The International Conference on Regional Climate-CORDEX 2023
(IITM HUB ICRC-CORDEX 2023)

Machine learning based lightning prediction over Eastern India

Dr. Javed Akhter* & Prof. Subrata Kumar Midya

Department of Atmospheric Sciences,
University of Calcutta, Kolkata-700019

*Email: akhterexpressju@gmail.com

Background

- Lightning is one of the severe atmospheric phenomena posing serious threats to human life and property. According to several studies (Illiyas et al., 2014 ;Singh and Singh, 2015; Yadava et al., 2020), lightning events in India cause deaths of more than 2000 people every year, which is about 9 percent of total deaths due to natural disasters.

- However, the limited availability of lightning data especially over the Indian Sub-continent is a major constraint for lightning study. Hence, the development of proxy lightning data from other atmospheric variables is very important for both analysis and prediction purposes.

- Recent advances in machine learning (ML) provide the opportunity to develop and predict lightning using various dynamic–thermodynamic atmospheric variables that influence lightning activity.

- Developing ML models with spatial data is a challenging task due to the presence of spatial auto-correlation (SAC) especially for gridded climate data. Use of spatial predictors using Moran's Eigenvector Maps (MEMs) is a useful approach to remove SAC and develop robust ML models.

Objectives

- Assessment of ML models in simulating spatial variability of lightning over Eastern India
- Incorporation of spatial predictors to remove spatial auto-correlation affecting ML models
- Comparison between non-spatial and spatial ML models

Data

Lightning data:

Flash rate density (FRD) data obtained from Low Resolution (2.5° × 2.5°) Monthly Climatology Time Series (LRMTS) v2.3 generated by Optical Transient Detector (OTD) and Lightning Imaging Sensor (LIS) onboard TRMM satellite [Data period: 1996-2013]

Reanalysis data:

Atmospheric predictors like Convective Available Potential Energy (CAPE), Convective Precipitation (CP), Total Column Water Vapour (TCWV), K-index(KI), Total Totals Index (TTI), 2m Air Temperature (t2m), Specific Humidity (SHUM), Relative Humidity (RHUM), Upper Air Temperature (TA) are taken from the ECMWF ERA5 monthly reanalysis dataset (0.25° × 0.25°).

SHUM,RHUM and TA has been used at 500,700 and 850 pressure levels.

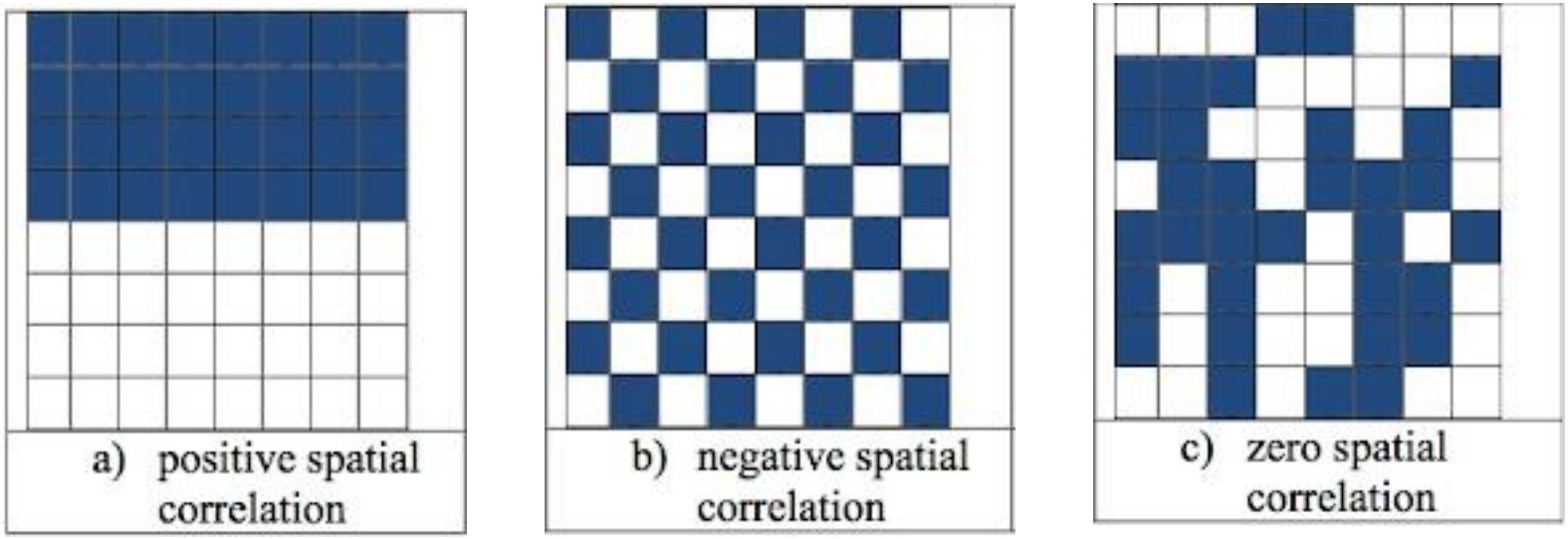
Spatial Autocorrelation

Spatial autocorrelation (SAC): Presence of systematic spatial variation in a mapped variable.

Positive spatial autocorrelation: Adjacent observations have similar data values

Negative spatial autocorrelation: Adjacent observations tend to have very contrasting values

Random spatial autocorrelation : Similar values are neither close nor distant from each other



a) positive spatial correlation

b) negative spatial correlation

c) zero spatial correlation

Spatial Autocorrelation

The presence of spatial autocorrelation is seen as posing a serious shortcoming for hypothesis testing and prediction (Lennon 2000, Dormann 2007b), because it **violates the assumption of independently and identically distributed (i.i.d.) errors** of most standard statistical procedures (Anselin 2002) and hence **inflates type I errors**, occasionally even **inverting the slope of relationships** from non-spatial analysis (Kuhn 2007).

Beale et al. (2007) found **precision of the standardized coefficients produced by the regression significantly decreased** when the residual autocorrelations were strong.

Spatial Autocorrelation

The Moran's I statistic for spatial autocorrelation is given as:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad (1)$$

where z_i is the deviation of an attribute for feature i from its mean $(x_i - \bar{X})$, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features, and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \quad (2)$$

Null hypothesis: The attribute being analyzed is randomly distributed among the features in the study area. When p-value returned is statistically significant, null hypothesis can be rejected.

Spatial Autocorrelation

Moran's Eigenvector Map (MEM)

The Moran eigenvector approach (Dray et al. 2006; Griffith and Peres-Neto 2006) involved the spatial patterns represented by maps of eigenvectors; by choosing suitable orthogonal patterns and adding them to a linear or generalized linear model, the spatial dependence present in the residuals can be moved into the model.

