Functional Median Polish, with Climate Applications

Marc G. Genton

Department of Statistics, Texas A&M University

Program in Spatial Statistics
(stat.tamu.edu/pss)

Institute for Applied Mathematics and Computational Sciences
(iamcs.tamu.edu)

Based on joint work with Ying Sun

May 11, 2012

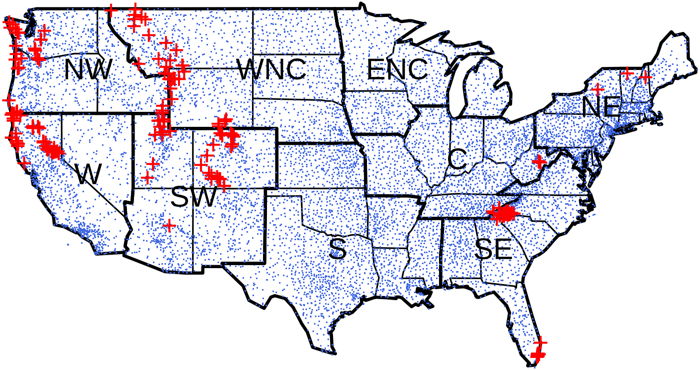
Functional Median Polish

Observations and Climate Models

- Observations:
  - provide a corroborating source of information about physical processes being modeled.
  - have methodological and practical issues due to uncertainties.
- Climate Models:
  - numerically solve systems of differential equations representing physical relationships in the climate system.
  - have huge uncertainties and biases.
- Scientific Questions:
  - How do we compare sources of variability in observations or climate model outputs? i.e. quantification of uncertainties?

Spatio-Temporal Precipitation Data

- Spatio-temporal precipitation data: annual total precipitation data for U.S. from 1895 to 1997 at 11,918 weather stations.
- Nine climatic regions for precipitation defined by National Climatic Data Center.
- Several areas of outliers detected by Sun and Genton (2011, 2012).



Analysis of Variance

- Analysis of Variance (ANOVA):
  - An important technique for analyzing the effect of categorical factors on a response.
  - It decomposes the variability in the response variable among the different factors.
  - A two-way additive model: for $i = 1, \dots, r$, $j = 1, \dots, c$,

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}.$$

- The ANOVA model can be fitted by arithmetic means (no outliers), or medians (robust).

ANOVA Model Fitting

- Fitted by means:
  - $\hat{\mu} = \bar{y}$ (grand effect),
  - $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}$ (row effect),
  - $\hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}$ (column effect).
- Fitted by medians:
  - Median polish (Tukey, 1970, 1977).
  - An iterative technique for extracting row and column effects in a two-way table using medians rather than means.
  - It stops when no more changes occur in the row and column effects, or changes are sufficiently small.

Median Polish Example

1. Original table: find row medians.
2. 1st iteration: subtract row medians, find column medians. Grand median in red, row effects in blue, column effects in green.
3. 2nd iteration: subtract column medians, find row medians,

$$
\begin{array}{ccc|c}
6 & 3 & 11 & 6 \\
3 & 2 & 4 & 3 \\
9 & 0 & 0 & 0
\end{array}
\rightarrow
\begin{array}{ccc|c}
0 & -3 & 5 & 6 \\
0 & -1 & 1 & 3 \\
9 & 0 & 0 & 0 \\
\hline
0 & -1 & 1 & \color{red}{3}
\end{array}
\rightarrow
\begin{array}{ccc|c||c}
0 & -2 & 4 & 0 & \color{blue}{3} \\
0 & 0 & 0 & 0 & \color{blue}{0} \\
9 & 1 & -1 & 1 & \color{blue}{-3} \\
\hline
0 & -1 & 1 & 0 & \color{red}{3}
\end{array}
\rightarrow
$$

4. subtract new row medians, add their medians to the grand median, find column medians.
5. Polished table: new row and column medians are zero after two iterations.

$$
\begin{array}{ccc||c}
0 & -2 & 4 & \color{blue}{3} \\
0 & 0 & 0 & \color{blue}{0} \\
8 & 0 & -2 & \color{blue}{-3} \\
\hline
0 & 0 & 0 & 0 \\
\hline
\color{green}{0} & \color{green}{-1} & \color{green}{1} & \color{red}{3+0}
\end{array}
\rightarrow
\begin{array}{ccc||c}
0 & -2 & 4 & \color{blue}{3} \\
0 & 0 & 0 & \color{blue}{0} \\
8 & 0 & -2 & \color{blue}{-3} \\
\hline
\color{green}{0} & \color{green}{-1} & \color{green}{1} & \color{red}{3}
\end{array}
$$

