INTERACTIVE VISUALIZATION OF PROBABILITY AND CUMULATIVE DENSITY FUNCTIONS

Kristin Potter,[1,*] Robert M. Kirby,[1] Dongbin Xiu,[2] & Chris R. Johnson[1]

[1]Scientific Computing and Imaging Institute, University of Utah,
Salt Lake City, Utah, 84112, USA.

[2]Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA.

Original Manuscript Submitted: 02/15/2012; Final Draft Received: 02/15/2012

The probability density function (PDF), and its corresponding cumulative density function (CDF), provide direct statistical insight into the characterization of a random process or field. Typically displayed as a histogram, one can infer probabilities of the occurrence of particular events. When examining a field over some two-dimensional domain in which at each point a PDF of the function values is available, it is challenging to assess the global (stochastic) features present within the field. In this paper, we present a visualization system that allows the user to examine two-dimensional data sets in which PDF (or CDF) information is available at any position within the domain. The tool provides a contour display showing the normed difference between the PDFs and an ansatz PDF selected by the user, and furthermore allows the user to interactively examine the PDF at any particular position. Canonical examples of the tool are provided to help guide the reader into the mapping of stochastic information to visual cues along with a description of the use of the tool for examining data generated from a uncertainty quantification exercise accomplished within the field of electrophysiology.

KEY WORDS: Visualization, probability density function, cumulative density function, generalized Polynomial Chaos, Stochastic Galerkin methods, Stochastic Collocation methods

[*] Correspond to: Kristin Potter, E-mail: kpotter@sci.utah.edu, URL: http://www.sci.utah.edu/∼kpotter

1. INTRODUCTION

In the past two decades, there has been a tremendous growth of interest within the computational science and engineering (CS&E) community concerning the topics of Validation and Verification (V&V) and Uncertainty Quantification (UQ) in the context of numerical simulation results. In 2011 alone, there have been nearly a dozen different workshops, symposia or conference sessions devoted to V&V and UQ. With the advent of such UQ computational techniques as stochastic finite element method [1] and generalized Polynomial Chaos method [2], there is an increasing need to convey UQ results in concise, informative ways. Visualization is the lens often through which scientists investigate their data. In response to the surge of the UQ focus within the simulation community, uncertainty visualization is considered one of the top visualization research problems by the scientific visualization community [3]. In this paper, we provide the mathematical and algorithmic description of a visualization system that can be used for exploring probability density functions (PDFs) and their corresponding cumulative density functions (CDFs). A special feature of our system is that it allows the user to propose a target ansatz PDF against which to present a contour plot of the normed differences between the ansatz and the data. The user can then interactively investigate regions of high deviation to understand the local PDF structure.

1.1 Related work

The display of PDFs and CDFs has a rich history in graphical data analysis. Commonly shown as simple function plots, the display plots either probability versus data value or value versus cumulative probability. The ubiquity of this presentation style makes these plots easy to read, and scientists can easily recognize canonical distribution types. Many other solutions to plotting this type of data have been established that rely on characteristics of distributions, including histograms, steam-and-leaf plots, percentile and quantile plots [4, 5]. A noteworthy example is the boxplot [6], which aggregates a distribution into its quartiles, allowing multiple distributions to be plotted side-by-side.

While these types of displays are prolific, plotting distributions in this way limits the display of multiple distributions as overlays or tables of plots, both of which become quickly cluttered, hard to read and, limit analysis tasks. In addition, for most complex types of data, such as the data we present here, there exists information, such as spatial domain, that is missing when using such techniques.

One example to displaying the spatial information within a data set shows concentration levels of groundwater dispersed through a 3D space as both a color mapped probability density function where location is plotted against concentration, and as a cumulative probability function where location is plotted against time [7]. In each case, the data are color mapped by probability. The user is given a slider to manually explore the data through animation of space, time, or concentration levels. This method resembles traditional 2D techniques for the presentation of distribution data, but incorporates elements of 3D to open exploration of this additional data characteristic.

Approaching the problem from a visualization, rather than graphical data analysis standpoint, opens up a large range of possible presentation techniques; however, the majority of these approaches are not designed with distribution data in mind. Rather than develop visualization approaches specific to distribution data, Luo, Kao, and Pang [8] systematically extend existing visualization methods by defining a set of mathematical operators to transform distribution data into formats appropriate for various visualization techniques. This allows for the direct application of traditional visualization techniques on the manipulated distributions.

In a similar vein, Potter *et al.* [9] look at a collection of data distributions as a volume of data. This allows for the application of volume rendering, isosurfacing and particle tracing of the gradient volume to explore the space of the data. The goals of this approach are less in data analysis but more in a general understanding of the data and guidance towards areas of interest.

A more global approach to analysis calculates various statistical measures including mean, median, standard deviation and kurtosis and encodes these measures through color mapping, surface deformation, and glyphs, displayed per-pixel [10]. The application provides a crosshair probe to allow the user to select pixels of interest and investigate clusters of data with similar statistics. The work is later extended to density estimate volume visualization [11].

Another method for understanding a collection of data distributions is clustering, which finds groups of similar distributions. Bordoloi *et al.* [12] use hierarchical spatial clusters to give a multi-resolution representation of the data distributions, giving the user the ability to interactively display a representative distribution for each cluster at multiple levels of detail. Chlan *et al.* [13] also use the idea of clustering, but develop a glyph to convey characteristics of the represented distributions, such as mean, standard deviation, and extent, as well as an understanding of the type of the distribution.

The difference between these previous approaches and the one presented here is that our goal is to provide an understanding of the collection of distributions through a global comparison measure, rather than displaying individual distributions. Our main focus is on the global display of all data distributions through meaningful measures of difference. We provide, as a secondary tool, an interactive visual display of the individual distributions to enhance local understanding. This approach allows for the quick identification of interesting areas of a data set, as well as an understanding of the characteristics of the underlying distributions across the spatial domain.

1.2 Outline

The paper is organized as follows. In Section 2, we lay out the mathematical details of the work. In Section 3, we present the implementation details necessary to replicate this work, with a description of the features that are available as part of our software package. In Section 4, we present our new methodology applied to several canonical examples on simple domains (to help demonstrate efficacy in easy-to-understand scenarios) and to simulation results of electric potential over a two-dimensional torso slice. We summarize our results in Section 5.

2. DESCRIPTION OF THE MATHEMATICS

In this section we lay the fundamental mathematical groundwork necessary for discussing our visualization system. With this groundwork in place, we can then provide specific implementation details as given in Section 3.

To begin, let us consider a stochastic field $u = u(x, t, \omega)$, which is usually the result of computation of a stochastic problem. Here x is the coordinate in a physical domain $D \subset \mathbb{R}^{\ell}, \ell = 1, 2, 3$, t is the temporal variable, and $\omega \in \Omega$ in a properly defined event space. Since most of our discussions will be based on any fixed location in physical space and time, we will suppress the notion of x and t whenever possible.

