Special Issue Paper

Environmetrics

Received: 5 January 2011, Revised: 30 August 2011, Accepted: 6 September 2011, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.1136

Adjusted functional boxplots for spatio-temporal data visualization and outlier detection†

Ying Sun[a] and Marc G. Genton[b]*

This article proposes a simulation-based method to adjust functional boxplots for correlations when visualizing functional and spatio-temporal data, as well as detecting outliers. We start by investigating the relationship between the spatio-temporal dependence and the 1.5 times the 50% central region empirical outlier detection rule. Then, we propose to simulate observations without outliers on the basis of a robust estimator of the covariance function of the data. We select the constant factor in the functional boxplot to control the probability of correctly detecting no outliers. Finally, we apply the selected factor to the functional boxplot of the original data. As applications, the factor selection procedure and the adjusted functional boxplots are demonstrated on sea surface temperatures, spatio-temporal precipitation and general circulation model (GCM) data. The outlier detection performance is also compared before and after the factor adjustment.

Keywords: functional data; GCM data; outlier detection; precipitation data; robust covariance; spatio-temporal data

1. INTRODUCTION

Functional data analysis is an attractive approach to study complex data in statistics. In many statistical experiments, the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire observed function rather than a string of numbers. There is also an interesting class of applications that can be characterized as random processes evolving in space and time in, for instance, environmental science, agriculture, climatology, meteorology and hydrology.

To analyze functional data, many model-based methods have been developed over the years, among which Ramsay and Silverman (2005) provided various parametric methods whereas Ferraty and Vieu (2006) developed detailed nonparametric techniques. For spatio-temporal data, one can imagine a random field $Z(\mathbf{s},t)$, $(\mathbf{s},t) \in \mathbb{R}^d \times \mathbb{R}$, observed at the space-time coordinates $(\mathbf{s}_1,t_1),\ldots,(\mathbf{s}_n,t_n)$. The spatio-temporal variable $Z(\mathbf{s},t)$ could stand for temperature, precipitation, wind speed or atmospheric pollutant concentrations, to name a few. Some recent literature points out the significance of the spatio-temporal modeling approach; see for example Cressie and Wikle (2011) and references therein.

Visualization methods can also help to display the data, highlight their characteristics and reveal interesting features. Sun and Genton (2011) proposed an informative exploratory tool, the functional boxplot, and its generalization, the enhanced functional boxplot, for visualizing functional data as well as detecting potential outliers. The functional boxplot orders functional data by means of band depth (López-Pintado and Romo, 2009). It allows for ordering a sample of curves from the center outwards and, thus, introduces a measure to define the centrality or outlyingness of an observation. Indeed, one can compute the band depths of all the sample curves and order them according to decreasing depth values. Suppose each observation is a real function $y_i(t)$, $i = 1,\ldots,n$, $t \in \mathcal{I}$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. Let $y_{[i]}(t)$ denote the sample curve associated with the ith largest band depth value. Then $y_{[1]}(t),\ldots,y_{[n]}(t)$ can be viewed as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. The implication is that a smaller rank is associated with a more central position with respect to the sample curves. The order statistics induced by band depth start from the most central sample curve and move outwards in all directions. Thus, it is straightforward to define a central region for functional data. More specifically, López-Pintado and Romo (2009) defined the band depth through a graph-based approach. The graph of a function $y(t)$ is the subset of the plane $G(y) = \{(t,y(t)) : t \in \mathcal{I}\}$. The band in $\mathbb{R}^2$ delimited by the curves $y_{i_1},\ldots,y_{i_k}$ is $B(y_{i_1},\ldots,y_{i_k}) = \{(t,x(t)) : t \in \mathcal{I}, \min_{r=1,\ldots,k} y_{i_r}(t) \leqslant x(t) \leqslant \max_{r=1,\ldots,k} y_{i_r}(t)\}$. Let J be the number of curves determining a band,

* *Correspondence to: Marc G. Genton, Department of Statistics, Texas A&M University, College Station, TX 77843-3143,USA. E-mail: genton@stat.tamu.edu*

a *Statistical and Applied Mathematical Sciences Institute, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709-4006, USA*

b *Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*

† *This article is published in Environmetrics as a special issue on Spatio-Temporal Stochastic Modelling (METMAV), edited by Wenceslao González Manteiga, University of Santiago de Compostela, Spain.*

where J is a fixed value with $2 \leqslant J \leqslant n$. If $Y_1(t), \ldots, Y_n(t)$ are independent copies of the stochastic process $Y(t)$ generating the observations $y_1(t), \ldots, y_n(t)$, the population version of the band depth for a given curve $y(t)$ with respect to the probability measure P is defined as $\mathrm{BD}_J(y, P) = \sum_{j=2}^{J} \mathrm{BD}^{(j)}(y, P) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\}$, where $B(Y_1, \ldots, Y_j)$ is a band delimited by j random curves. The sample version of $\mathrm{BD}^{(j)}(y, P)$ is defined as $\mathrm{BD}_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_j \leqslant n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\}$, where $I\{\cdot\}$ denotes the indicator function. Then, the sample band depth of a curve $y(t)$ is $\mathrm{BD}_{n,J}(y) = \sum_{j=2}^{J} \mathrm{BD}_n^{(j)}(y)$. López-Pintado and Romo (2009) also proposed a modified band depth that replaces the aforementioned indicator function by a function, which measures the proportion of time that a curve $y(t)$ is in a band. This yields a more flexible ordering of the curves in the sample and tends to prevent depth ties.

In the classical boxplot, the box itself represents the middle 50% of the data. By analogy, the 50% central region in the functional boxplot can be defined by extending the concept of central region introduced by Liu *et al.* (1999) to functional data. The band delimited by the α proportion ($0 < \alpha < 1$) of deepest curves from the sample is used to estimate the α central region. In particular, the sample 50% central region is

$$C_{0.5} = \{(t, y(t)) : \min_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t) \leqslant y(t) \leqslant \max_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t)\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$. The envelope of the 50% central region represents the box in a classical boxplot and is the analog to the "inter-quartile range" (IQR). It gives useful indication of the spread of the 50% most central curves.

For functional boxplots, based on the center outwards ordering induced by band depth for functional data, the descriptive statistics are: the envelope of the 50% central region, the median curve and the maximum non-outlying envelope. In addition, potential outliers can be detected in a functional boxplot by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. The outer region (the "fence") is obtained by inflating the inner region (the "envelope") by 1.5 times the height of the 50% central region. Any curves crossing the fences are flagged as potential outliers.