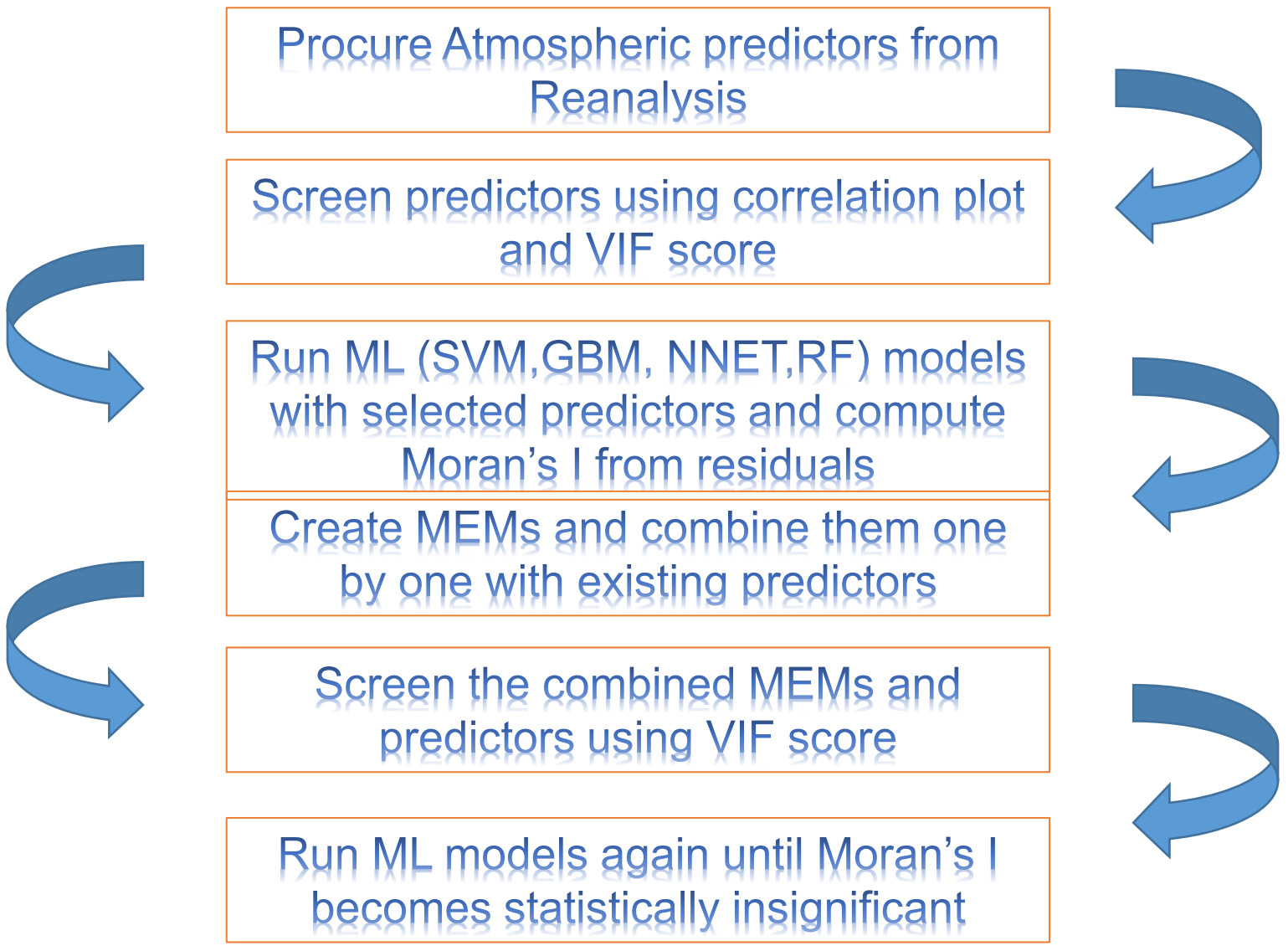
MEMs are a set of spatial weight matrix eigenvectors into a regression model specification to capture SAC (Griffith,2003). Eigenvectors can be extracted from a doubly centered spatial weights matrix $\mathbf{C}$,which can be expressed as follows:

$$\mathbf{MCM}=(\mathbf{I}-\mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I}-\mathbf{1}\mathbf{1}^T/n)$$

where $\mathbf{I}$ is an n -by-n identity matrix, $\mathbf{1}$ is a n-by-1 vector of ones, n is the number of areal units (Grid points), T is the matrix transpose operator.

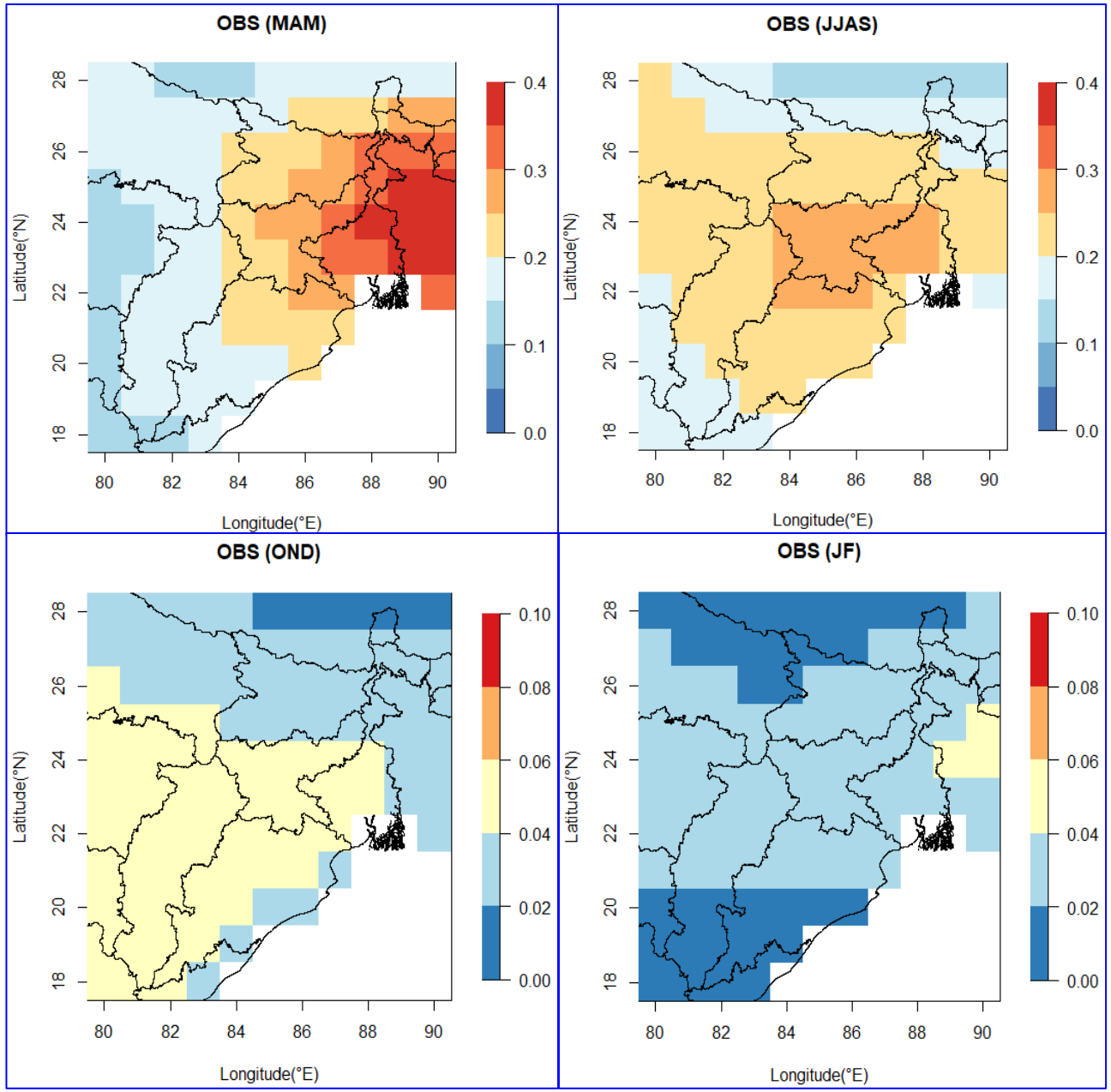
A subset of these eigenvectors was included as independent variables in a model specification and captures SAC so that a linear regression did not suffer from a violation of the independence assumption that is caused by SAC (Griffith,2003). This subset can be identified from a candidate eigenvector set with a stepwise regression procedure (Chun et al., 2019).

