Functional Boxplots

Ying Sun and Marc G. Genton

Texas A&M University

September 17, 2010

Functional Boxplots

1. Informative Exploratory Tool
2. Band Depth for Functional Data
3. Functional Boxplots v.s. Classical Boxplots
4. Simulation Studies
5. Applications
6. Discussion

Functional Data

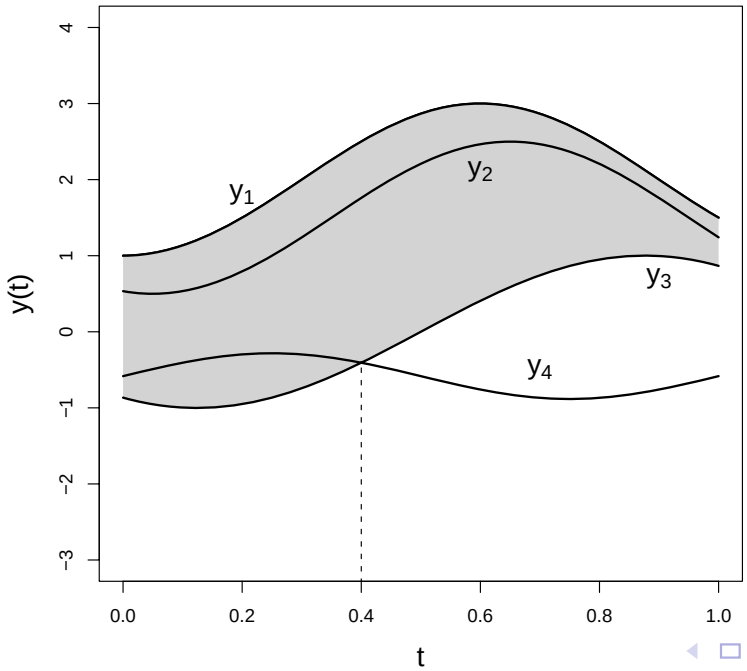
- The goal of this paper is
  - to develop functional boxplots for visualizing functional data.
- To do this, we need to
  - know the features of functional data,
  - generalize order statistics or ranks to the functional context.

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.
- Completely contains the curve y_2, but only partly contains 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).

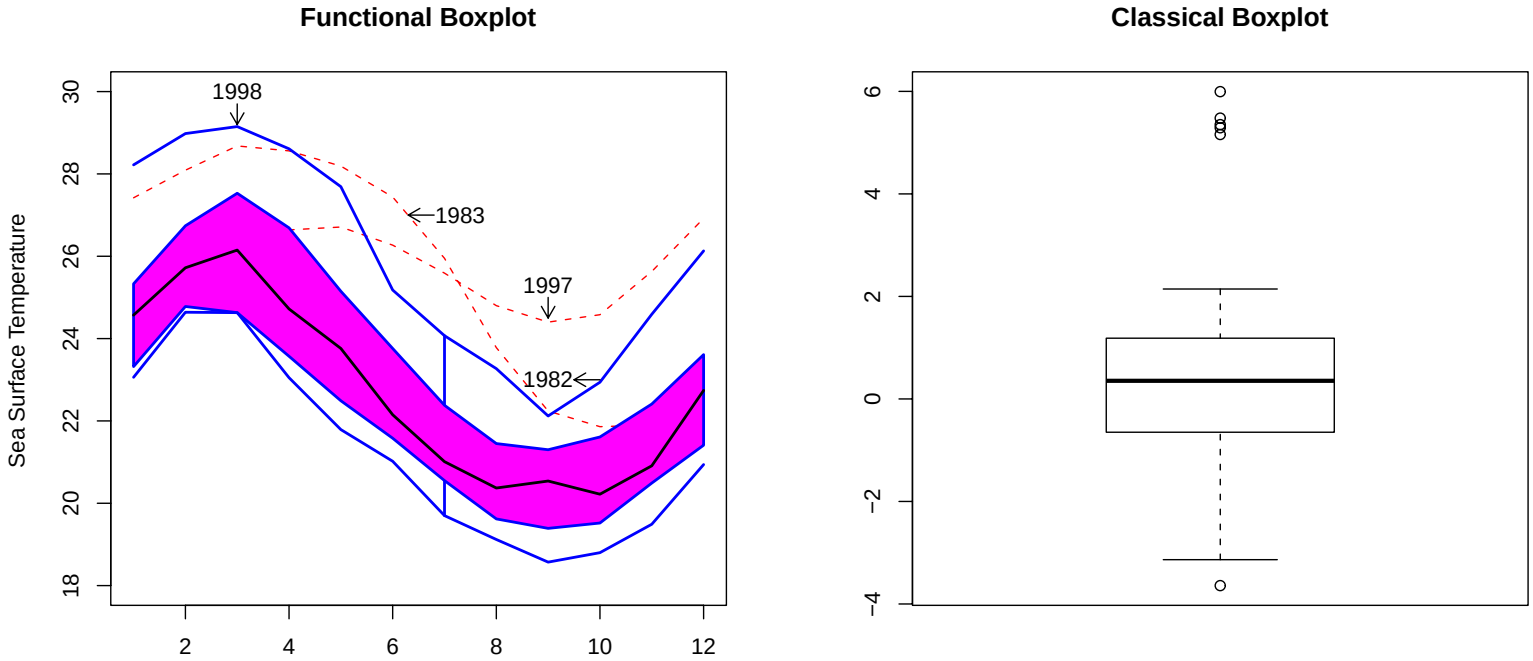
$$BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\},$$

