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

Subhabrata Majumdar (majum010@umn.edu), Lindsey Dietz (diet0146@umn.edu), and Snigdhansu Chatterjee (chatterjee@stat.umn.edu)
University of Minnesota Twin Cities, School of Statistics

Introduction

Objective: Selection of important predictors behind Indian Monsoon rainfall, and using them to build a predictive model.

Challenges for covariate selection:

- Several sources of variability, e.g. variation across years and weather station;
- Potentially heteroskedastic error structure;
- Linearity or other regression assumptions are not guaranteed to hold and are hard to verify;
- Huge number of possible models (2^p) for even moderate number of predictors (p).

Our solution: Use a novel model selection criterion based on data depth that works on a wide range of models, and **selects important predictors by comparing only $p+1$ models**.

Data and modelling

Annual median observations for 1978-2012;

Fixed covariates **($\mathbf{x}_{p\times 1}, p = 35$)**

(A) Station-specific: (from 36 weather stations across India) Latitude, longitude, elevation, maximum and minimum temperature, tropospheric temperature difference (ΔTT), Indian Dipole Mode Index (DMI), Niño 3.4 anomaly;

(B) Global:

- u-wind and v wind at 200, 600 and 850 mb;
- 10 indices of Madden-Julian Oscillations: 20E, 70E, 80E, 100E, 120E, 140E, 160E, 120W, 40W, 10W;
- Teleconnections: North Atlantic Oscillation (NAO), East Atlantic (EA), West Pacific (WP), East Pacific-North Pacific (EPNP), Pacific/North American (PNA), East Atlantic/Western Russia (EAWR), Scandinavia (SCA), Tropical/Northern Hemisphere (TNH), Polar/Eurasia (POL);
- Solar Flux;
- Land-Ocean Temperature Anomaly (TA).

Random effects **($\boldsymbol{\Gamma}$)**: Random intercept by year;

Linear Mixed Model (LMM):

Y: log of annual median rainfall at a weather station (WS);

$$\text{Level 1:}\quad Y_{\text{WS,year}}|\boldsymbol{\Gamma} = \boldsymbol{\gamma} \overset{\text{ind}}{\sim} N(\theta_{\text{WS,year}}, \sigma^2);$$

$$\theta_{\text{WS,year}} = \mathbf{x}^T_{\text{WS,year}}\boldsymbol{\beta} + \gamma_{\text{year}};$$

$$\text{Level 2:}\quad \Gamma_{\text{year}} \overset{\text{i.i.d}}{\sim} N(0, \tau^2)$$

