A Machine Learning Approach for Probabilistic Multi-Model Ensemble Predictions of Indian Summer Monsoon Rainfall

Nachiketa Acharya

Department of Meteorology and Atmospheric Science,
The Pennsylvania State University

Collaborator: Kyle Hall

IRI, Columbia University

PennState
College of Earth
and Mineral Sciences
Department of Meteorology and Atmospheric Science

IWM-7, 22-26 March 2022

Motivation

Various Approaches used for the Seasonal Forecasting of Indian Monsoon

- **Empirical/ Statistical model**
  - Multiple Regression
  - Canonical Correlation Analysis
  - Artificial Neural Network
  - Discriminant Analysis
  - Ex: IMD Operational LRF till 2020

– based on historical observed data for the predictand (e.g. rainfall, temperature) and for relevant predictors (e.g. SST, Snow Cover. Surface air temperature et

- **Dynamical/ Numerical Model**
  - SST Forced Atmospheric General Circulation models (AGCMs)
  - Ex: IMD SFM (experimental)
  - Coupled General Circulation Models (CGCMs)
  - Ex: IITM CFS

– using prognostic physic

2-tiered systems (first pre then climate).

1-tiered systems (predict atmosphere together)

- **Hybrid Model (Statistical + Dynamical)**
  - Statistical rescaling of dynamical model simulations
  - Ex. IITD – ERPS
  - Multi-Model Ensemble Forecasting
  - Ex: IMD's Current MME Forecasting System

Use of Dynamical Models for the Seasonal Forecasting of Indian Monsoons:

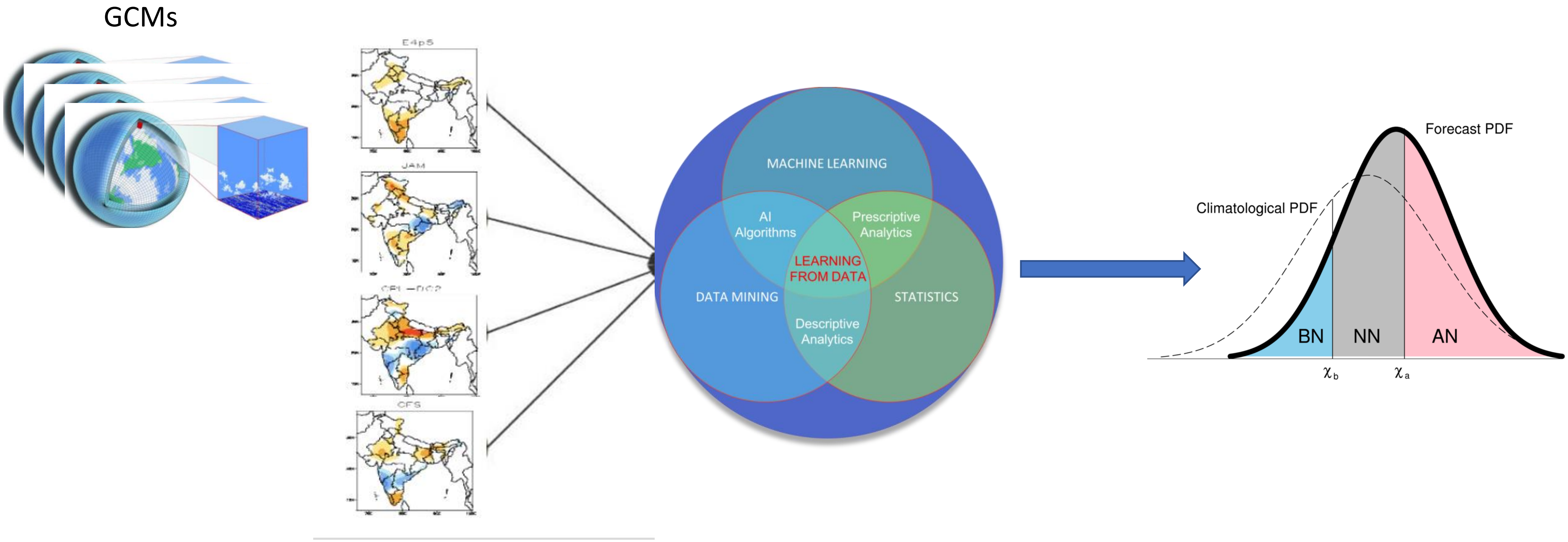
- In 2004, IMD implemented experimental dynamical forecasting system for the southwest monsoon rainfall using the seasonal forecast model (SFM) of the Experimental Climate Prediction Center (ECPC), USA. But the skill of these experimental forecasts based on SFM was very limited.
- In 2012, coupled forecasting system (CFS) was implemented at IITM, Pune on research model under monsoon mission.
- In January 2015, the SFM model was used to generate seasonal forecast outlook for south Asia under WMO Regional Climate Center (RCC)
- In 2016, monsoon mission CFS (MMCFS) model was used to generate ENSO & IOD bulletin and replace SFM to prepare seasonal forecast outlook for south Asia.
- In 2016, MMCFS was started to use to issue Temperature Forecast Outlook for AMJ season for the first time. Temperature forecast outlook for DJF was started in 2016 and that for MAM in 2017.
- In 2017, The high resolution (T382L64) monsoon mission CFS (MMCFS) was transferred to IMD, Pune.
- From 2017, IMD started to use MMCFS to generate experimental forecast for monsoon season along with SEFS
- In 2021 IMD implemented MME methodology based on coupled models

Courtesy: D.S.Pai's talk at IWM-7

Highlights of today’s discussion

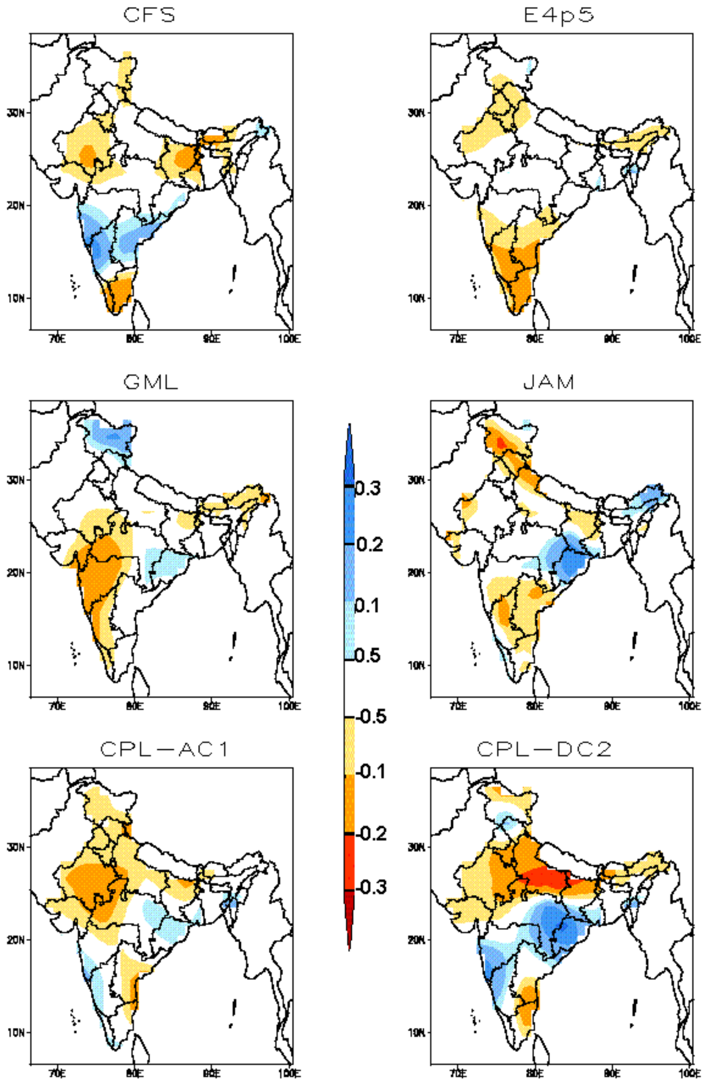
Construction of Probabilistic Multi Model Ensemble prediction for ISMR