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

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

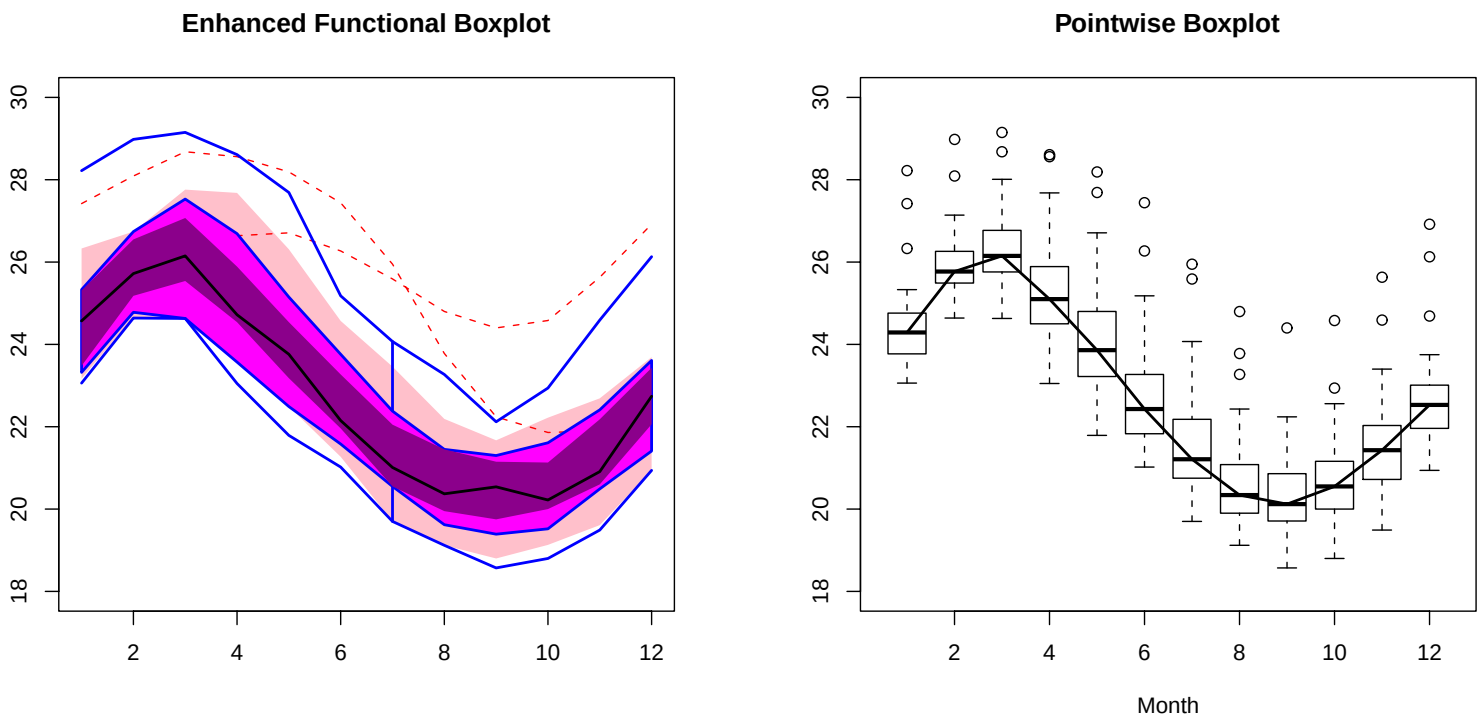
Functional Boxplots v.s. Classical Boxplots

- Comparing functional boxplots to classical boxplots.
- Functional boxplot: median, 50% central region, maximum non-outlying region, empirical outlier rule.