Considering that when each curve is simply a point, the functional boxplot degenerates to a classical boxplot, Sun and Genton (2011) suggested the constant factor 1.5 as in a classical boxplot, but left to the user the possibility of modifying it. However, for functional data, there will be necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations will be spatially correlated as well. The outlier detection performance may be affected by the dependence in time and space. Therefore, in this article, we investigate the relationship between the dependence and the constant factor, and then propose a method to adjust the factor in a functional boxplot. This leads to an adjusted functional boxplot. Febrero *et al.* (2007, 2008) also considered outlier detection in functional data by depth measures but they did not account for the temporal or spatio-temporal correlation in the data and their method is quite different from the functional boxplot approach. In the finite dimensional case, Becker and Gather (1999, 2001) and Hubert *et al.* (2008) also discussed some model-based outlier detection rules for robust estimators.

Classical boxplots were first introduced by Tukey (1970) and Tukey (1977, pp. 39–43) in exploratory data analysis. In a classical boxplot, outliers can be detected by the 1.5 times IQR empirical rule. Here, the constant factor 1.5 can be justified by a standard normal distribution. Let Q_1 and Q_3 be the first and third quartiles of the standard normal distribution, respectively. The fences determined by $L_1 = Q_1 - 1.5\times$ IQR and $L_2 = Q_3 + 1.5\times$ IQR are -2.698 and 2.698. Then, the probability of being detected as an outlier is 0.7%. If we change the factor to 2, then the probability that a value is an outlier is only 0.07%. Therefore, in the functional boxplot, we would like to adjust the value of the constant factor on the basis of the dependence such that the probability of detecting no outliers is 99.3%, when actually, no outliers are present. It is clear that in a functional boxplot, the factor adjustment is crucial for outlier detection because it determines the percentage of detected outliers. However, the adjustment involves a certain amount of computation, thus, it is not necessary if one only wants to visualize and compare functional or spatio-temporal data.

This article is organized as follows. Section 2 illustrates how the dependence in time and space affects the outlier detection performance of functional boxplots. Then, the new method to select the constant factor in a functional boxplot is proposed in Section 3. The adjusted functional boxplots are demonstrated on applications to space–time datasets in Section 4, and a discussion is provided in Section 5.

2. SIMULATION STUDIES

To illustrate how the dependence in time and space affects the outlier detection performance of the functional boxplots, simulation studies are conducted under different spatio-temporal covariance models reviewed by Gneiting *et al.* (2007).

2.1. Data generation

We consider data drawn from a zero-mean, stationary spatio-temporal Gaussian random field $Z(\mathbf{s}, t)$, where $(\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}$. Let $C(\mathbf{h}, u) = \mathrm{cov}\{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)\}$ be the covariance function between any two observations whose locations are apart by a vector $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$ and a time span $u = |t_1 - t_2|$. Then, $C(\mathbf{h}, 0)$ and $C(\mathbf{0}, u)$ are purely spatial and purely temporal covariance functions, respectively. We aim at seeing how the strength of the correlation in time, in space, or in both, affects the outlier detection performance of the functional boxplot with the constant factor $F = 1.5$. We consider the following isotropic correlation models:

1. A purely temporal correlation function of Cauchy type,

$$C_T(\mathbf{h}, u) = (1 + a|u|^{2\alpha})^{-1} I\{\mathbf{h} = \mathbf{0}\}, \tag{1}$$

where $\alpha \in (0, 1]$ controls the strength of the temporal correlation, and $a > 0$ is the scale parameter in time. We set $a = 1$ and let α vary from 0.1 to 0.9.

2. A purely spatial correlation function of the form

$$C_S(\mathbf{h}, u) = [(1-\nu)\exp(-c\|\mathbf{h}\|) + \nu I\{\mathbf{h} = \mathbf{0}\}]I\{u = 0\}, \tag{2}$$

where $c > 0$ controls the strength of the spatial correlation, and $\nu \in (0, 1]$ is a nugget effect. We set $\nu = 0.05$ and let c vary from 0.1 to 2.

3. A space–time separable correlation function of the form

$$C_{\mathrm{SEP}}(\mathbf{h}, u) = C_S(\mathbf{h}, 0)C_T(\mathbf{0}, u), \tag{3}$$

which is the product of the purely temporal correlation function (1) and the purely spatial correlation function (2). Here, we consider combinations of α and c, where each has three levels, $\alpha = 0.1, 0.5, 0.9$ and $c = 0.1, 1, 2$.

4. A fully symmetric but generally non-separable correlation function

$$C_{\mathrm{FS}}(\mathbf{h}, u) = \frac{1-\nu}{1+a|u|^{2\alpha}}\left[\exp\left\{-\frac{c\|\mathbf{h}\|}{(1+a|u|^{2\alpha})^{\beta/2}}\right\} + \frac{\nu}{1-\nu}I\{\mathbf{h} = \mathbf{0}\}\right], \tag{4}$$

where $0 \leqslant \beta \leqslant 1$ controls the non-separability. It reduces to the separable model (3) when $\beta = 0$. Here, we set $\beta = 1$, the most non-separable version of this model, $\nu = 0.05$ and consider the same combinations of α and c as in model (3).

5. A general stationary correlation model

$$C_{\mathrm{STAT}}(\mathbf{h}, u) = \frac{(1-\nu)(1-\lambda)}{1+a|u|^{2\alpha}}\left[\exp\left\{-\frac{c\|\mathbf{h}\|}{(1+a|u|^{2\alpha})^{\beta/2}}\right\} + \frac{\nu}{1-\nu}I\{\mathbf{h} = \mathbf{0}\}\right] + \lambda\left(1 - \frac{1}{2v}|h_1 - vu|\right)_+, \tag{5}$$

where $q_+ = \max(q, 0)$ and h_1 is the first component of the spatial separation vector $\mathbf{h} = (h_1, h_2)'$. Here $0 \leqslant \lambda \leqslant 1$ controls the asymmetry. Again, we set $a = 1$, $\beta = 1$ and $\nu = 0.05$, then let $\lambda = 0.5$, $v = 0.05$ and consider the same combinations of α and c as in model (3).

The functional data $X_i(t) = Z(\mathbf{s}_i, t)$, $i = 1, \ldots, n$, are generated without any outliers from $Z(\mathbf{s}, t)$ with each of the correlation models above at $n = 100$ locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ on a grid of size 10×10 of the unit square with the grid spacing $1/9$ and $p = 50$ equally spaced time points in $[0, 1]$. With 1,000 replications, we compute the proportion of times that functional boxplots with the constant factor $F = 1.5$ detect no outliers. Thus, a proportion much smaller than 1 is an indication of bad outlier detection performance.