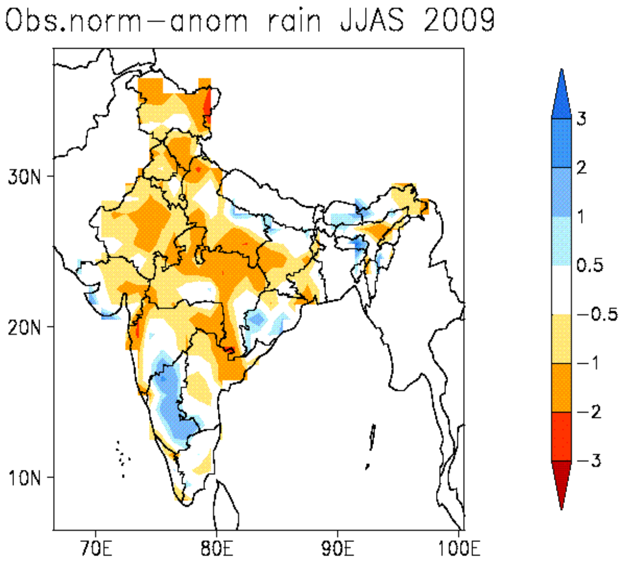
- What/How is PMME?
- How machine learning models can be used for constructing PMME



Why we need MME?

Example: Prediction of summer (JJAS) monsoon rainfall for India in 2009 in May.





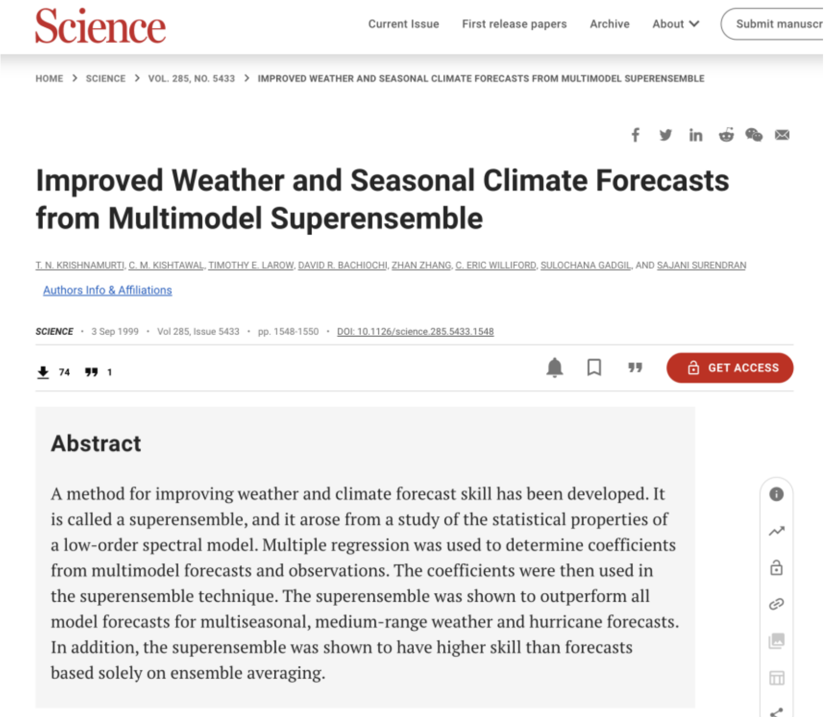
The purpose of the multi-model ensemble (MME) is to improve the skill of forecasts from individual GCM.

Probabilistic Multi Model Ensemble prediction:

- There are two common approaches to make a MME, viz., combining the individual ensemble forecasts with equal weights, or weighted according to their prior performance.

- Weighted MME: Mostly linear combination of GCM using multiple linear regression (MLR).

- Majority of the studies focus on deterministic forecast.



Science | Current Issue | First release papers | Archive | About | Submit manuscr

HOME > SCIENCE > VOL. 285, NO. 5433 > IMPROVED WEATHER AND SEASONAL CLIMATE FORECASTS FROM MULTIMODEL SUPERENSEMBLE

Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble

T. N. KRISHNAMURTI, C. M. KISHTAWAL, TIMOTHY E. LAROW, DAVID R. BACHIOCHI, ZHAN ZHANG, C. ERIC WILLIFORD, SULOCHANA GADGIL, AND SAJANI SURENDRAN

Authors Info & Affiliations

SCIENCE • 3 Sep 1999 • Vol 285, Issue 5433 • pp. 1548-1550 • DOI: 10.1126/science.285.5433.1548

74 | 1 | GET ACCESS

Abstract

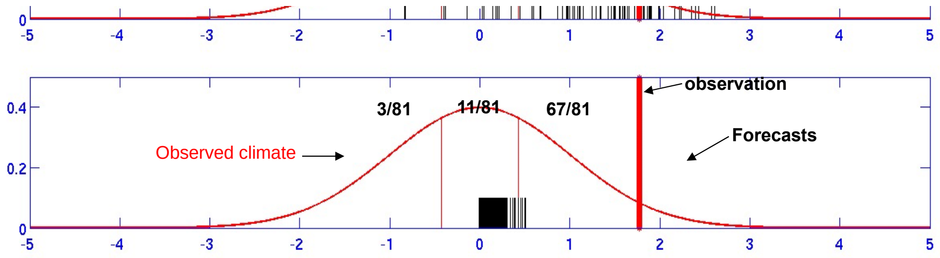
A method for improving weather and climate forecast skill has been developed. It is called a superensemble, and it arose from a study of the statistical properties of a low-order spectral model. Multiple regression was used to determine coefficients from multimodel forecasts and observations. The coefficients were then used in the superensemble technique. The superensemble was shown to outperform all model forecasts for multiseasonal, medium-range weather and hurricane forecasts. In addition, the superensemble was shown to have higher skill than forecasts based solely on ensemble averaging.