The **(cumulative) distribution function** of u is defined as

$$F_u(s) = \text{Prob}(u \leq s), \qquad s \in \mathbb{R}. \tag{1}$$

If u is continuously distributed, which is the case we are considering here, its **probability density function** (PDF), f_u, exists and satisfies

$$F_u(s) = \int_{\infty}^{s} f_u(y)dy, \tag{2}$$

and (if f is continuous at s)

$$f_u(s) = \frac{dF_u(s)}{ds}. \tag{3}$$

2.1 Distances between probability distributions

To alert the viewer to regions of interest within a dataset, we seek to display not just the PDF or CDF directly at any particular point, but rather to display the "distance" between the distributions found in the data and some ansatz distribution posited by the viewer. For two probability distributions, there exist various ways to measure the distance between them. Here we list a few common ones. Let $f(s)$ and $g(s)$ be two PDFs, and d the distance between them.

- L^1 distance:

$$d_L(f,g) = \int_{-\infty}^{\infty} |f(s) - g(s)| ds. \tag{4}$$

- Hellinger distance:

$$d_H^2(f,g) = \frac{1}{2}\int_{-\infty}^{\infty} \left(\sqrt{f(s)} - \sqrt{g(s)}\right)^2 ds. \tag{5}$$

  Note this is written in its squared form.

- Kullback-Leibler divergence:

$$d_{KL}(f,g) = \int_{-\infty}^{\infty} f(s) \log \frac{f(s)}{g(s)} ds. \tag{6}$$

  Note this distance is not symmetric. One could adopt a symmetric version by using $d_{KL}(f,g) + d_{KL}(g,f)$.

In this work, we will primarily display the L^1 and Hellinger distances, although other choices for distance can easily be implemented within the system we provide.

2.2 Deriving distribution functions from polynomial chaos simulations

Here we pay special attention to the generalized polynomial chaos (gPC) method, because it is one of the most widely used stochastic simulation techniques in practical applications. In gPC, the stochastic solution field u is usually expressed in term of multi-variate orthogonal polynomials.

$$u(x, t, \omega) = \sum_{|\mathbf{i}|=0}^{P} \hat{u}_{\mathbf{i}}(x, t)\Phi_{\mathbf{i}}(Z(\omega)), \tag{7}$$

where P is the order of the expansion. Here $Z(\omega) = (Z_1, \ldots, Z_N)$ is a random vector consisting of N independent components. These random variables are used to parametrize the inputs of the underlying stochastic system. Their probability distributions are prescribed prior to the simulation. $\mathbf{i} = (i_1, \ldots, i_N)$ is multi-index with $|\mathbf{i}| = i_1 + \cdots + i_N$. $\Phi_{\mathbf{i}}(Z)$ are N-variate orthogonal polynomials satisfying

$$\int \Phi_{\mathbf{i}}(y)\Phi_{\mathbf{j}}(y) f_Z(y) dy = \delta_{\mathbf{i},\mathbf{j}}, \tag{8}$$

where $f_Z(y)$ is the PDF of the random vector Z, for $y \in \mathbb{R}^N$, and the Kronecker delta function satisfies $\delta_{\mathbf{i},\mathbf{j}} = 1$ if $\mathbf{i} = j$, and $\delta_{\mathbf{i},\mathbf{j}} = 0$ otherwise. The orthogonality relation (8) establishes a connection between the type of the orthogonal polynomials and the PDF of Z. For example, Gaussian PDF in the orthogonality defines the Hermite polynomials, uniform PDF defines the Legendre polynomials, *etc.* Such connections were recognized and systematically studied in [2].

The key quantities in the gPC expansion (7) are the expansion coefficients $\hat{u}$. These are quantities of physical space and time, and their evaluations require full-scale numerical simulations. The computations of the coefficients usually can be accomplished by two type of approaches. One is a stochastic Galerkin method and the other is a stochastic collocation method. Their implementation will depend on the underlying stochastic problem. Each has its own advantages and disadvantages. Here we will not devote more discussions on the details of Galerkin and collocation. Interested readers are referred to [14].

Once the gPC expansion (7) is obtained, it is straightforward to derive the statistical properties of the solution u. This is because the expression (7) is of an analytical form. The quantities we are interested in are the PDF and CDF of the solution u. While it is possible, in principle, to derive the distributions of u analytically based on the distribution of Z, the procedure is usually of little practical meaning because the derived expression is not of an explicit closed form. In practice, it is usually more straightforward to conduct the following operations to estimate the PDF.

i. Generate a large number of samples of the random vector Z. That is, draw independent samples $Z^{(1)}, \ldots, Z^{(M)}$ from the distribution of f_Z, where $M \gg 1$ is the total number of samples.

ii. For each $m = 1, \ldots, M$, evaluate the gPC expansion and obtain the solution ensemble

$$u^{(m)} = \sum_{|\mathbf{i}|=0}^{P} \hat{u}_{\mathbf{i}}(x,t)\Phi_{\mathbf{i}}(Z^{(m)}), \qquad m = 1, \ldots, M. \tag{9}$$

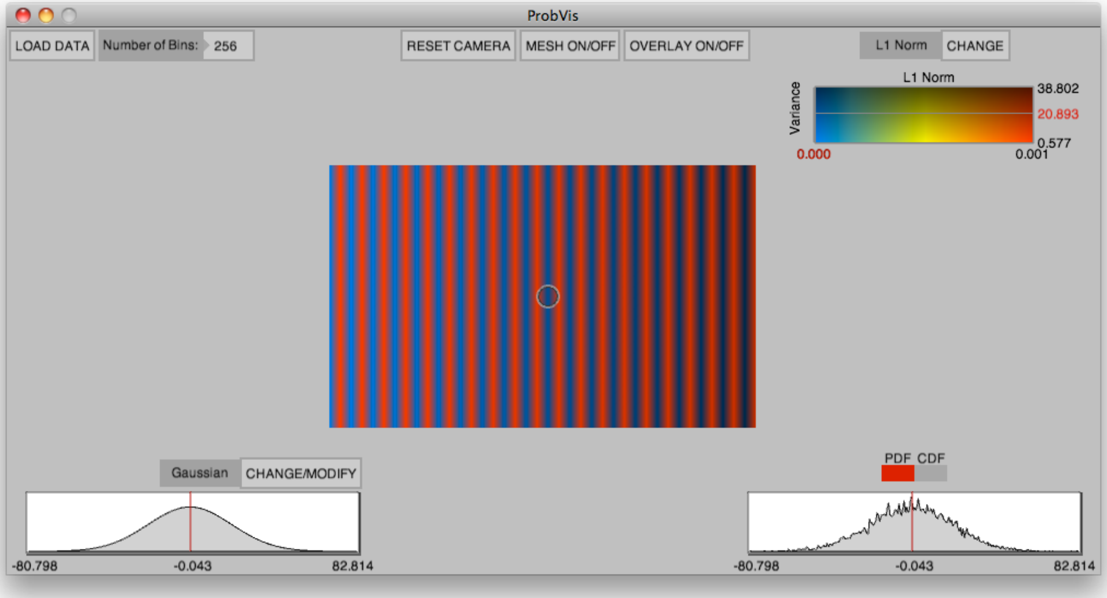
Note this step requires only evaluations of a polynomial expression repetitively. No simulation of the underlying stochastic system is required.

iii. Based on the solution ensemble $\{u^{(m)}\}_{m=1}^{M}$, estimate the PDF of u. This can be done in various ways, with the most popular choice being the kernel density estimation [15, 16].