Functional Median Polish

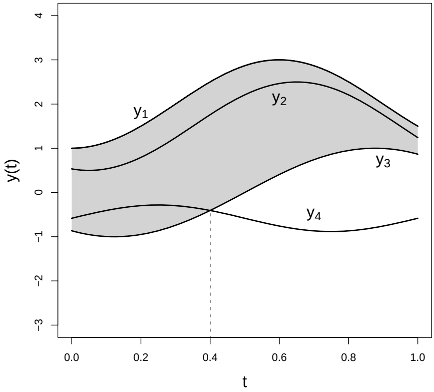
- Observe functional data at each combination of two categorical factors.
- Examine their effects: functional row or column effects.
- $y_{ijk}(x) = \mu(x) + \alpha_i(x) + \beta_j(x) + \epsilon_{ijk}(x)$, where $i = 1, \ldots, r$, $j = 1, \ldots, c$, $k = 1, \ldots, m_{ij}$.
- Constraints: $\text{median}_i\{\alpha_i(x)\} = 0$, $\text{median}_j\{\beta_j(x)\} = 0$ and $\text{median}_i\{\epsilon_{ijk}(x)\} = \text{median}_j\{\epsilon_{ijk}(x)\} = 0$ for all k.
- x can be time for curves or spatial index for surfaces/images.
- Iterative procedure sweeping out column and row medians.
- One-way functional ANOVA can be done in a similar way.
- Need to order functional data.

Multivariate Ordering

- Basic ideas of depth in functional context
  - provides a method to order sample curves according to decreasing depth values,
  - $y_{[1]}$: the deepest (most central or median) curve,
  - $y_{[n]}$: the most outlying (least representative) curve,
  - $y_{[1]}, \ldots, y_{[n]}$: start from the center outwards.
- Usual order statistics: ordered from the smallest sample value to the largest.

Band Depth for Functional Data

- López-Pintado and Romo (2009) introduced the band depth (BD) concept through a graph-based approach.
- Grey area: band determined by two curves, y_1 and y_3.
- Contains the curve y_2, but does not contain y_4.



Band Depth for Functional Data

- Population version of $BD^{(2)}$:

$$BD^{(2)}(y, P) = P\{G(y) \subset B(Y_1, Y_2)\}.$$

  - $G(y)$: graph of the curve y,
  - $B(Y_1, Y_2)$: band delimited by 2 random curves.

- The band could be delimited by more than 2 random curves,

$$BD_J(y, P) = \sum_{j=2}^{J} BD^{(j)}(y, P).$$

Sample Band Depth

- Population level: $BD^{(j)}(y, P)$ is a probability.
- Sample version of $BD^{(j)}(y, P)$

$$BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\},$$

  - $I\{\cdot\}$: the indicator function,
  - fraction of the bands completely containing the curve y.
- Sample BD: $BD_{n,J}(y) = \sum_{j=2}^{J} BD_n^{(j)}(y)$.

Modified Band Depth

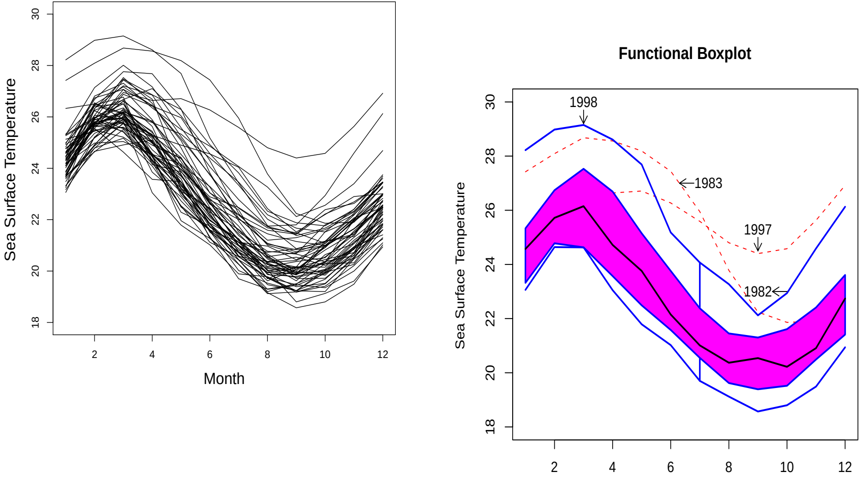
- López-Pintado and Romo (2009) also proposed a more flexible definition, the modified band depth (MBD).