Depth-based model selection

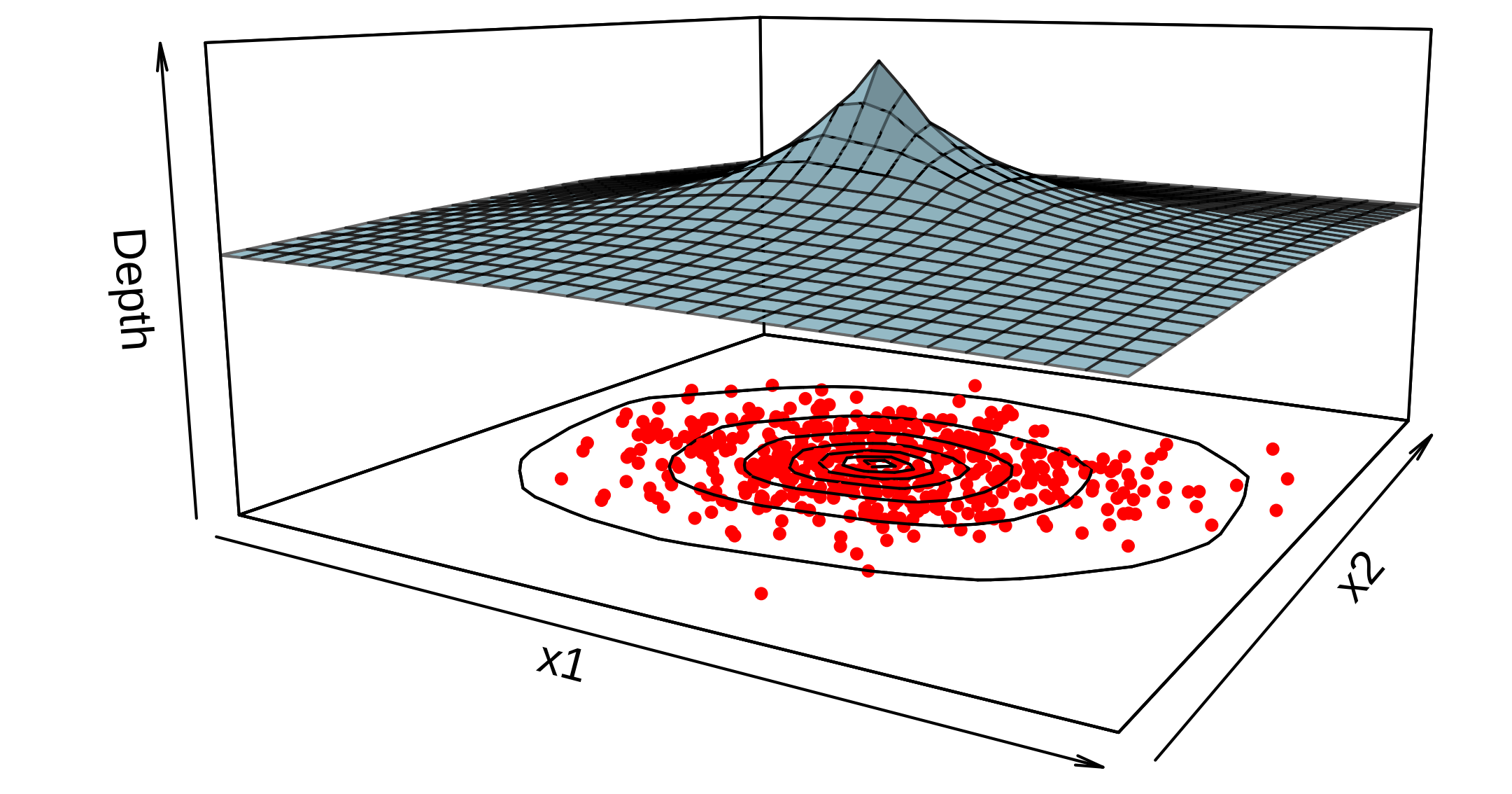

Figure 1: Samples from bivariate normal and their depths: points away from center have less depth while those close to center have more depth

Data depth

A nonparametric, scalar measure of centrality for a point $\mathbf{x}$ in sample space with respect to a data cloud $\mathbf{X}$ or probability distribution F: denoted by $D(\mathbf{x}, \mathbf{X})$ or $D(\mathbf{x}, F)$, respectively [Zuo and Serfling, 2000].

The selection criterion

In any regression setup, consider estimators of the coefficient $\boldsymbol{\beta}$ based on a sample of size n having elliptical sampling distributions F_n centered at $\boldsymbol{\beta}$ that approach unit mass at $\boldsymbol{\beta}$ as $n \to \infty$.

For a candidate model, uniquely specified by its non-zero index set α, define

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

where $\tilde{\boldsymbol{\beta}}_\alpha$ is estimate of truncated coefficient vector $\boldsymbol{\beta}_\alpha$, concatenated with 0 at indices not in α.

Suppose α_0 is the smallest correct model. Then:

- For any correct model, i.e. when $\alpha \supseteq \alpha_0$, we have $C_n(\alpha) = C(\alpha)$, i.e. depends only on α;
- For any wrong model, $C_n(\alpha) \to 0$ as $n \to \infty$;
- Among correct models, $C(\alpha)$ maximizes at $\alpha = \alpha_0$, and decreases monotonically as superfluous variables are added;
- In a sample setup, we use bootstrap to estimate $\tilde{\boldsymbol{\beta}}_\alpha$ and F_n [Majumdar and Chatterjee, 2015+].

The one-step algorithm

1. **For large enough n, Calculate C_n for full model;**
2. **Drop a predictor, calculate C_n for the reduced model;**
3. **Repeat for all p predictors;**
4. **Collect predictors dropping which causes C_n to decrease. These are the predictors in the smallest correct model.**

