Identifying Driving Factors Behind Indian Monsoon Precipitation using Model Selection based on Data Depth

Subhabrata Majumdar

Lindsey Dietz

Snigdhansu Chatterjee

University of Minnesota, School of Statistics

Outline

- Introduction
- Methods
- Data and implementation
- Results
- Future work

Introduction: the problem

- Previous studies on Indian Monsoon precipitations have highlighted importance of using random effects in modelling (Dietz and Chatterjee, 2014 and 2015);
- No study has been done to collectively identify important factor influencing precipitation levels, i.e. model selection in this scenario;
- The reasons are non-robustness of likelihood based methods, strict assumptions of conventional methods of model selection, potentially heteroskedastic error structure

Introduction: our solution

- We introduce a novel one-step model selection technique for general regression estimators, and implement it in a linear mixed model setup to identify important predictors affecting Indian Monsoon precipitation;

- This technique correctly identifies the set of non-zero values in the true coefficient (of length p) by comparing only $p + 1$ models.
- Wild bootstrap used to estimate the selection criterion;
- Rolling validation done for 10 testing years for comparison of full and reduced models

Methods: Data depth

- The depth of a point $\boldsymbol{x} \in \mathbb{R}^p$, is any scalar measure of its centrality with respect to a data cloud $\mathbf{X}$ (or equivalently the underlying distribution *F*) (Zuo and Serfling, 2000)

Depth-based model selection

- Assume estimates $\widehat{\boldsymbol{\beta}}_n$ of a coefficient vector $\boldsymbol{\beta}$ which follow asymptotically elliptical sampling distributions F_n that approach unit mass at $\boldsymbol{\beta}$ as $n \to \infty$;
- Example: in multiple linear regression, $F_n \equiv N(\boldsymbol{\beta}, (\mathrm{X}'\mathrm{X})^{-1})$;
- Specify a candidate model by α: the set of non-zero indices;

- The model selection criterion is defined as:

$$C_n(\alpha) = \mathrm{E}\left[D\left(\widetilde{\boldsymbol{\beta}}_\alpha, F_n\right)\right]$$

Where $\widetilde{\boldsymbol{\beta}}_\alpha$ is the estimate of truncated coefficient $\boldsymbol{\beta}_\alpha$ concatenated with 0 at indices not in α.

Depth-based model selection

- When α contains all non-zero indices in the true model, we have $C_n(\alpha) = C(\alpha)$ for any n. Also it maximizes at smallest correct model, say α_0, and decreases monotonically as zero indices are added to α_0 one-by-one;

- Otherwise $C_n(\alpha) \to 0$ as $n \to \infty$ (Majumdar and Chatterjee, 2015+);

- C_n is estimated in a general sample setup by bootstrap.

One-step model selection using C_n

- Calculate C_n for full model;
- Drop a predictor, calculate C_n for the reduced model;
- Repeat for all p predictors;
- Collect predictors dropping which causes C_n to decrease. These are the very predictors in the smallest correct model.

Methods: Linear Mixed Model

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

- $\mathbf{Y}_{n\times 1}$ is the vector of response, $X_{n\times p}$ the matrix of predictors, $\boldsymbol{\beta}_{p\times 1}$ the coefficient vector;
- $Z_{n\times k}$ is random effect design matrix; $\boldsymbol{\gamma}_{k\times 1}$ random effect vector;
- $\boldsymbol{\gamma} \sim N(\mathbf{0}, \Sigma)$ with Σ positive-definite;

Data

- Daily rainfall levels from 36 weather stations during 1978-2012;
- Station-specific variables: latitude, longitude and elevation;
- Local variables: max and min temperature, tropospheric temperature difference, *u*- and *v*- winds at 200, 600 and 850 mb, Nino 3.4 anomaly and Indian Dipole Mode Index (DMI);
- Global variables: 10 indices of Madden-Julian Oscillation (MJO), 9 northern hemisphere teleconnection indices, solar flux levels and land-ocean Temperature Anomaly (TA);

- **Total 35 predictors**

Implementation

- For all variables, station-wise median taken for each year;
- All predictors as fixed effects, random intercept due to year;