$$BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\},$$

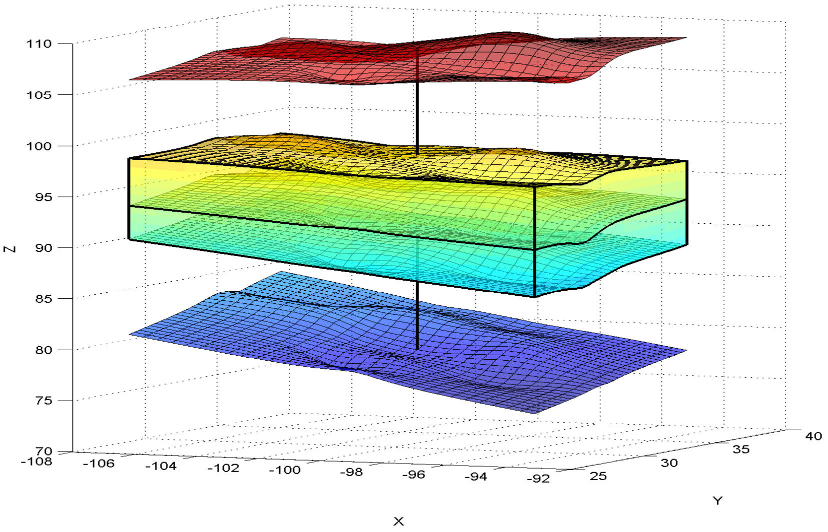
$$MBD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} \lambda_r\{A(y; y_{i_1}, \ldots, y_{i_j})\}.$$

- $\lambda_r\{A(y; y_{i_1}, \ldots, y_{i_j})\}$ measures the proportion of time that a curve y is in the band.

Functional Boxplots (Sun and Genton, 2011, 2012)

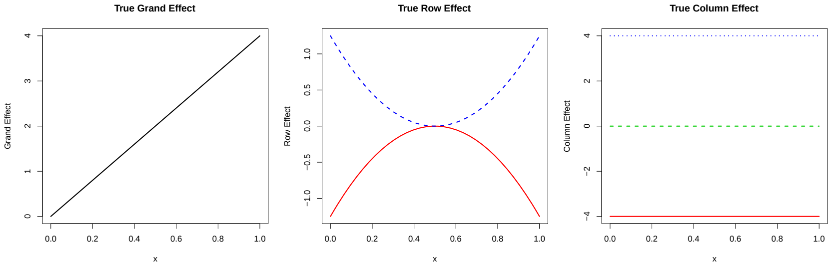


Surface Boxplot



True Model

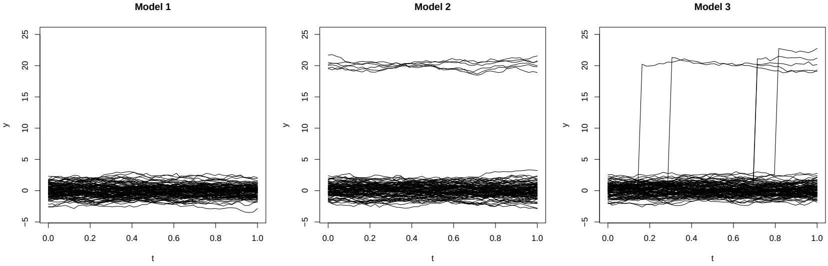
- Generate data from a true model with $r = 2$, $c = 3$, and $m = 100$ curves in each cell at $p = 50$ time points.



