Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Visualization and Outlier Detection for Probability Density Function Ensembles

Alexander C. Murph[1] | Justin D. Strait[1] | Kelly R. Moran[1] | Jeffrey D. Hyman[2,3] | Philip H. Stauffer[2]

[1]Statistical Sciences (CCS-6), Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, New Mexico, USA

[2]Energy and Earth System Science (EES-16), Earth and Environmental Sciences Division, Los Alamos National Laboratory, New Mexico, USA

[3]Department of Geology and Geological Engineering, Colorado School of Mines, Colorado, USA

Correspondence
Corresponding author Alexander C. Murph, Los Alamos National Lab, Los Alamos, NM
Email: murph@lanl.gov

Abstract

Exploratory data analysis (EDA) for functional data – data objects where observations are entire functions – is a difficult problem that has seen significant attention in recent literature. This surge in interest is motivated by the ubiquitous nature of functional data, which are prevalent in applications across fields such as meteorology, biology, medicine, and engineering. Empirical probability density functions (PDFs) can be viewed as constrained functional data objects that must integrate to one and be non-negative. They show up in contexts such as yearly income distributions, zooplankton size structure in oceanography, and in connectivity patterns in the brain, among others. While PDF data are certainly common in modern research, little attention has been given to EDA specifically for PDFs. In this paper, we extend several methods for EDA on functional data for PDFs and compare them on simulated data that exhibit different types of variation, designed to mimic that seen in real-world applications. We then use our new methods to perform EDA on the breakthrough curves observed in gas transport simulations for underground fracture networks.

KEYWORDS
probability densities, functional boxplot, band depth, discrete fracture networks

1 | INTRODUCTION

Uncertainty quantification (UQ) for simulation sciences often uses *ensemble* data to assess the variation in a simulated output. Ensembles are collections of the same simulation output taken at different parameter settings, initial conditions, or instances of stochastic phenomena [1]. Large-scale simulations are attractive surrogates for complicated physical phenomena that can be difficult or impossible to directly observe, such as fluid (liquid and gas) particle transport through several different types of membranes or fracture systems [2, 3]. The times until breakthrough for a collection of particles can be naturally represented as a probability density function (PDF), which can be viewed as a non-negative, constrained functional data object that integrates to one. The primary aim of this paper is to develop, review, and compare tools for the UQ of an ensemble of PDFs, focusing on exploratory data analysis (EDA) and on identifying and visualizing outliers.

This project was strongly motivated by the need to analyze and detect outliers in an ensemble of breakthrough curves produced from gas particle transport simulations through subsurface fractured media using a discrete fracture network (DFN) model. As pointed out in [4], the breakthrough curves calculated from these transport simulations exhibit wide variability even at identical input values; visualizing breakthrough curves at identical input locations helps in assessing this underlying variability in stochastic transport simulations. An illustrative EDA on such an ensemble is included in this paper; it is meant to provide an example to users in how to apply these tools on real data, as well as highlighting their practical utility.

Much of the early research on data analysis for functional data focused on modeling, clustering, and forecasting, with little attention given to EDA tasks such as visualization and outlier detection [5, 6, 7]. A stronger emphasis on visualization came at the turn of the century, with the phase-plane plot [8], the rug plot [9], the singular value decomposition plot [10], the rainbow plot, the functional bagplot, and the functional highest density region boxplot [11]. Of these methods, only the functional

Abbreviations: PDF, probability density function; NEXT ONE.

bagplot and the functional highest density region boxplot detect outliers, which they do by ordering the curves using the first two principal components from a robust Principal Component Analysis (PCA) on the entire dataset, mapping this ordering back to the functional space, then applying an Inter-Quartile Range (IQR) rule. Outlier detection for functional data has been studied without visualization using likelihood ratio tests and smoothed bootstrapping [12, 13].

Since the developments in [11], boxplots have been applied several times to functional data, with much success. In [14], they were used to highlight regions of reasonable daily wind speeds in West Texas in 2008. In [15], the functional boxplot reveals in what ways COVID mortality rates vary across regions in Italy, highlighting the "usual" rates and visualizing the significant differences observed in the outlier regions. If previous work is any indication, the functional boxplot has strong potential for visualizing ensembles of PDFs. However, as pointed out in an experiment in [16], the curve-ordering methods suggested in [17] applied directly to PDF ensembles may lead to nonsensical results. In this experiment, the methods in [17] chose a median PDF curve that was a clear outlier.

Much like in the univariate setting, determining the mean, first and third quartiles, and outliers for functional data using a boxplot requires some method of ordering the data. Such an ordering for multivariate or functional data is typically done using *data depth*, which numerically quantifies how representative of – or "central" to – the data as a whole a single observation is [1]. There exist several notions of data depth, many of which are reviewed and compared in [13]. Given a choice of depth, the "most central" point, defined as the point with the highest depth, is regarded as the higher-dimensional analogue of the median point in the univariate setting, and functions are then ordered from the median outward according to their depth. In the case of the bagplot and the highest density region boxplot, Tukey depth [18] and the value of a kernel density estimator were used on the first two principal components to order the curves, although these plots could be produced by alternative ordering methods. One complaint on these ordering methods is that they require mapping to a lower-dimensional space, rather than working directly on the functional space. This concern is addressed in [17], where boxplots are constructed according to the band depth and modified band depth measures developed in [19].

Extensions of the functional boxplot from [17] have been applied and developed for special types of object-oriented functional data. These typically rely on modifying the notion of band depth to properly account for the structure of the functions being studied. In [1], band depth is generalized to the *contour band depth*, which is used to calculate depth for isocontours. An analogous development is made in [20] for parameterized curves. The authors in [21] create visualizations for trajectory curves (multivariate processes mapped to a space without the time axis) using the simplicial band depth from [22]. Numerous geometric extensions for functions which exhibit both phase and amplitude variability have been proposed; in [23], the authors perform optimization to obtain separate simultaneous boxplots for phase and amplitude. A more explicit use of the functional shape distribution via the notion of *elastic depth* is developed by [24]. Finally, [25] extends geometrically-obtained functional boxplots to construct shape and orientation boxplots for open and closed curves.

While visualizing several types of functional data and detecting outliers have been heavily researched topics in the past two decades, there has been relatively little attention given to ensembles of PDFs, even though they are extremely common functional data objects showing up in, for instance, yearly income distributions [26], zooplankton size structure in oceanography [27], connectivity patterns in the brain [28], in structural health monitoring data [29], and in many geoscience simulation applications [30, 31, 32]. The authors in [33] discuss methods for EDA for ensembles of PDFs, but focus on different ways to calculate a Fréchet mean and do not address outliers. They visualize the first and second modes of variation of a Functional Principal Component Analysis (FPCA) on an ensemble of PDFs as a tool to analyze their variance.

As far as the authors are aware, the only paper to give direct treatment to outliers in PDF ensembles is [29]. In [29], the authors classify outliers under several different types of data orderings, often transforming the PDF to an unconstrained data type (such as the log quantile density of [28]), applying an $\mathcal{L}^2$ distance from the estimated median as the method for determining depth, and then applying an IQR rule. This work gives special attention to two different types of anomalous behavior a PDF might exhibit: a horizontal-shift outlier and a shape outlier (see Figure 1 for examples of each). The several methods for discovering outliers suggested in [29] are each designed to catch one of these two types of outliers.

This paper addresses two gaps in the literature on EDA for PDF ensembles. First, while [29] discusses several methods to determine data depth for ensembles of PDFs, only one of these methods uses a metric defined directly on the space of probability distributions. Thus, we apply several such metrics to determine data depth, and compare them to two methods from [29] (the PDF metric and one other method, described in Section 2.1) via a simulation study. Second, the authors have not identified any instances where the functional boxplot has been applied to PDF ensembles, perhaps partly due to the issues highlighted in the experiment in [16]. We update the functional boxplot method in [17] using the several notions of data depth studied in this paper.