Hence, whether given directly-sampled simulation data obtained through Monte Carlo methods or given implicitly sampled data from the stochastic Galerkin or collocation approaches, we can now build a discrete representation of the PDF or CDF as a histogram with a user-specified bin size. This is in fact the presumed input of our visualization system: a (discrete) histogram representing the PDF of our function of interest given as each point (for instance, each vertex of our mesh) in physical space.

3. IMPLEMENTATION DETAILS

With our mathematical fundamentals now in place, in this section we present the implementation details and the corresponding visualization software system, ProbVis.



Number of Bins: 256

Figure 1: An overview of the ProbVis system using synthetic data which alternates between a Gaussian and Uniform distribution across the x-axis. The spatial domain of the data is shown centrally. A color map encodes the difference measure and the precise value of the measure is shown as crosshairs in the colorbar (upper right). A pointer allows for the investigation of individual data points, which are displayed at the bottom right as a PDF or CDF, and the comparator distribution is shown at the bottom left.

3.1 Overview

We have created a visualization tool called ProbVis for exploring differences between distributions across a spatial domain. That is, we have a single data distribution at each point across a spatial mesh. The goal of ProbVis is to be able to quickly understand the variations across the spatial field and further explore the data through a series of interactions. To this end, we have defined a distance measure that incorporates two distinct comparison characteristics. We encode this measure using a two-dimensional color map, coloring each point of the spatial domain with this measure. We then provide a visualization of the data distribution at a point selected by the user, as well as control over the distribution against which all data distributions are compared. A screenshot of the ProbVis system can be seen in Figure 1.

Figure 1 shows a screenshot of the ProbVis system displaying an exemplary data set which modulates between Uniform and Gaussian distributions defined across a rectangular grid. The data set is displayed centrally and color mapped based on the currently selected distance function. In this image, the L1 Norm is being used as the distance function and a Gaussian distribution is used as the canonical comparison. Thus, where the data is color mapped blue the distance measure is close to zero, indicating the data is a Gaussian at that location, and conversely, red indicates the locations of Uniform distributions. At the center of the image is a small circle which is controlled by the user and is used to select specific locations within the data and reflect that location into the sub-display on the bottom, right. Here, a traditional plot of the distribution function is shown, with mean, minimum and maximum notated. As the user moves the circle picker, the data is updated in this singular display.

3.2 Comparing distributions

As previously described, in order to compare distributions, we have decided to employ both the probability distribution function (PDF) and the cumulative distribution function (CDF). A PDF describes the probability of a random variable taking a particular value within an interval. A CDF describes the probability that the random variable will be less than or equal to a particular value. We incorporate both ways of looking at a data distribution because, while interrelated, some scientific fields prefer to look at data in one way rather than the other.

3.2.1 Formulation of the PDF

We use a histogram to estimate the PDF of the incoming data. To facilitate flexibility, we allow the user to select the number of bins to use for the histogram, which controls the size and number of features exposed in the distribution estimation. The calculation of the histogram iterates through each sample point (obtained directly from Monte Carlo type sampling or implicitly through evaluation of the stochastic Galerkin or collocation expansions, as discussed in Section 2.2) and determines in which bin the sample point lies by transforming the point from the interval in which the data lies into the histogram space which is controlled by the number of bins. Then, the number of points in each bin is counted and divided by the number of bins. This value is used as a density estimate of the data distribution.

3.2.2 Formulation of the CDF

To estimate the CDF, we begin from the histogram estimation of the PDF, as described above. For each position in the interval in which the original data exists, we sum the probabilities of the PDF and divide by the number of bins. We use the same number of bins as the histogram and again allow user control over this parameter.

3.2.3 Comparator Distributions

To evaluate the similarity of a collection of distributions defined across a spatial field, we compare each distribution to a canonical distribution, and use a measure of difference between the canonical and data distributions as the measure of similarity between each of the data distributions. By default, we allow the user to choose between a uniform, normal or beta comparator distribution, however the system can be extended to use any distribution. To form an appropriate comparison distribution, parameters are chosen by finding related statistics from the original data distributions.

Uniform The uniform distribution is a distribution in which all intervals of the same length, within the distribution's support, are equally likely. Because there are no assumptions or restrictions enforced on the data, in order to form an appropriate comparator distribution, we normalize the uniform distributions by using an interval of support from the data distributions. Thus, at each point in the spatial domain, a uniform distribution is generated using an interval taken from the data distribution at that point. The uniform distribution is estimated by calculating the PDF:

$$f_u(s) = \begin{cases} \frac{1}{b-a} & \text{for} \quad \mathrm{a} \le s \le \mathrm{b} \\ 0 & \text{for} \quad \mathrm{a} > s \text{ or } \mathrm{b} < s \end{cases}$$

where a and b denote the left and right extents of the interval respectively. Alternatively, one could specify mean and variance or midpoint and half-length of the interval. All three of these specifications uniquely determine the uniform distribution.

Normal A normal, or Gaussian, distribution describes a first approximation to a real-valued random variable that clusters around a single mean value. To form a normal distribution against which to compare, we take the mean, given by μ, and standard deviation, given by σ, of the original data distribution and use those values in the calculation of the PDF as follows:

$$f_u(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(s-\mu)^2}{2\sigma^2}}$$

Using the mean and standard deviation from the original data ensures that the mean of the generated Gaussian is the same as the mean of the data and that the standard deviation is contained within the same interval on which the data is defined.

Beta The beta distribution is a class of distributions defined on the (0,1) interval and controlled by two positive shape parameters. The PDF of the beta distribution is given by:

$$f_u(s) = \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}$$

where α and β are shape parameters greater than zero and B is the beta function defined as:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

The shape parameters describe the look of the distribution and are derived from our data distributions. To estimate the shape parameters, we use method-of-moments estimates [17],

$$\alpha = \mu(\frac{\mu(1-\mu)}{v} - 1)$$

$$\beta = (1-\mu)(\frac{\mu(1-\mu)}{v} - 1)$$

where μ is the sample mean and v is the sample variance, or σ^2.