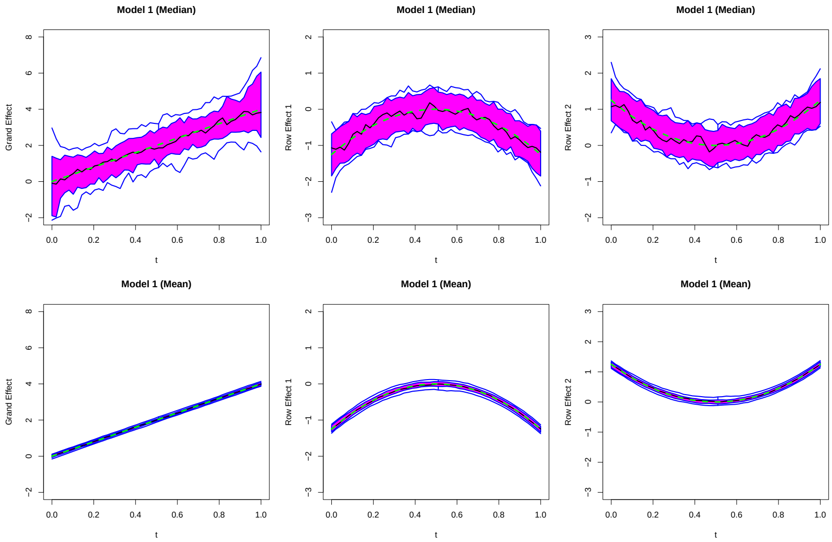
- Introduce outliers through a Gaussian process $\epsilon_{ijk}(t)$.
- Replications: 1,000.

Outlier Models

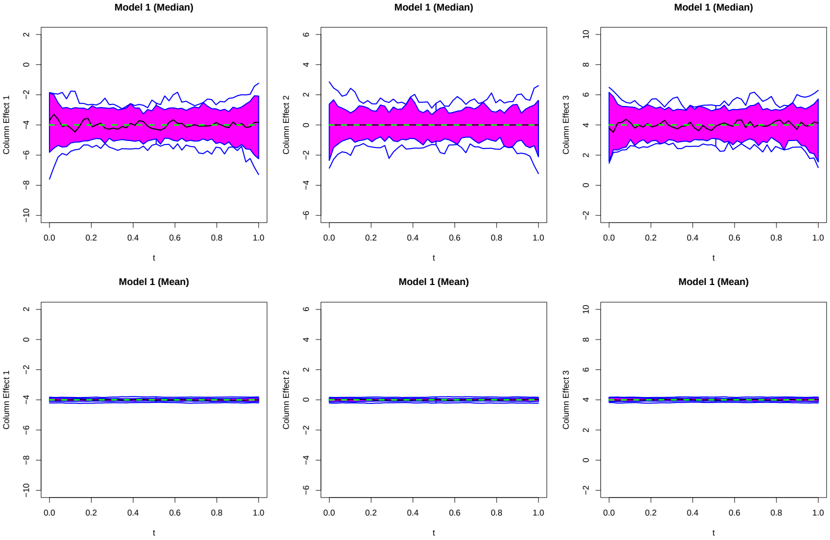
- Model 1: $\epsilon_{ijk}(t) = e_{ijk}(t)$, where $e_{ijk}(t) \sim GP(0, \gamma)$ with $\gamma(t_1, t_2) = \exp\{-|t_2 - t_1|\}$.
- Model 2: $\epsilon_{ijk}(t) = e_{ijk}(t) + c_{ijk}K$, where c_{ijk} is 1 with prob q_{ij} and 0 with prob $1 - q_{ij}$, q_{ij} is different for each cell.
- Model 3: $\epsilon_{ijk}(t) = e_{ijk}(t) + c_{ijk}K$, if $t \geq T_{ijk}$ and $\epsilon_{ijk}(t) = e_{ijk}(t)$, if $t < T_{ijk}$, where $T_{ijk} \sim U(0, 1)$.