- Bootstrap scheme- separate wild bootstraps on estimated random effects and residuals:

$$\mathbf{Y}_b = X\widehat{\boldsymbol{\beta}} + ZV_b\widehat{\boldsymbol{\gamma}} + U_b\widehat{\boldsymbol{\varepsilon}}$$

Where $U_b = \mathrm{diag}(u_{1b}, \ldots, u_{nb}) \sim N(0, n^{0.2})$ iid,

$V_b = \mathrm{diag}(v_{1b}, \ldots, v_{kb}) \sim N(0, k^{0.2})$ iid

Results

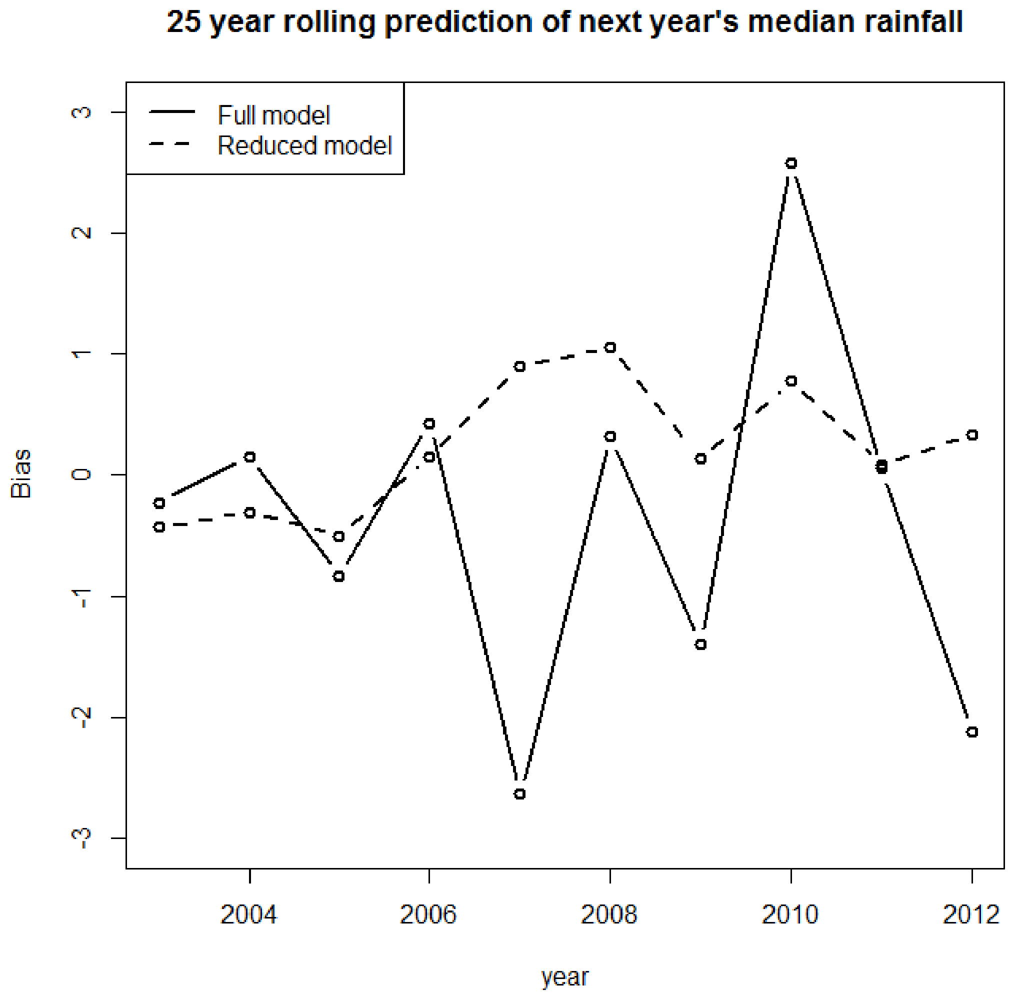
- Among 35 predictors considered, 21 get selected by our model selection procedure;
- All selected variables have documented effect on Indian Monsoon, e.g. elevation, longitude, max temperature, Nino3.4 etc;
- Temperature Anomaly (TA) has a large influence: several MJO indices are selected when we start from a full model with everything but TA, but get masked and are dropped in favor of TA when it is in the full model