Data-driven parameters A choice made in this work is to not allow the user to change the parameters of the canonical distributions. While our first intuition was to give the user control over these parameters, two problems to this approach arose. The first problem is that when the data at each spatial point is defined across drastically different domains, it is not clear how to specify appropriate parameters for the comparison distributions that will not result in large variations in the difference measure. For example, when using the uniform distribution, the choice of support should be similar to the data values, otherwise, for samples of the uniform distribution outside its support, the comparison will only be the distance of the data value to zero. Similarly for the Gaussian, centering the mean near the mean of the data will ensure that the difference will measure how close to the shape of a Gaussian the data is, rather than emphasize the difference between the means. While this can be viewed as "normalising" in that we are now comparing against similarly appropriate distributions with identical means, variances and supports, going back to allowing the user to modify the underlying parameters again is not straightforward. In this case, we could easily bring up a dialog box to allow parameter tweaks, but the question then arises - are these tweaks reflected globally to all comparison distributions, or is this applied only locally to the current distribution under the pointer? If global, how should the changes to the local distribution be reflected to the rest of the canonical distributions? If only local, how does changing a single comparison distribution change the difference measure across the entire spatial domain? Because the goal of this work is to view the differences between distributions across a spatial domain, we reject the idea of manipulating the parameters of a single, solitary distribution. Likewise, reflecting parameter manipulations across an entire collection of distributions seems inappropriate because it is not clear as to how to reflect those changes. Thus, we have decided that simply deriving an appropriate comparator distribution is the most appropriate choice, and rely on the local views of the comparison and data distributions to provide insights into the local nature of the data.

3.2.4 Shape Measure

The first method we use to compare distributions is a shape measure. Here, we want to determine what the difference is in the shape of the distribution. For this, we use discretizations of the L^1 and Hellinger distances defined in Section

2.1. The discretized L^1 distance is defined as

$$\tilde{d}_L(f,g) = \sum_{i=0}^{n} \frac{|f_i - g_j|}{n}$$

where f and g are the distributions (defined as a PDF or CDF) and n is the number of samples.

Similarly, the discretized Hellinger distance is defined as

$$\tilde{d}_H(f,g) = \sqrt{\left(\sum_{i=0}^{n} \frac{(f_i - g_j)^2}{2n}\right)}$$

To calculate these distances, we compare each bin of the histograms of the data distribution and the canonical distribution. We sum the difference and divide by the number of bins to get a single distance value for each distribution across the spatial domain.

3.2.5 Interval Size

Another measure of comparison is the size of the interval over which the data distribution is defined. A distribution with a larger interval will have a larger range within which the variable value lies, and thus the probability of a particular point is diminished. To evaluate the size of the distributions, we find the minimum and maximum values of each distribution and define the measure as

$$r_i = max_i - min_i.$$

We use both the shape and interval measures to quantify the difference between distributions.

3.3 Visualization

The goal of the visualization is to quickly compare the data distributions across the spatial domain. Our approach leverages a color map to convey the difference measure such that areas of similarity all retain the same color, while regions of difference quickly stand out. This piece of the visualization provides a way to display of all distributions simultaneously, which is suitable for 2D presentations such as publications. In the stand-alone application, we also provide interactive features. These include a pointer which can be moved to each grid point in the spatial domain, and the corresponding data distribution is displayed as a traditional PDF or CDF plot. We also provide the user with the ability to change the comparison distribution or the number of bins used to calculate the histogram.

3.3.1 Color mapping

To express the difference measure discussed above, we generate a surface based on the spatial domain of the data, and color map each point according to its associated differences. Our difference measure is composed of two values, a shape measure and an interval measure. We use a two-dimensional color map to simultaneously display these values across the domain, as shown on the left side of Figure 2. The color map encodes the shape measure across color and the interval length is displayed as a change in value of the shape measure color. This leads to darker (more black) colors that have a longer interval length and lighter (less black) colors with a short interval. This approach corresponds to the idea that the longer the interval the less strong the probability of any particular point within that interval, and thus the darker, or less emphasized, the color of that point.

While the simultaneous display of both values of the difference measure is concise, it can be problematic to precisely identify variations in color versus variations in value. This can be seen in Figure 2, left, where the increasing darkness of the blue lines is somewhat subtle. While this data set is regularized in that variance is increasing from

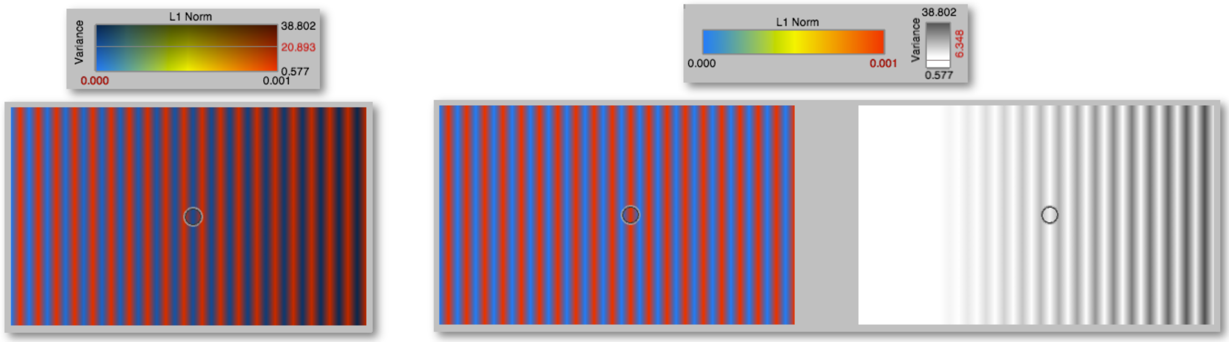


Figure 2: Left, two-dimensional color map displaying the two values, of the difference measure. The color of the position encode the value of the shape measure, while the trending towards black expressed the size of the interval. On the right, the two distance measures are displayed using two separate color maps. Again, color is used to show the shape measure and brightness is used to show variance, however the measures are plotted separately.

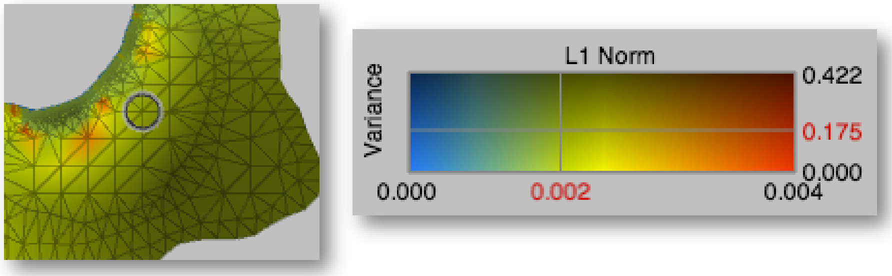


Figure 3: The colorbar, right, displays the range of the two-dimensional color map as well as a set of crosshairs to explicitly display the distance measure at the user-selected point, as shown on the left.

left to right, other data sets may not have such structure and detecting subtle differences between brightness may be difficult. To alleviate this problem, we have provided a toggle to switch from an overlaid display to a side-by-side display where the shape and interval measures are shown separately, as seen in Figure 2, right. Similarly, to the simultaneous display, shape is encoded though color, and interval length as a grayscale ramp which corresponds to changes in value.

One of the problems with using color maps is the inadequacy of the color map as a tool to effectively interpret quantitative data values [18]. Color maps convey a general idea of the data value; a viewer sees a color in the data space and subsequently matches that color in a colorbar. From the legend of the colorbar, the user must roughly guess the precise quantitative value. To facilitate a precise quantitative understanding, we have given the user a pointer which can be moved around the spatial domain. The position of the pointer is then reflected as crosshairs in the color map, as shown in Figure 3. From this exploration, the user can access the precise values of the difference measure.