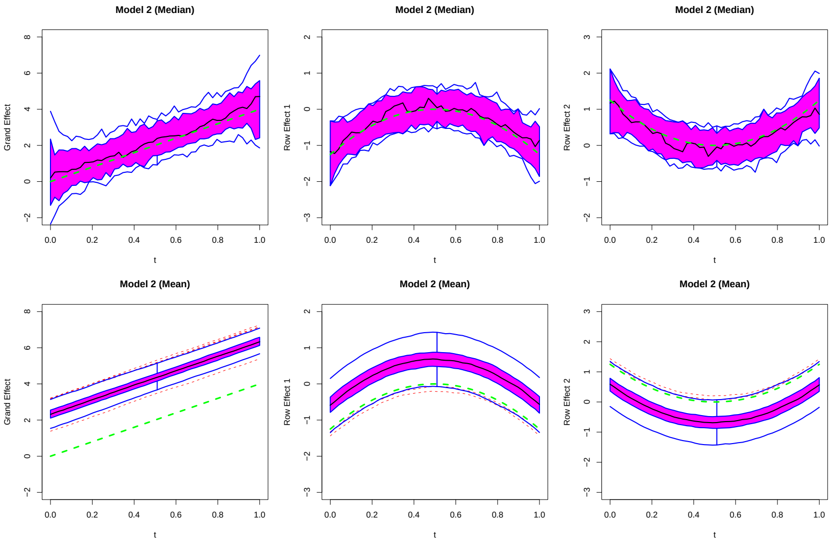
Simulations: Model 1



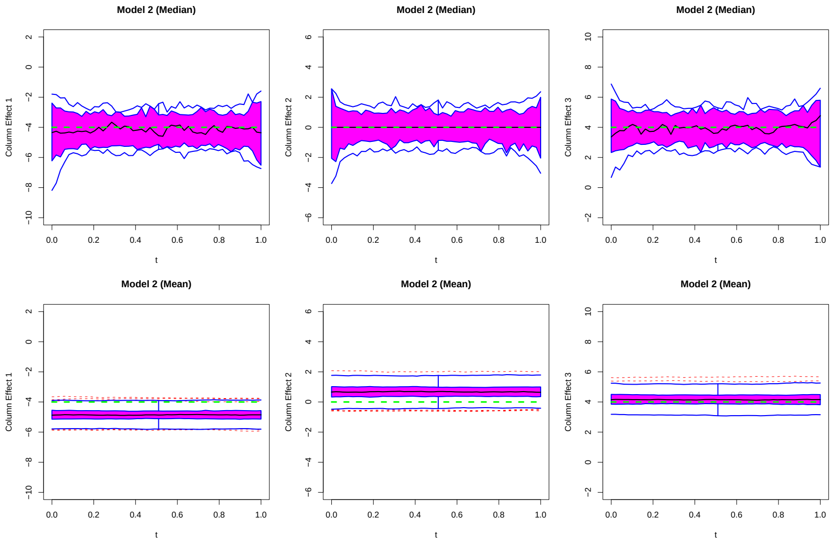
Simulations: Model 1



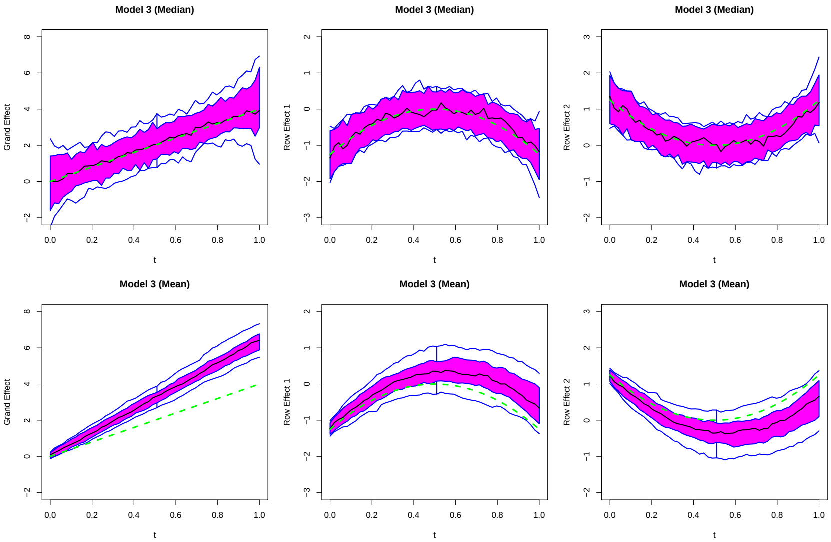
Simulations: Model 2



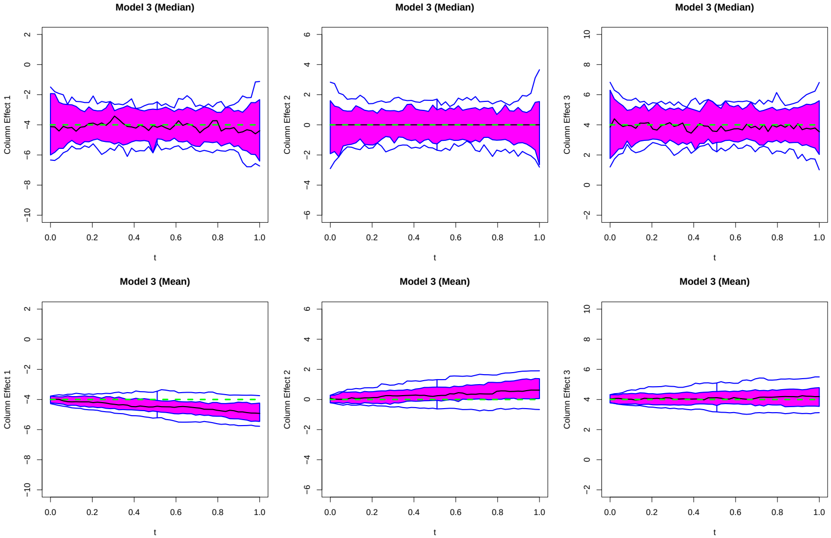
Simulations: Model 2



Simulations: Model 3

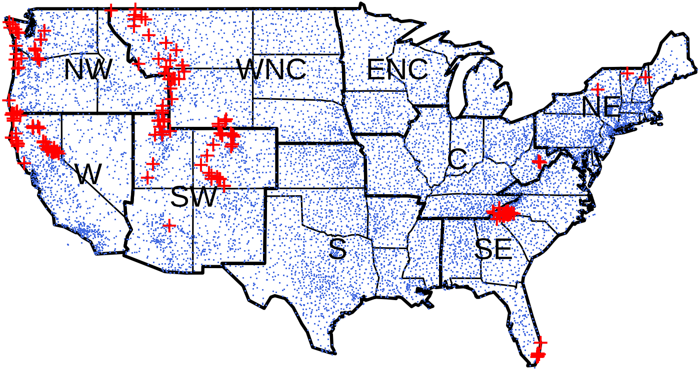


Simulations: Model 3

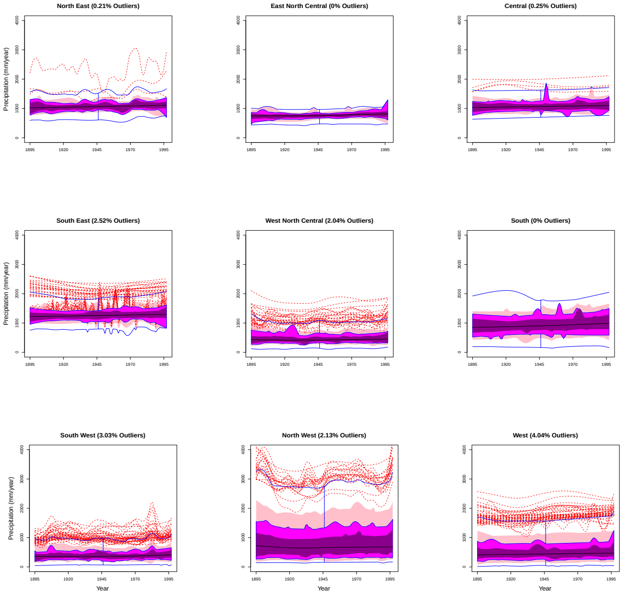