Methodology



Results

Lightning Climatology (1996-2013)

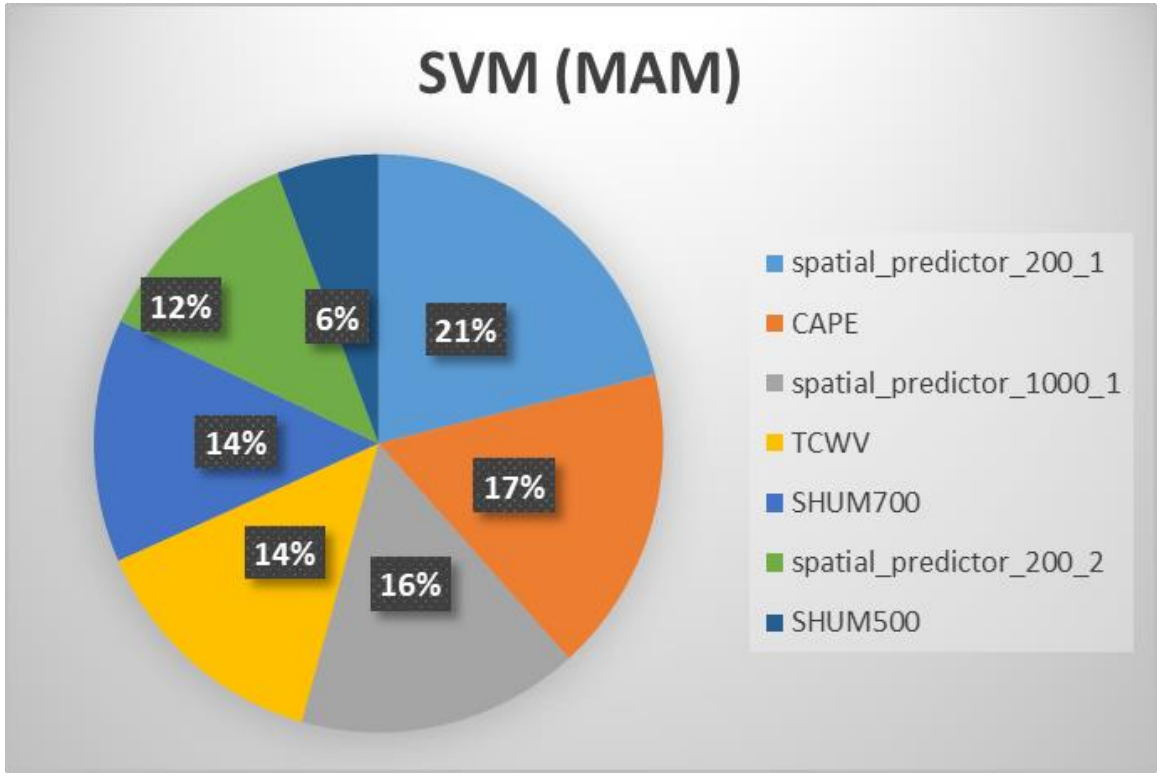


Results

Support Vector Machine (SVM)

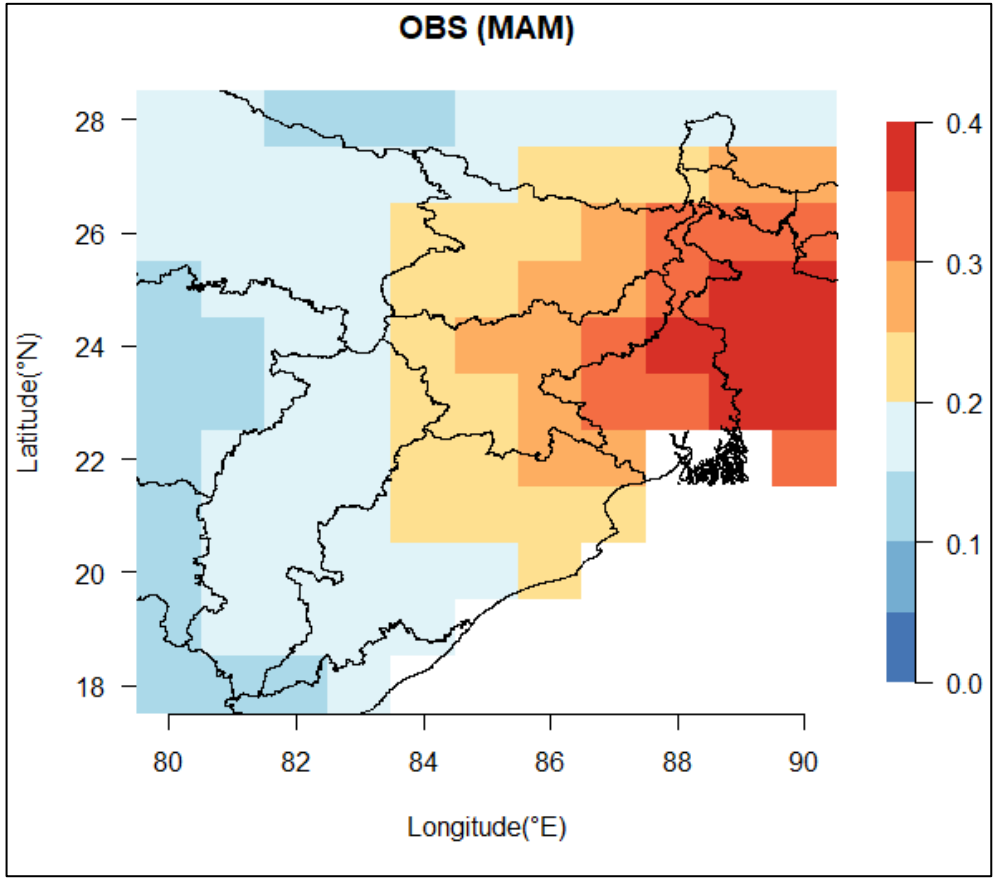
	SVM	Spatial-SVM
Moran's I	**0.022**	**0.002**
p value	**0.00**	**0.26**
No. of MEMs	-	**3**
R2	**0.88**	**0.97**
RMSE	**0.02**	**0.01**
MAE	**0.02**	**0.01**