Probabilistic Multi Model Ensemble: Traditionally

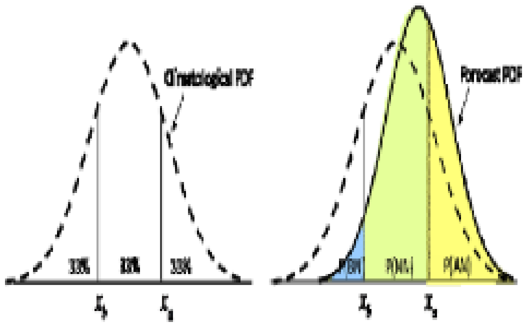
Probability from GCM

Non-parametric

Parametric



$$P(W|\,GCM_i) = \frac{m_i}{M}$$



$$X = \beta + \varepsilon$$

where

X is the forecast to be given,

β is the potentially predictable signal

ε is the error part

$$P_x(AN\,|\,\beta,\sigma_e) = P[(X > x_a)\,|\,\beta,\sigma_e]$$
$$= 1 - F_N\left[\frac{x_a - \beta}{\sigma_e}\right]$$
$$= F_N\left[\frac{\beta - x_a}{\sigma_e}\right]$$

$$P_x(BN\,|\,\beta,\sigma_e) = P[(X < x_b)\,|\,\beta,\sigma_e]$$
$$= F_N\left[\frac{x_b - \beta}{\sigma_e}\right]$$

Probabilistic Multi-Model Ensemble (PMME) by Non-parametric

$$P(\Omega) = \sum_{i=1}^{N} w_i P(\Omega \mid GCM_i)$$

W_i = Weights

Common Approaches

PMME1 $w_i = \frac{1}{N}$ N = Total number of GCM

PMME2 $w_i = \frac{\sqrt{M_i}}{\sum_{i=1}^{N} \sqrt{M_i}}$ M_i is the total number of ensemble members of ith GCM

Different approaches

PMME3 $w_i = \frac{b_i^{0.5}}{\sum_{i=1}^{N} b_i^{0.5}}$

b_i is the regression coefficient of the ith GCM

PMME4 $w_i = \frac{SNR_i^{0.5}}{\sum_{i=1}^{N} SNR_i^{0.5}}$ $SNR = \frac{\text{Variance of Ensemble Mean}}{\text{Variance of Intraensemble Deviation}}$

SNR_i is the signal-to-noise ratio of the ith GCM

PMME5 $w_i = \frac{r_i^{0.5}}{\sum_{i=1}^{N} r_i^{0.5}}$

r_i is the inverse of the root-mean square error of the ith GCM

Limitation:

1. some poor model bring down the skill of PMME.
2. More members doesn't mean more skillful

Note: The simple monotonic function can be considered as the weights proportional to some power of the regression coefficients. Hence, larger regression coefficients would be ‘‘better’’ than those with smaller regression coefficients and should therefore be assigned larger weights. The best choice for the power has been empirically estimated as 0.5

Acharya et al (2013)

Major Limitation of non-parametric method:

Some of the GCMs seasonal hindcast or subseasonal reforecast ensembles **contain fewer ensemble members, straightforward counting of the ensemble members exceeding a chosen threshold to determine probabilities can lead to large errors**.

Probabilistic Multi-Model Ensemble (PMME) by Parametric

$\beta =$ Deterministic MME (superensemble)

$$X = \beta + \varepsilon$$

$\sigma_\varepsilon =$??

Probabilistic Forecast

Ensemble Spread (ES)

Uncertainty represents by Ensemble spread which is calculated as the variance of ensemble members for a particular year or average of year-to-year variance of ensemble members.

$$\sigma_t^2 = \frac{1}{M}\sum_{i=1}^{M}\left(f_t^i - \overline{f}_t\right)$$

Error Residual (ER)

Uncertainty represents by Root Mean Square Error (RMSE).

$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(X_t - \overline{f}_t\right)^2$$

Correlation (CR)

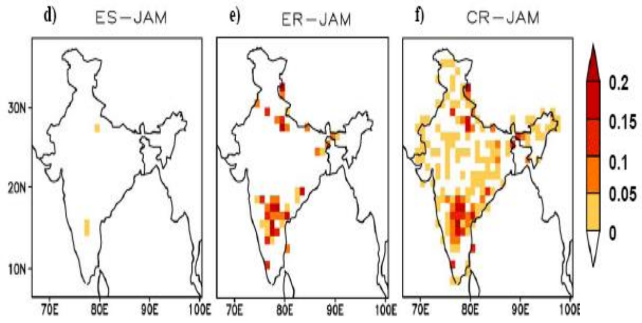
Uncertainty is considered as the function of correlation between observation and signal (β)

$$r = \frac{\mathrm{cov}(Y,\beta)}{\sqrt{\sigma_X^{\ 2}\ \sigma_\beta^2}}$$

$$= \frac{\sigma_\beta}{(\sqrt{(\sigma_\beta^{\ 2} + \sigma_\varepsilon^2})}$$

$$\Rightarrow (\sigma_\beta^{\ 2} + \sigma_\varepsilon^2) = \frac{\sigma_\beta^2}{r^2}$$

$$\Rightarrow \sigma_\varepsilon^2 = \sigma_\beta^2\left(\frac{1}{r^2} - 1\right)$$



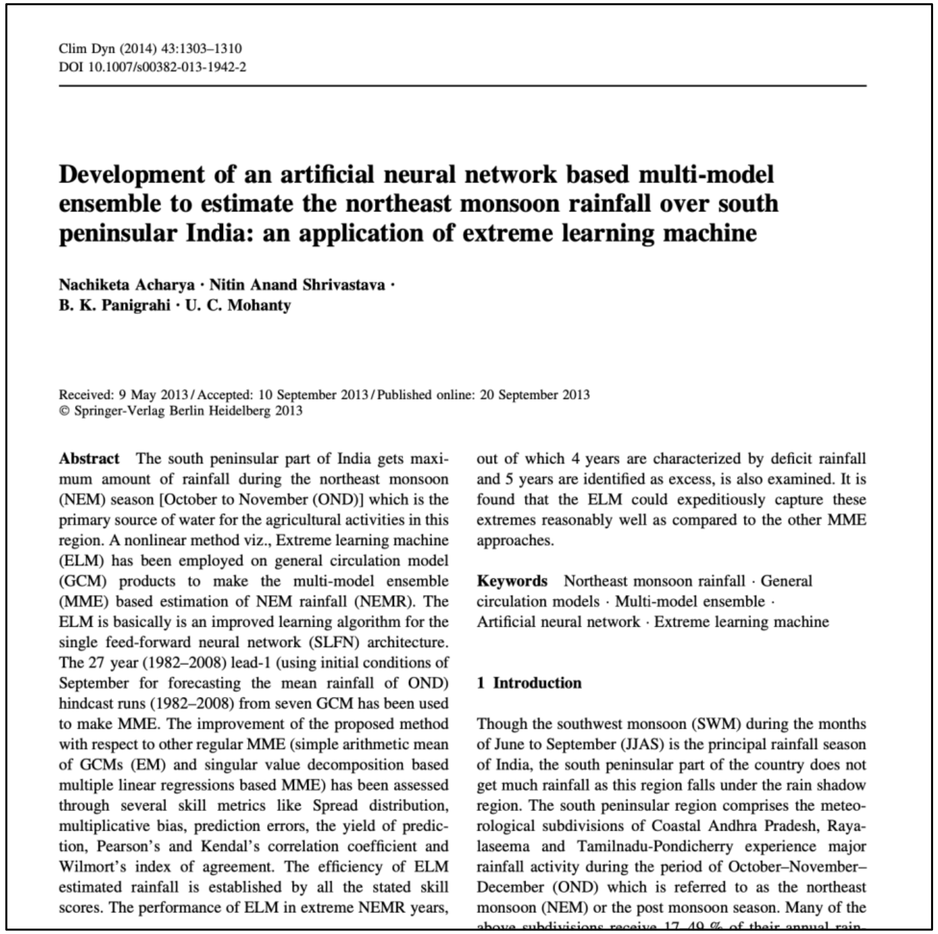
Acharya et al (2014)