2.2. Numerical results

Tables 1 and 2 summarize the simulation results. In the purely temporal model (1), the larger the value of α, the stronger the temporal dependence is when $u < 1$. For a fixed constant factor $F = 1.5$ in the functional boxplot, Table 1 shows that the proportion of times that the functional boxplot correctly detects no outliers decreases as α increases. In other words, the stronger the correlation in time, the worse the outlier detection performance is. Similarly, in the purely spatial model (2), the smaller the value of c, the stronger the spatial dependence is. For all the values of c in Table 1, the proportions of correctly detecting no outliers are close to 1. This is an evidence that the constant factor 1.5 is too large when spatial correlation exists because usually, spatially correlated curves are more concentrated than independent ones.

Table 2 provides the proportion of times that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers for each combination of α and c under the separable, symmetric but non-separable and the general stationary spatial–temporal correlation models. It also shows that the proportion of correctly detecting no outliers decreases as α or the temporal dependence increases. For each value of α under different correlation models, because the strongest spatial correlation ($c = 0.1$) makes curves most concentrated, with the fixed constant factor $F = 1.5$ in functional boxplots, the proportions of correctly detecting no outliers for $c = 0.1$ are always the largest among those for $c = 0.1, 1, 2$. When the dependence in time is relatively large, $\alpha = 0.5, 0.9$, all the proportions under the separable correlation model are greater, hence better outlier detection performance, than those under either the symmetric but non-separable or the general stationary correlation models. This suggests that the interaction or the separability between spatial and temporal dependence has an effect on the outlier detection performance in a functional boxplot. To investigate the possible effect of the asymmetry in a correlation model, we compare the proportion of a functional boxplot correctly detecting no outliers under the symmetric but non-separable and the general stationary correlation models. When $\alpha = 0.5, 0.9$ and $c = 1$, the two proportions under the symmetric correlation model are larger than those under the general stationary one. However, there are still several cases where the general stationary model shows a better outlier detection performance.

It is now clear that the adjustment of the constant factor in functional boxplots is necessary when spatio-temporal correlations exist. Next, we propose a method for selecting the factor F.

3. SELECTION OF THE ADJUSTMENT FACTOR

The simulation studies in Section 2 have shown that the constant factor $F = 1.5$ gives different probability coverages under different spatio-temporal correlation models. In other words, we should choose the value of the factor F by controlling the probability of detecting no outliers to be 99.3% when, actually, no outliers are present.

To demonstrate the performance of outlier detection in the adjusted functional boxplot, we consider three outlier models, which have been studied by Sun and Genton (2011) along with two covariance models from our simulation studies, the purely spatial covariance function (2) and the general stationary covariance function (5). For the purely spatial model, we set $\nu = 0.05$ and let $c = 0.1$. For the

Table 1. The proportion of times (p) that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers under the purely temporal and the purely spatial correlation models with 1,000 replications and $n = 100$ curves

Temporal	α	0.1	0.3	0.5	0.7	0.9
	p	0.998	0.974	0.938	0.839	0.745
Spatial	c	0.1	0.5	1	1.5	2
	p	1	0.999	1	1	1

Table 2. The proportion of times that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers under the separable, symmetric but non-separable and the general stationary spatio–temporal correlation models with 1,000 replications and $n = 100$ curves

		Separable			Symmetric			Stationary		
α	c	0.1	1	2	0.1	1	2	0.1	1	2
0.1		1	0.997	0.993	0.995	0.990	0.994	0.995	0.994	0.995
0.5		1	0.980	0.961	0.956	0.945	0.952	0.957	0.943	0.950
0.9		1	0.942	0.908	0.901	0.854	0.836	0.900	0.833	0.844

general stationary model, we set $a = 1$, $\beta = 1$ and $\nu = 0.05$, then let $\lambda = 0.5$, $v = 0.05$ and consider the worst case in Table 2 where $\alpha = 0.9$ and $c = 1$. The models with outlying curves are:

1. Model 1 includes a symmetric contamination: $Y_i(t) = X_i(t) + c_i \sigma_i K$, where c_i is 1 with probability 0.1 and 0 with probability 0.9, K is a contamination size constant, which is equal to 2 for the purely spatial covariance model and 6 for the general stationary model. σ_i is a sequence of random variables independent of c_i taking values 1 and -1 with probability 1/2;
2. Model 2 is partially contaminated: $Y_i(t) = X_i(t) + c_i \sigma_i K$, if $t \geqslant T_i$ and $Y_i(t) = X_i(t)$, if $t < T_i$, where T_i is a random number generated from a uniform distribution on [0, 1];
3. Model 3 is contaminated by peaks: $Y_i(t) = X_i(t) + c_i \sigma_i K$, if $T_i \leqslant t \leqslant T_i + \ell$, and $Y_i(t) = X_i(t)$ otherwise, where T_i is random from a uniform distribution on $[0, 1 - \ell]$ and $\ell = 3/49$.

Under the purely spatial covariance model (2), the selected adjustment factor is $F = 1.0$, whereas $F = 2.2$ for the general stationary model. To compare with the functional boxplots without adjustment, we consider the distributions of two quantities in Sun and Genton (2011): p_c, the percentage of correctly detected outliers (number of correctly detected outliers divided by the total number of outlying curves), and p_f, the percentage of falsely detected outliers (number of falsely detected outliers divided by the total number of non-outlying curves). The simulation results are summarized in Table 3. For the purely spatial covariance model, the adjusted functional boxplots improve the outlier detection performance by keeping a low false detection rate $\hat{p}_f$, whereas increasing the correct detection rate $\hat{p}_c$, owing to the smaller adjustment factor $F = 1.0 < 1.5$. For the general stationary covariance model, the outlier detection performance is improved in the adjusted functional boxplots by keeping a high correct detection rate $\hat{p}_c$, whereas reducing the false detection rate $\hat{p}_f$, owing to the larger adjustment factor $F = 2.2 > 1.5$.

Sharing the same idea as in the simulations, we propose first to estimate the covariance matrix of the data to generate observations without any outliers. Then we use the simulations described in Section 2 to choose the constant factor F such that the proportion of times that a functional boxplot detects no outliers is 99.3%. Finally, we apply this adjusted factor to the functional boxplot on the original data and detect outliers. When estimating the covariance matrix, robust techniques are needed because outliers may exist in the original data. We use a componentwise estimator of a dispersion matrix proposed by Ma and Genton (2001), on the basis of a highly robust estimator of scale, Q_n. This estimator is location-free and has already been successfully used in the context of variogram estimation (Genton, 1998) in spatial statistics, and autocovariance estimation (Ma and Genton, 2000) in time series. Ma and Genton (2001) only studied the robust estimator when the sample size $n > p$, but the robust estimator can be also computed for $n \leqslant p$ because the dispersion matrix is estimated componentwisely. There are also many other robust estimators that could be used; some of which are based on the minimization of a robust scale of Mahalanobis distances. For example, the minimum volume ellipsoid and minimum covariance determinant estimators (Rousseeuw, 1984, 1985). However, their computation can be challenging. Alternatively, a more rapid orthogonalized Gnanadesikan–Kettenring estimator was proposed by Maronna and Zamar (2002) for high dimensional datasets.