Implementation

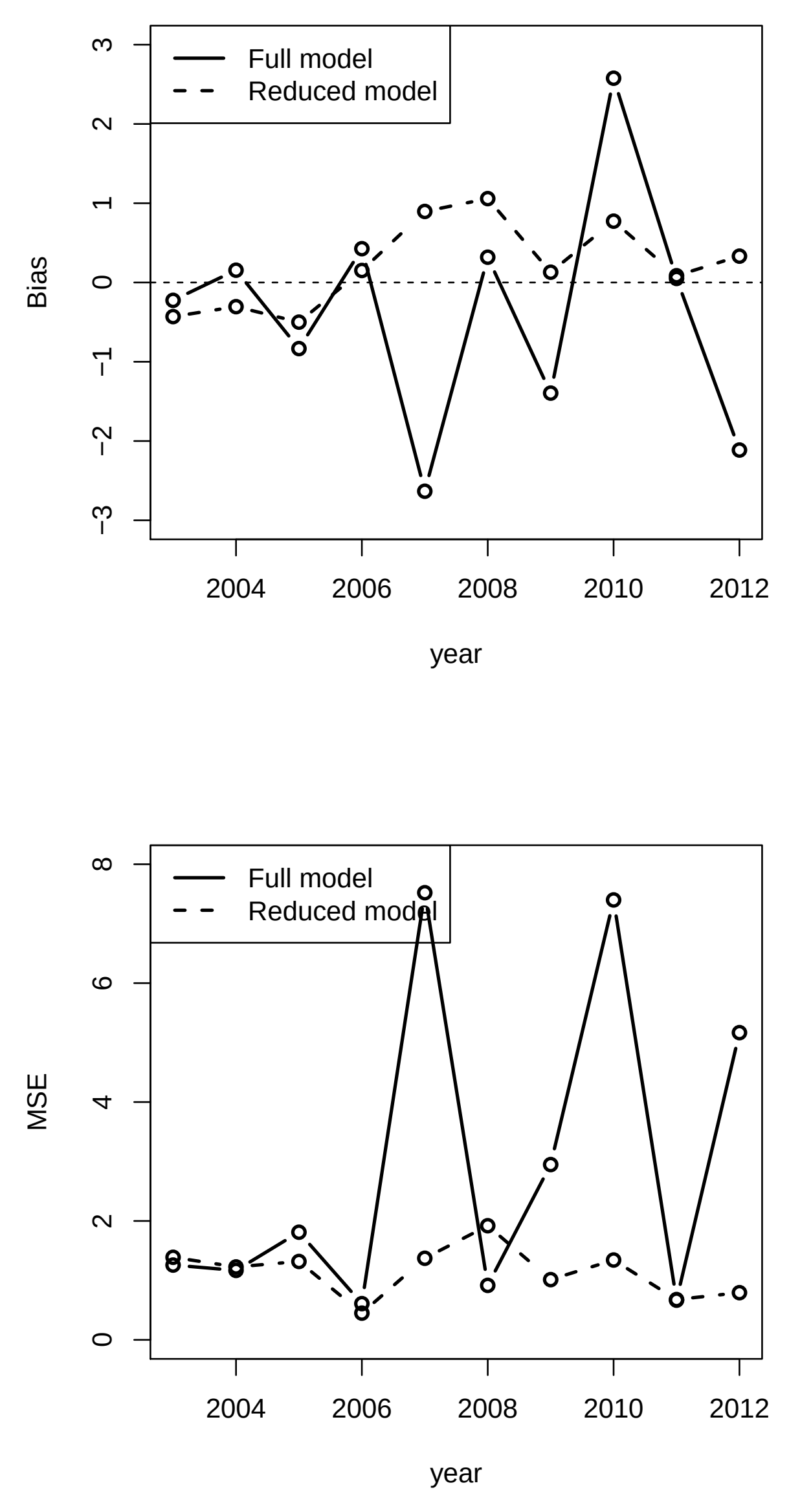

Figure 2: Bias and MSE of rolling predictions

Bootstrap scheme

- Wild bootstrap [Mammen, 1993];
- Say n is total number of observations, k is number of years;
- Start with estimators from initial LMM: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\epsilon}}$;
- Generate $u^b_{\text{WS,year}} \overset{\text{i.i.d}}{\sim} N(0, n^{0.2}), v^b_{\text{year}} \overset{\text{i.i.d}}{\sim} N(0, k^{0.2})$;
- Get 'new' observations: $Y^b_{\text{WS, year}} = \mathbf{x}^T_{\text{WS, year}}\hat{\boldsymbol{\beta}} + v^b_{\text{year}}\hat{\gamma}_{\text{year}} + u^b_{\text{WS, year}}\hat{\epsilon}_{\text{WS, year}}$