3.3.2 Display of PDF and CDF

The color maps described above display the difference measure between data distributions and a canonical distribution in a global way; every distribution is represented as the two values of the difference measure, and these values are concurrently displayed. This type of view shows general trends, clusters and discontinuities across the data space, which leads to the need to more fully investigate features of the data. Thus, the user needs a way to start exploring the individual data distributions. Unfortunately, displaying all of the data distributions at once leads to massive visual clutter and an unreadable display, a leading reason for our visualization approach to aggregate the distributions through a distance measure.

Because of the complexity of showing detailed information about each data distribution, we have chosen to provide the user with a pointer to select individual data locations. As stated above, the location of this pointer is reflected in the colorbar giving precise values of the difference measure. In addition, we display the PDF (or CDF) of that single data distribution, as well as the PDF (or CDF) of the comparison distribution. This can be seen in the lower left and right hand corners of Figure 1.

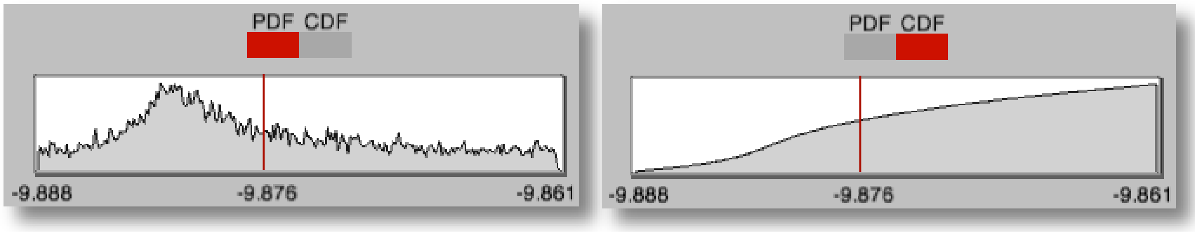


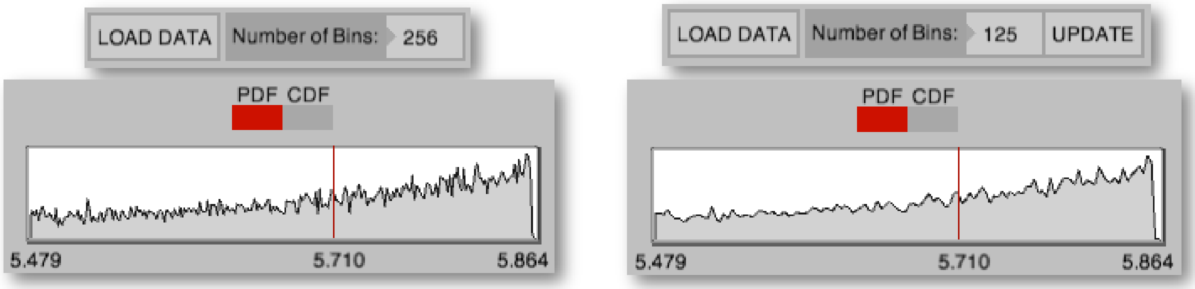
Figure 4: The individual data distributions can be plotted as either a PDF (left) or a CDF (right).

Traditional displays of PDFs and CDFs plot probability (or accumulated probability) versus location as a graph. We use this approach as well, showing an individual data distribution as either a PDF or CDF plot. The user can toggle between the two, and this choice is reflected in both the display of the data distribution as well as the comparative distribution. The difference between a PDF and a CDF of the same data set can be seen in Figure 4. In this image, the PDF of the data set is shown on the left. The user can see the density of the distribution is highest just left of the mean value. On the right, the CDF of the same data is show. Here, the user can see the accumulation of density across the data values.

3.4 Interaction

We have designed ProbVis to be a general tool for the exploration of a collection of data distributions. To this end, we provide the user with a variety of interaction devices to allow them to investigate their data in a manner in which they are comfortable.

3.4.1 Histogram estimation



Number of Bins: 256

Number of Bins: 125

Figure 5: When the number of bins is changed, an update button appears allowing the user to re-calculate the underlying histogram which is used as an approximation to the PDF.

The PDF and CDF are continuous functions describing the behavior of a distribution. In order to represent these functions computationally, we must approximate them through some sort of kernel density estimator. In this prototype, we use the histogram as our approximation. The formulation of the histogram sorts the data into buckets where a data point falls into a bucket if its value lies within the interval range of that bucket; the interval of a bucket being an equally sized partition of the data domain. The value of the histogram at each bucket is then taken to be the number of data samples within the bucket over the total number of samples. This formulation is highly reliant on the number of buckets. As shown in Figure 5, as the number of bins estimating a distribution decreases, the smoother the histogram approximation. Because of this sensitivity, we have given the user the ability to change the number of bins used for the histogram estimation. This is particularly important because the size of features within data sets change and an arbitrary number of bins may miss key characteristics of the data.

3.4.2 Comparative distributions

To allow for the investigation of general data distributions, we provide the user with the ability to choose the form of the comparative distribution. Three canonical distribution types are provided; these are displayed in Figure 6. A uniform, Gaussian, and beta distribution can be chosen and the user is relied upon to decide which is the most appropriate. The choice to include these particular distributions in ProbVis is semi-arbitrary, these types of distributions

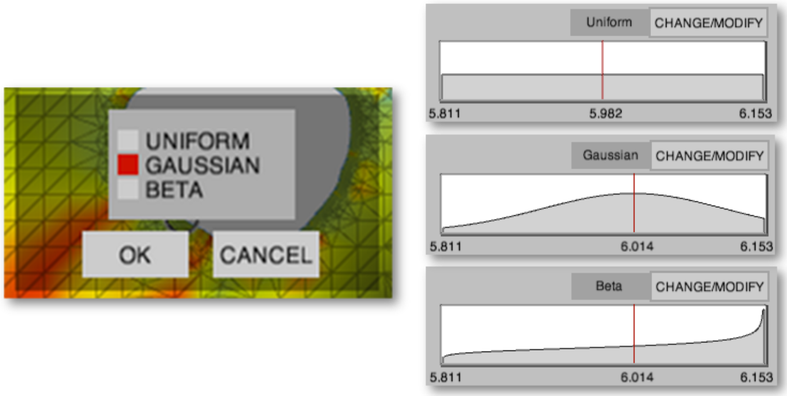


Figure 6: The user is given options to choose the canonical distribution against which the data is compared.

are commonly used to describe simulation results, however they are by no means the only distributions users are interested in. In fact, the ProbVis system supports the use of other canonical distributions through small extensions to the source code.