To reduce the computational burden and simplify the covariance matrix estimation, we only generate a small number of curves, $n = 100$, without any outliers at p time points from the model $Z(\mathbf{s}, t) = g(\mathbf{s}, t) + e(\mathbf{s}, t)$, with mean $g(\mathbf{s}, t) = 0$, $(\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}$. Here, $e(\mathbf{s}, t)$

Table 3. The mean and standard deviation (in the parentheses) of the percentage $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplots, the adjusted functional boxplots under the purely spatial and the general stationary covariance models with 1,000 replications, 100 curves for models 1, 2, and 3

Spatial	Model 1		Model 2		Model 3	
	$F = 1.5$	$F = 1.0$	$F = 1.5$	$F = 1.0$	$F = 1.5$	$F = 1.0$
$\hat{p}_c$	99.9(0.7)	100(0)	85.6(13.9)	96.8(6.6)	34.3(18.3)	75.0(17.6)
$\hat{p}_f$	0(0)	0.006(0.085)	0(0)	0.003(0.060)	0(0)	0(0)
Stationary	Model 1		Model 2		Model 3	
	$F = 1.5$	$F = 2.2$	$F = 1.5$	$F = 2.2$	$F = 1.5$	$F = 2.2$
$\hat{p}_c$	100(0)	100(0)	88.2(13.3)	88.2(13.3)	99.9(1.0)	99.7(2.5)
$\hat{p}_f$	0.143(0.582)	0.012(0.153)	0.146(0.572)	0.007(0.085)	0.001(0.036)	0(0)

is a Gaussian random field with mean zero and covariance function estimated from the standardized original data, hence with marginal variance 1. For simplicity, we let the trend $g(\mathbf{s}, t)$ be 0 and the marginal variance be 1 because they do not affect the values of band depth, hence, the order of these curves. In addition, for spatio-temporal data, to reduce the dimension of the spatio-temporal covariance, we only estimate the covariances at certain distances and time lags depending on the simulation design.

4. APPLICATIONS

4.1. Sea surface temperatures

A dataset of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean was used by Hyndman and Shang (2010) and Sun and Genton (2011) to demonstrate the functional bagplot and the functional boxplot, respectively. In this case, each curve represents one year of observed sea surface temperatures from January 1951 to December 2007. The functional boxplot with the constant factor $F = 1.5$ in Sun and Genton (2011) detected two potential outliers: the years 1983 and 1997. In addition, the year 1982 from September to December and the year 1998 from January to June were viewed as being part of the maximum envelope. These 57 annual temperature curves show similarity in shape because they share a common mean function. Therefore, we detrend them first by subtracting the sample median at each time point and then check the correlations between the curves. Because the correlations are not statistically significant, we assume that these annual temperature curves are independent copies of each other and estimated the 12×12 covariance matrix in time. In simulations, by generating $n = 100$ curves at $p = 12$ time points from a Gaussian process with zero mean and estimated covariance function, the coverage probabilities for different values of the constant factor are listed in Table 4 with 1,000 replications. We select the constant factor to be 1.8 because when $F = 1.8$, the coverage probability is 0.995, close to 99.3%. After adjusting the constant factor, the functional boxplot still detects two El Niño years as outlier candidates.

4.2. Spatio-temporal precipitation

The functional boxplot can summarize information from complex data, such as space–time datasets. Sun and Genton (2011) illustrated this aspect by visualizing the observed annual total precipitation data for the coterminous US from 1895 to 1997, provided by the Institute for Mathematics Applied to Geosciences (*http://www.image.ucar.edu/Data/US.monthly.met/*). There are 11,918 stations reporting precipitation at some time in this period. The observations are time series with $p = 103$ yearly precipitation observations, or one curve, at each spatial location. Functional boxplots were applied to nine climatic regions for precipitation in the US, defined by the National Climatic Data Center and the percentages of detected potential outliers for each region were reported. Sun and Genton (2011) noticed that for spatio-temporal data, the precipitation curves are not independent but spatially correlated. Therefore, the percentages of potential outliers might not be correct because of the spatial correlations.

By taking the spatio-temporal correlation into account, we adjust the constant factor in the functional boxplots again by simulations. For each climatic region, in the simulation, we generate spatio-temporal data from a zero-mean Gaussian random field at $n = 100$ locations on a

Table 4. The coverage probabilities for different values of the constant factor F with $n = 100$ curves at $p = 12$ time points and 1,000 replications in simulations for the sea surface temperatures example

Factor F	1.2	1.3	1.4	1.5	1.6	1.7	**1.8**	1.9	2.0
Coverage	0.768	0.859	0.922	0.956	0.979	0.988	**0.995**	1.000	1.000

Note: The selected factor is in bold font.

grid of size 10×10 with the grid spacing 1/9 at $p = 30$ time points. Here, the distance unit is kilometer and the time unit is year. Then, we estimate the $3,000 \times 3,000$ covariance matrix from the standardized original data and obtain the coverage probabilities for different values of the constant factor with 1,000 replications. The results are summarized for each region in Table 5. Each component of the $3,000 \times 3,000$ covariance matrix is estimated by the Q_n-based procedure of Ma and Genton (2001). To reduce the computational effort, for each combination of time lag and distance, we estimate the covariance element by randomly selecting pairs of the irregularly spaced locations that are close to the distances on the 10×10 grid under the assumption of stationarity.

With the concept of central regions, Sun and Genton (2011) generalized the functional boxplot to an enhanced functional boxplot where the 25% and 75% central regions are provided as well in addition to the 50% central region. For the spatio-temporal precipitation data, the nine adjusted enhanced functional boxplots in Figure 1 can still reveal information about the different annual precipitation characteristics for different climatic regions, but with less potential outliers than previously detected by Sun and Genton (2011). The percentage of detected outliers for each region is summarized in Table 6.

4.3. General circulation model

A general circulation model (GCM) is a climate model of the general circulation of a planetary atmosphere or ocean. It uses complex computer programs to simulate the earth's climate system and allows us to look into the earth's past, present and future climate states. Here, we consider precipitation data generated from the National Center for Atmospheric Research-Community Climate System Model Version 3.0 (Collins *et al*., 2006, and references therein), which was run given scenarios from the Intergovernmental Panel on Climate Change (IPCC)'s Special Report on Emission Scenarios; see IPCC (2000) and Ammann *et al*. (2010). Functional boxplots can be used to visually compare the annual precipitation produced by the GCM with the real observations from weather stations considered in the previous section.