Major Limitation of parametric method:

- Two stage modeling: one for mean of the distribution (deterministic) and then for variance (convert deterministic to probabilistic)
- Assumption of Normal distribution not valid for S2S scale with less hindcast period.

Other option: Machine Learning??

Why machine learning models can be used for constructing PMME?



Clim Dyn (2014) 43:1303–1310
DOI 10.1007/s00382-013-1942-2

Development of an artificial neural network based multi-model ensemble to estimate the northeast monsoon rainfall over south peninsular India: an application of extreme learning machine

Nachiketa Acharya · Nitin Anand Shrivastava · B. K. Panigrahi · U. C. Mohanty

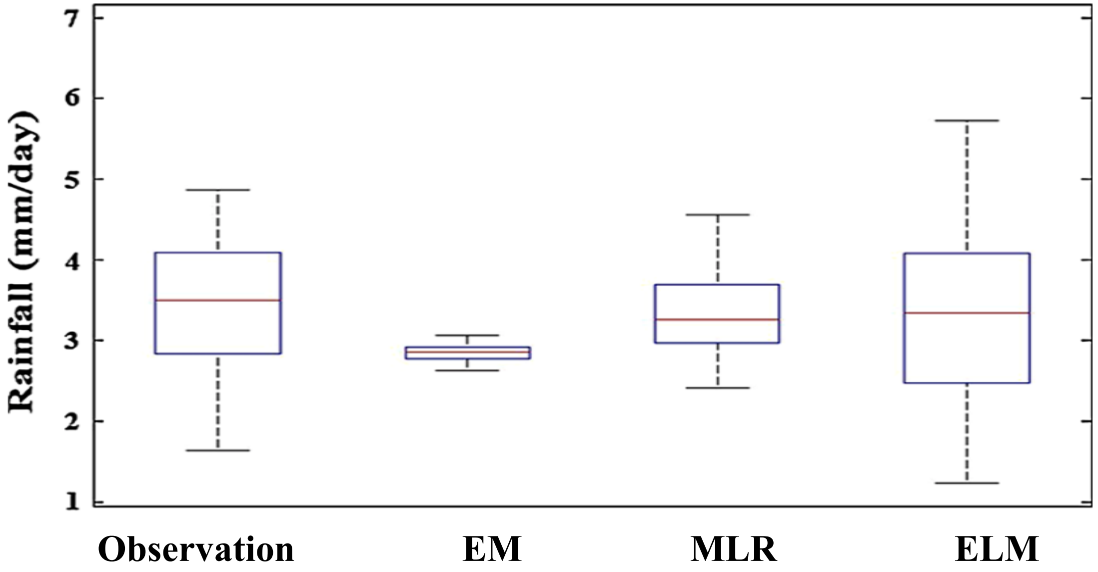
Received: 9 May 2013 / Accepted: 10 September 2013 / Published online: 20 September 2013

Abstract The south peninsular part of India gets maximum amount of rainfall during the northeast monsoon (NEM) season [October to November (OND)] which is the primary source of water for the agricultural activities in this region. A nonlinear method viz., Extreme learning machine (ELM) has been employed on general circulation model (GCM) products to make the multi-model ensemble (MME) based estimation of NEM rainfall (NEMR). The ELM is basically is an improved learning algorithm for the single feed-forward neural network (SLFN) architecture. The 27 year (1982–2008) lead-1 (using initial conditions of September for forecasting the mean rainfall of OND) hindcast runs (1982–2008) from seven GCM has been used to make MME. The improvement of the proposed method with respect to other regular MME (simple arithmetic mean of GCMs (EM) and singular value decomposition based multiple linear regressions based MME) has been assessed through several skill metrics like Spread distribution, multiplicative bias, prediction errors, the yield of prediction, Pearson's and Kendal's correlation coefficient and Wilmort's index of agreement. The efficiency of ELM estimated rainfall is established by all the stated skill scores. The performance of ELM in extreme NEMR years, out of which 4 years are characterized by deficit rainfall and 5 years are identified as excess, is also examined. It is found that the ELM could expeditiously capture these extremes reasonably well as compared to the other MME approaches.

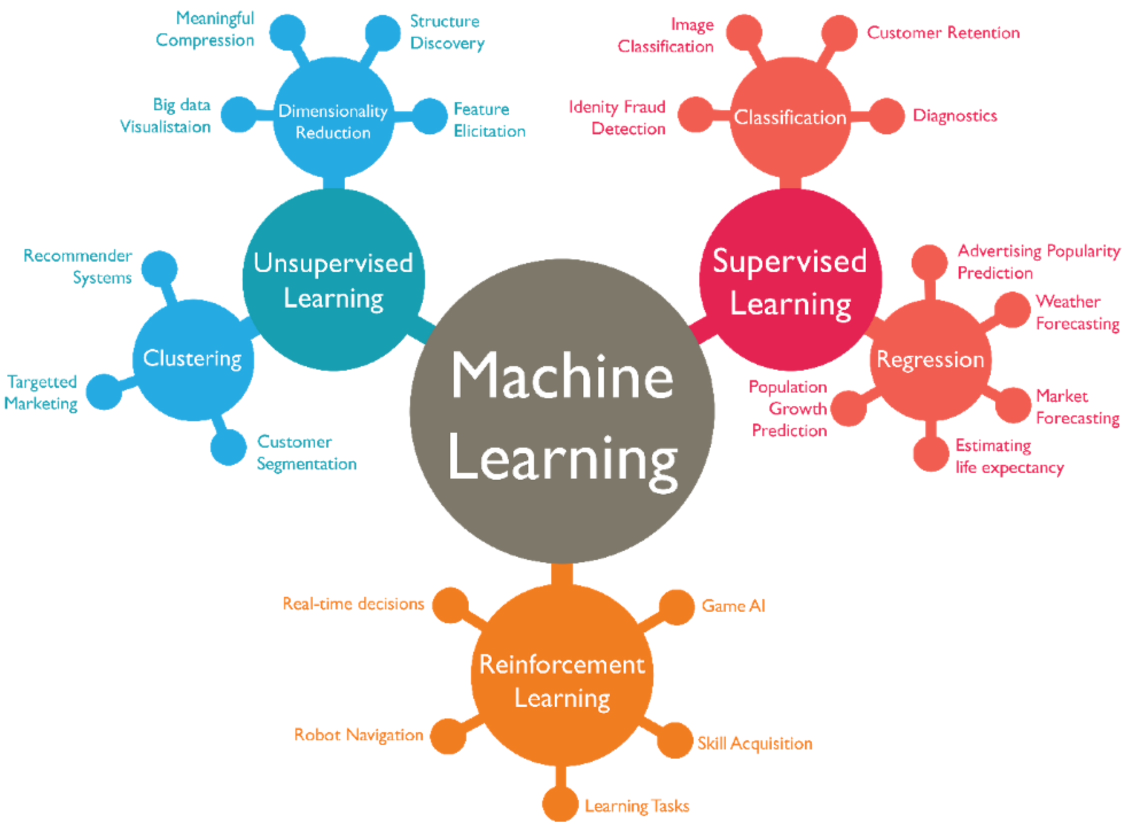
Keywords Northeast monsoon rainfall · General circulation models · Multi-model ensemble · Artificial neural network · Extreme learning machine