Spatio-Temporal Precipitation Data

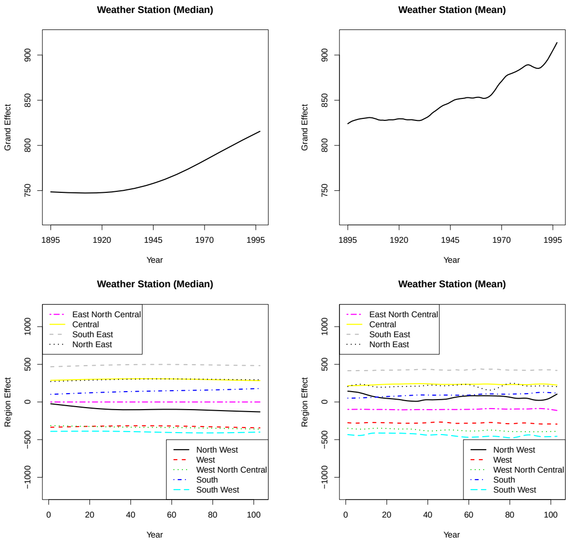
- Spatio-temporal precipitation data: annual total precipitation data for U.S. from 1895 to 1997 at 11,918 weather stations.
- Nine climatic regions for precipitation defined by National Climatic Data Center.
- Several areas of outliers detected by Sun and Genton (2011, 2012).