Enhanced Functional Boxplots

- The enhanced functional boxplot is a natural extension.
- Different from pointwise boxplots.

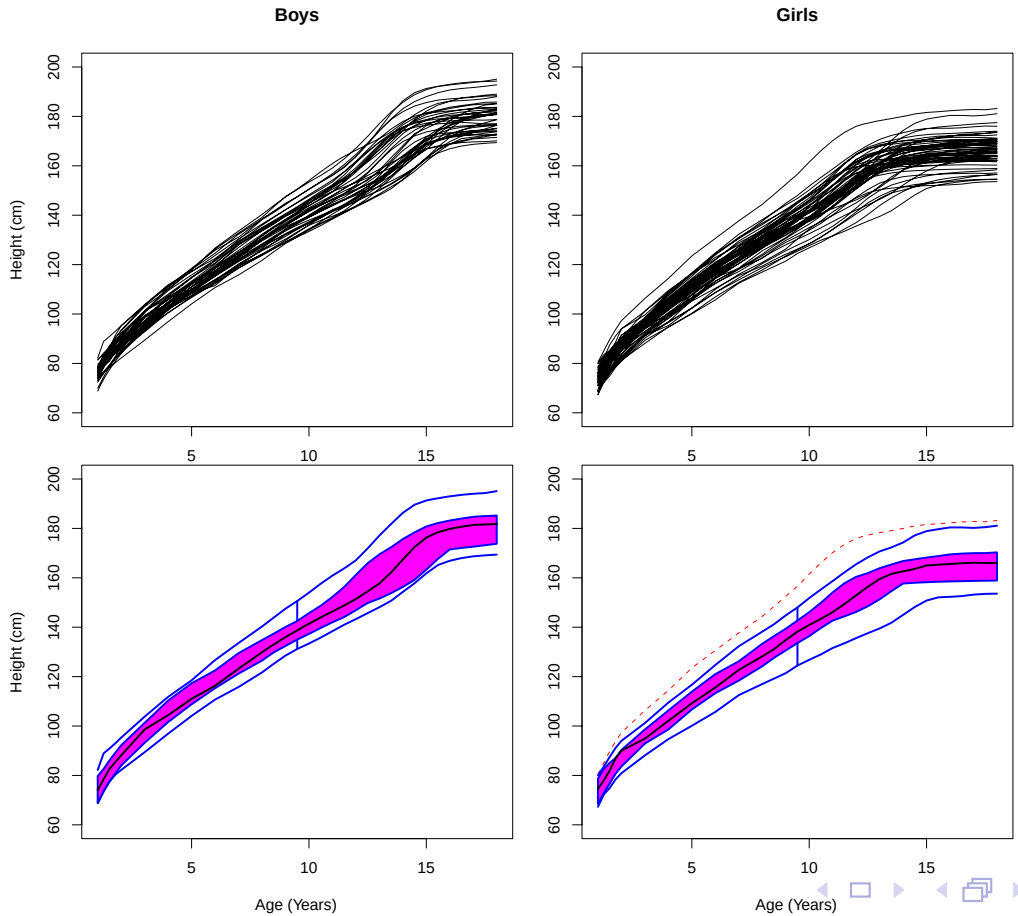


Functional Boxplot v.s. Functional Bagplot

- Functional boxplot:
  - uses MBD to order functional data,
  - visualizes functional data directly in the functional space,
  - a strong analog to the classical boxplot.
- Functional bagplot:
  - proposed by Hyndman and Shang (2010),
  - based on the first two robust principal component scores,
  - applies the bivariate bagplot (Rousseeuw et al., 1999),
  - maps the features of the bagplot into the functional space.