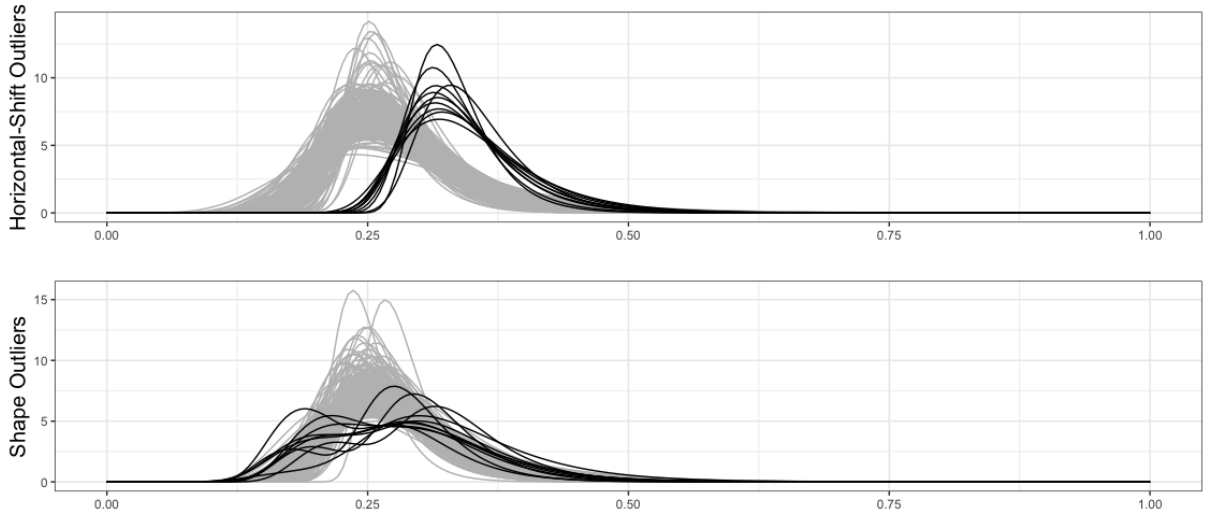

FIGURE 1 Examples of Horizontal-Shift and Shape Outliers, in black.

We study this visual tool's ability to accentuate outliers of various types. Taking what we learn from this simulation study, we apply these tools to analyze an ensemble of PDFs produced from gas transport simulations at Los Alamos National Laboratory.

2 | PDF DISTANCES AND FUNCTIONAL DEPTH

Several of the methods considered here for determining data depth require a pre-processing of the PDF data to some other functional data type. When using a metric to order these functions, we first estimate the median function $m(t)$, then use the metric to order the functions from the median outwards. Let $\Phi = \{\phi_i(t)\}_{i=1,\dots,n}$ be a set of functions with compact unit support indexed by a variable $t \in \mathcal{I} \subset \mathbb{R}$, which in this context will either be a set of PDFs or some transformation of a set of PDFs. When the functions in Φ and the metric chosen forms a Hilbert space, the median can be estimated according to its cross-sections (i.e. the point-wise median),

$$m(t) = \text{median}(\phi_1(t), \dots, \phi_n(t)), \tag{1}$$

where the median operator chooses the median value of the set of functions at a fixed t. The cross-sectional median function $m(t)$ will also belong to the Hilbert space, and is thus treated as the overall median for the remainder of this manuscript.

When the elements in Φ and the metric do not form a Hilbert space, the function $m(t)$ may not be representative of this set. For example, if Φ describes a set of PDFs, the function $m(t)$ need not also be a PDF. When $m(t)$ still belongs to the metric space defined by the metric on $\mathcal{I}$, the overall median is taken to be the element of Φ with the smallest distance from $m(t)$. When $m(t)$ does not belong to this space, we propose two different methods for determining a median: finding the element in Φ that minimizes the $\mathcal{L}^2$ distance from $m(t)$ and treating this element as the overall median, or selecting the overall median to be the function from Φ with the minimal summed distance to all other functions in Φ. The former method, which is suggested in [29], approximates a (perhaps more appropriate) metric on the functional space with the $\mathcal{L}^2$ metric on $\mathcal{I}$, which may or may not be a suitable approximation. We will further assess the use of this approximation in Section 3. The latter method corresponds to the geometric median of Φ, which is a strong choice for the median value, but comes with the additional computational burden of calculating every pairwise distance. Calculating this many distances is on the order of $\mathcal{O}(n^2)$; this is in contrast to the cross-section median, which is only on the order of $\mathcal{O}(n)$.

In addition to the several metrics we consider, we apply the band depth and modified band depth from [19] on the space of log quantile densities (LQDs) in Section 2.3 to determine data depth directly. For this scenario, the overall median is chosen to be the function in Φ with the largest depth value.

For some of the outlier detection methods discussed in [29], an initial horizontal alignment step is performed on the PDFs. In this step, some centrality feature of each PDF (such as the mean, median, or mode) in an ensemble is aligned along the horizontal axis. This is meant to help expose shape outliers by removing horizontal variability. Thus, in addition to assessing the use of the

two median calculations, we assess the use of this horizontal alignment operation in Sections 3 & 4. For this simulation study, and for the subsequent application, we perform this operation by aligning the median values to the midpoint of the collective support of the PDF ensemble, and then renormalizing the PDFs.

2.1 | Non-Probability metrics

The authors in [28] argue that since the space of PDFs is not a linear space, meaning it is not closed under addition and scalar multiplication, it may be more convenient to map a set of PDFs to linear functional space when performing distance calculations. In [29], two such transformations are considered as a pre-processing step before for ordering and detecting outliers in PDF ensembles. In this section, we review and assess these two methods.

2.1.1 | $\mathcal{L}^2$ Distance on the LQD Space

The Log Quantile Density (LQD) transformation, introduced by [28], is a commonly used transformation for analyzing PDFs [34, 35, 36]. Let f be a PDF, and Q be its corresponding quantile function ($Q = F^{-1}$ where $F(x) = \int_0^x f(x)dx$ is the cumulative distribution function (CDF)). The LQD is defined as:

$$\psi(x) = -\log\{q(x)\} = -\log\{f(Q(x))\}, \tag{2}$$

where $q = \frac{d}{dx}Q$. Since the space of LQDs is linear, $\mathcal{L}^2$ distance is a natural choice for performing the median and depth calculations.

The authors in [29] point out that using $\mathcal{L}^2$ on the LQD space, which we refer to as simply "the LQD method" for the subsequence, naturally centers the PDFs, and is thus better for identifying shape outliers than horizontal-shift outliers. For instance, let $c \in [0, 1]$ and note that the horizontally-shifted PDF ($f(x + c)$) maps to a vertically-shifted quantile function ($Q(x) + c$). Thus, both the shifted and non-shifted PDF map to the same quantile function, $q(x) = \frac{d}{dx}(Q(x) + c) = \frac{d}{dx}(Q(x))$. Due to this invariance property, the LQD method is recommended for finding shape outliers over horizontal-shift outliers.

It is recommended in [29] to perform a normalization on the LQD space, $\psi_{\text{norm}}(x) = \psi(x)/\int_0^1 \psi(x)dx$, which makes the outlying curves more distinguishable. This normalization works for the cross-section median approach, since this will estimate an overall median that is part of the original set. However, this does not work when using the geometric median, since it is not immediately clear how to reverse this transformation for a curve that is created directly on the LQD space. For this reason, we normalize the LQDs only when using the cross-section median.