For these GCM data, there are 256×128 cells over the whole globe with a resolution of 1.406×1.406 degrees, or around 156×156 km. The observations from weather stations are much denser, but only for the coterminous US. To make the weather station observations comparable to the GCM data, we match them by longitude and latitude first, and then average the observations from weather stations within each cell, which leads to 473 cells in total, hence 473 annual precipitation curves for the coterminous US.

For the coterminous US precipitation, the functional boxplots with the constant factor $F = 1.5$ for weather station and GCM data are shown in the top panel of Figure 2 with the percentage of detected outliers. Now, we estimate the $3,000 \times 3,000$ spatio-temporal covariance matrix from the standardized original data by the same simulation design as in Section 4.2 and the coverage probabilities for different values of the constant factor with 1,000 replications are summarized in Table 7 for both weather station and GCM data. The adjusted functional boxplots with percentage of detected outliers are shown in the bottom panel of Figure 2.

Figure 2 shows that the precipitation data produced by the GCM roughly capture the overall pattern of the US precipitation. The two functional boxplots of weather station and GCM data coincide on the median curves, and the maximum annual precipitation for some wet locations is also on the same level. However, the narrower 50% central region in the functional boxplot of the GCM data indicates that the first 50% most representative precipitation curves have less variability, which leads to a relatively large percentage of outliers. Moreover, both functional boxplots are skewed to the right (i.e., to large precipitation), but the one for GCM does not produce as low annual precipitation as the real observations from weather stations for some dry locations.

The maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots with respect to either the whole US or each of the climatic regions are shown in Figure 3. For the whole US, the outliers, denoted by red plus signs, are around the North of the US for GCM data, but the only one outlier detected by the functional boxplot is located at the North West for weather station data. The

Table 5. The coverage probabilities for different values of the constant factor $F = 1.4, 1.5, \ldots, 2.2$ with $n = 100$ curves at $p = 30$ time points and 1,000 replications in simulations for the precipitation application

Region	1.4	1.5	1.6	1.7	**1.8**	**1.9**	**2.0**	**2.1**	2.2
NE	0.927	0.951	0.971	0.985	0.987	**0.993**	0.994	0.995	0.996
ENC	0.896	0.941	0.958	0.972	0.985	0.989	**0.993**	0.994	0.995
C	0.926	0.951	0.970	0.984	0.988	**0.994**	0.995	1.000	1.000
SE	0.926	0.948	0.974	0.979	0.988	**0.993**	0.996	0.999	1.000
WNC	0.893	0.944	0.966	0.973	0.984	0.990	**0.992**	0.995	0.996
S	0.902	0.934	0.959	0.976	0.983	0.987	0.989	**0.993**	0.994
SW	0.920	0.948	0.972	0.979	0.987	0.992	**0.993**	0.995	0.995
NW	0.911	0.939	0.962	0.974	0.983	0.987	0.992	**0.993**	0.995
W	0.911	0.949	0.974	0.988	**0.993**	0.994	0.996	0.999	0.999

NE, North East; ENC, East North Central; C, Central; SE, South East; WNC, West North Central; S, South; SW, South West; NW, North West; and W, West. The selected factors are in bold font.

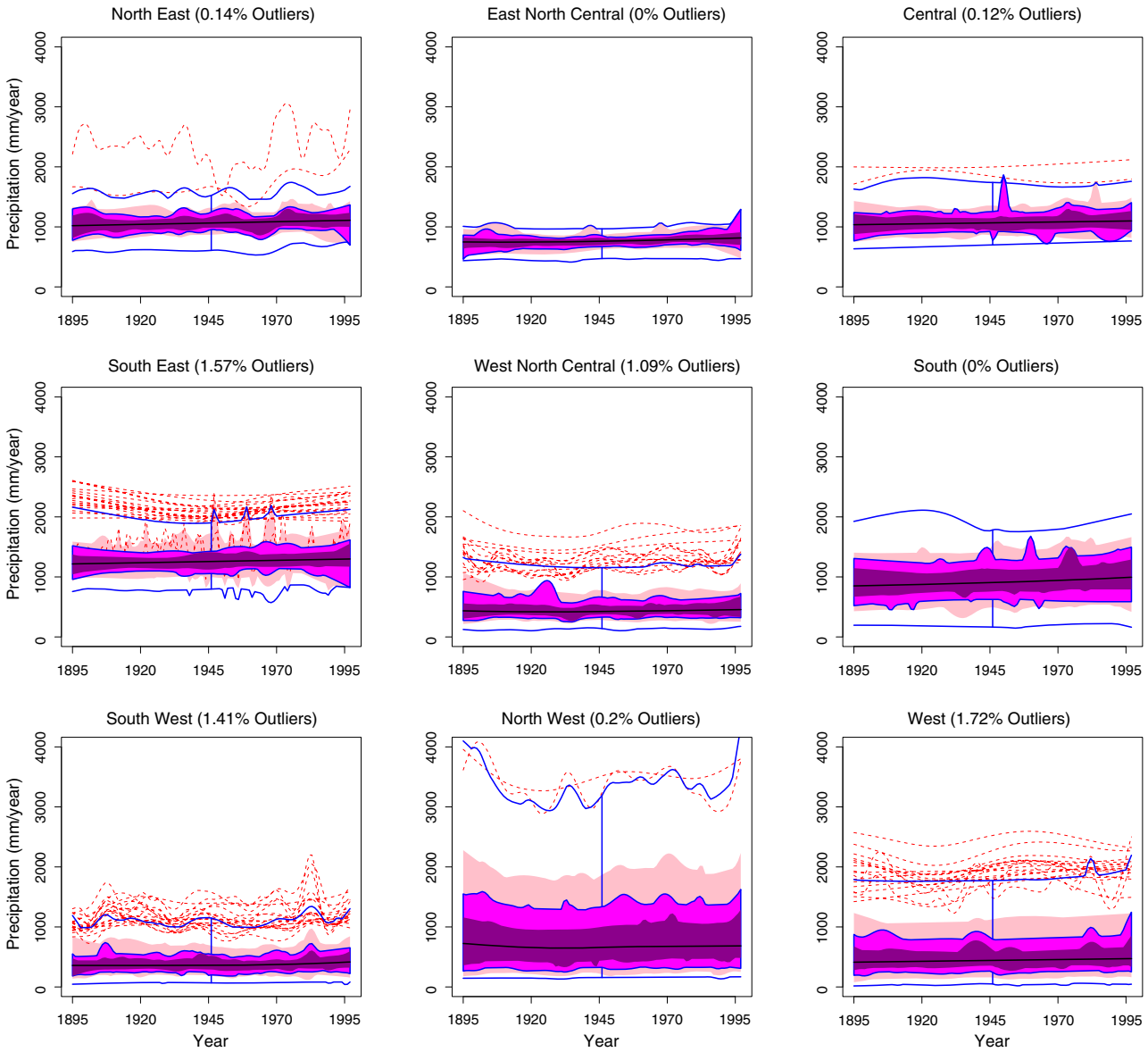

Figure 1. Adjusted enhanced functional boxplots of observed yearly precipitation over the nine climatic regions for the coterminous US from 1895 to 1997. Dark magenta, magenta and pink denote the 25%, 50% and 75% central regions, respectively, and the outlier rule is the adjusted constant factor times the 50% central region. The percentage of detected outliers in each climatic region is provided.