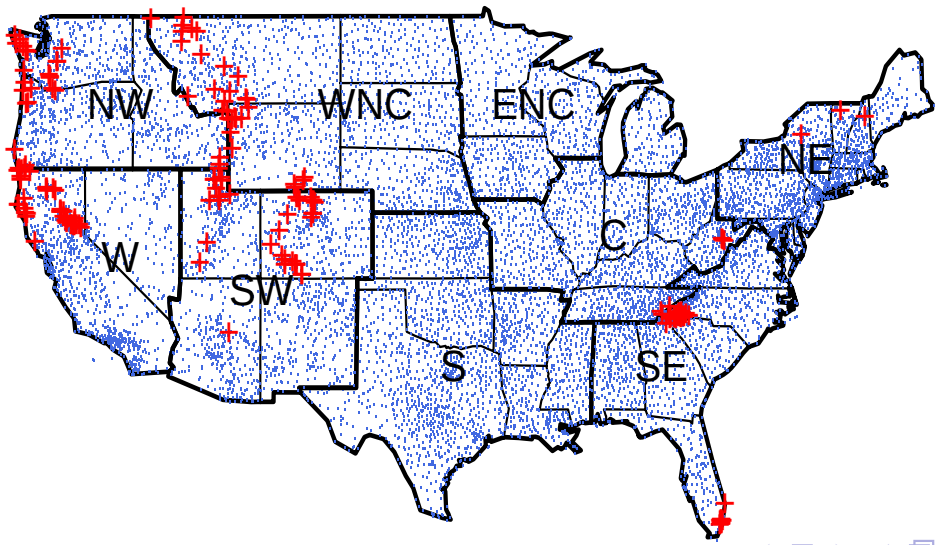
Children Growth Data

- The heights of 54 girls and 39 boys.



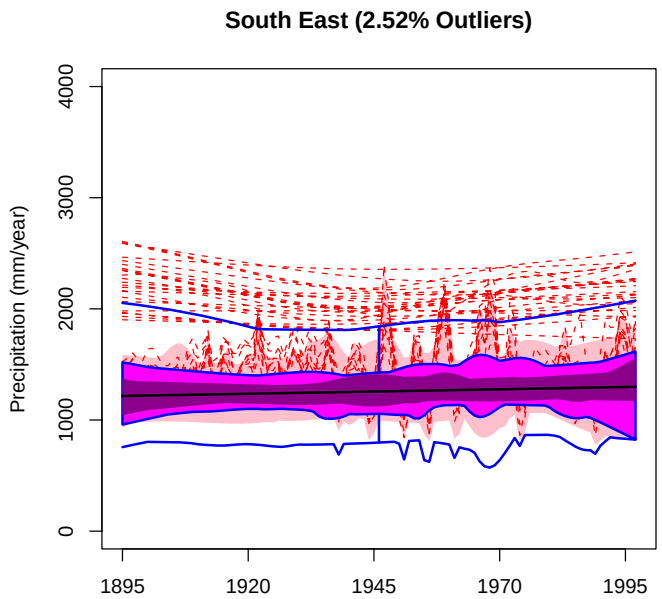
Spatio-Temporal Precipitation Data

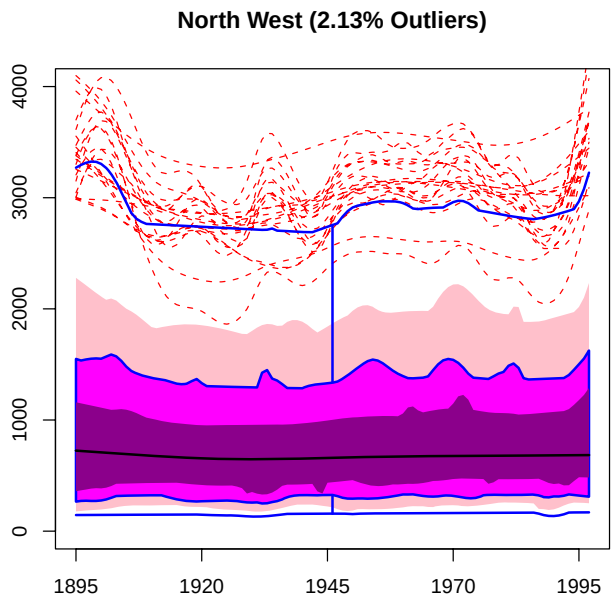
- Spatio-temporal precipitation data: annual total precipitation data for U.S. from 1895 to 1997 at 11,918 weather stations.
- Nine climatic regions for precipitation defined by National Climatic Data Center.
- Four areas of outliers detected by functional boxplots.