2.1.2 | Bayes Distance

The Bayes Space on a compact interval I, $\mathfrak{B}(I)$, is defined as the equivalence class of positive functions with support I that are equivalent under scalar multiplication. This space is a Hilbert space under the linear operations,

1. $(f \otimes g)(x) = \frac{f(x)g(x)}{\int_I f(t)g(t)dt}$,
2. $(b \odot f)(x) = \frac{f(x)^b}{\int_I f(t)^b dt}$,

with inner product operation,

$$\langle f, b\rangle_{\mathfrak{B}} = \frac{1}{2c}\int_I\int_I \log\left\{\frac{f(t)}{f(s)}\right\}\log\left\{\frac{g(t)}{g(s)}\right\}\, dt\, ds,$$

for $f, g \in \mathfrak{B}(I)$ and $c = \int_I f(x)dx$. Let $\mathfrak{B}^2(I)$ be the subset of functions in $\mathfrak{B}(I)$ that have square-integrable logarithms. As shown in [37], $\mathfrak{B}^2(I)$ is a metric space according to the Bayes metric,

$$d_{\mathfrak{B}}(f, g) = \|f \oplus (-1 \odot g)\|_{\mathfrak{B}},$$

where $\|f\|_{\mathfrak{B}} = \sqrt{\langle f, f\rangle_{\mathfrak{B}}}$. An ensemble of PDFs can be trivially embedded in $\mathfrak{B}^2(I)$ (see [16]), and thus the Bayes metric is a natural notion of distance between PDFs.

Consider the Centered Log Ratio (CLR) map on a function $f \in \mathfrak{B}^2(I)$,

$$\nu_f(x) = \log\{f(x)\} - c^{-1} \int_I \log(\{f(t)\})\, dt, \tag{3}$$

where c is defined as above. As argued in [38], $\nu(x)$ is an isometry to the $\mathcal{L}^2$ space of square-integrable functions. Thus, for computational ease, calculations on Bayes space are done using equivalence $d_{\mathfrak{B}}(f,g) = d_{\mathcal{L}^2}(\nu_f, \nu_g)$.

It is recommended in [29] to shift the ensemble of PDFs horizontally until some feature of each PDF (such as the mode) are aligned in the horizontal direction. As mentioned previously, we will study the merits of using or not using this alignment operation in later sections.

2.2 | Probability metrics

In this section, we review four distances defined directly on the space of probability distributions, which we propose as tools to determine the median value and detect outliers for a PDF ensemble. Note that since these metrics are defined on the space of the PDF data, the cross-section median from (1) calculated directly on this space need not be a PDF. As such, we apply both of the methods discussed above to determine a median from the set of PDFs. For all of the following metrics, let f, g be PDFs with some compact support I, and let F, G be their corresponding CDFs. The metrics considered are:

1. Wasserstein Distance: $d_W(f,g) = \sqrt{\int_0^1 \left(F^{-1}(q) - G^{-1}(q)\right)^2\, dq}$;
2. Total-Variation Distance: $d_{TV}(f,g) = \sup_{\iota \subseteq I} \left| \int_\iota (f(t) - g(t))\, dt \right|$;
3. Hellinger Distance: $d_H(f,g) = \sqrt{\int_I \left(\sqrt{f(t)} - \sqrt{g(t)}\right)^2\, dt}$;
4. Fisher-Rao Distance: $d_{FR}(f,g) = \cos^{-1}\left(\int_0^1 \sqrt{f(t)} * \sqrt{g(t)}\, dt\right)$,

where Total-Variation Distance is equivalent to one half of the $\mathcal{L}^1$ distance between the PDFs as a consequence of Scheffe's Theorem [39]. The Wasserstein distance is computed between the quantile functions (the inverse CDFs), hence the integral being with respect to quantiles q from 0 to 1.

The first two distances, Wasserstein and Total-Variation, each compare location of the area under the curves between two PDFs. The Wasserstein distance can be thought of as the minimum "cost" of moving the area under one curve until it aligns completely with the area under the second curve [40]. Similarly, the Total-Variation distance is equivalent to the one half of the absolute area between the two curves. We suspect that these metrics should each be able to identify both horizontal-shift outliers as well as shape outliers, since each of these outlier types may lead to abnormal differences in the area under PDFs in the ensemble.

The last two distances, Hellinger and Fisher-Rao, are respectively the extrinsic and intrinsic distances between the PDFs when mapped to the positive orthant of a unit sphere on Hilbert space [41, 42]. This map is simply the square-root operation, and Hellinger is the $\mathcal{L}^2$ distance under this operation (i.e., the chord distance), while Fisher-Rao is the geodesic distance under this operation.

2.3 | Band Depth and Modified Band Depth

The visualization used in this paper, the functional boxplot, was initially developed in [17] using the band depth (BD) and modified band depth (MBD) from [19] to order the curves. As pointed out earlier, these visualizations can be produced using any notion of depth or distance between the curves.

Loosely speaking, the BD and MBD of a curve f each measure the proportion of curve pairs in the ensemble which "surround" f. Thus, the curve with the largest depth value is deemed the most "central" in the ensemble. Let $\mathcal{I}$ be a finite set of discrete points in the compact interval I, where I is the support of f. Define the graph of f as the pairs $G(f) = \{(t, f(t)) : t \in \mathcal{I}\}$: the values of the function in $\mathbb{R}^2$ evaluated at the points in $\mathcal{I}$. The band between two curves g_1, g_2 is defined as

$$B_2(g_1, g_2) = \{(t, x(t)) : t \in \mathcal{I}, \min_{r=1,2} g_r(t) \leq x(t) \leq \max_{r=1,2} g_r(t)\}.$$

The sample BD for f is the fraction of pairs of curves in the entire ensemble whose band fully contains f. So, for an ensemble of curves $\{g_1, \ldots, g_n\}$,

$$BD(f) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} I\{G(f) \subseteq B(g_{i_1}, g_{i_2})\}.$$

A more flexible definition of BD is also developed in [19], called the modified band depth (MBD), that measures the overall proportion of time that a given curve exists within the bands created by the curves in the ensemble. Define the set,

$$A_2(f; g_1, g_2) = \{t \in \mathcal{I} : \min_{r=1,2} g_r(t) \le f(t) \le \max_{r=1,2} g_r(t)\}.$$

The MBD is simply the averaged Lebesgue measure of this set for every pair of functions in the ensemble over the Lebesgue measure on the entire set,

$$MBD(f) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \frac{\lambda(A(f; g_1, g_2))}{\lambda(\mathcal{I})}.$$

Note that when a function is not inside a band B_2 at *any* time, there is no contribution to BD. The MBD allows for partial coverage by including the proportion of times a graph of a function is covered by a given band. Note also that the definitions for BD and MDB above focus on sets created by pairs of curves, B_2 and A_2. The general definition for each of these quantities defines these sets for an arbitrary subset of curves, B_J and A_J for $2 \le J \le n$. This extension comes with a great computational burden, so we (and several other authors who use this depth measure) focus only on $J = 2$ [17].

While these approaches have become one of the most popular methods of calculating depth, they are inadequate measures when directly applied to PDFs. For instance, in an experiment in [16], band depth applied directly to a PDF ensemble is shown to select a shape outlier as the median curve. To examine the merits of BD and MBD in this study, we first map the PDFs to their corresponding LQDs before calculating their depth. As we will see in Section 3, this is a strong method for detecting shape outliers, but shows poor performance for horizontal-shift outliers.

3 | COMPARISON OF METHODS

We compare each of the methods in Section 2 by way of two simulation studies. In each study, we investigate a different type of outlier: the horizontal-shift outlier or the shape outlier. Each simulation repeats the following steps 100 times:

1. Simulate 190 'central' Gumbel PDFs using location parameter $(0.25 + a_I)$ and scale parameter $(0.05 + b_i)$ and 10 outlier curves. The outlier curves are drawn in one of two ways:
   (a) horizontal-outliers are drawn from Gumbel PDFs using location parameter $(0.32 + a_i)$ and scale parameter $(0.05 + b_i)$;
   (b) shape outliers are drawn from the mixture $\alpha f_1 + (1 - \alpha) f_2$, where f_1, f_2 are Gumbel PDFs with scale parameters both set to $(0.05 + b_i)$ and location parameters set to $(0.2 + a_i)$ and $(0.3 + a_i)$, respectively.