Prediction

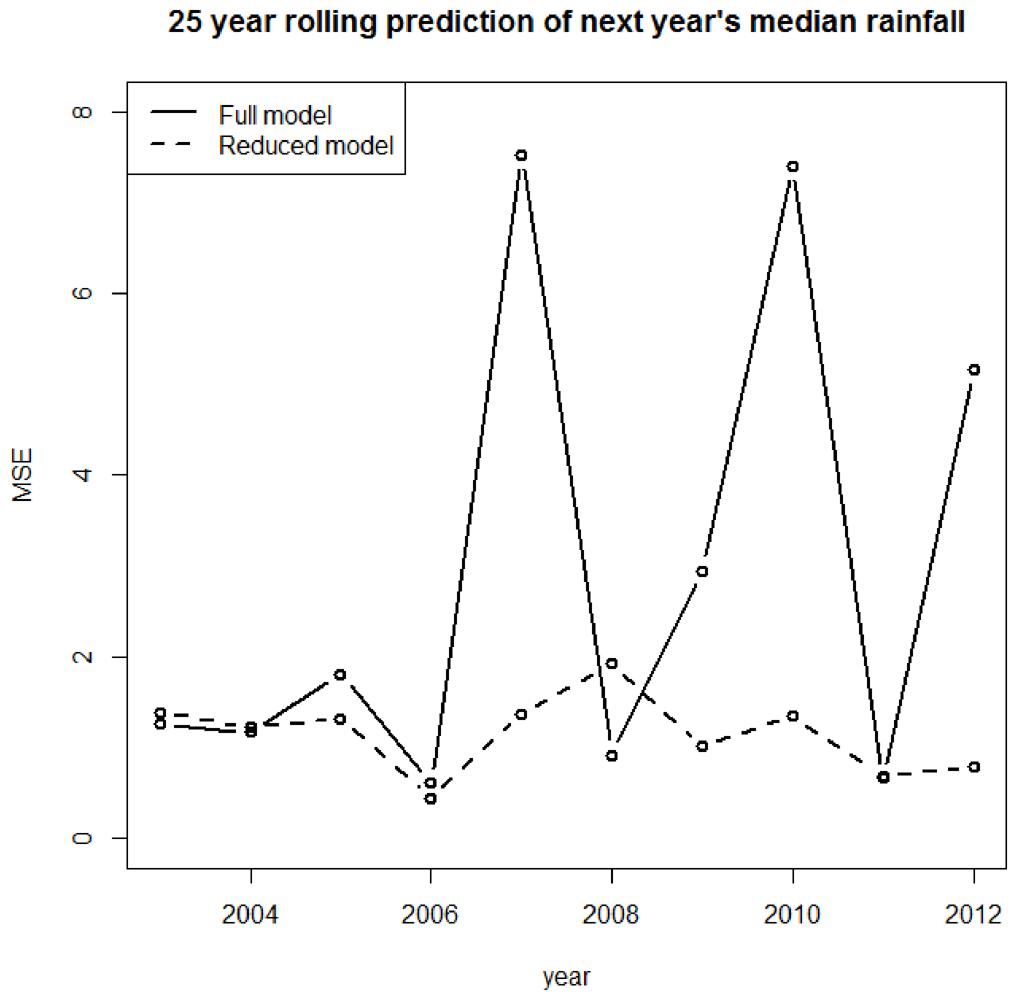
- Testing years 2003 to 2012

- 25 year rolling validation scheme: train using past 25 yrs data (e.g. 1978-2002 for 2003, 1979-22003 for 2004, …);