3.4.3 Shape measures

We have implemented two difference measures for shape - the L^1 and Hellinger distances as defined in Section 2.1. The user can switch between the two measures – the results of which can be seen in Figure 7. We have chosen these measures because they are standard measures of difference and the user may be more familiar or comfortable with one versus the other. While our choice of Hellinger and L^1 distance was motivated by wanting to compare PDFs, other distance measures may be more desirable for other applications or data sets. For example, a user may be more interested in measures focusing on divergence rather than distance and thus use a measure from the family of contrast measures rather than the distance metrics we have implemented thus far. Again, the software system can be easily extended to support these other measures of shape to enable domain-specific exploration.

3.5 Implementation

The ProbVis system presented in this paper is implemented using the Processing programming language [19], which encapsulates a Java-based environment for fast prototyping. All of the graphics are implementing using OpenGL

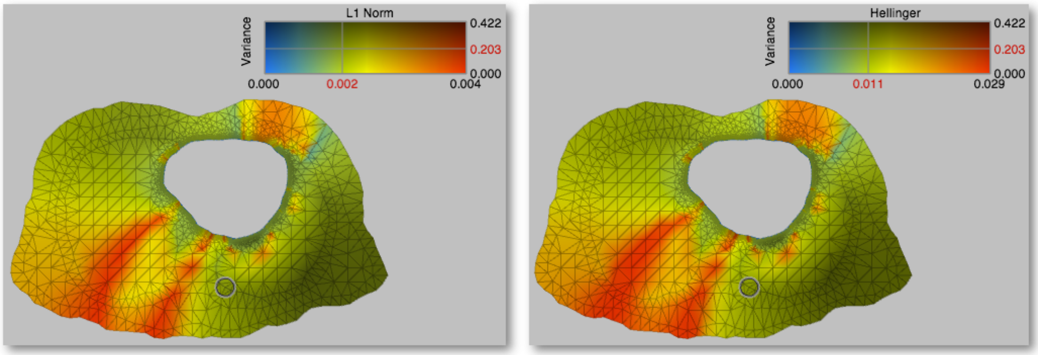


Figure 7: Visualizations using the L_1 (left) and the Hellinger (right) distances.

libraries. The software and data is freely available at http://www.sci.utah.edu/research/visualization/422-uncertainty-vis.html.

4. RESULTS

In this section we present a demonstration of the features of our density function system. We first present a collection of canonical examples on regular domains to help demonstrate particular features of our visual mapping and how they are to be interpreted by the user. We then present images generated by our software system when applied to a application in electrophysiology. This example involves the solution of the elliptic bioelectric forward problem on an unstructured triangular finite element mesh in which traditional linear finite elements are used for the discretization in space and gPC stochastic collocation is used to represent the stochastic variation. Note the underlying stochastic problem and solution technique is not very important in our demonstration. The visualization requires only the solution fields expressed by finite element approximation and the gPC approximation (with corresponding probability density functions).

4.1 Canonical Examples

To help the reader understand the mapping of stochastic information to visual cues as discussed in Section 3, we have constructed a two-dimensional rectangular mesh over which we have specified a function with known PDF. We have constructed two such examples, one to demonstrate variation in the shape and one in the interval measure.

Spatial Domain We have constructed a two-dimensional rectangular mesh over which we define a collection of distributions. To create the mesh, we first define a set of points regularly latticed across the spatial domain. We then triangulate the set of points by choosing four neighboring points and creating two triangles.

Shape To demonstrate the shape measure (either the L_1 Norm or the Hellinger distance) we have created a data set which is a linear blend from a Gaussian to a Uniform distribution. This is demonstrated in Figure 8. The left image in the figure shows the overlay view with the pointer towards the left side of the spatial domain. The dark blue color under the pointer shows a large value in the variance direction, but a small value (0.0) in L_1. As seen in the PDF display, the data looks very much like a Gaussian. On the right, the pointer is on the right side of the spatial domain. This indicates that the data distribution is less of a Gaussian than the previous pointer location. In addition, and as seen in the PDF display, the distribution is close to uniform, thus the variance of the data is very low.

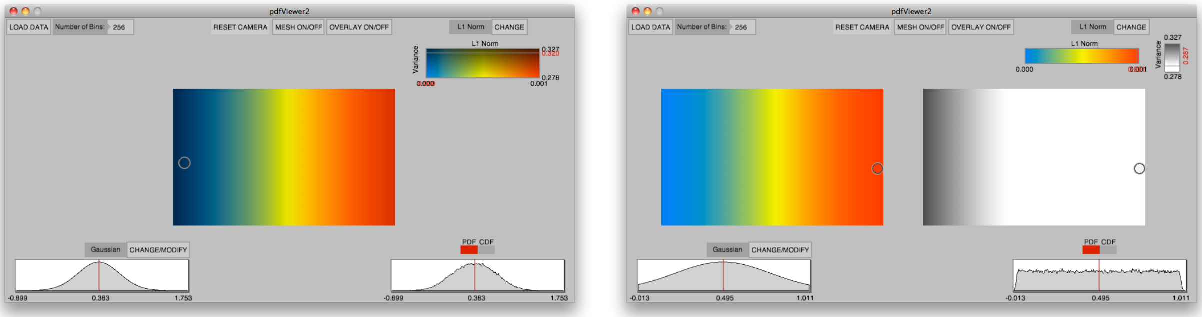

Number of Bins: 256

Figure 8: Demonstrative data showing variation in shape across the spatial domain. The data is a linear blend from Gaussian on the left (left image) to uniform on the right (right image).

Interval The interval measure evaluates the strength of the probabilities based on the variation in the data samples. To demonstrate this, we show a uniform distribution at each location in the spatial domain (Figure 9), however we increase the size of the interval width with x-axis. Thus, the shape measure returns a value of 0.0, indicating a Uniform distribution (left side of image), however the interval measure (shown on the right) clearly displays the increase in interval width as the increase of black in grayscale color map. This is also demonstrated in the data from Figures 1 and 2. The data in these images alternates between a Gaussian or uniform distribution and increases the interval length along the x-axis. Thus, for both data sets, as the interval width increases along the x-axis, the value of the color decreases indicating more uncertainty in the difference measure.