   In the above, the values a_i and b_i are each independent draws from the normal distribution $\mathcal{N}(0, 0.01)$ for $i \in \{1, \ldots, 200\}$ and the mixture scale α is drawn from a Beta(3, 3) distribution.
2. Outliers are then calculated for every possible method discussed in Section 2: $\mathcal{L}^2$ Distance on the LQD space (L2 on LQD), Bayes Distance (Bayes), Wasserstein Distance (Wasserstein), Total-Variation Distance (TV), Hellinger Distance (Hellinger), Fisher-Rao Distance (Fisher-Rao), Band Depth (Band Depth) and Modified Band Depth (Mod. Band Depth).

The results from a single iteration of this simulation study are visualized in Figure 2. In this single iteration, we already see validation for some of the comments made on the different methods in Section 2. Horizontal-shift outliers are consistently captured by the PDF metrics for both median calculations when the PDFs are not centered, with Wasserstein distance having the lowest false positive rate. The $\mathcal{L}^2$ distance on the LQD space, and the BD & MBD methods perform poorly for horizontal-shift outliers. These three methods perform better when identifying shape outliers, which is expected since these measures are all calculated on the LQD space, which removes horizontal variability. The LQD distance seems to perform better using the geometric median when detecting shape outliers, which is different than the cross-section median estimate used in [29].

We validate these initial observations over the entire simulation study in Figures 3 & 4. For the horizontal-shift outliers in Figure 3, we see that the probability metrics all perform strongly. The best average accuracy, best true positive rate, and lowest average false positive rate are all enjoyed by the Wasserstein distance using the geometric median without centering the PDFs. While the Band Depth approach does have a lower False Positive Rate, it would appear to mostly achieve this by not flagging

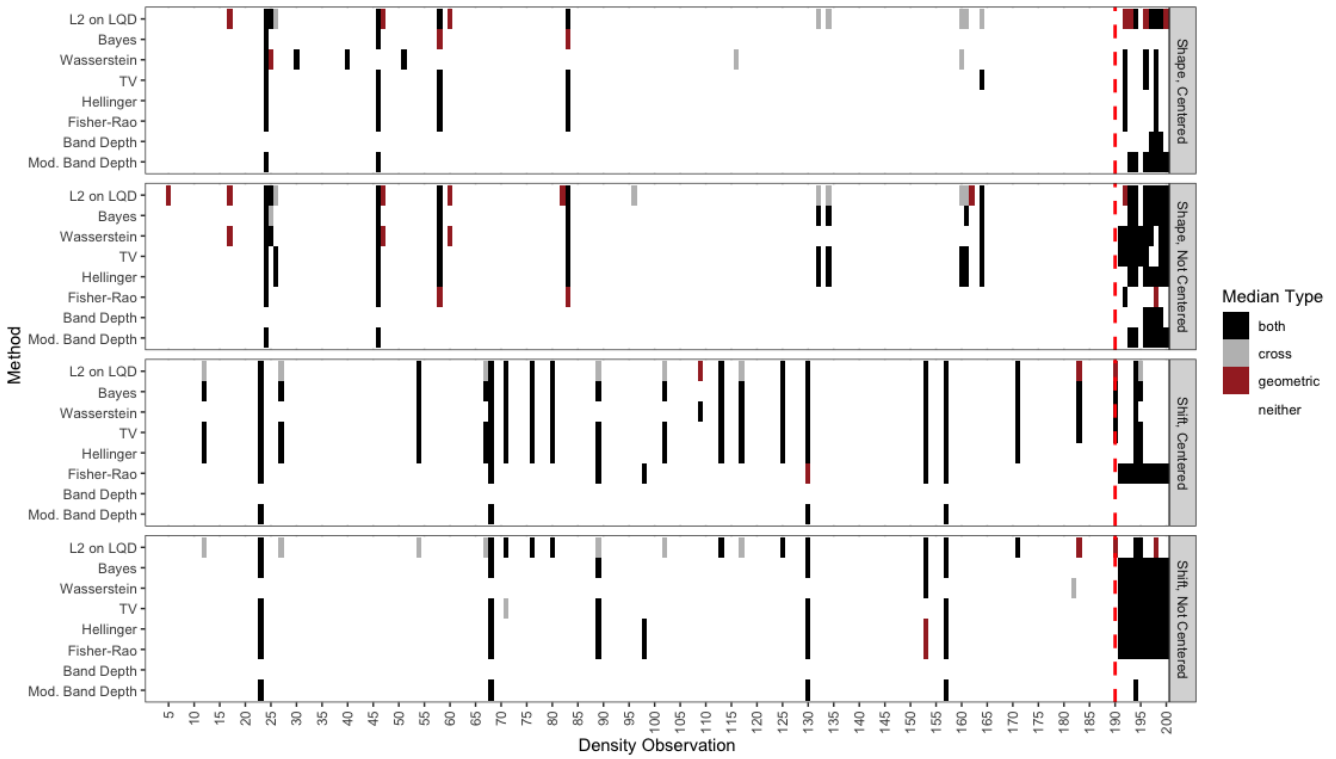

FIGURE 2 The 200 PDFs from Figure 1, using the methods from Section 2 to classify outliers. The last 10 observations correspond to the black curves in Figure 1: the outliers. These curves are separated by a dashed red line. A filled in cell indicates that the particular method detected an outlier, with the color depending on which median estimate led to that PDF being detected as an outlier. A strong method marks (by coloring in) the 10 curves to the right of the dotted line and does not mark the 190 curves to the left.

any curves as outliers. Wasserstein using the cross-section median seems to have around the same average performance, with a slightly wider spread. Based on these results, we would recommend using the Wasserstein distance with the geometric median without centering the PDFs when it is computationally feasible, and substituting the geometric median with the cross-section median whenever data size is an issue.

In Figure 3, we examine the overall performance of each method in detecting shape outliers. There are less clear distinctions on which method is best for this type of outliers. The Wasserstein distance continues to enjoy a strong performance on shape outliers whenever the PDFs are centered. Somewhat surprisingly, the MBD with either median has the best overall average accuracy whenever the PDFs are not centered. This may be because the MBD is calculated after transforming the PDFs into LQDs, which is already equivalent to aligning the PDFs horizontally. Across all methods, the ability to classify shape outliers is less accurate, suggesting that these outlier detection tools should be used in conjunction with subject-matter expertise to ensure no data are erroneously treated as outliers.

Based on these results, our recommendation is to use both Wasserstein distance with the geometric median and the MBD method using the cross-section median, each without centering. The former should identify horizontal-shift outliers and the most pronounced shape outliers, while the latter should be used to identify more subtle shape outliers. We further recommend to try Wasserstein distance with the geometric median and centering, since this method also seems to have strong performance for shape outliers, and can be seen as an intermediary between the two aforementioned approaches.

With the three recommended methods, we assess the functional boxplot of [17] as a method for visualizing PDF ensembles, using the same curves that are summarized in Figure 2. In Figure 5, we draw the functional boxplot for each outlier type and for each of the three recommended methods. In the functional boxplot, the median curve is in black, the center 50% of data are colored yellow, the data outside of this center 50% that are not outliers are shaded light blue, and the outlier curves are visualized with dotted vermilion lines. Note that the colors used here are different than the original colors used in [17]; this was done to make the visualization more color-blind friendly.