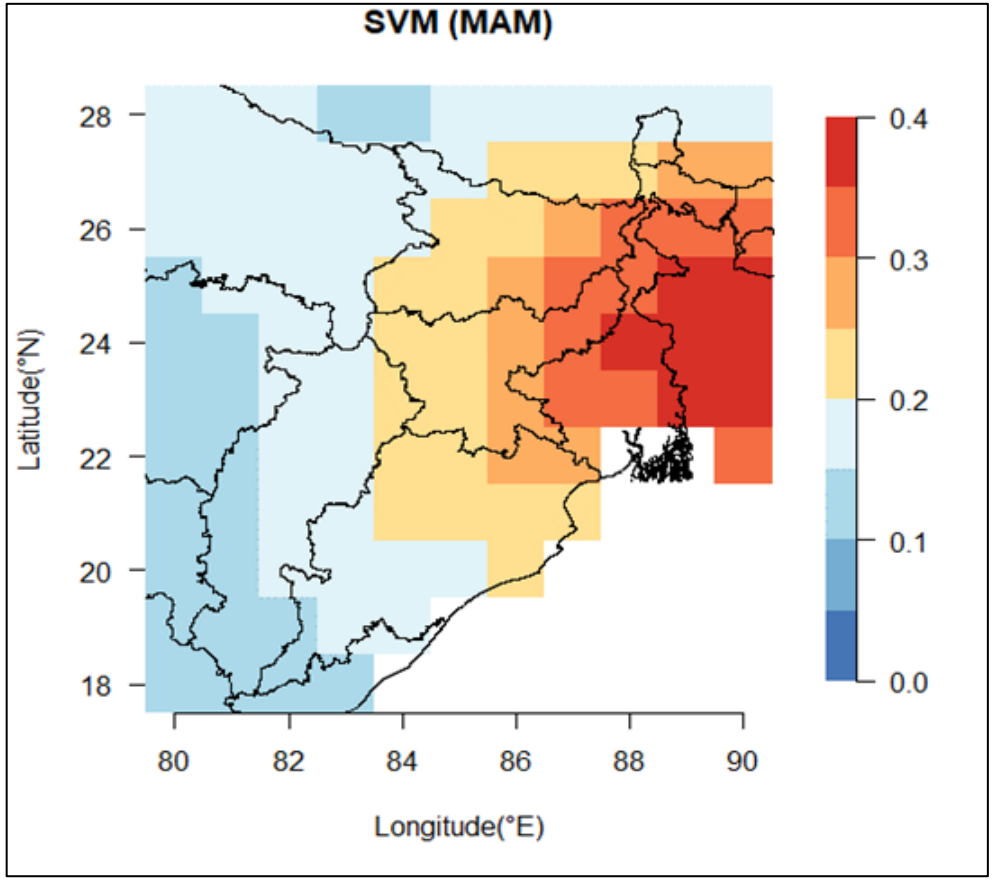
Training period: 1996-2010



Non-spatial: sigma = 0.5 and C = 8
Spatial: sigma = 0.1 and C = 8

Testing period: 2011-2013





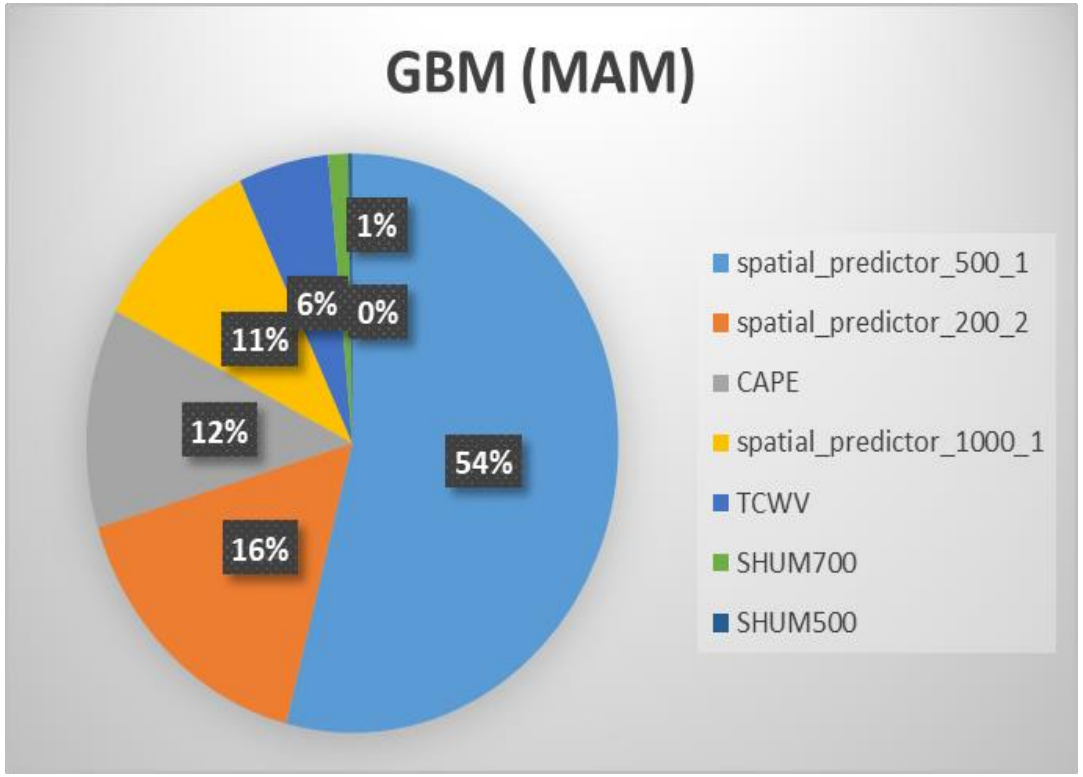
R	0.97
d	0.98
NSE	0.91
RSR	0.29

Results

Gradient Boosting Machine (GBM)

Training period: 1996-2010

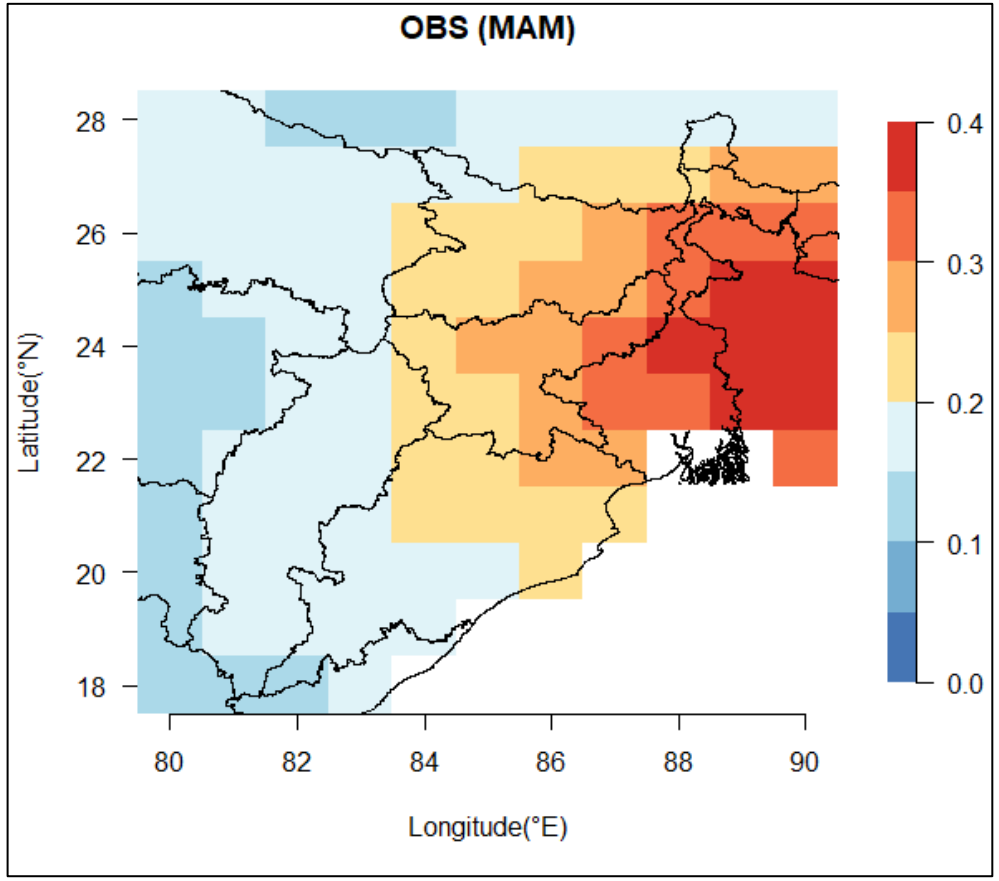
	GBM	Spatial-GBM
Moran's I	**0.02**	**-0.02**
p value	**0.004**	**0.18**
No. of MEMs	-	**3**
R2	**0.82**	**0.93**
RMSE	**0.03**	**0.02**
MAE	**0.02**	**0.01**