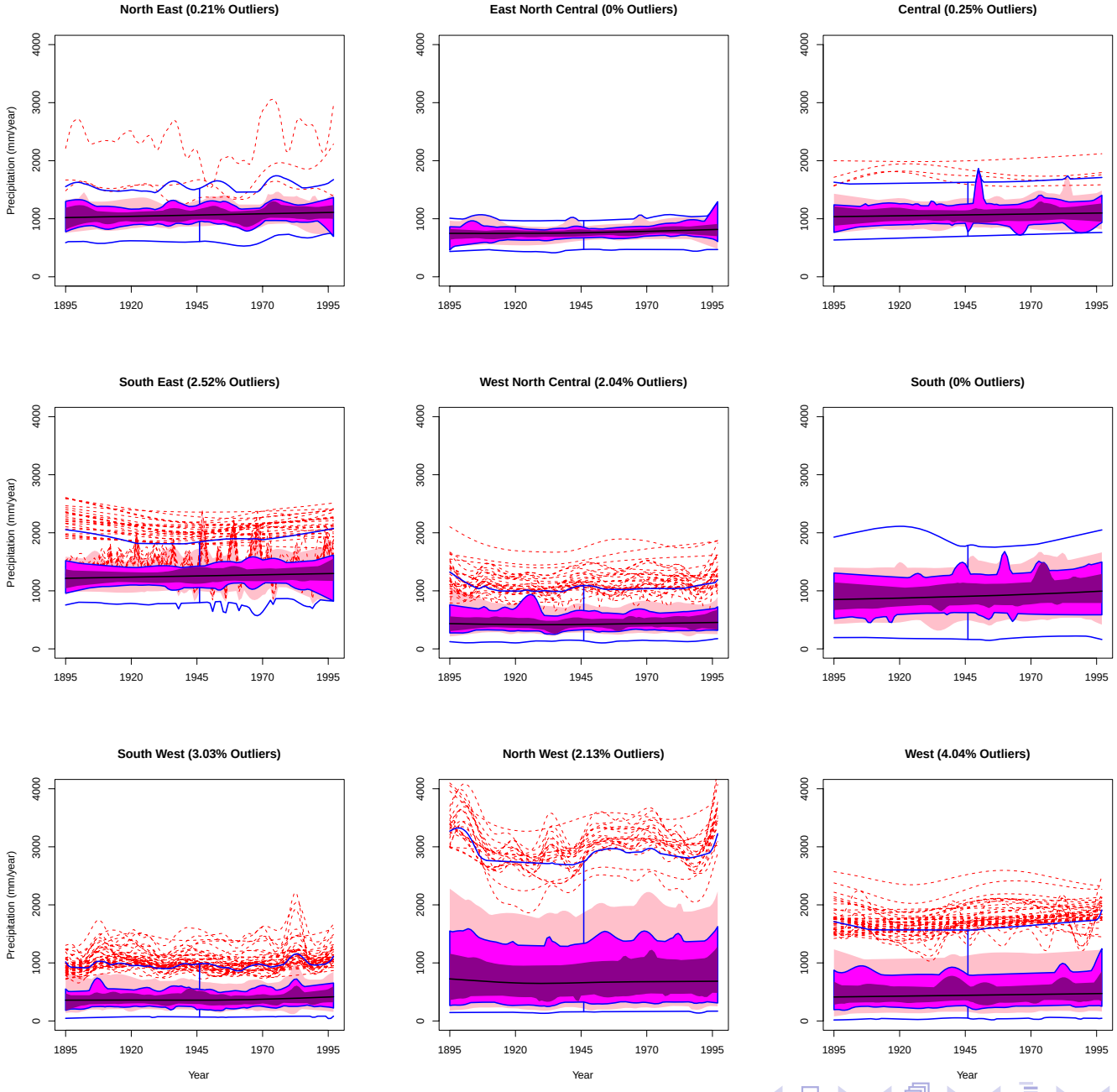
South East and North West

- Outliers at the southmost tip in Florida show an oscillatory pattern, which captures the hurricane effect.
- Outliers along the west coast have higher precipitation.
- Easy to compare two functional boxplots.