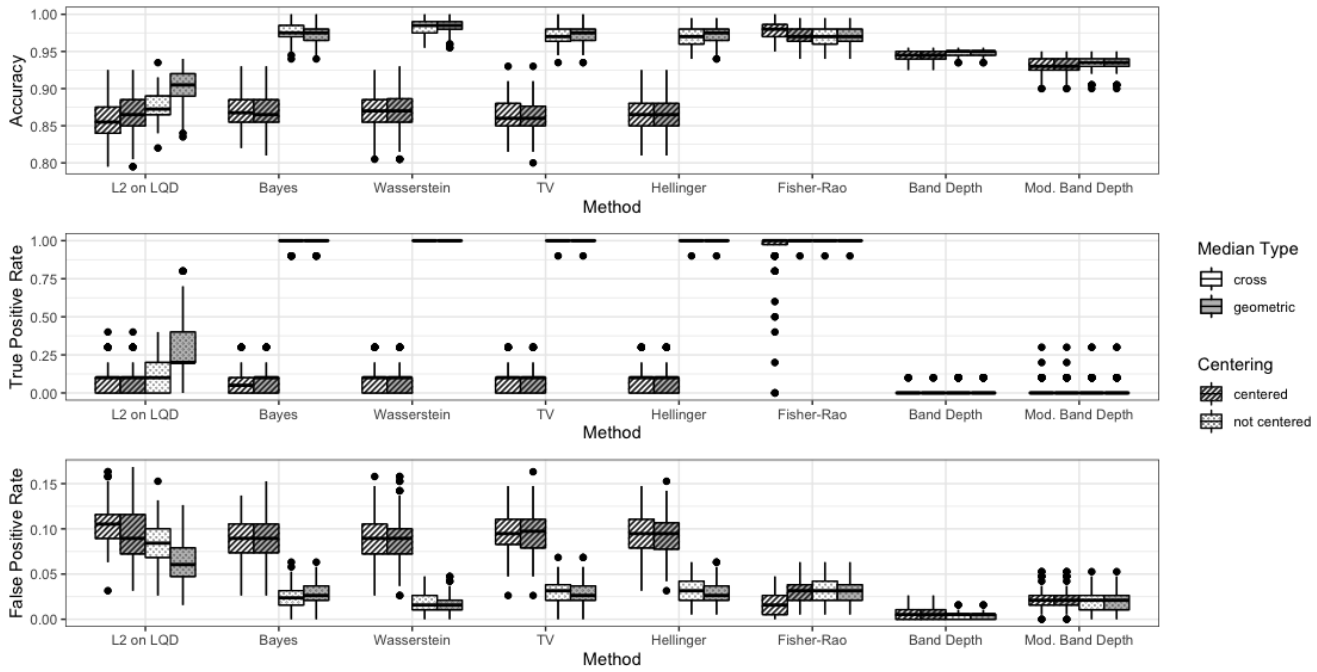

FIGURE 3 Boxplots of classification metrics across the simulation study for the horizontal-shift outlier data. Higher values are better for accuracy and True Positive Rate; lower values are better for False Positive Rate. For every classification metric, we consider the cross-sectional and geometric medians, and whether to center the PDFs.

We identify both strengths and weaknesses of the functional boxplot applied to PDF ensembles. A strength of this visualization is that it permits a user to see how significantly different an outlier is to the central curves, and in what way. For instance, for the data with horizontal-shifted outliers, in the functional boxplot using the geometric median under the Wasserstein distance without centering (the top-left plot in Figure 5), most of the outlier PDFs identified are those which are horizontally shifted to the right. There is one PDF classified as an outlier since it is heavier-tailed (i.e., less prominent mode) compared to the more central PDFs. A researcher using this visual may focus on investigating the significant shift to the right rather than the curve that is just barely classified as an outlier due to heavier tails.

A weakness of this visualization is that the outlines of the shaded regions need not be valid PDFs. This issue is especially pronounced in the functional boxplot for Wasserstein distance using the geometric median with centering on the data with horizontal-shift outliers (the middle-left and and bottom-left plots in Figure 5). The outline of the large shaded yellow regions in each do not describe PDF curves, and may be misleading when trying to visualize a region of valid PDFs.

The functional boxplots applied to PDF data highlight the effects of focusing on shape outliers with a method that is better designed to detect shape outliers. For the MBD and Wasserstein distance without centering methods (the middle-left and bottom-left plots in Figure 5), horizontal variation is not a significant factor in driving outlier detection, so these regions are significantly enlarged in the horizontal direction.

4 | APPLICATION

Modeling the flow of fluids (liquids and gases) and associated solutes through low-permeability rock is the focus of numerous civil and industrial engineering applications, such as in aquifer management, hydrocarbon extraction, and the long-term storage of spent nuclear fuel [43, 44, 45, 46, 47, 48, 2, 49, 50]. Studies have shown that this subsurface transport primarily occurs within interconnected networks of fractures [51]. Since direct observation of such fracture networks in subsurface rock is generally infeasible, large-scale, stochastic simulations are generally used for analysis [2, 3]. In these simulators, a set amount of particles are walked through the system and the time each takes to exit the system is recorded. A PDF is a natural data type to represent the distribution of particle breakthrough times.

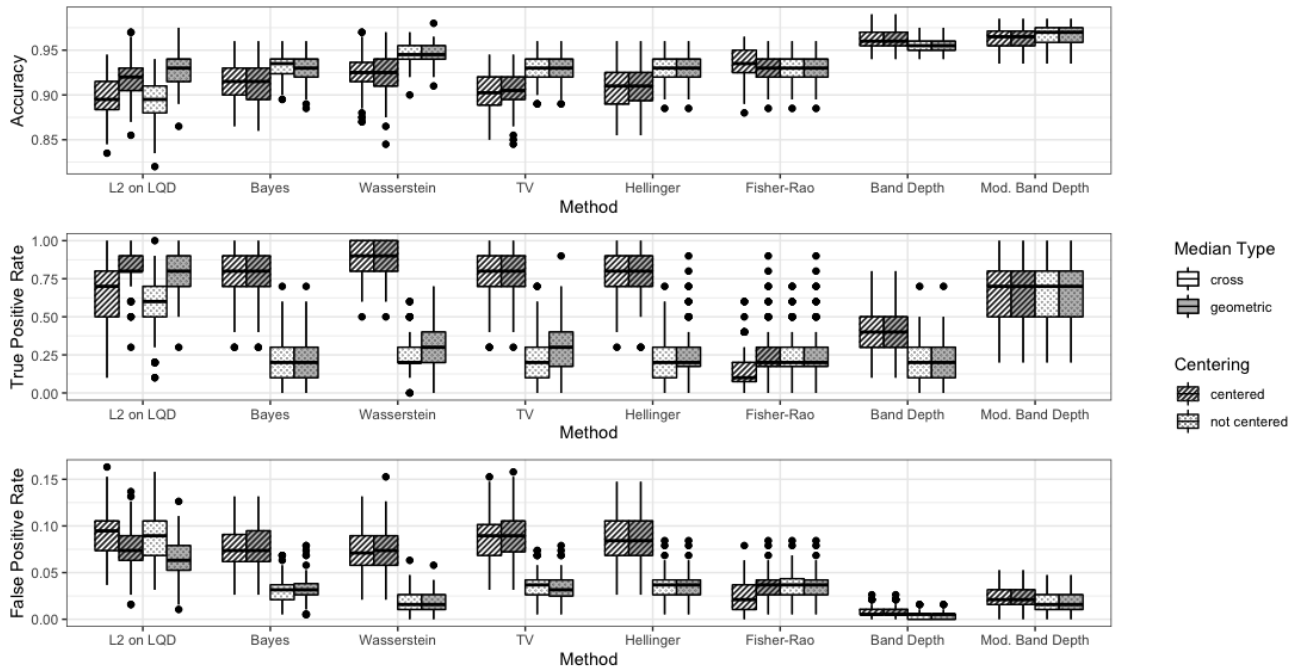

FIGURE 4 Boxplots of several classification metrics across the simulation study for the shape outlier data. For every classification metric, we consider the cross-sectional and geometric medians, and whether to center the PDFs.

Several studies have acknowledged the numerous layers of uncertainty involved in simulating flow through fractured media using the DFN simulators [52, 53, 54, 55, 31, 56, 57, 51]. Visualizing breakthrough curves and classifying outliers are major interests for researchers looking to understand the variation observed in breakthrough PDF data for ensembles of DFN simulation runs.