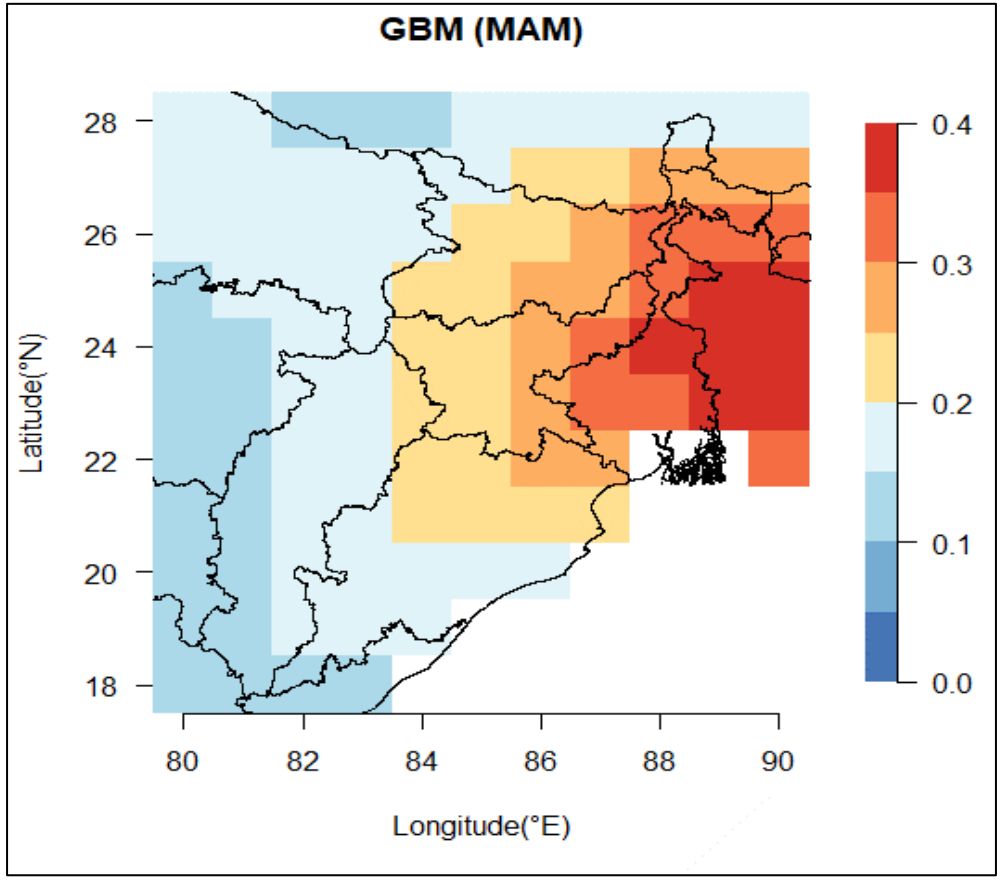


Non-Spatial: interaction.depth = 3, n.trees = 2000,shrinkage = 0.01,n.minobsinnode = 5
Spatial: n.trees = 2000, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 5

Results

Testing period: 2011-2013





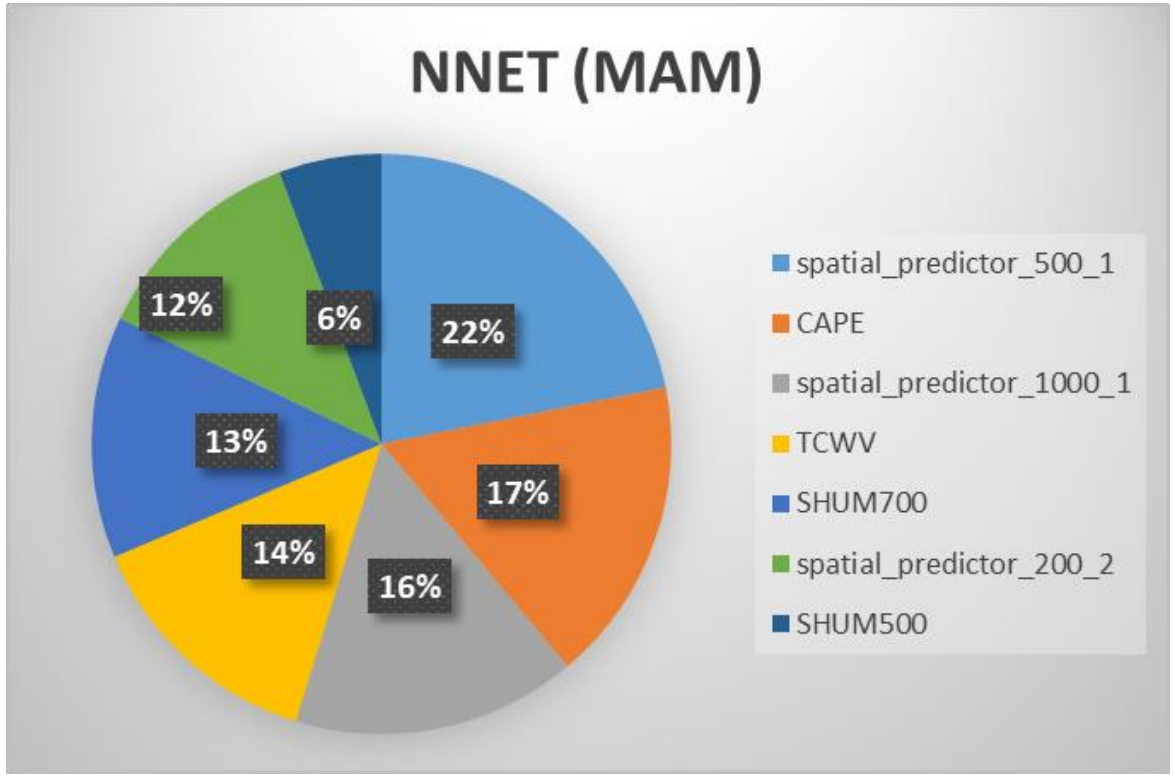
R	0.98
d	0.98
NSE	0.93
RSR	0.27

Results

Artificial Neural Network (NNET)

Training period: 1996-2010

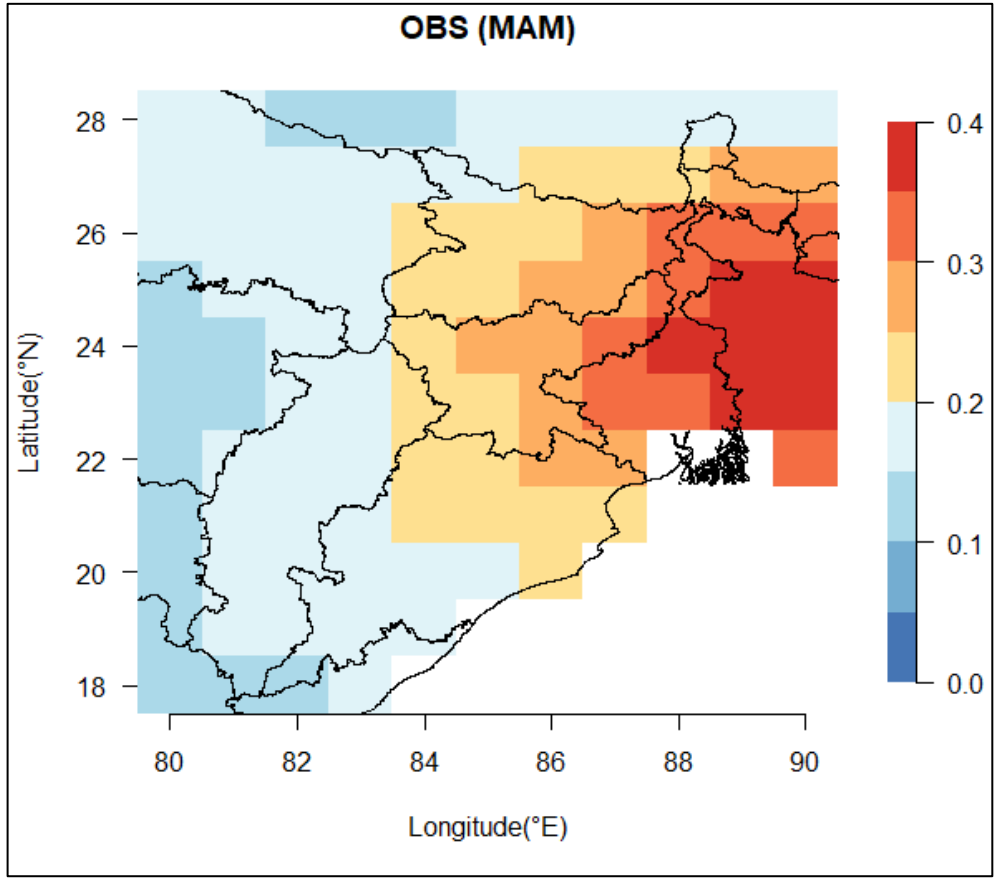
	NNET	Spatial-NNET
Moran's I	**0.11**	**0.001**
p value	**0**	**0.29**
No. of MEMs	**-**	**3**
R2	**0.78**	**0.98**
RMSE	**0.03**	**0.01**
MAE	**0.02**	**0.007**