1 Introduction

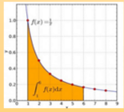
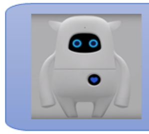
Though the southwest monsoon (SWM) during the months of June to September (JJAS) is the principal rainfall season of India, the south peninsular part of the country does not get much rainfall as this region falls under the rain shadow region. The south peninsular region comprises the meteorological subdivisions of Coastal Andhra Pradesh, Rayalaseema and Tamilnadu-Pondicherry experience major rainfall activity during the period of October–November–December (OND) which is referred to as the northeast monsoon (NEM) or the post monsoon season. Many of the above subdivisions receive 17–49 % of their annual rain-



How machine learning models can be used for constructing PMME?



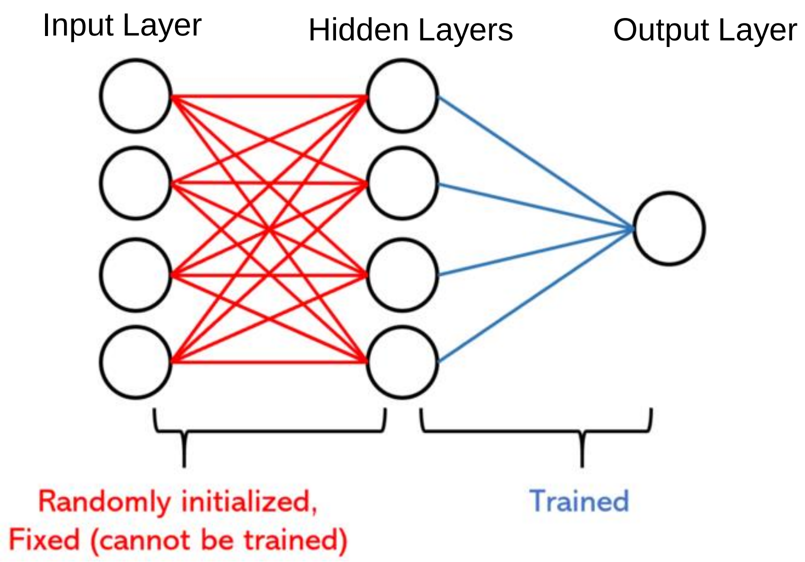
"*Statistical Learning* theory is a framework for Machine Learning drawing from the fields of statistics and functional analysis."

	 Statistics	+  Machine Learning	= $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$ Statistical Learning
Subfield of...	Mathematics	Computer Science (AI)	Statistics & Machine Learning
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data	Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy	Building statistical models for prediction; understanding data
Prior assumptions about data	Some knowledge about population usually required	None	Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data	Usually applied to high dimensional data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed: stats is basis for algorithms	Knowledge of statistics and ML required

Extreme Learning Machine: A “Generalized” SLFN

❖To overcome such shortcomings, recently, a novel learning algorithm for single-hidden-layer feed forward neural networks (SLFN) called extreme learning machine (ELM) has been proposed by Huang et al., (2008).

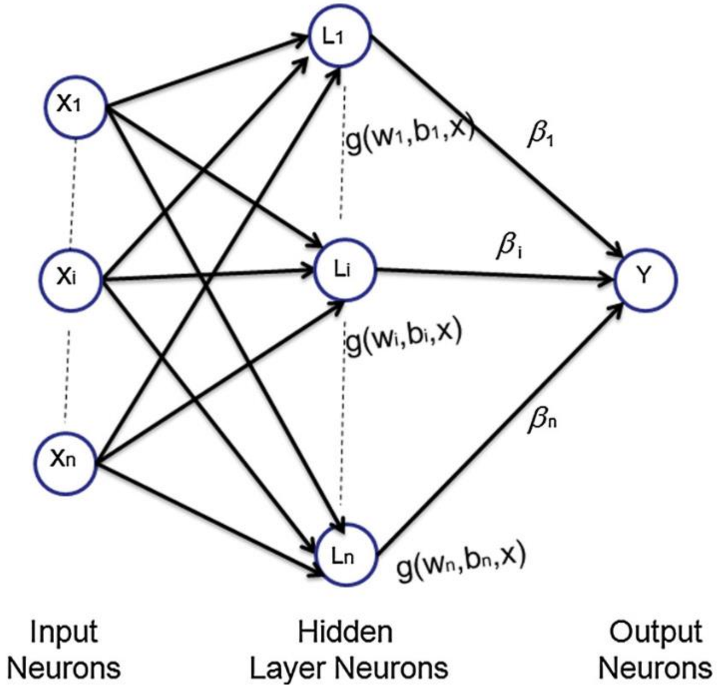
❖The basic principle which distinguishes ELM from the traditional SLFN is that all the parameters (input weights and hidden layer biases) are not required to be tuned.



❖After randomly choosing the input weights and the hidden layer biases, SLFNs can be simply considered as a linear system. The output weights which link the hidden layer to the output layer of this linear system can now be analytically determined through Moore-Penrose (MP) generalized inverse of the hidden layer output matrices.

❖This simplified approach makes ELM thousands of times faster than that of traditional SLFN. ELM also avoids many difficulties faced by gradient-based learning methods such as stopping criteria, learning rate, learning epochs, local minima, and the over- tuning problems.