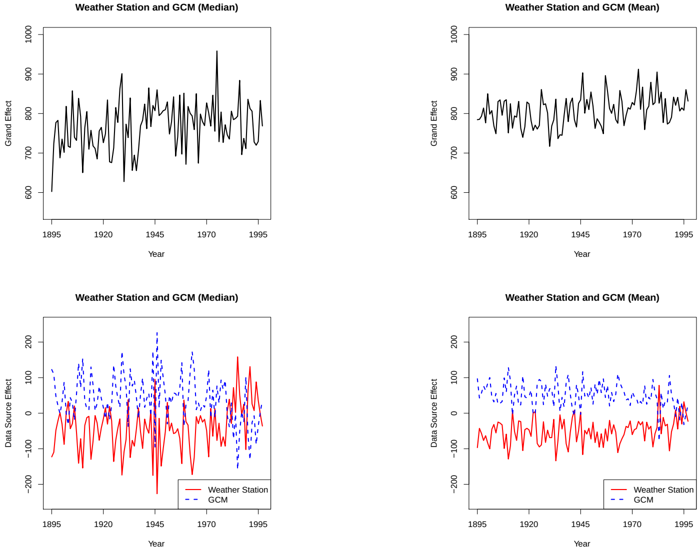
Functional Boxplots for Nine Climatic Regions



Climatic Region Effects



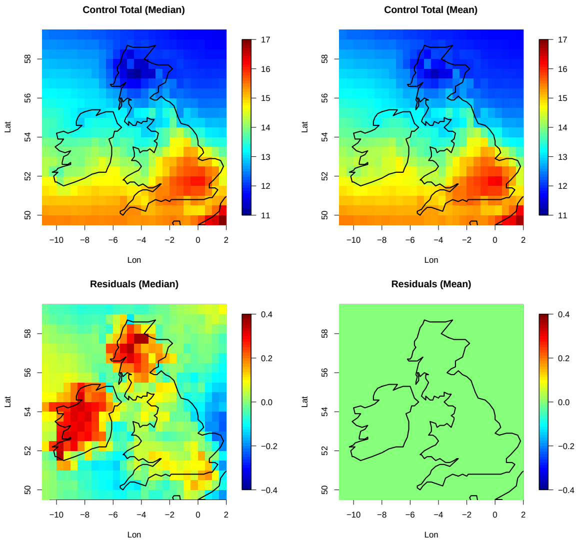
Weather Station and GCM Effects



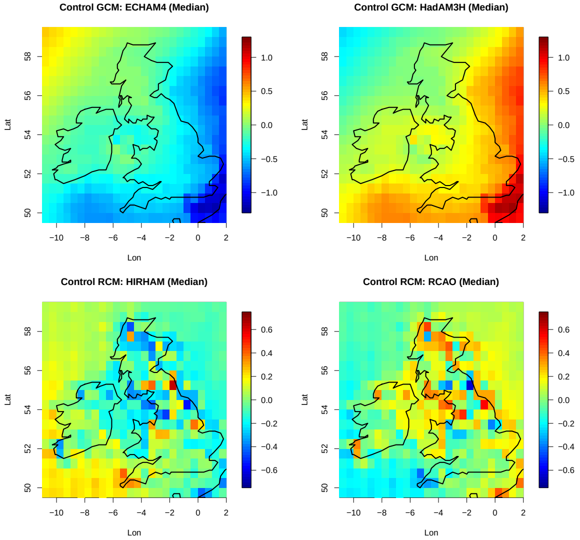
UK and Ireland Temperatures

- Regional Climate Model (RCM): has higher resolution, covers a limited area of the globe.
- The boundaries of RCM are driven by variables output from a global climate model (GCM).
- Question: how much variability in the model output is from RCM and how much is due to the boundary conditions from GCM.
- Functional ANOVA: Kaufman and Sain (2010) proposed a mean-based Bayesian framework for spatial data.
- Data: PRUDENCE project (Christensen, Carter, and Giorgi 2002), consists of control runs (1961-1990).
- Two factors: RCM (HIRHAM and RCAO), GCM (ECHAM4 and HadAm3H).

Grand Effect and Residuals



GCM and RCM Effects



Discussion

- Functional Median Polish: robust functional ANOVA fitted by functional median.
- Band depth: graph-based nonparametric ordering for functional/image data (e.g. median image).
- The functional median polish algorithm does not guarantee to yield the least L_1-norm residuals.
- Fink (1988) proposed a rather complex modification of the classical procedure that converges to the least L_1-norm residuals.