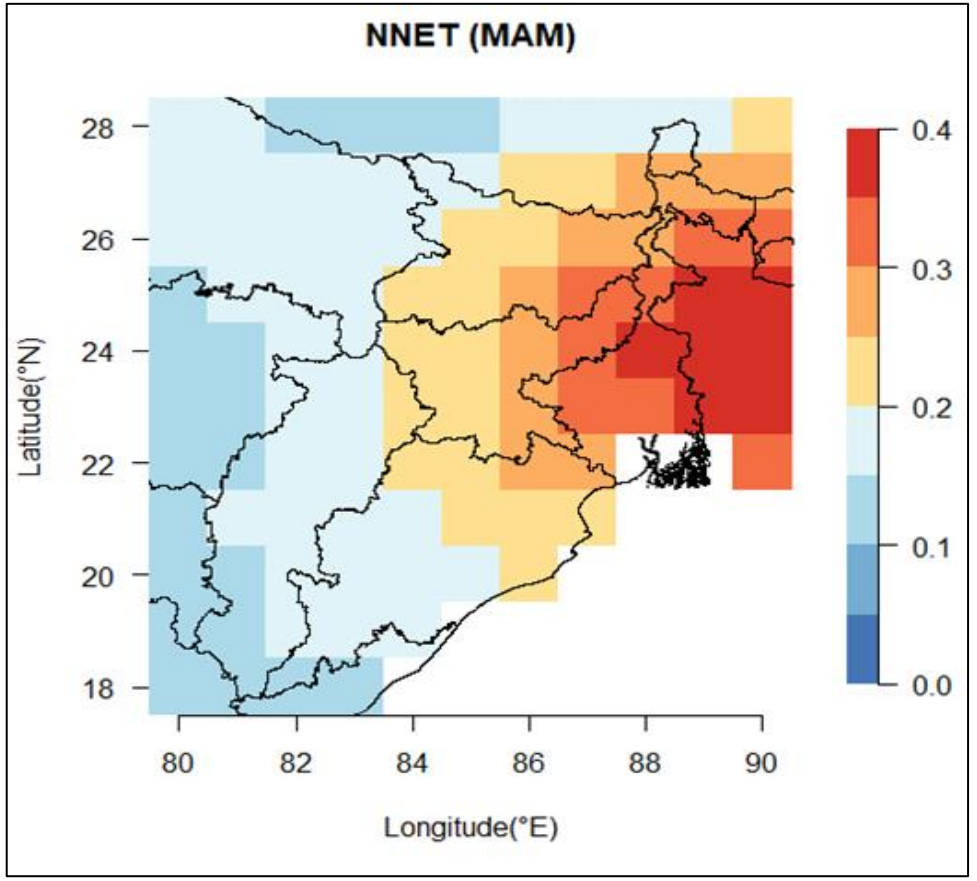


Non-spatial: size = 5, decay = 0.01
Spatial: size = 14, decay = 0.001

Results

Testing period: 2011-2013





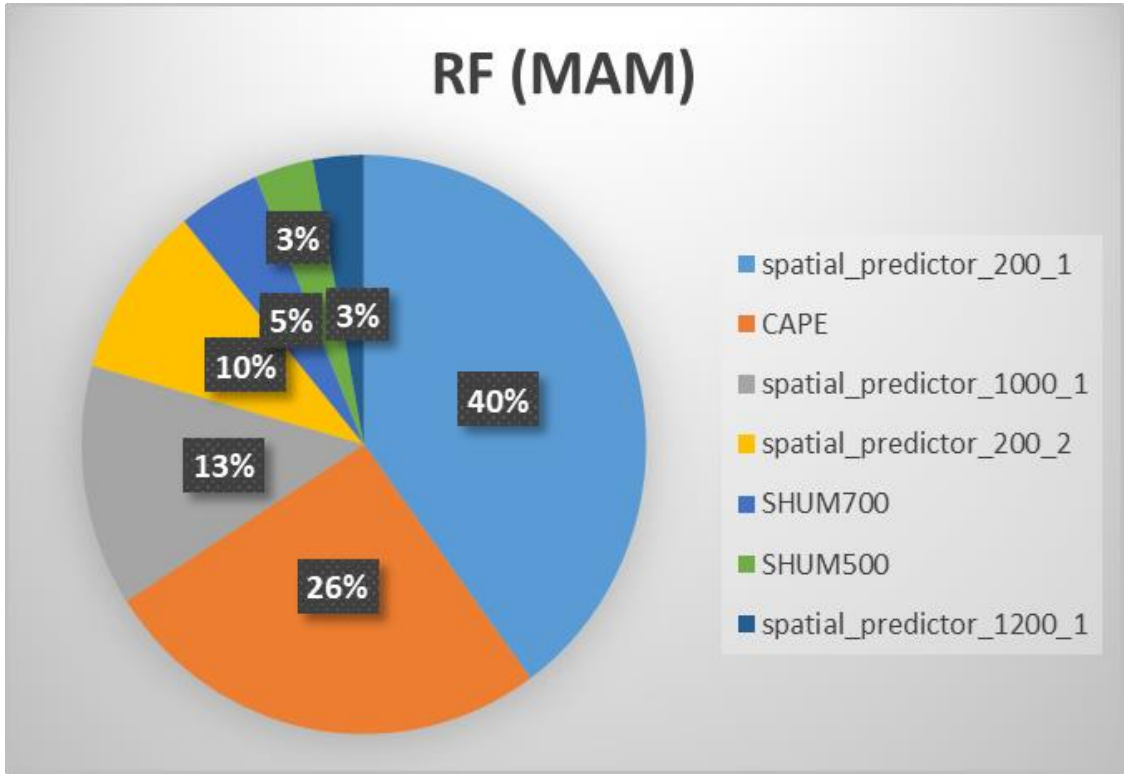
R	0.97
d	0.98
NSE	0.92
RSR	0.28

Results

Random Forest (RF)

Training period: 1996-2010

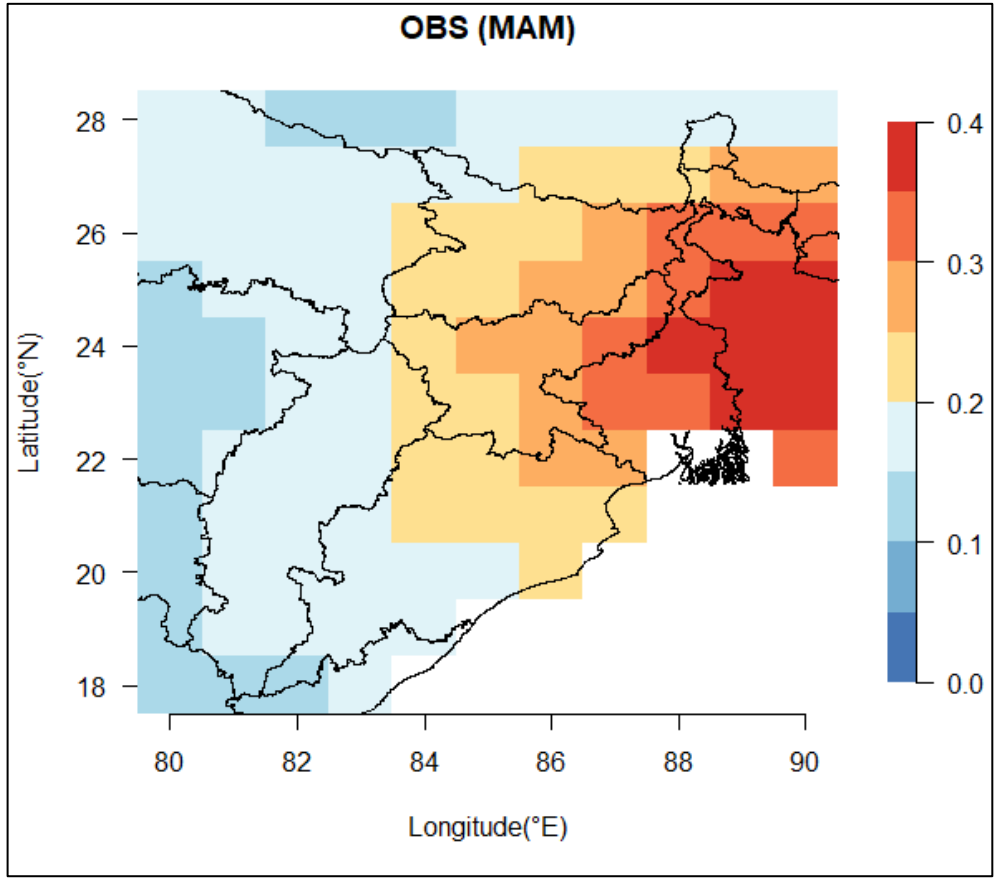
	RF	Spatial-RF
Moran's I	**0.14**	**0.06**
p value	**0**	**0**
No. of MEMs	-	**4**
R2	**0.80**	**0.92**
RMSE	**0.03**	**0.02**
MAE	**0.02**	**0.01**