We analyze an ensemble of PDFs collected over 300 runs of a DFN simulator. For our simulations, we use the DFNWORKS computational suite to resolve flow and gas transport through semi-generic fracture networks and record breakthrough times. These simulations are all run at identical input parameter values, but with different random seeds [58, 59].

The Discrete Fracture Network (DFN) used to produce the PDFs involves three orthogonal fracture families in a 300Lx100Wx100H 3D environment. The simulation places fractures into this environment until a P_{32} value of 0.05 is achieved. The value p_{32} is the sum of surface area of each fracture divided by the volume of the domain. Variation between the hydraulic properties of the fractures is introduced via the following relationship between the fracture radius with the transmissivity:

$$\log(T) = \log(\alpha \cdot r^{\beta}). \tag{4}$$

This equation is referred to as a correlated relationship [60]. For each of the three families, α is set to 10^{-5} and β is set to 0.5. The radius of each fracture is drawn from a truncated power law distribution, with upper and lower cutoffs (r_u; r_0) and exponent γ:

$$p_r(r, r_0, r_u) = \frac{\gamma}{r_0} \frac{(r/r_0)^{-1-\gamma}}{1-(r_u/r_0)^{-\gamma}}. \tag{5}$$

For this application, γ is set to 1.8, r_u is set to 30, and r_0 is set to 10. These parameter choices were deemed suitable by subject matter experts.

Using the findings from our simulation study, we consider three methods for visualizing these data and classifying outliers. To identify anomalous horizontal-shifts, we use Wasserstein distance directly on the space of PDFs, and identify the central point using the geometric median. Since we observed similar performance for this method with and without centering the PDFs, we apply both methods here. To identify anomalous behavior in the shape of our curves, we transform the curves to the LQD space, and calculate the MBD of each curve using as the observed PDF with the smallest $\mathcal{L}^2$ distance from the cross-section median in (1) as the central point. The PDFs were not centered for this third method.

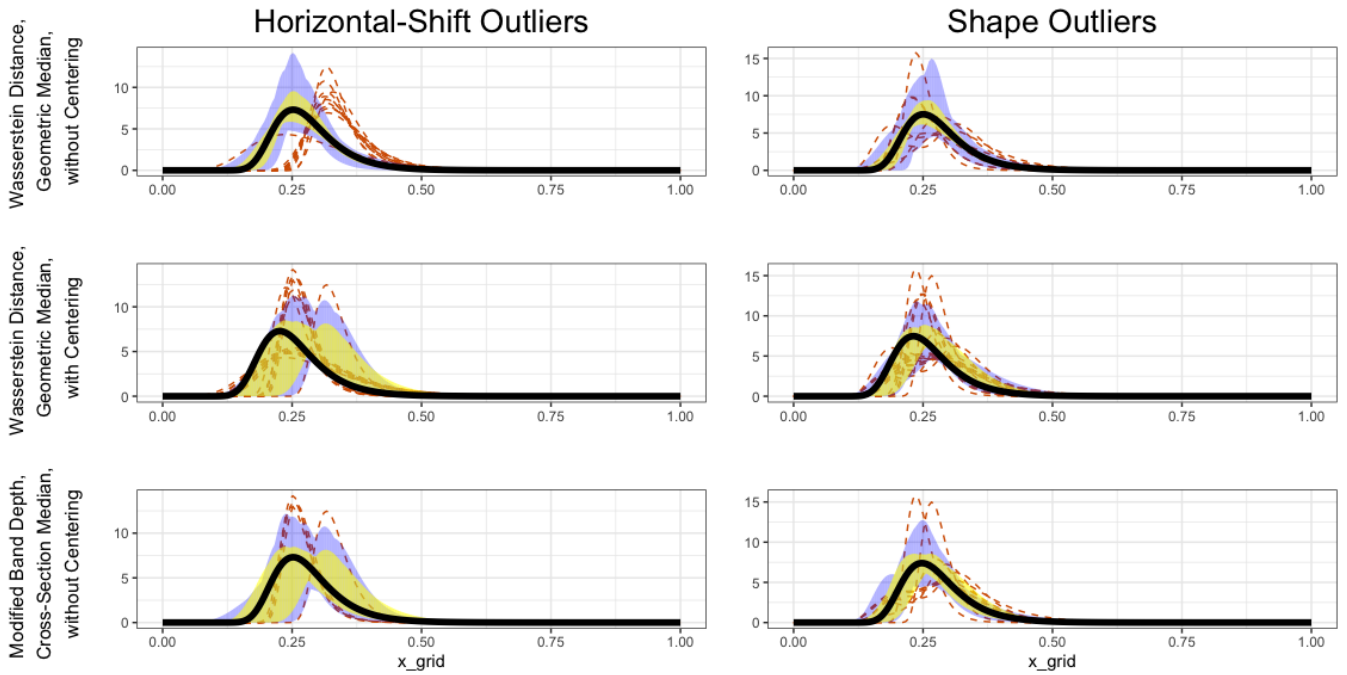

FIGURE 5 The functional boxplots of [17] applied to the simulation study iteration visualized in Figure 2 for the three recommended methods. The median curve is in black, the center 50% of data are colored magenta, the data outside of this center 50% that are not outliers are shaded light blue, and the outlier curves are visualized with dotted red lines.

Three functional boxplots, each according to one of the three methods used for this application, are shown in Figure 6. The top plot, which uses the Wasserstein distance method without centering, isolates several horizontal-shift outliers. In the center plot, this same method is applied with the set of centered PDFs, which appears to isolate several shape outliers. The bottom plot, which uses the MBD method without centering the PDFs, isolates fewer shape outliers than the previous method.

As expected from the simulation study, the Wasserstein distance method without centering (in the top plot) is the best at identifying potential horizontal-shift outliers. This same method with the centered PDFs (in the middle plot) also appears to be a strong method for identifying potential shape outliers. The MBD method without centering (in the bottom plot) is also a strong method for identifying potential shape outliers, but it appears to be a more conservative method as it identifies far fewer outliers than the method in the middle plot.

Once a set of outliers has been identified, detailed probing of the particular realization in the ensemble can be performed to link these observables with geostructural information, such as network topology, geometry, or hydrological properties. This assessment may also lead to further analyses and visualizations of the curves depending on the initial findings. For instance, if a small set of strong outliers are identified with MBD and understood in the context of geostructural information, a further study might attempt the method in the middle plot of Figure 6 to determine if there are additional outliers that can be understood in this context.

5 | CONCLUSIONS

In this paper, we have assessed several different methods for ordering ensembles of PDF data and determining outliers. This analysis compares these methods to two of the approaches outlined in [29] (which are discussed in Section 2.1), and applies these methods to the functional boxplot of [17] as means to study the usefulness of this visualization for PDF ensembles. The methods discussed in this paper consider several distance metrics defined directly on the space of PDFs to order the data, which have seen surprisingly little use in modern literature for outlier detection with PDF ensembles. We also assess two different approaches for determining the most central point, and whether or not the PDFs should be initially centered, as recommended in [29].