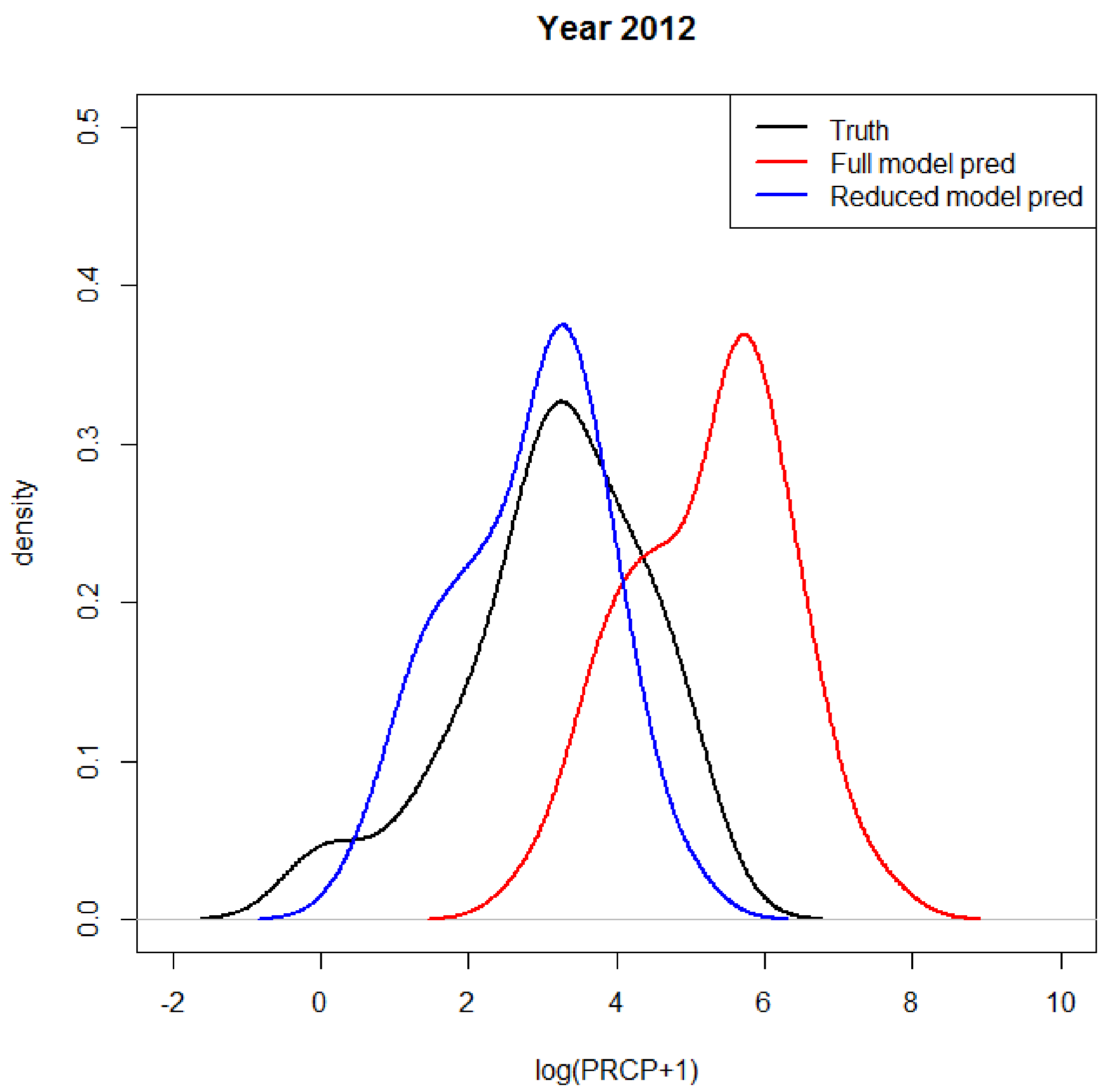
- Reduced models predictions have less bias across training years compared to full model predictions, as well as less mean squared error (MSE);