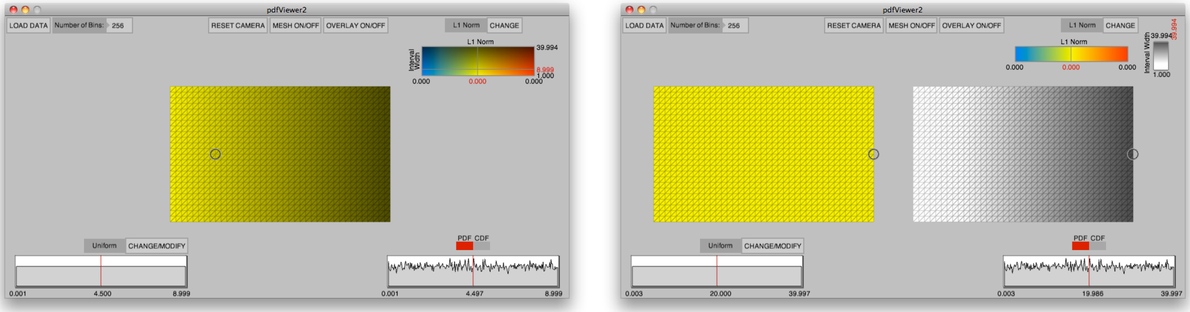


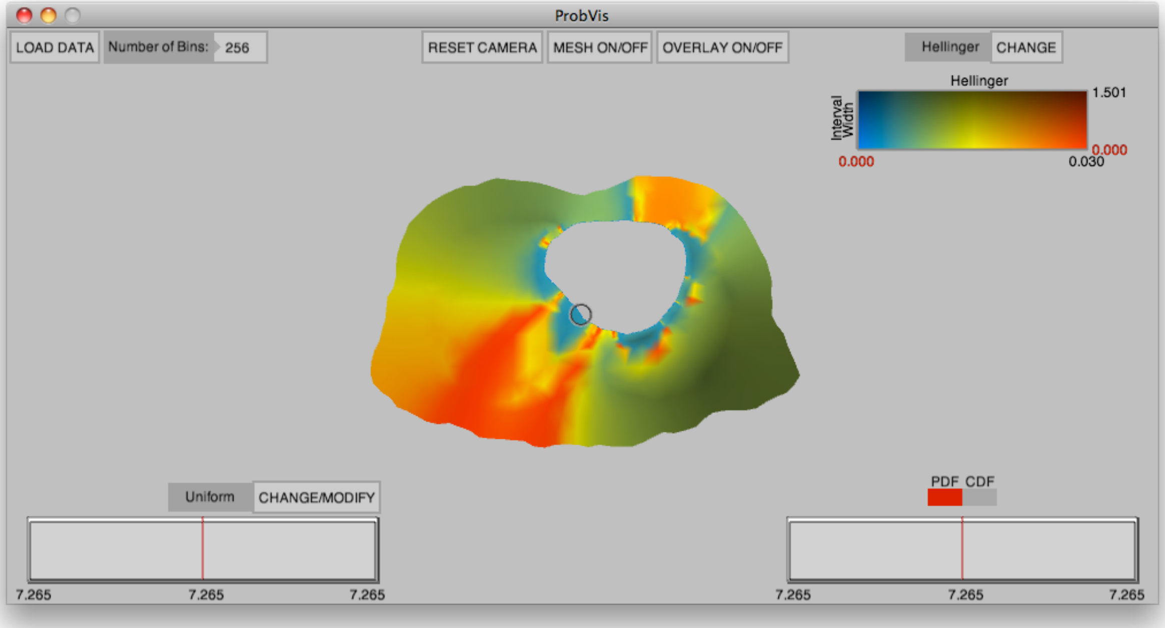
Figure 9: Variation in interval width. A uniform distribution is shown across the spatial domain, with variations in interval width increasing along the x-axis.

4.2 PDF Visualization of an Electrophysiology Simulation

This example is based upon the data set used in [20] in which we were interested in solving the bioelectric forward problem. The data set consists of a triangular finite element method obtained through the segmentation of MRI data. There are 618 vertices and 1071 triangles in the computational mesh. To replicate the results in [20] (which employed the gPC Galerkin approach) the gPC collocation approach was used with 9 quadrature points in the stochastic direction. Only perturbations with respect to a single uniformly-distributed random variable are considered.

Figures 10, 11 and 12 show the results of our visualization of this data using the three canonical distributions. By changing the type of the canonical comparator distribution it is easy to identify regions with particular distribution types. For example, Figure 10 highlights, in blue, the area around the hole in the middle representing the heart. In this area, the data samples all have the same value, as shown in the PDF display, and thus, the data distribution is best represented by a uniform distribution. As we move away from the heart, the data change distribution type. Figure 11 shows an area where the distributions closely resemble a normal distribution. This type of interaction inspired the inclusion of the beta distribution. As we moved the pointer around to each of the data distributions we noticed that the PDFs reminded us of the beta distribution. Thus, we added this comparator distribution to satiate our own curiosity, the results of which are shown in Figure 12. We expect the exploration of other data sets to generate the need for more comparison distributions, and thus we have made this possible through a simple extension of our source code.

Through the use of this tool, we are able to explore the large bioelectric data set. Previous visualizations of this data have used separate color maps of mean and standard deviation, however such visualizations assume a Gaussian distribution across the entire spatial domain. Our tool exposes this assumption as false and shows where the data is Gaussian and where it diverges from Gaussian. In addition, our tool elucidates on where the distributions are similar in shape, as well as interval, and thus having similar responses to the simulation. By using this tool we are able display a much larger amount of information on the data than before and thus tease out relationships between areas of the data that had previously been undiscovered.



Number of Bins: 256

Figure 10: Visualization of electrophysiology simulation data using the Hellinger distance and a uniform comparison distribution.