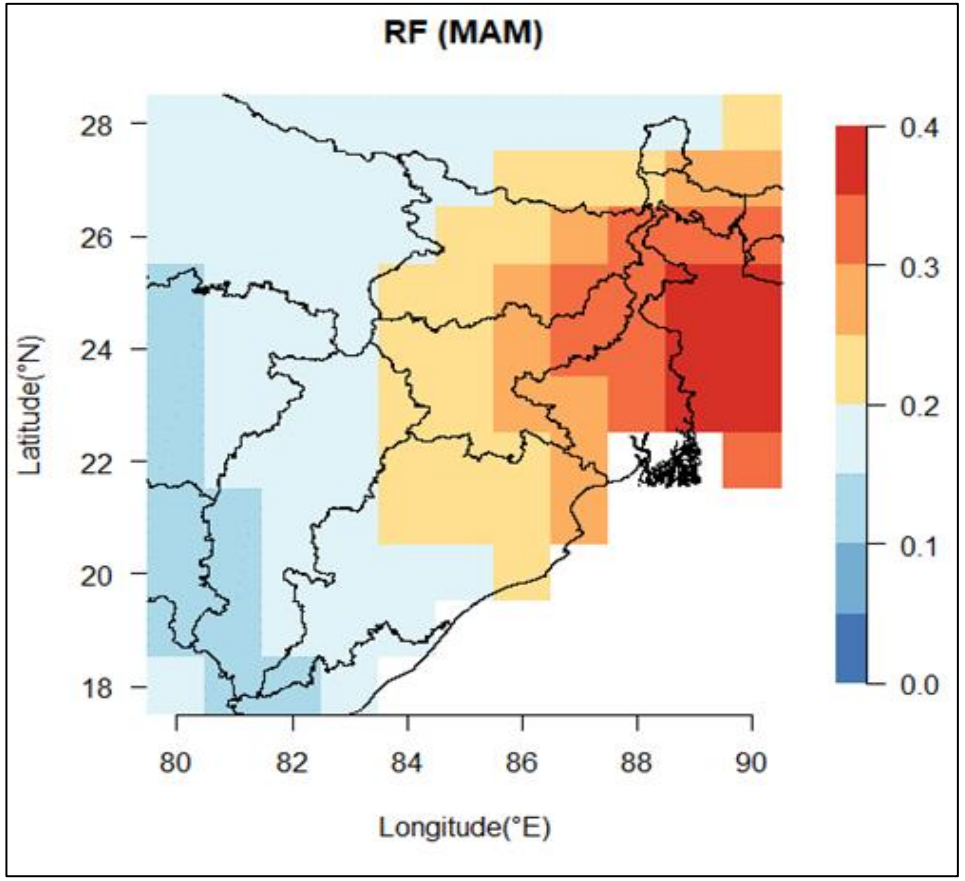


Spatial: mtry=4,ntree=1000
Non-spatial: mtry=4,ntree=1000

Results

Testing period: 2011-2013

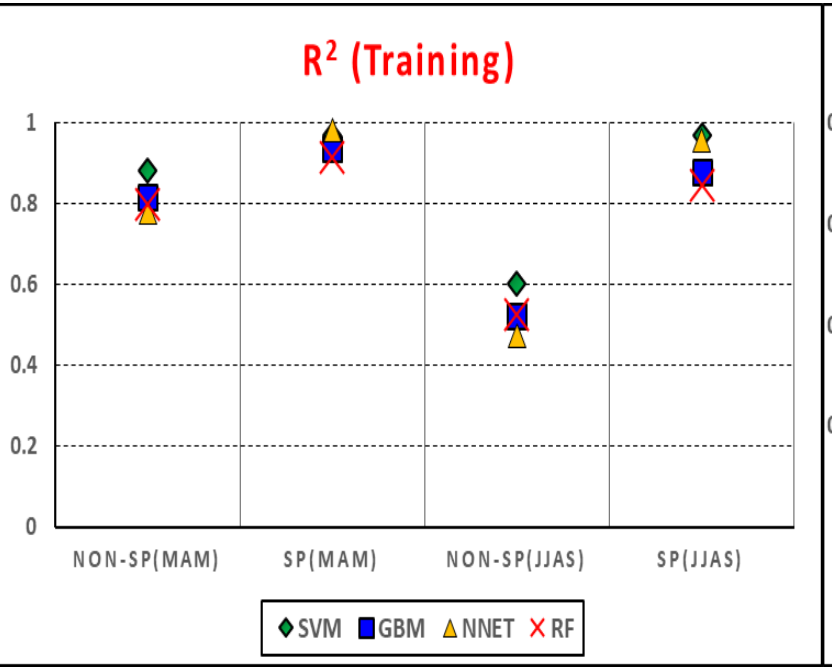


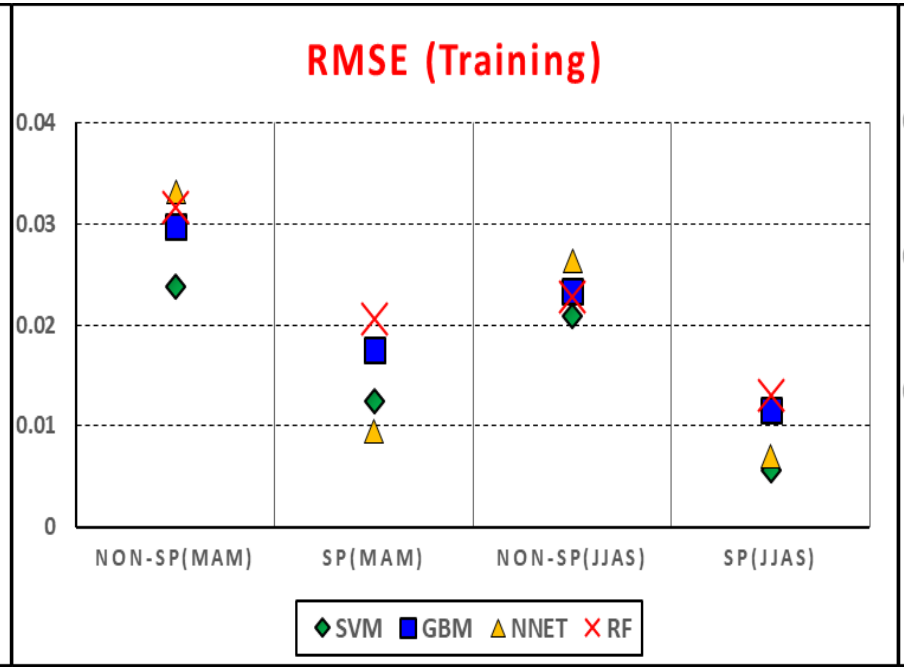


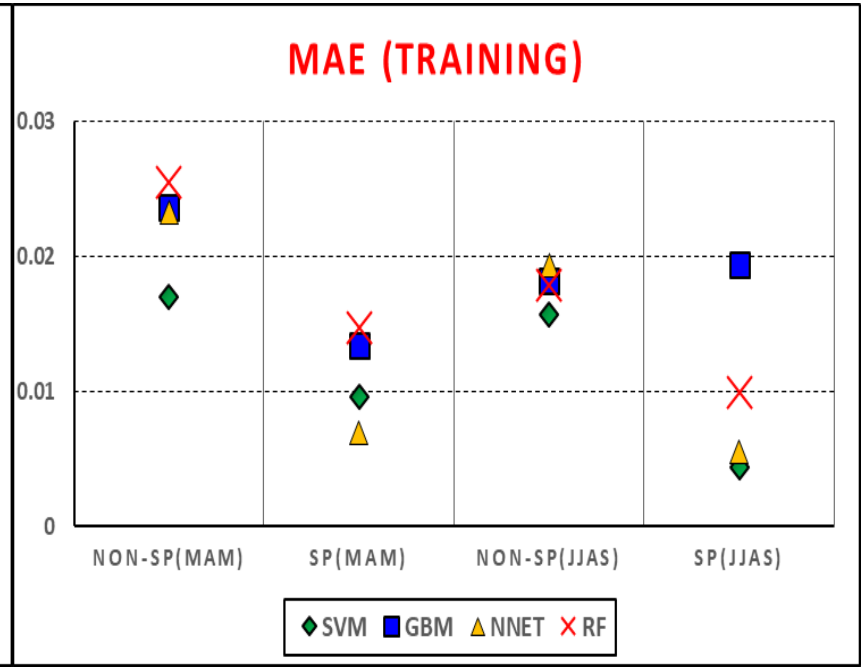
R	0.98
d	0.98
NSE	0.93
RSR	0.27

Results

Comparison among 4 models during two seasons:

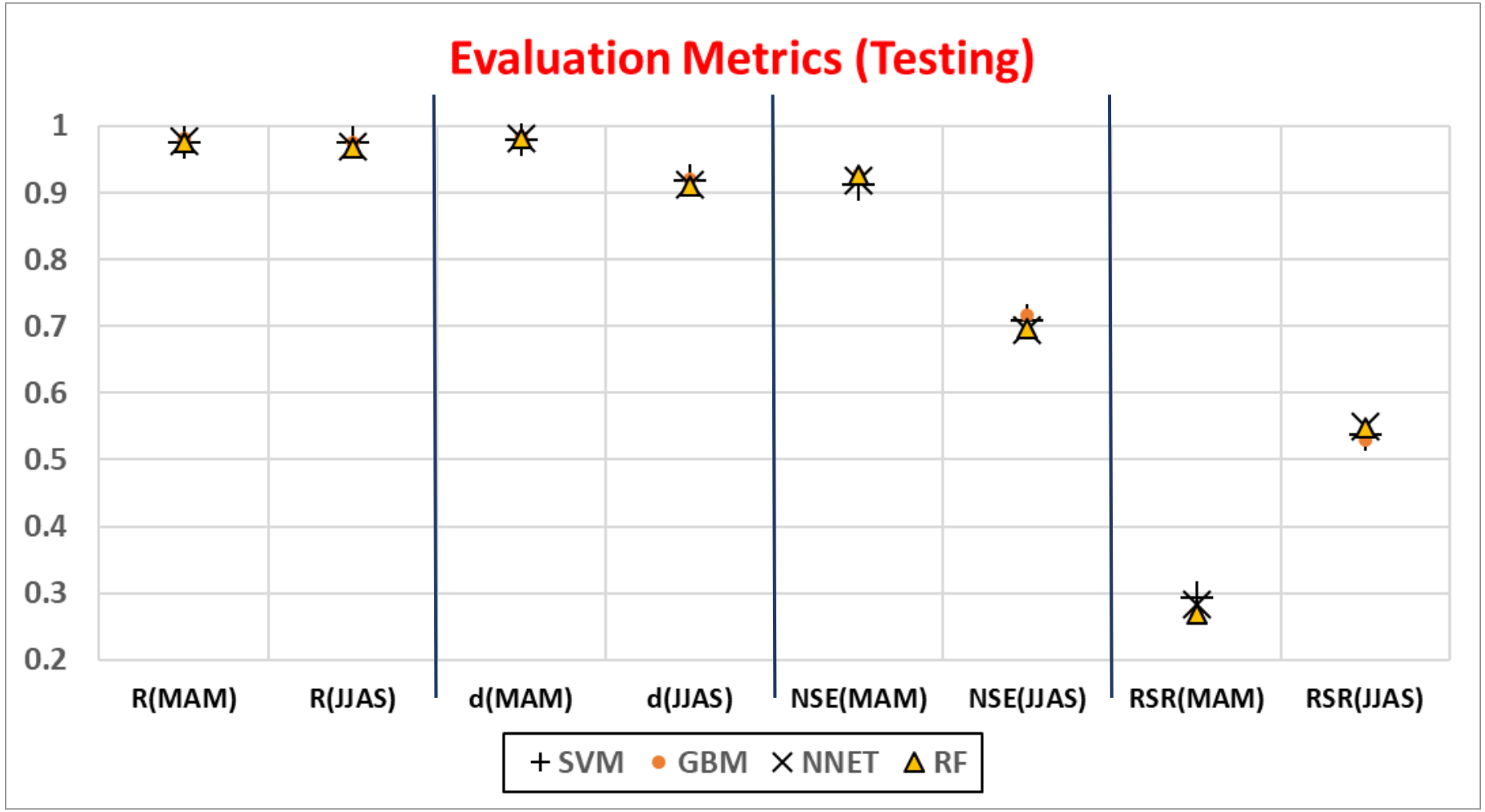






Results

Comparison among 4 models during two seasons:



Conclusions

- ML models have shown quite good fitting in pre-monsoon lightning. But, results are not satisfactory in monsoon during training period. However, error indices have shown comparatively lesser value during monsoon. Overall, SVM shown better performance than other three models in both seasons.
- **Residuals of Non-spatial ML models have shown significant spatial auto-correlation as suggested by Moran's I statistics. Incorporation of spatial predictors (MEMs) have reduced the spatial auto-correlation and made it statistically non-significant (except for RF).**
- Improved skills of Spatial Models have been found compared to their non-spatial counterparts during training period.
- **During testing period, spatial models have shown good agreement with observed lightning data. Better NSE and RSR values have been obtained in MAM compared to JJAS.**
- Spatial NNET model has been found to be best model in both seasons during both training and testing period.

References

- Lennon, J. J. 2000. Red-shifts and red herrings in geographical ecology. Ecography 23: 101-113.