Table 6. Comparison of outlier detection percentages for each climatic region before and after adjustment of the constant factor

Region	NE	ENC	C	SE	WNC	S	SW	NW	W
Before	0.21	0	0.25	2.52	2.04	0	3.03	2.13	4.04
After	0.14	0	0.12	1.57	1.09	0	1.41	0.20	1.72

NE, North East; ENC, East North Central; C, Central; SE, South East; WNC, West North Central; S, South; SW, South West; NW, North West; and W, West.

maps also show that the precipitation data produced by GCM do not capture the characteristics of the observed precipitation from weather stations well. From weather stations, the West of the US overall has a lower precipitation than the East, and the higher precipitation locations are along the west coast and the South East. This pattern is hard to see from the GCM, and the higher precipitation locations appear in the

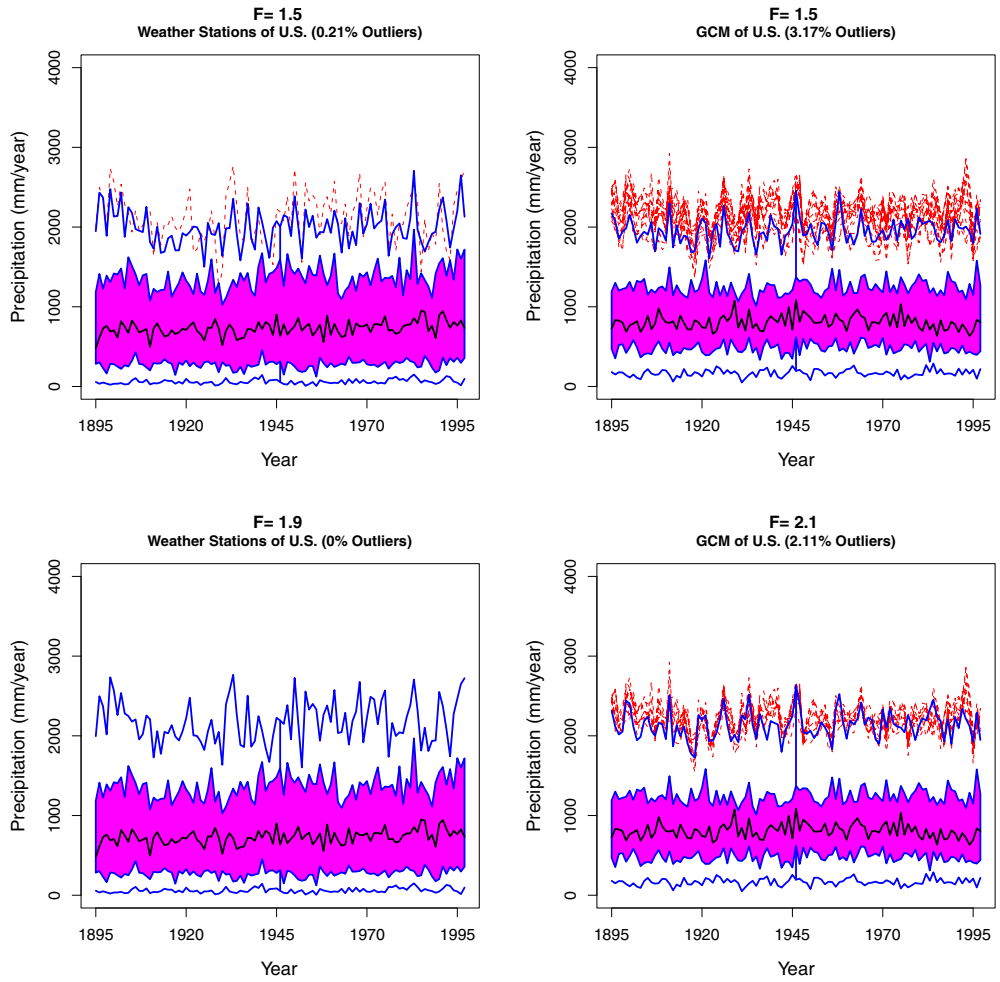

Figure 2. Top panels: the functional boxplots of weather station and GCM data with the constant factor $F = 1.5$ for the coterminous US precipitation. Bottom panels: the adjusted functional boxplots of weather station and GCM data with the adjusted constant factor F for the coterminous US precipitation.

Table 7. The coverage probabilities for different values of the constant factor $F = 1.4, 1.5, \ldots, 2.2$ with $n = 100$ curves at $p = 30$ time points and 1,000 replications in simulations for both weather station and general circulation model (GCM) data. The weather station and the GCM past are for the time period from 1970 to 1997. The GCM future is for the time period from 2070 to 2097

Source	1.4	1.5	1.6	1.7	1.8	**1.9**	**2.0**	**2.1**	2.2
Weather Stations	0.904	0.945	0.968	0.979	0.987	**0.993**	0.997	0.998	0.999
GCM Past	0.914	0.950	0.971	0.979	0.985	0.990	0.992	**0.993**	0.996
GCM Future	0.912	0.947	0.966	0.981	0.988	0.991	**0.994**	0.998	0.998

GCM, general circulation model.
Note: The selected factors are in bold font.

North shown as outliers. As can be expected, the detected outliers with respect to each climatic region are different from those for the whole US. For climatic regions, the potential outliers are still in the North of the US for GCM data with a larger percentage, but are located along the West coast and the Rocky Mountain area for weather station data.