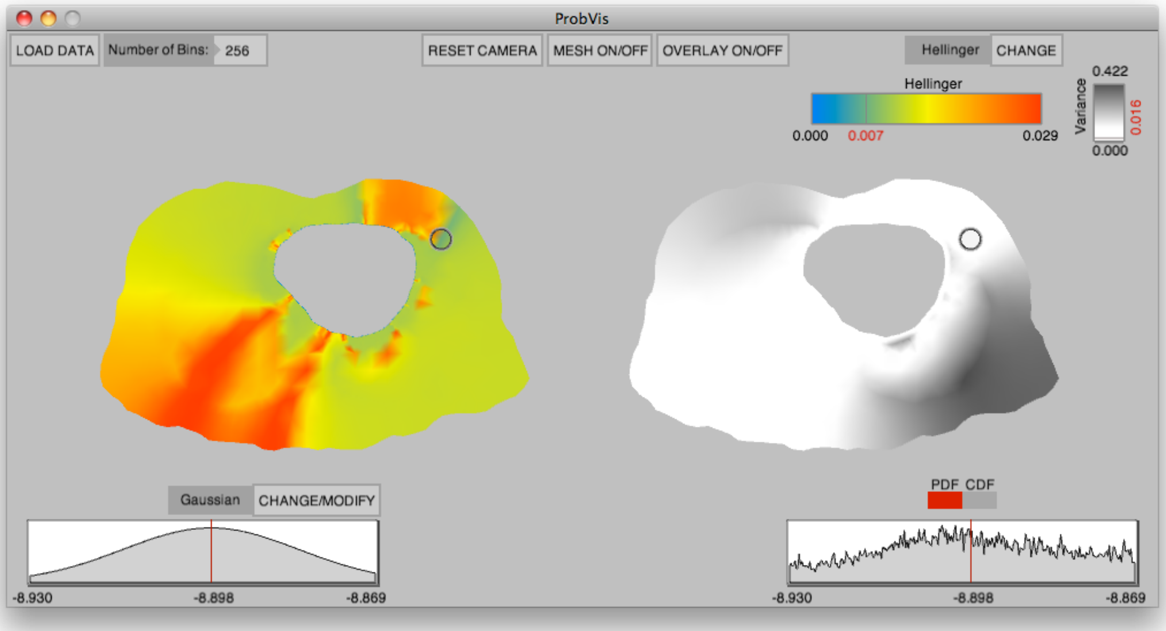
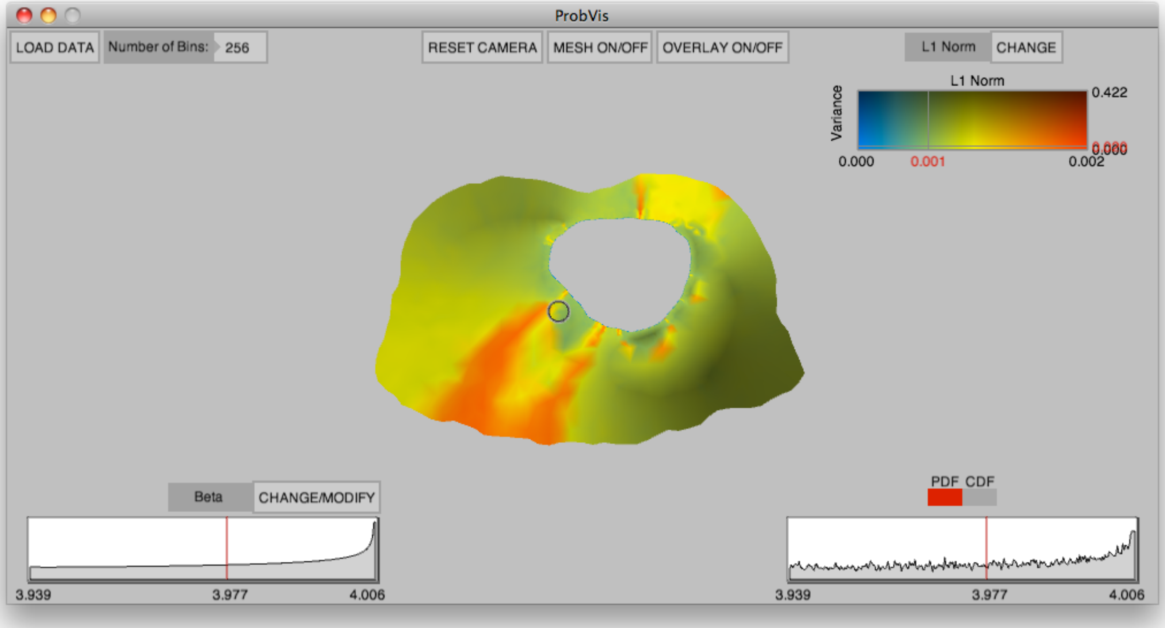

Number of Bins: 256

Figure 11: Visualization of electrophysiology simulation data using the Hellinger distance and a normal comparison distribution.

5. SUMMARY

In this paper we have presented the mathematical formulation and implementation details of a software system designed for displaying probability density functions over two-dimensional (spatial) domains. Although the concept of the PDF, and its normal visualization as a histogram, is very familiar, it is very challenging to construct visualization methodologies that allow the user to interpret “correlations” (in the sense of interdependency) between the PDFs of a function at different spatial locations. The purpose of this software effort was to provide an exploratory tool that (1) provided through contouring of normed differences of the PDFs of the function against a specified or optimally computed ansatz and (2) allowed the user to then interactively explore the field and the particular PDFs available at any particular data point.

The mathematical extension of this work to three-dimensional fields is straightforward; however, the many visualization issues such as glyph occlusion will need to be addressed in future work. This work provides an example



Number of Bins: 256

Figure 12: Visualization of electrophysiology simulation data using the L^1 distance and a beta comparison distribution.

of effective interaction between the UQ and visualization communities in attempting to solve a specific mathematical abstraction and the visualization needed.

ACKNOWLEDGMENTS

This is a collaborative research project based on work supported in part by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST), and supported under NSF IIS-0914564 (Kirby) and NSF IIS-0914447 (Xiu) and through DOE NETL DE-EE0004449 (Johnson/Kirby) and NIH 2P41 RR0112553-12 (Johnson). Infrastructure support provided through NSF-IIS-0751152.

REFERENCES

1. Ghanem, R. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, NY, 1991.

2. Xiu, D. and Karniadakis, G., The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.*, 24:619–644, 2002.

3. Johnson, C., Top scientific visualization research problems, *IEEE Computer Graphics and Applications*, 24(4):13–17, July/Aug 2004.

4. Cleveland, W., *The Elements of Graphing Data*, Wadsworth Advanced Books and Software, 1985.

5. Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P., *Graphical Methods for Data Analysis*, Wadsworth, 1983.

6. Tukey, J., *Exploratory Data Analysis*, Addison-Wesley, 1977.

7. McKinnon, A. E. and Raymond, E., Visualising the probability distribution function of uncertain data: application to stochastic modelling of ground water solute transport, In *Proceedings of the 2001 Asia-Pacific symposium on Information visualisation - Volume 9*, pp. 139–142, 2001.

8. Luo, A., Kao, D., and Pang, A., Visualizing spatial distribution data sets, In *VISSYM '03: Proceedings of the symposium on Data visualisation 2003*, pp. 29–38, 2003.

9. Potter, K., Krüger, J., and Johnson, C., Towards the visualization of multi-dimensional stochastic distribution data, In *Proceedings of The International Conference on Computer Graphics and Visualization (IADIS) 2008*, 2008.

10. Kao, D., Dungan, J. L., and Pang, A., Visualizing 2d probability distributions from eos satellite image-derived data sets: a case study, In *Proceedings Visualization*, pp. 457–561, 2001.

11. Kao, D., Luo, A., Dungan, J. L., and Pang, A., Visualizing spatially varying distribution data, In *Proceedings of the Sixth International Conference on Information Visualisation, 2002*, pp. 219–225, 2002.

12. Bordoloi, U. D., Kao, D. L., and Shen, H.-W., Visualization techniques for spatial probability density function data, *Data Science Journal*, 3:153–162, 2004.

13. Chlan, E. B. and Rheingans, P., Multivariate glyphs for multi-object clusters, In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pp. 141–148, 2005.

14. Xiu, D., *Numerical methods for stochastic computations*, Princeton Univeristy Press, Princeton, New Jersey, 2010.

15. Rosenblatt, M., Remarks on some nonparametric estimates of a density function, *Ann. Math. Stat.*, 27:832–837, 1956.

16. Parzen, E., On estimation of a probability density function and mode, *Ann. Math. Stat.*, 33:1065–1076, 1962.

17. Fielitz, B. D. and Myers, B. L., Estimation of parameters in the beta distribution, *Decision Sciences*, 6(1):1–13, 1975.

18. Cleveland, W. S. and McGill, R., Graphical perception: Theory, experimentation, and application to the development of graphical methods, *Journal of the American Statistical Association*, (387):531–554, 1984.

19. Fry, B. and Reas, C. Processing programming language. http://processing.org/.

20. Geneser, S., MacLeod, R., and Kirby, R., Application of stochastic finite element methods to study the sensitivity of ECG forward modeling to organ conductivity, *IEEE Journal on Biomedical Engineering*, 55(1):31–40, 2008.