Mathematical Background of ELM



$$f_L(x) = \sum_{i=1}^{L} \beta_i g_i(x) = \sum_{i=1}^{L} \beta_i g(w_i * x_j + b_i), j = 1, ..., N$$

- *L* is a number of hidden units
- *N* is a number of training samples
- β is weight vector between *i* th hidden layer and output
- *w* is a weight vector between input and hidden layer
- *g* is an activation function
- *b* is a bias vector
- *x* in an input vector

Mathematical Background of ELM

$$T = H\beta$$

$$H = \begin{bmatrix} g(w_1 * x_1 + b_1) & \ldots & g(w_L * x_1 + b_L) \\ \vdots & \ldots & \vdots \\ g(w_1 * x_N + b_1) & \ldots & g(w_L * x_N + b_L) \end{bmatrix}_{N \times L}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}$$

- m is a number of outputs
- H is called Hidden Layer Output Matrix
- T is a training data target matrix

Learning algorithm

1. Randomly assign weight w_i and bias b_i, $i = 1, \ldots L$
2. Calculate hidden layer output **H**
3. Calculate output weight matrix $\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T}$.
4. Use $\hat{\beta}$ to make a prediction on new data $T = H\hat{\beta}$

Limitation of traditional ELM for probabilistic prediction:

Output is deterministic and classify for only binary cases.

Solution: Probabilistic Output Extreme Learning Machine (PO-ELM) (Wong et al 2019)

Discard the linear output function of ELM.

ELM	PO-ELM
$T = H\beta$	$\mathbf{T} = \dfrac{1}{1+\exp(-\mathbf{H}\boldsymbol{\beta})}.$
$\boldsymbol{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T}.$	$\boldsymbol{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\log(\mathbf{1}-\mathbf{T}) - \log(\mathbf{T}))$

Limitation of PO-ELM:

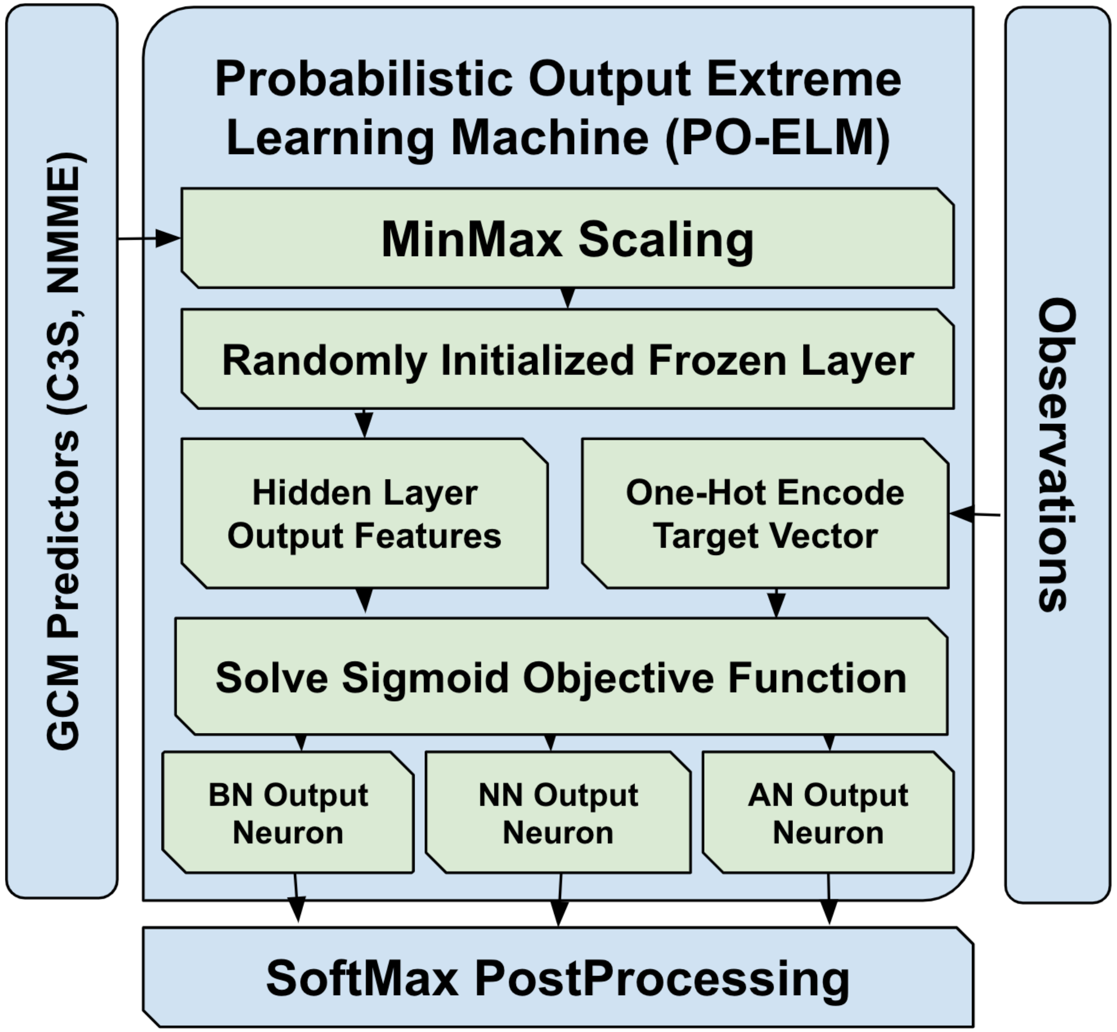
- Log function on the observed category.
- Probabilities for PO-ELM do not add to one rather a binary probability.
- Not for multi-label classification.

Our Contributions:

- We modify it by introducing "softmax" function on the output probabilities from PO-ELM and give observed category as 0.0001 or 0.9999 to prevent undefined behavior in logs in POELM fitting.

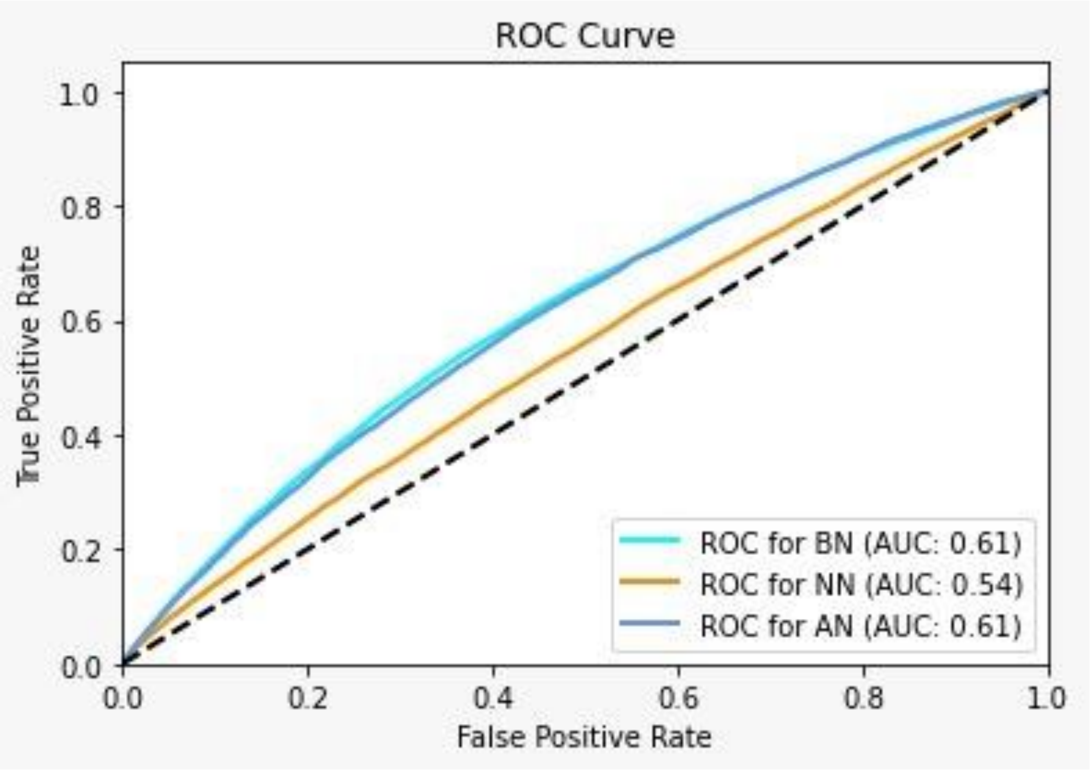
Acharya and Hall (in preparation)