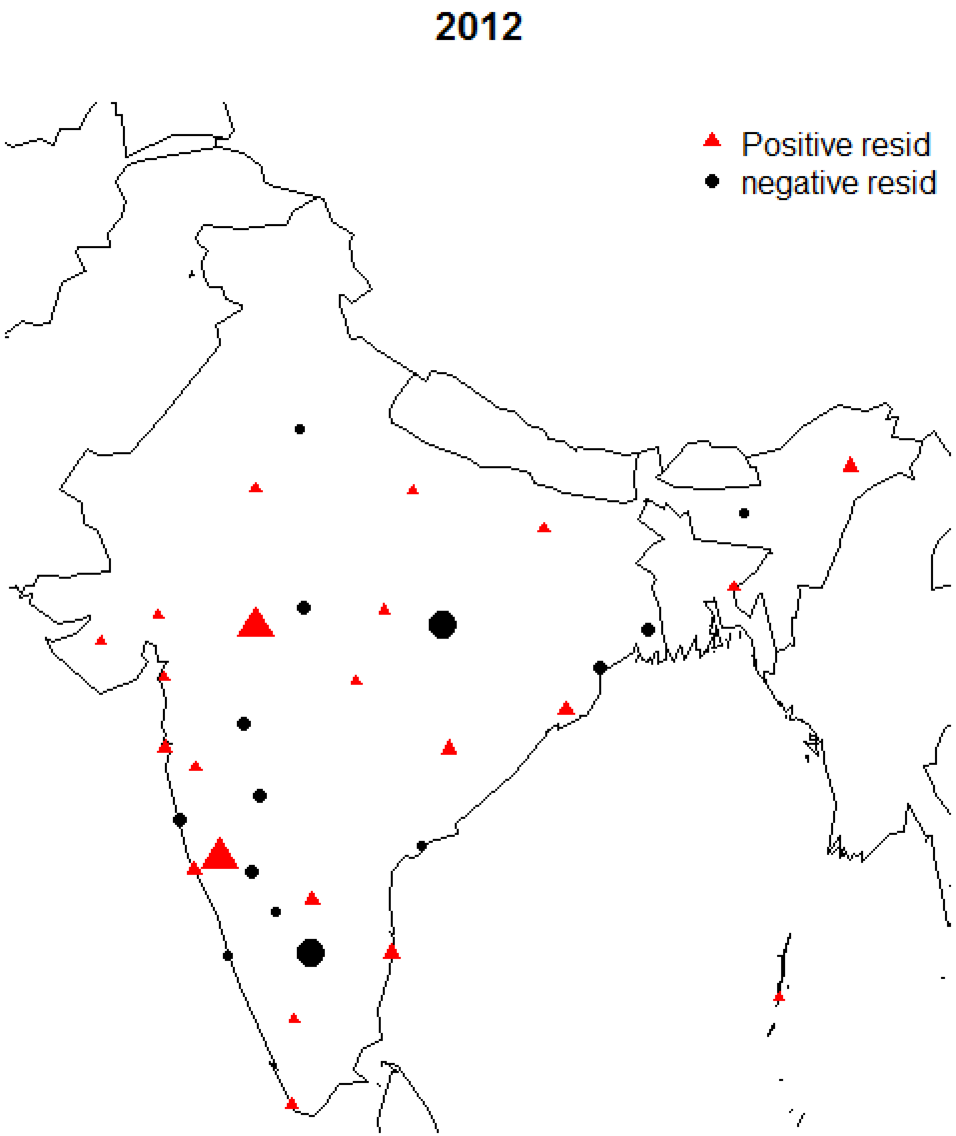


Bias comparison



MSE comparison





2012 predictions

Future work

- Investigating spatio-temporal dependency patterns;
- Detailed studies into algorithmic efficiency issues and different bootstrap schemes;
- Further development of theoretical properties of the proposed model selection tool

References

1. L. R. Dietz and S. Chatterjee, "Logit-normal mixed model for Indian monsoon precipitation," *Nonlin. Processes Geophys.*, vol. 21, pp. 939–953, sep 2014;

2. L. R. Dietz and S. Chatterjee, "Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models," in *Machine Learning and Data Mining Approaches to Climate Science: Proceedings of the Fourth International Workshop on Climate Informatics* (V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, eds.), pp. 239–246, Springer, 2015;

3. Y. Zuo and R. Serfling, "General notions of statistical depth function," *Ann. of Statist.*, vol. 28-2, pp. 461–482, 2000;

4. S. Majumdar and S. Chatterjee, "A model selection criterion for regression estimators based on data depth." Working paper, 2015+.