The GCM also simulates precipitation for the future. The future runs of the GCM were under the IPCC A2 scenario (Ammann *et al.*, 2010) after the year 2020, which considers a continued increase of atmospheric green house gases and the associated warming throughout the 21st century. To compare the past precipitation with the future rainfall, we use the adjusted functional boxplots to visualize the spatio-temporal

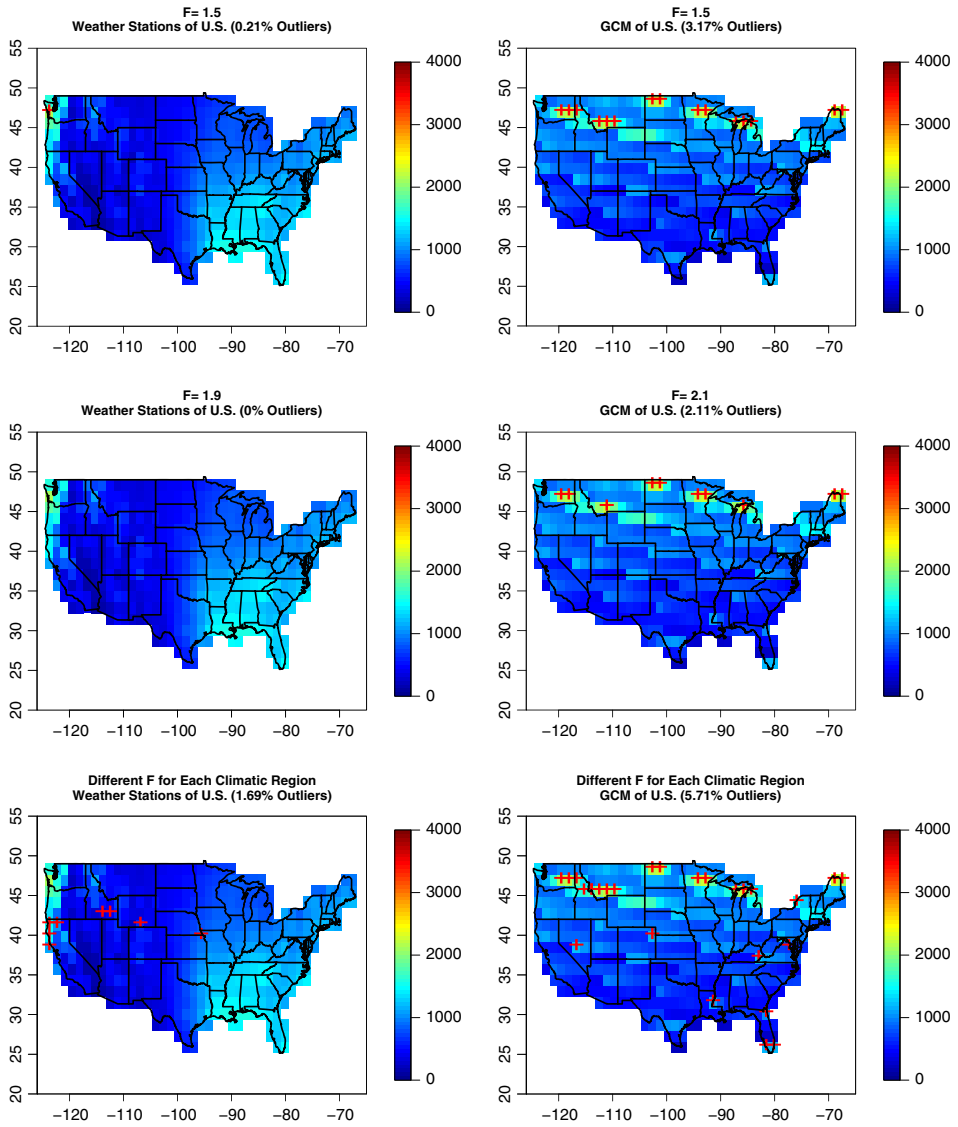

Figure 3. Top panels: the maps of weather station and general circulation model (GCM) data where outliers are detected by the functional boxplots with the constant factor $F = 1.5$ for the coterminous US precipitation. Middle panels: the maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots for the coterminous US precipitation. Bottom panels: the maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots for each climatic region. The colors of each cell denote the averaged annual precipitation (mm/year) and the red plus signs indicate the detected outliers.

precipitation for the time period from 1970 to 1997 and for the future period from 2070 to 2097. For the past, the functional boxplots of weather station and GCM data are shown in the top and middle panels of Figure 4. The bottom panels of Figure 4 show the functional boxplots of GCM data for the future. The future runs of the GCM produce a little wider 50% central region than the past runs do, thus a smaller outlier percentage, but both have narrower 50% central regions hence larger outlier percentages compared with the weather station data. We can also see that the median curves of the past and future runs from GCM are higher than that from weather stations, and the minimum precipitation is also higher than the real observations.

5. DISCUSSION

This article has focused on how to adjust the functional boxplot proposed by Sun and Genton (2011) for correlations to perform functional and spatio-temporal data visualization and outlier detection. In a functional boxplot, potential outliers can be detected by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. However, for functional data, there is necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations are spatially correlated as well. The simulation studies in Section 2 showed that the outlier detection performance is obviously affected by the dependence in time and space. Therefore, to correct the outlier detection performance, the constant factor of the empirical outlier rule is important. The factor $F = 1.5$ in a classical boxplot can be justified by a standard normal distribution, because it leads to a probability of 99.3% for correctly detecting no outliers. Following this