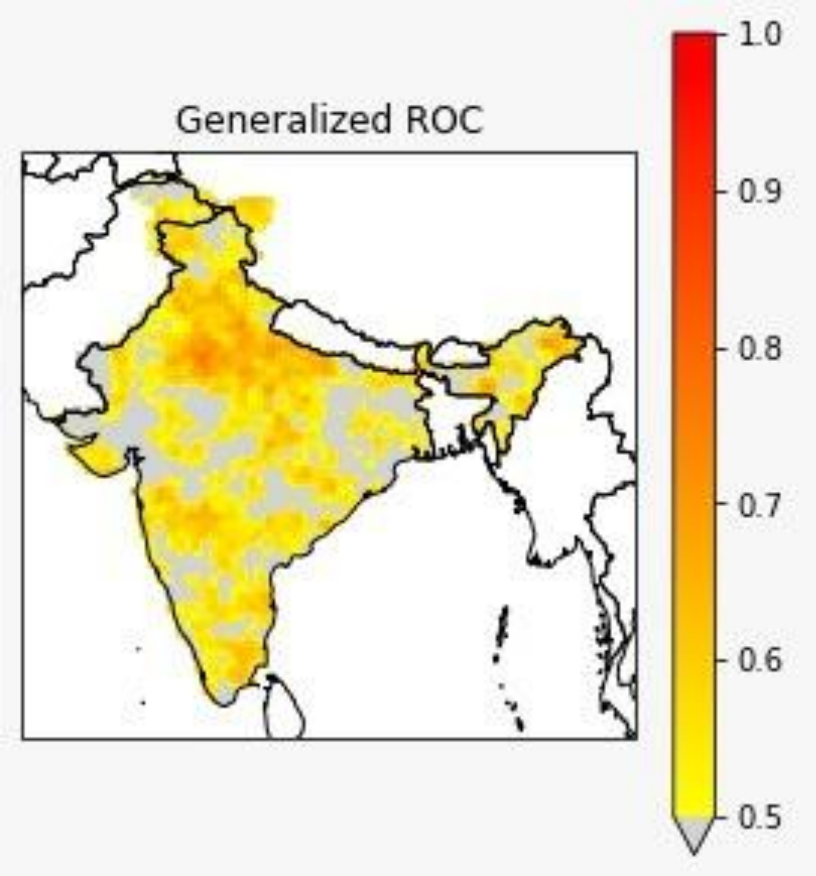
Architecture of PO-ELM based PMME



Courtesy: Kyle Hall

Results

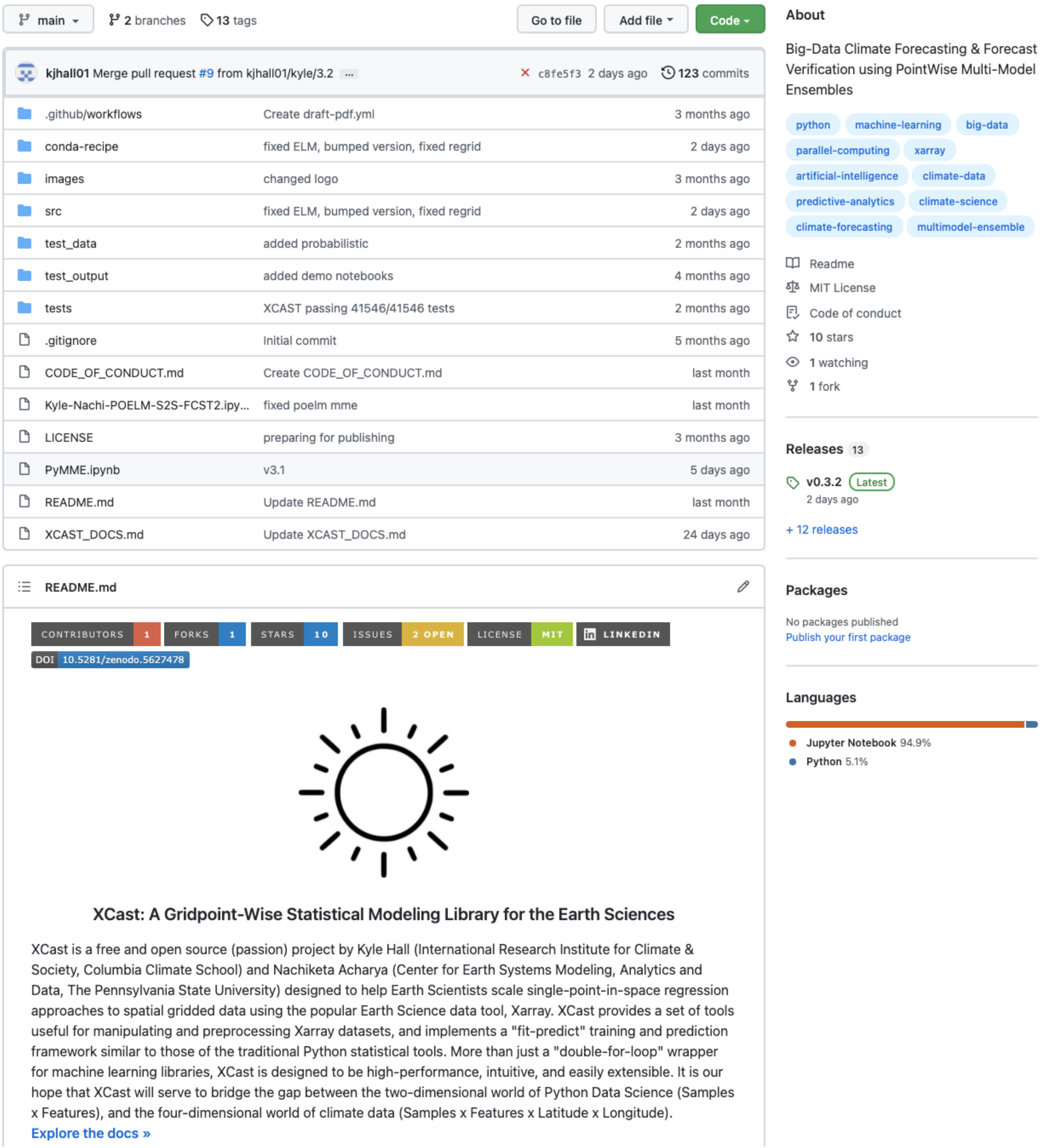

ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC for BN (AUC: 0.61)
ROC for NN (AUC: 0.54)
ROC for AN (AUC: 0.61)