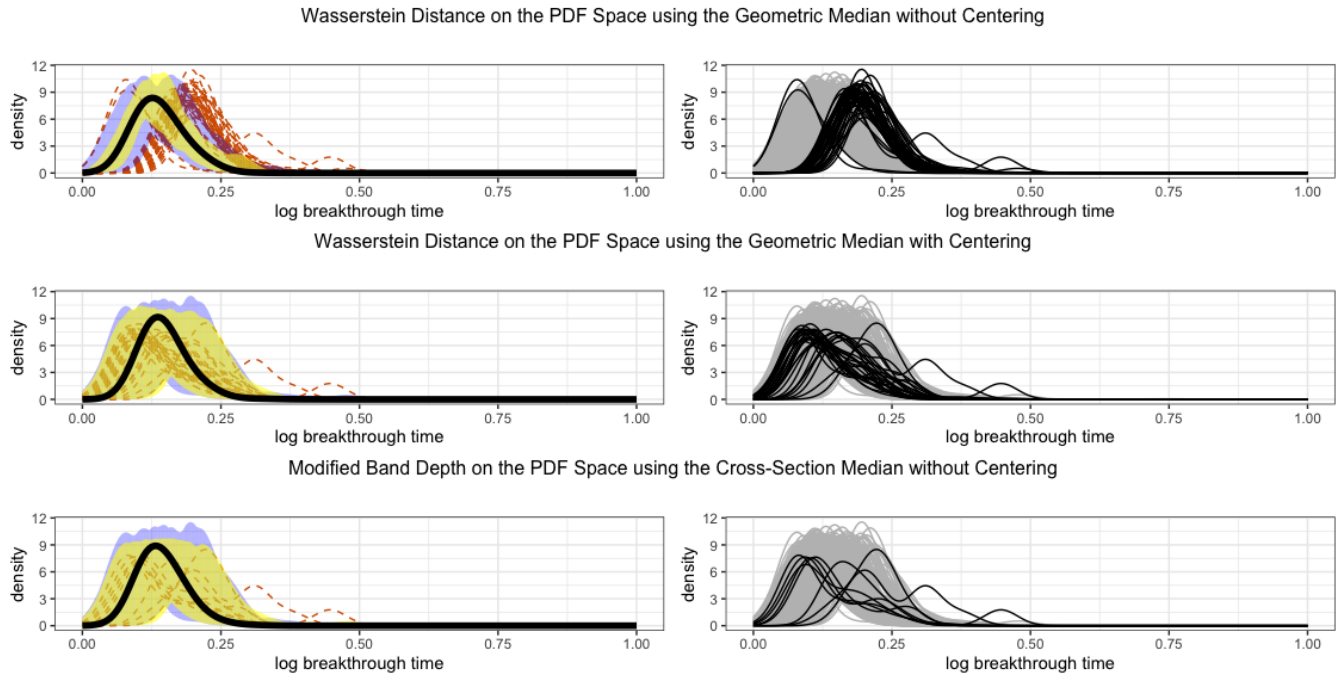

FIGURE 6 Data application on DFNWORKS simulations using the best-performing methods from the simulation study. The functional boxplots on the left are created according to the method outlined in [17]. The median curve is in black, the center 50% of data are colored yellow, the data outside of this center 50% that are not outliers are shaded light blue, and the outlier curves are visualized with dotted vermilion lines. On the right, we show all the PDF curves for the ensemble, with the outlier curves colored in black.

Our findings show that Wasserstein distance with the geometric median is a strong choice for both horizontal-shift and shape outlier detection, centering or not centering the PDFs depending on which type of outlier on which one wishes to focus. For a parsimonious approach to shape outlier detection, our findings indicate that the MBD approach with the cross-section median without centering the PDFs is a strong approach. These recommendations are used to analyze an ensemble of PDF breakthrough curves calculated from DFN simulations at Los Alamos National Lab.

As a part of this work, we have developed an R package to order an arbitrary ensemble of PDFs, determine outliers, and visualize these PDFs via a functional boxplot for every method seen in Figures 3 & 4. This package, called DEBOINR (DENSITY BOXPLOTS IN R), is under review at Los Alamos National Lab and should be available on the Comprehensive R Archive Network (CRAN) in 2024.

FINANCIAL DISCLOSURE

Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20220019DRDen. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

CONFLICT OF INTEREST

The authors have no non-financial or other financial competing interests to declare that are relevant to the content of this article other than the aforementioned declared funding source.

REFERENCES

1. Whitaker RT, Mirzargar M, Kirby RM. Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE Transactions on Visualization and Computer Graphics.* 2013;19(12):2713-2722. doi: 10.1109/TVCG.2013.143
2. Neuman S. Trends, prospects and challenges in quantifying flow and transport through fractured rocks. *Hydrogeol. J.*. 2005;13(1):124–147.
3. Zhang D. *Stochastic Methods for Flow in Porous Media: Coping With Uncertainties*, 2002.
4. Murph AC, Strait JD, Moran KR, Hyman JD, Viswanathan HS, Stauffer PH. Sensitivity Analysis in the Presence of Intrinsic Stochasticity for Discrete Fracture Network Simulations. On ArXiv.; 2023.