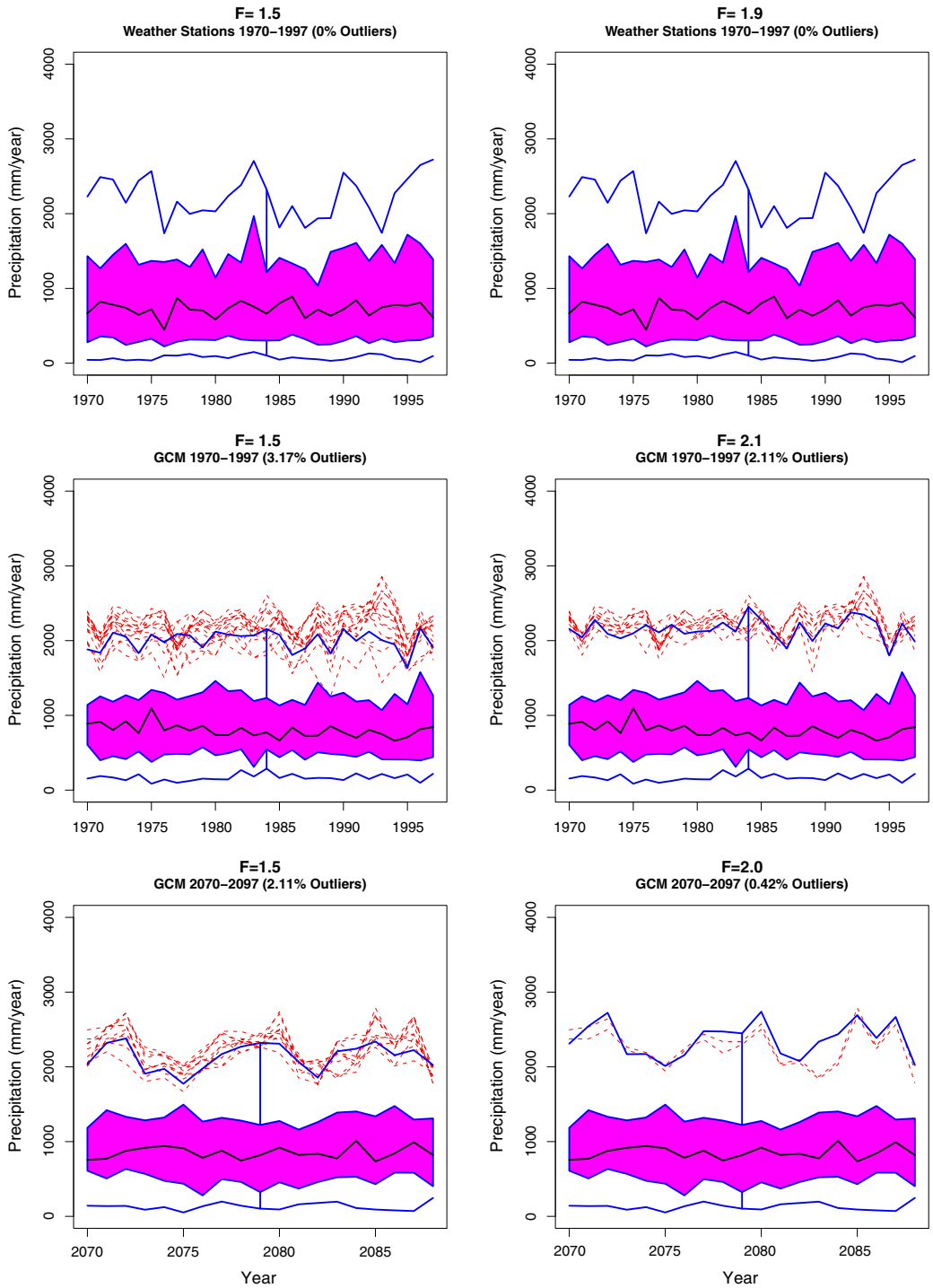

Figure 4. Top panels: the functional boxplot and the adjusted functional boxplot of weather station data for the coterminous US precipitation from 1970 to 1997. Middle panels: the functional boxplot and the adjusted functional boxplot of general circulation model (GCM) data for the coterminous US precipitation from 1970 to 1997. Bottom panels: the functional boxplot and the adjusted functional boxplot of GCM data for the coterminous US precipitation from 2070 to 2097 under Intergovernmental Panel on Climate Change A2 scenario.

idea, we proposed a simulation-based method to select this constant factor for a functional boxplot by controlling the probability of detecting no outliers to be 99.3% when actually no outliers are present. Then, how to estimate the covariance function, especially for spatio-temporal data, is also important, and robust techniques are needed when considering the potential presence of outliers in the original data.

As applications, we used our method to adjust the functional boxplots for sea surface temperatures, spatio-temporal precipitation and GCM data. In fact, all the selected factors were greater than 1.5, which agrees with the simulation results in Section 2. The interpretation is

that a positive correlation leads to a larger variability, therefore, the extreme observations may not be outliers but may be due to the positive correlation in time and space.

For spatio-temporal data, we have viewed the information as a temporal curve at each spatial location. Sun and Genton (2011) also proposed an alternative to treat such data as a spatial surface at each time. In this case, it would lead to a three-dimensional surface boxplot with similar characteristics as the functional boxplots. Similarly, for outlier detection, the fences are obtained by the 1.5 times the 50% central region rule. Any surfaces crossing the fences are flagged as outlier candidates. Therefore, for the surface functional boxplot in $\mathbb{R}^3$, the constant factor can be also adjusted by the simulation-based method described in this article and leads to an adjusted surface functional boxplot.

Acknowledgements

This research was partially supported by NSF grants DMS-1007504, DMS-1100492, and Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the Guest Editor, two referees and Noel Cressie for helpful comments, as well as Caspar M. Ammann for providing the GCM data.

REFERENCES

Ammann CM, Washington WM, Meehl GA, Buja L, Teng H. 2010. Climate engineering through artificial enhancement of natural forcings: magnitudes and implied consequences. *Journal of Geophysical Research* **115**: D22109. DOI:10.1029/2009JD012878.

Becker C, Gather U. 1999. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94**: 947–955.

Becker C, Gather U. 2001. The largest nonindentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Computational Statistics & Data Analysis* **36**: 119–127.

Collins WD, et al. 2006. The community climate system model version 3 (CCSM3). *Journal of Climate* **19**: 2122–2143. DOI: 10.1175/JCLI3761.1.

Cressie N, Wikle C. 2011. *Statistics for Spatio-Temporal Data*. Wiley: New York.

Febrero M, Galeano P, González-Manteiga W. 2007. Functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics* **22**: 411–427.

Febrero M, Galeano P, González-Manteiga W. 2008. Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics* **19**: 331–345.

Ferraty F, Vieu P. 2006. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag: New York.

Genton MG. 1998. Highly robust variogram estimation. *Mathematical Geology* **30**: 213–221.

Gneiting T, Genton MG, Guttorp P. 2007. Geostatistical space-time models, stationarity, separability and full symmetry. In *Statistics of Spatio-Temporal Systems*, Finkenstaedt B, Held L, Isham V (eds). Chapman & Hall/CRC Press; 151–175.

Hubert M, Rousseeuw PJ, Van Aelst S. 2008. High-breakdown robust multivariate methods. *Statistical Science* **23**: 92–119.

Hyndman RJ, Shang HL. 2010. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* **19**: 29–45.

Intergovernmental Panel on Climate Change (IPCC). 2000. Special report on emission scenarios. In *A special report of Working Group III of the Intergovernmental Panel on Climate Change*, Nakicenovic N, Swart R (eds). Cambridge Univ. Press: Cambridge, U.K; 612.

Liu RY, Parelius JM, Singh K. 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics* **27**: 783–858.

López-Pintado S, Romo J. 2009. On the concept of depth for functional data. *Journal of the American Statistical Association* **104**: 718–734.

Ma Y, Genton MG. 2000. Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis* **21**: 663–684.

Ma Y, Genton MG. 2001. Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis* **78**: 11–36.

Maronna RA, Zamar RH. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* **50**: 295–304.

Ramsay JO, Silverman BW. 2005. *Functional Data Analysis*, 2nd ed. Springer: New York.

Rousseeuw PJ. 1984. Least median of squares regression. *Journal of the American Statistical Association* **79**: 871–881.

Rousseeuw PJ. 1985. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Vol. B, Maddala GS, Rao CR (eds). Elsevier: Amsterdam; 101–121.

Sun Y, Genton MG. 2011. Functional boxplots. *Journal of Computational and Graphical Statistics* **20**: 316–334.

Tukey JW. 1970. *Exploratory Data Analysis. (Limited Preliminary Edition)*, Vol. 1, Ch. 5. Addison-Wesley: Reading, MA.

Tukey JW. 1977. *Exploratory Data Analysis*. Addison-Wesley: Reading, MA.

