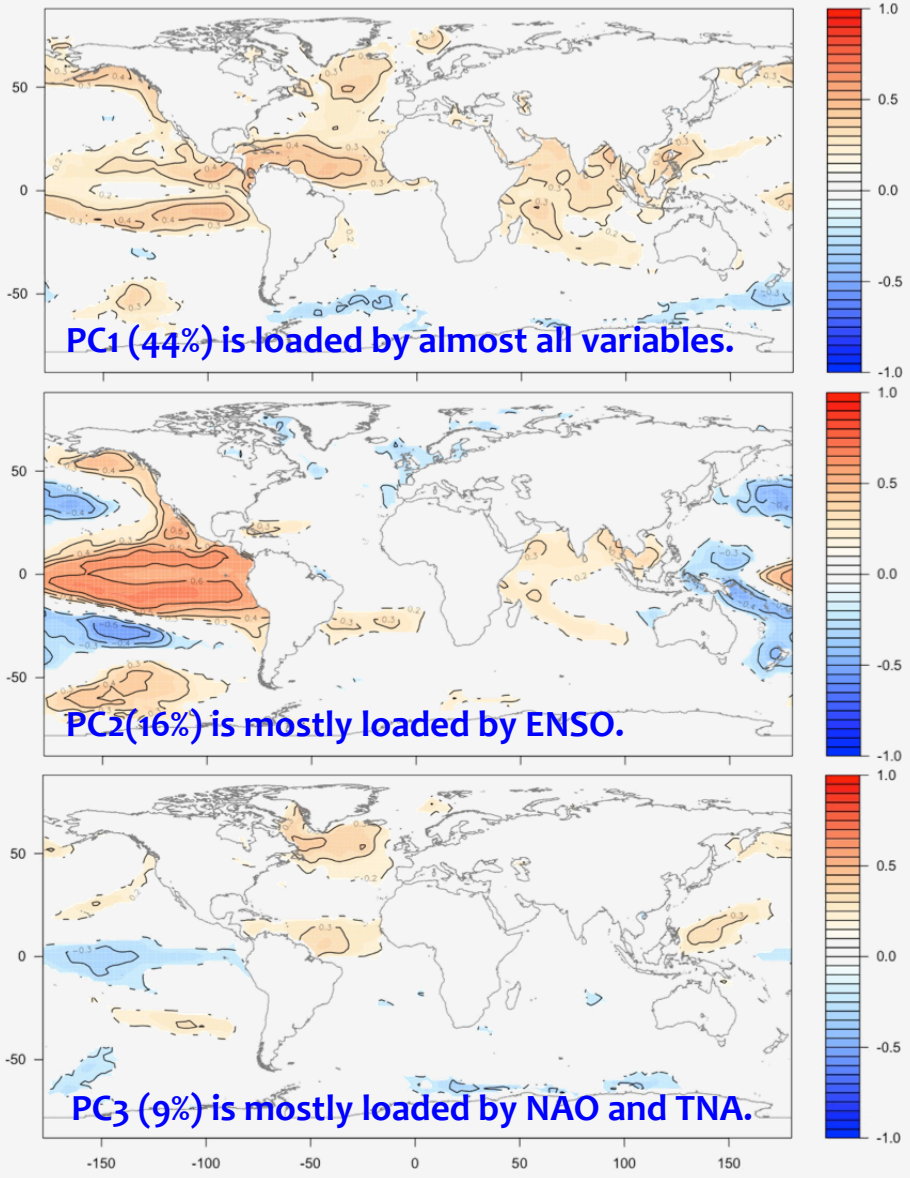


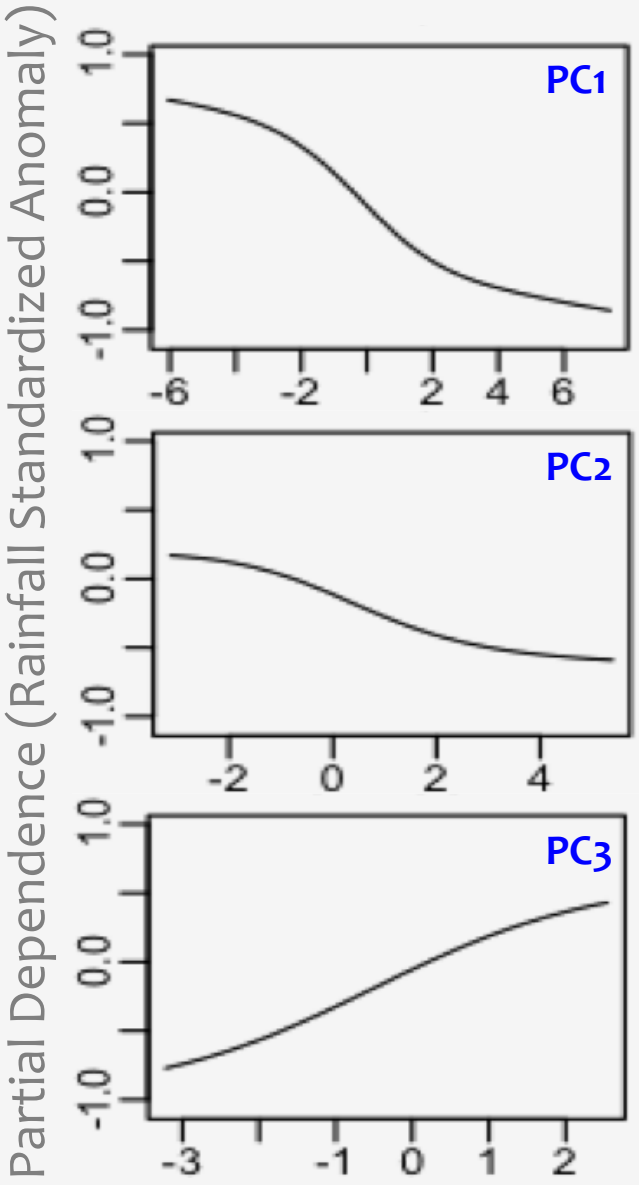




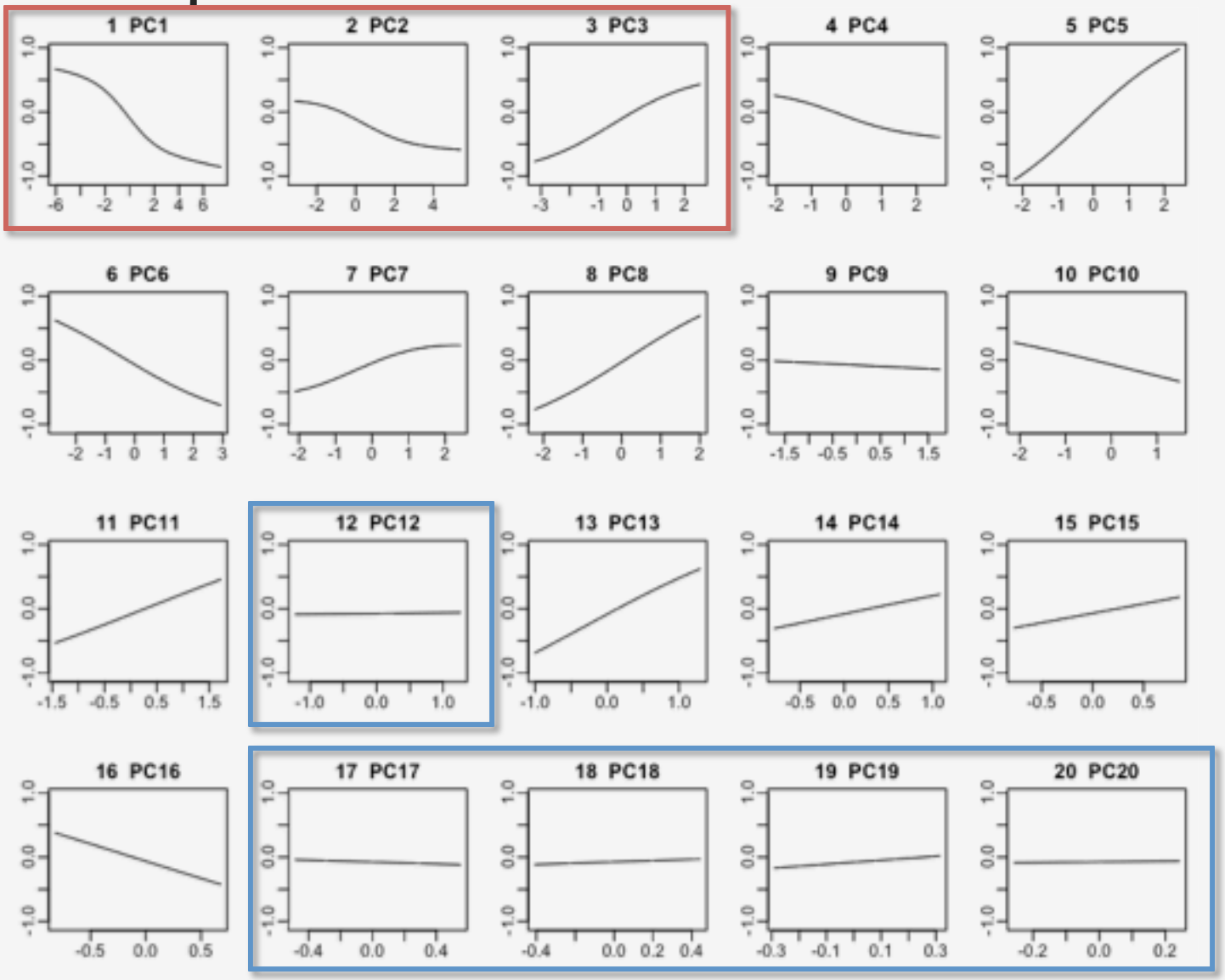


Regression Patterns and Partial Dependence





Partial Dependence Plots



Conclusions

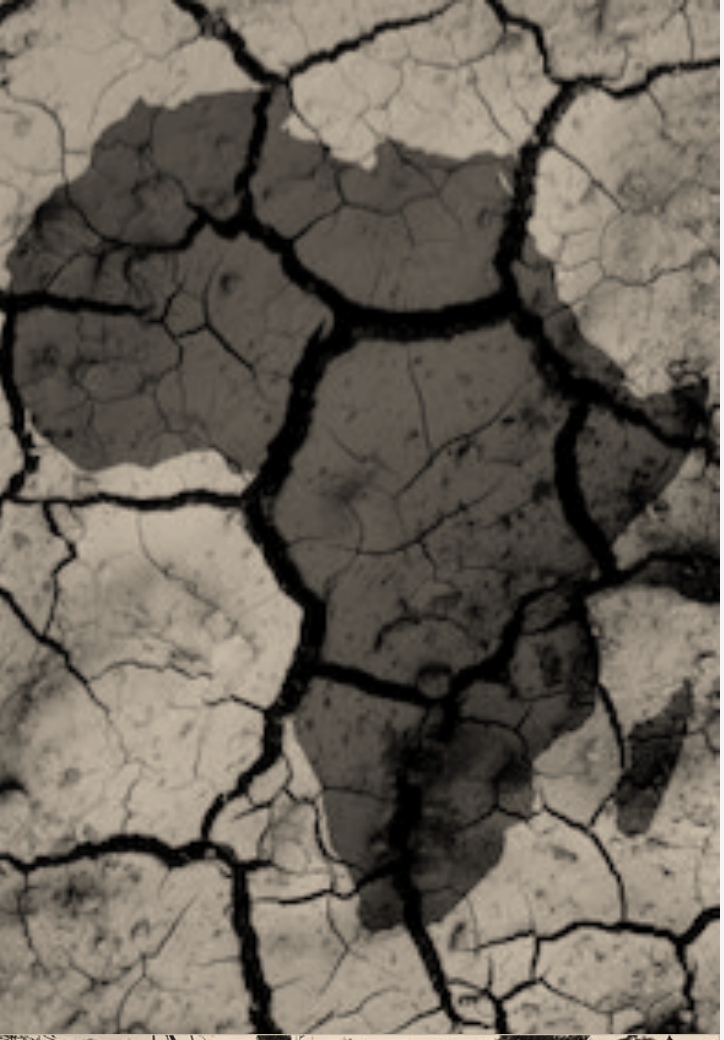
- The artificial neural network (ANN) model outperforms all other statistical models in terms of predictive accuracy and, relatively, goodness-of-fit.

- ANNs to uncover nonlinear interactions; Some variables have nonlinear effects on the rainfall anomalies response while other variables have linear or even constant/zero effects.

- The leading principal components are the most important variables and they also have had the nonlinear effects and large influences on the Sahel rainfall anomalies.

Future Work

- Extension to other regions in Africa…
- Understanding the drivers and mechanisms of precipitation variability and its representation in climate models…
- Statistical vs. dynamical downscaling…



QUESTIONS?