Generalized ROC
1.0
0.9
0.8
0.7
0.6
0.5

How to do: Xcast

https://github.com/kjhall01/xcast/

- Xcast is designed to help Earth Scientists scale single-point-in-space regression approaches to spatial gridded data using the popular Earth Science data tool.
- XCast provides a set of tools useful for manipulating and preprocessing Xarray datasets and implements a "fit-predict" training and prediction framework similar to those of the traditional Python statistical tools.
- XCast will serve to bridge the gap between the two-dimensional world of Python Data Science and the four-dimensional world of climate data.

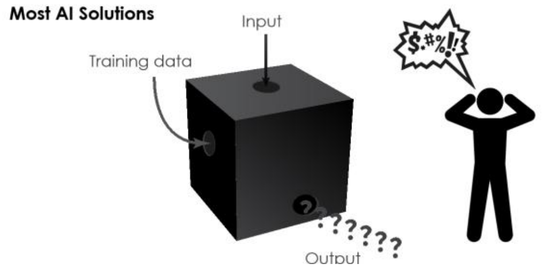


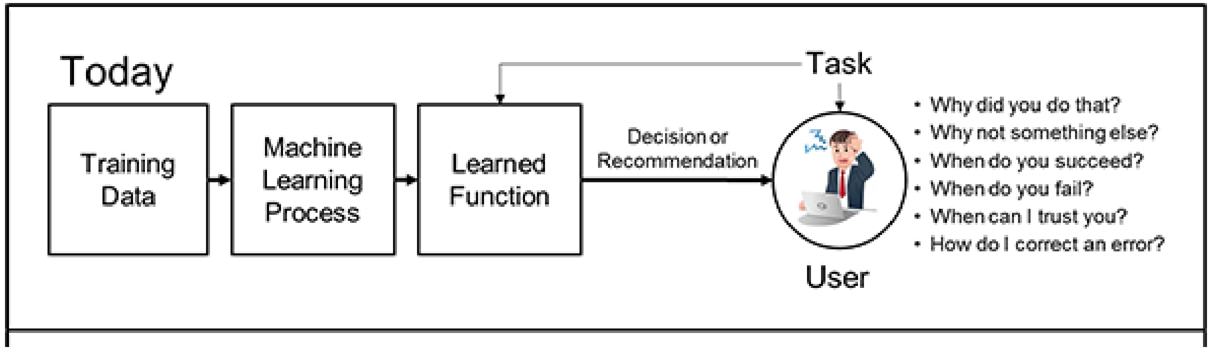
Hall and Acharya,2021

Remarks

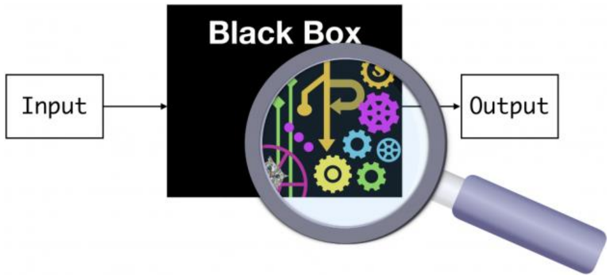
- “PMME” is a simple concept, but it has many layers, and I took a step-by-step approach.
- Non-parametric, parametric are various way to create “PMME” but each have its limitations.
- Machine Learning based method can be used but with “white box” approach.

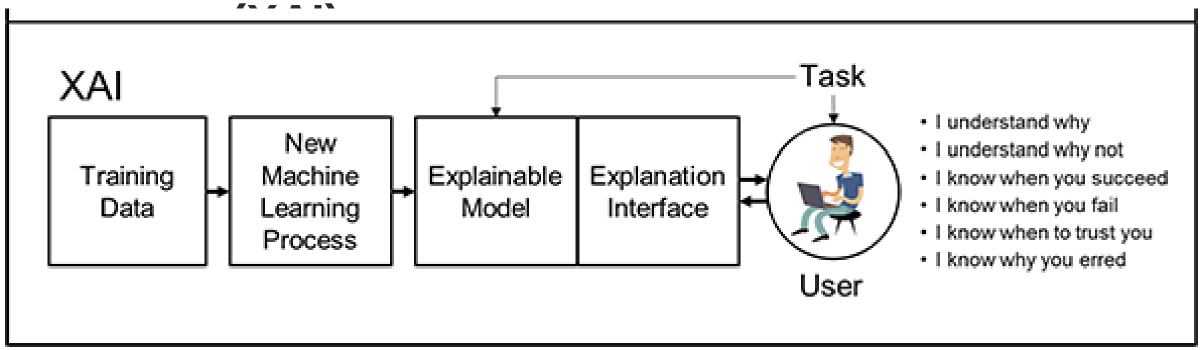
Take Home Messages:

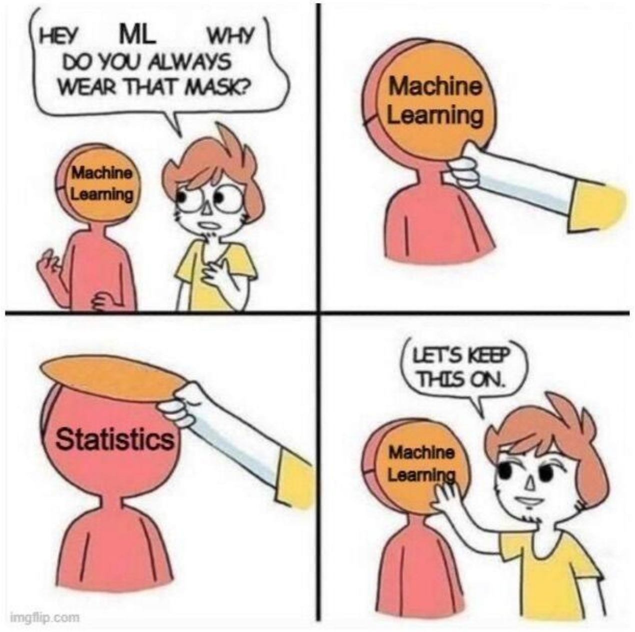




Explainable Artificial Intelligence (XAI)







Thank you!

ny Questions/future collaborations?

dr.nachiketaacharya@gmail.com