- Anselin, L. 2002. Under the hood: issues in the specification and interpretation of spatial regression models. Agricult. Econ. 17: 247-267
- BealeC. M. et al. 2007. Red herring remain in geographical ecology: a reply to Hawkins et al. –Ecography 30: 845–847.
- Chun, Y.; Griffith, D.A.; Lee, M.; Sinha, P. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. J. Geogr. Syst. 2016, 18, 67–85.
- Dormann, C. F. 2007b. Effects of incorporating spatialautocorrelation into the analysis of species distribution data. Global Ecol. Biogeogr. 16: 129-138
- Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. “Spatial Modeling: A Comprehensive Framework for Principle Coordinate Analysis of Neighbor Matrices (PCNM).” Ecological Modelling 196: 483–93.
- Griffith, D.A. Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization; Springer: Berlin, Germany, 2003
- Griffith, D. A., and P. R. Peres-Neto. 2006. “Spatial Modeling in Ecology: The Flexibility of Eigenfunction Spatial Analyses.” Ecology 87: 2603–13.
- Illiyas F T, Mohan K, Mani S K and Pradeepkumar A P 2014 Lightning risk in India Challenges in disaster compensation; Economic and Political Weekly Jun7 23-27, https://www.jstor.org/stable/24479604
- Kuhn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. ! Div. Distrib. 13: 66-69.
- Singh O and Singh J 2015 Lightning fatalities over India 1979–2011; Meteorol. Appl. 22(4) 770-8, https://doi.org/10.1002/met.1520
- Yadava P K, Soni M, Verma S, Kumar H, Sharma A and Payra S 2020 The major lightning regions and associated casualties over India; Nat. Hazards 101(1) 217-29, https://doi.org/10.1007/s11069-020-03870-8