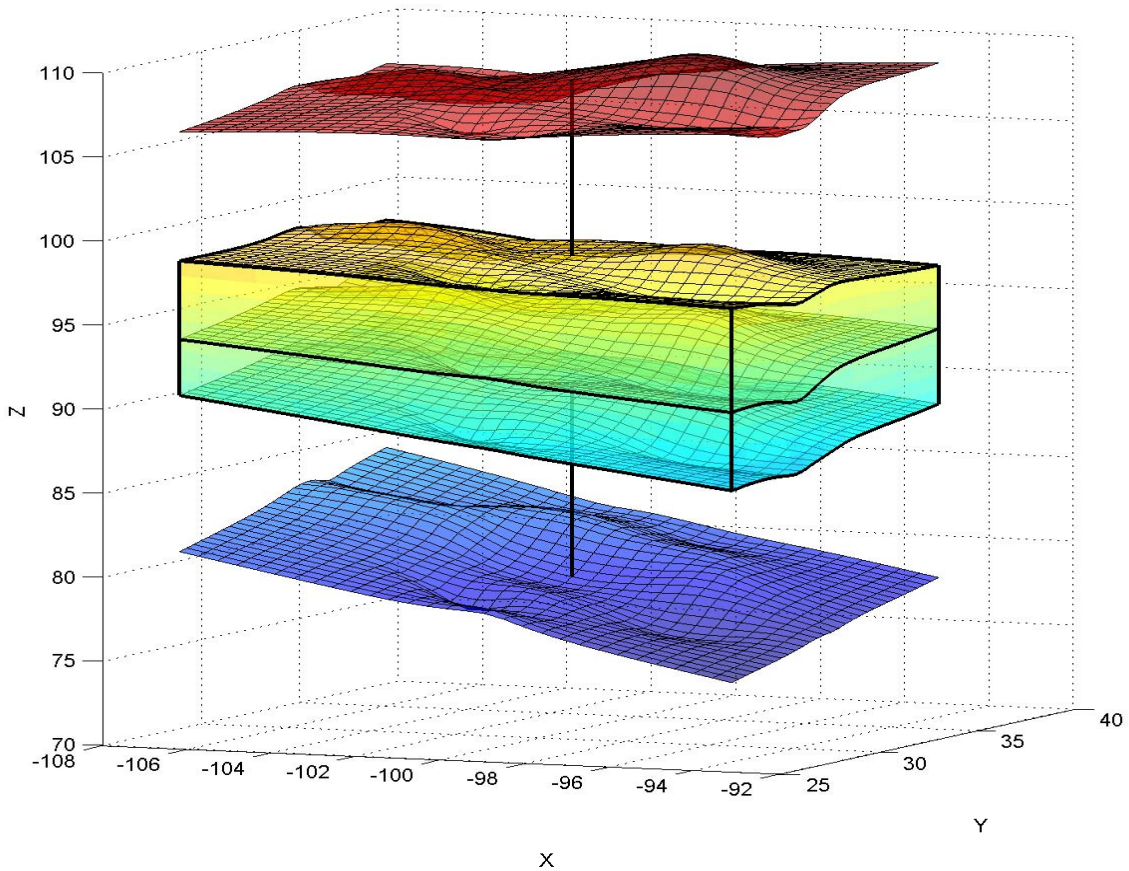
Functional Boxplots for Nine Climatic Regions



Surface Boxplots for Spatio-temporal Data

- We have viewed the information as a temporal curve at each spatial location.
- An alternative: treat the dataset as a spatial surface at each time.
- Define a volume-based surface band depth for a surface S by counting the proportion of surface bands determined by J different surfaces ($2 \leq J \leq n$) in $\mathbb{R}^3$, containing S.
- Lead to a three-dimensional surface boxplot.

Surface Boxplot



Conclusions

- BD: graph-based nonparametric ordering for functional data.
- Functional boxplot: visualize complex datasets, detect outliers.
- Precipitation data: provided by the Institute for Mathematics Applied to Geosciences (*http://www.image.ucar.edu/Data/US.monthly.met/*).
- R command: fbplot (*http://www.stat.tamu.edu/∼sunwards/publication.html*).