5. Ramsay JO, Silverman BW. *Functional Data Analysis*. Springer, 2005.
6. Ferraty F, Vieu P. *Nonparametric Functional Data Analysis: Theory and Practice*. 51. Springer, 2006
7. Fitzenberger B, Koenker R, F. Machado JA. *Economic Applications of Quantile Regression*. Springer, 2002
8. Ramsay JO, Ramsey JB. Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics.* 2002;107(1):327-344. Information and Entropy Econometricsdoi: https://doi.org/10.1016/S0304-4076(01)00127-0
9. Hyde V, Jank W, Shmueli G. Investigating Concurrency in Online Auctions Through Visualization. *The American Statistician.* 2006;60(3):241–250.
10. Zhang L, Marron JS, Shen H, Zhu Z. Singular Value Decomposition and Its Visualization. *Journal of Computational and Graphical Statistics.* 2007;16(4):833–854.
11. Hyndman RJ, Shang HL. Rainbow Plots, Bagplots, and Boxplots for Functional Data. *Journal of Computational and Graphical Statistics.* 2010;19(1):29-45. doi: 10.1198/jcgs.2009.08158
12. Febrero M, Galeano P, González-Manteiga W. A functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics.* 2007;22(3):411–427. doi: 10.1007/s00180-007-0048-x
13. Febrero M, Galeano P, González-Manteiga W. Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics.* 2008;19(4):331-345. doi: https://doi.org/10.1002/env.878
14. Xie L, Gu Y, Zhu X, Genton MG. Short-Term Spatio-Temporal Wind Power Forecast in Robust Look-ahead Power System Dispatch. *IEEE Transactions on Smart Grid.* 2014;5(1):511-520. doi: 10.1109/TSG.2013.2282300
15. Boschi T, Di Iorio J, Testa L, Cremona MA, Chiaromonte F. Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy. *Scientific Reports.* 2021;11(1):17054. doi: 10.1038/s41598-021-95866-y
16. Xinyi Lei ZC, Li H. Supplementary Materials for "Functional Outlier Detection for Density-Valued Data with Application to Robustify Distribution-to-Distribution Regression". *Technometrics.* 2023;65(3):351-362. doi: 10.1080/00401706.2022.2164063
17. Sun Y, Genton MG. Functional Boxplots. *Journal of Computational and Graphical Statistics.* 2011;20(2):316-334. doi: 10.1198/jcgs.2011.09224
18. Tukey JW. Mathematics and the Picturing of Data. *Proceedings of the International Congress of Mathematicians, Vancouver.* 1975;2:523-531.
19. López-Pintado S, Romo J. On the Concept of Depth for Functional Data. *Journal of the American Statistical Association.* 2009;104(486):718–734.
20. Mirzargar M, Whitaker RT, Kirby RM. Curve Boxplot: Generalization of Boxplot for Ensembles of Curves. *IEEE Transactions on Visualization and Computer Graphics.* 2014;20(12):2654-2663. doi: 10.1109/TVCG.2014.2346455
21. Yao Z, Dai W, G. Genton M. Trajectory functional boxplots. *Stat.* 2020;9(1):e289. e289 sta4.289doi: https://doi.org/10.1002/sta4.289
22. López-Pintado S, Sun Y, Lin JK, Genton MG. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification.* 2014;8(3):321–338. doi: 10.1007/s11634-014-0166-6
23. Weiyi Xie KB, Sun Y. A Geometric Approach to Visualization of Variability in Functional Data. *Journal of the American Statistical Association.* 2017;112(519):979-993. doi: 10.1080/01621459.2016.1256813
24. Trevor Harris BL, Shand L. Elastic Depths for Detecting Shape Anomalies in Functional Data. *Technometrics.* 2021;63(4):466-476. doi: 10.1080/00401706.2020.1811156
25. W. Xie SK. Visualization and Outlier Detection for Multivariate Elastic Curve Data. *IEEE Transactions on Visualization and Computer Graphics.* 2020;26:3353-3364.
26. Kneip A, Utikal KJ. Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association.* 2001;96(454):519-542. doi: 10.1198/016214501753168235
27. Nerini D, Ghattas B. Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis.* 2007;51(10):4984-4993. doi: https://doi.org/10.1016/j.csda.2006.09.028
28. Petersen A, Müller HG. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics.* 2016;44(1):183–218.
29. Xinyi Lei ZC, Li H. Functional Outlier Detection for Density-Valued Data with Application to Robustify Distribution-to-Distribution Regression. *Technometrics.* 2023;65(3):351-362. doi: 10.1080/00401706.2022.2164063
30. Stansberry AR, Sweeney MR, Hyman JD, et al. Fracture Network Influence on Rock Damage and Gas Transport Following an Underground Explosion. tech. rep., Los Alamos National Lab; Los Alamos, NM, USA: 2023. LA-UR-23-28644.
31. Strait JD, Moran KR, Hyman JD, Viswanathan HS, Sweeney MR, Stauffer PH. Fracture Network Flow Prediction with Uncertainty using Physics-Informed Graph Features. *Computational Geosciences.* 2023(published online). doi: https://doi.org/10.1007/s10596-023-10256-9
32. Stauffer PH, Lu Z. Quantifying Transport Uncertainty in Unsaturated Rock using Monte Carlo Sampling of Retention Curves. *Vadose Zone Journal.* 2012;11(4):vzj2011.0171. doi: 10.2136/vzj2011.0171
33. Petersen A, Zhang C, Kokoszka P. Modeling Probability Density Functions as Data Objects. *Econometrics and Statistics.* 2022;21:159-178. doi: https://doi.org/10.1016/j.ecosta.2021.04.004
34. Petersen A, Muller HG. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics.* 2016.
35. Chen Z, Bao Y, Li H, Spencer BF. LQD-RKHS-based distribution-to-distribution regression methodology for restoring the probability distributions of missing SHM data. *Mechanical Systems and Signal Processing.* 2019;121:655-674. doi: https://doi.org/10.1016/j.ymssp.2018.11.052
36. Kokoszka P, Miao H, Petersen A, Shang HL. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting.* 2019;35(4):1304-1317. doi: https://doi.org/10.1016/j.ijforecast.2019.05.007
37. Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V. Hilbert Space of Probability Density Functions Based on Aitchison Geometry. *Acta Mathematica Sinica.* 2006;22(4):1175–1182. doi: 10.1007/s10114-005-0678-2
38. Boogaart v. dKG, Egozcue JJ, Pawlowsky-Glahn V. Bayes Hilbert Spaces. *Australian & New Zealand Journal of Statistics.* 2014;56(2):171-194. doi: https://doi.org/10.1111/anzs.12074
39. Tsybakov AB. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2009.
40. Ollivier Y, Pajot H, Villani C. *Optimal Transport : Theory and Applications*. Cambridge, UNITED KINGDOM: Cambridge University Press, 2014.
41. Srivastava A, Jermyn I, Joshi S. Riemannian Analysis of Probability Density Functions with Applications in Vision. In: IEEE. 2007:1-8
42. Miyamoto HK, Meneghetti FCC, Costa SIR. On Closed-Form expressions for the Fisher-Rao Distance. *arXiv preprint arXiv:2304.14885.* 2023. doi: 2304.14885
43. Follin S, Hartley L, Rhén I, et al. A methodology to constrain the parameters of a hydrogeological discrete fracture network model for sparsely fractured crystalline rock, exemplified by data from the proposed high-level nuclear waste repository site at Forsmark, Sweden. *Hydrogeol. J..* 2014;22(2):313–331.

44. Hyman J, Jiménez-Martínez J, Viswanathan H, et al. Understanding hydraulic fracturing: a multi-scale problem. *Phil. Trans. R. Soc. A.* 2016;374(2078):20150426.
45. Jenkins C, Chadwick A, Hovorka SD. The state of the art in monitoring and verification—ten years on. *Int. J. Greenh. Gas. Con.*. 2015;40:312–349.
46. Kueper BH, McWhorter DB. The behavior of dense, nonaqueous phase liquids in fractured clay and rock. *Ground Water.* 1991;29(5):716–728.
47. Middleton R, Gupta R, Hyman JD, Viswanathan HS. The shale gas revolution: Barriers, sustainability, and emerging opportunities. *Applied Energy.* 2017;199:88–95.
48. National Research Council . *Rock fractures and fluid flow: contemporary understanding and applications*. National Academy Press, 1996.
49. Selroos JO, Walker DD, Ström A, Gylling B, Follin S. Comparison of alternative modelling approaches for groundwater flow in fractured rock. *J. Hydrol.*. 2002;257(1-4):174–188.
50. VanderKwaak J, Sudicky E. Dissolution of non-aqueous-phase liquids and aqueous-phase contaminant transport in discretely-fractured porous media. *J. Contam. Hydrol.*. 1996;23(1-2):45–68.
51. Viswanathan HS, Ajo-Franklin J, Birkholzer JT, et al. From Fluid Flow to Coupled Processes in Fractured Rock: Recent Advances and New Frontiers. *Reviews of Geophysics.* 2022;60(1):e2021RG000744. e2021RG000744 2021RG000744doi: https://doi.org/10.1029/2021RG000744
52. Bonnet E, Bour O, Odling NE, et al. Scaling of fracture systems in geological media. *Reviews of Geophysics.* 2001;39(3):347–383.
53. Berkowitz B. Characterizing flow and transport in fractured geological media: A review. *Advances in Water Resources.* 2002;25:861-884. doi: 10.1016/S0309-1708(02)00042-8
54. The National Academies of Sciences, Engineering, and Medicine . *Characterization, modeling, monitoring, and remediation of fractured rock*. National Academies Press, 2021.
55. Hyman JD, Jiménez-Martínez J, Gable CW, Stauffer PH, Pawar RJ. Characterizing the Impact of Fractured Caprock Heterogeneity on Supercritical CO_2 Injection. *Transport in Porous Media.* 2019:1–21.
56. Osthus D, Hyman JD, Karra S, Panda N, Srinivasan G. A Probabilistic Clustering Approach for Identifying Primary Subnetworks of Discrete Fracture Networks with Quantified Uncertainty. *SIAM/ASA Journal on Uncertainty Quantification.* 2020;8(2):573–600.
57. O'Malley D, Karra S, Hyman J, Viswanathan HS, Srinivasan G. Efficient Monte Carlo with graph-based subsurface flow and transport models. *Water Resources Research.* 2018;54(5):3758–3766.
58. Hyman JD, Karra S, Makedonska N, Gable CW, Painter SL, Viswanathan HS. dfnWorks: A discrete fracture network framework for modeling subsurface flow and transport. *Comput. Geosci.*. 2015;84:10–19.
59. Hyman JD, Gable CW, Painter SL, Makedonska N. Conforming Delaunay Triangulation of Stochastically Generated Three Dimensional Discrete Fracture Networks: A Feature Rejection Algorithm for Meshing Strategy. *SIAM J. Sci. Comput.*. 2014;36(4):A1871–A1894.
60. Hyman JD, Aldrich G, Viswanathan H, Makedonska N, Karra S. Fracture size and transmissivity correlations: Implications for transport simulations in sparse three-dimensional discrete fracture networks following a truncated power law distribution of fracture size. *Water Resour. Res.*. 2016;52(8):6472–6489. doi: 10.1002/2016WR018806

SUPPORTING INFORMATION

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.