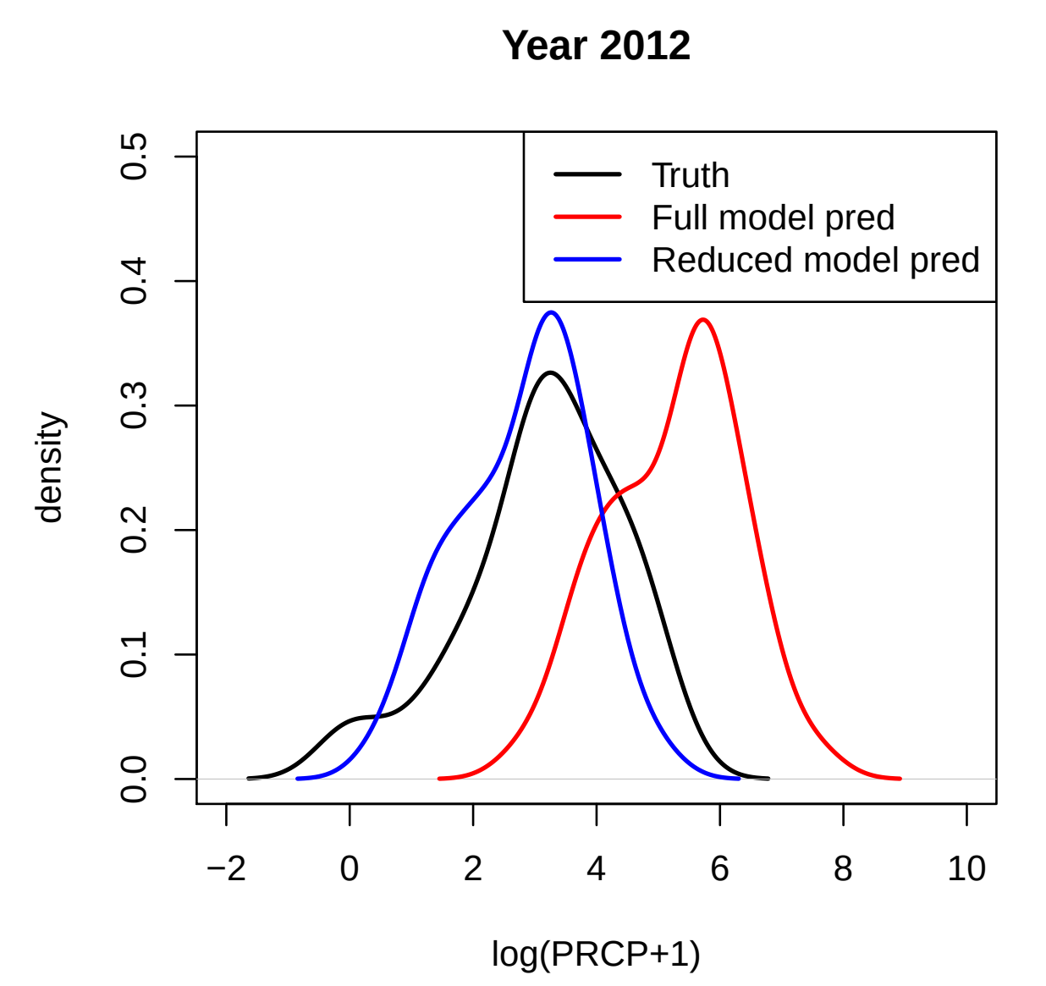

Figure 3: Density plots of 2012 predictions and truth

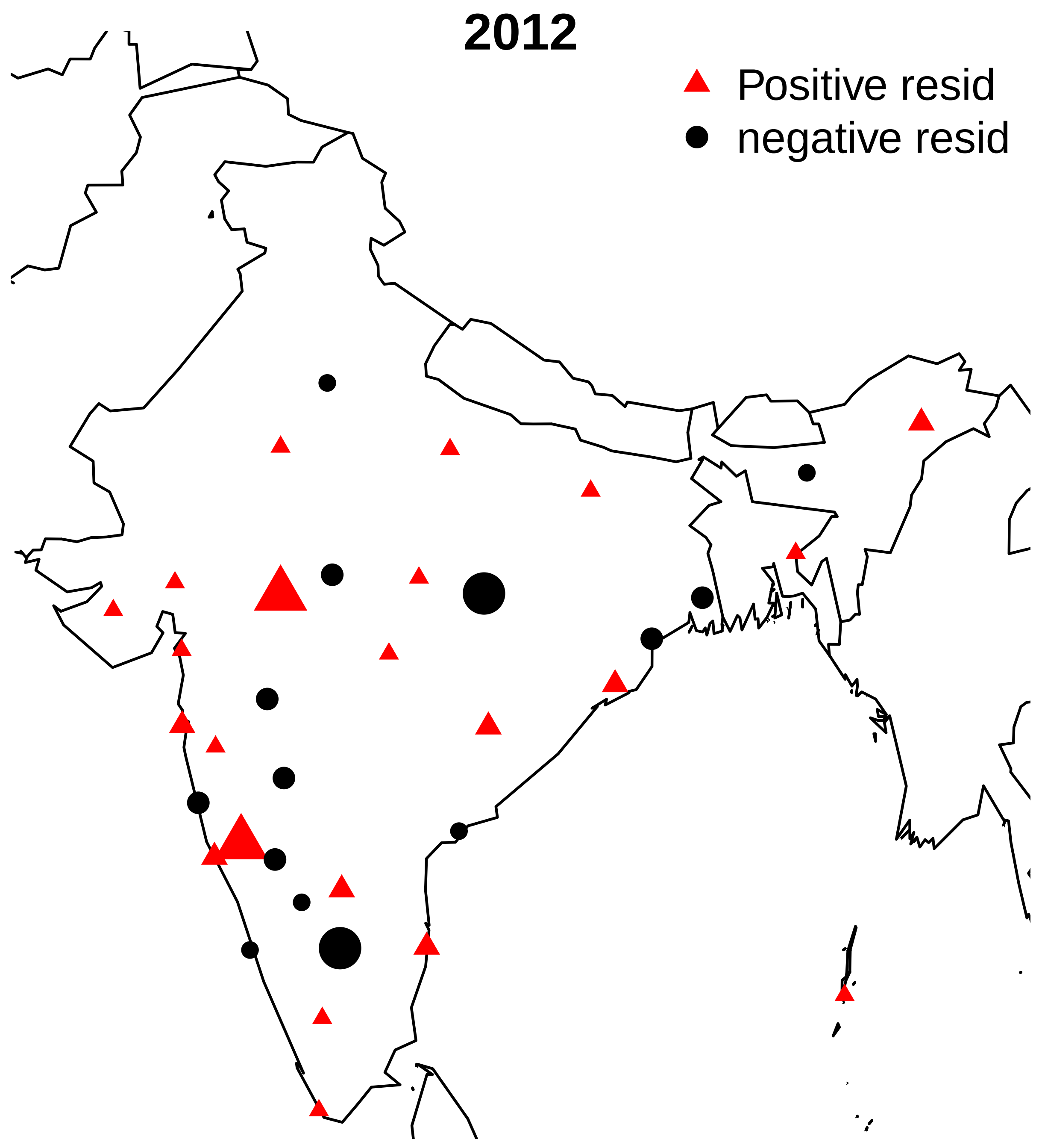

Figure 4: Stationwise residuals for 2012

Dropped var	C_n
Max temp	0.1311765
Elevation	0.156789
Temp Anomaly	0.1744005
ΔTT	0.2107309
Niño34	0.2407754
v-wind 850 mb	0.2409109
POL	0.2434999
Solar Flux	0.2437457
EPNP	0.2440997
MJO 120W	0.2443009
Longitude	0.2453031
u-wind 850 mb	0.246112
TNH	0.2471282
EA	0.2477579
u-wind 600 mb	0.2479662
Latitude	0.2503148
u-wind 200 mb	0.2506321
NAO	0.2519853
DMI	0.2520626
MJO 20E	0.2523445
<None>	0.2532943
v-wind 200 mb	0.2537721
EAWR	0.254549
v-wind 600 mb	0.255402
WP	0.2557333
Min temp	0.2558768
MJO 160E	0.2559359
PNA	0.2569771
MJO 140E	0.2580453
MJO 120E	0.2615192
SCA	0.2616763
MJO 40W	0.2623092
MJO 70E	0.2630892
MJO 100E	0.2648561
MJO 10W	0.2655411
MJO 80E	0.2732275

Table 1: Ordered values of C_n from bootstrap

Discussion

- All selected variables (marked in blue in Table 1) have documented effects on Indian monsoon;
- EPNP teleconnection and 120W MJO are both selected: both deal with same longitudinal region;
- Interesting variables: Solar Flux and Polar/Eurasia teleconnection (POL): an indicator of Eurasian snow cover;
- TA has a large influence. Several MJO indices, particularly 80E and 40W, are selected when starting from a full model with everything but TA, but are dropped in favor of TA when it is included in the full model;
- Reduced model predictions have consistently less bias and are more stable across testing years (Figs. 2 and 3). Also there are no spatial patterns in residuals (Fig. 4).

GitHub: `https://github.com/shubhobm/Climate-indian-monsoon`

Acknowledgement: NSF grant IIS-1029711;

References:

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

E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.*, 21-1:255–285, 1993.

Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Statist.*, 28-2:461–482, 2000.