Uncertainty-aware Multidimensional Ensemble Data Visualization and Exploration

Haidong Chen, Song Zhang, *Senior Member, IEEE,* Wei Chen, *Member, IEEE,* Honghui Mei, Jiawei Zhang, Andrew Mercer, Ronghua Liang, *Member, IEEE*, and Huamin Qu, *Member, IEEE*

Abstract—This paper presents an efficient visualization and exploration approach for modeling and characterizing the relationships and uncertainties in the context of multidimensional ensemble datasets. Its core is a novel dissimilarity-preserving projection technique that characterizes not only the relationships among the mean values of the ensemble data objects but also the relationships among the distributions of ensemble members. This uncertainty-aware projection scheme leads to an improved understanding of the intrinsic structure in an ensemble dataset. The analysis of the ensemble dataset is further augmented by a suite of visual encoding and exploration tools. Experimental results on both artificial and real-world datasets demonstrate the effectiveness of our approach.

Index Terms—Ensemble visualization, uncertainty quantification, uncertainty visualization, multidimensional data visualization

1 Introduction

Recent developments in scientific simulation research have resulted in an increased role for scientific simulations and analysis. One common way to study uncertainty is the ensemble simulation, which employs stochastic initial conditions or multiple parameterizations to produce an ensemble of simulation outcome. For example, in the numerical weather simulation, both initial conditions and simulation parameters (e.g., cumulus schemes and microphysics schemes) can be perturbed in an ensemble forecast simulation. The number of ensemble runs ranges from dozens to thousands. All ensemble members of a single data entry are called an ensemble data object in this paper, e.g., the ensemble numerical weather forecast at a geospatial location. The ensemble members of an ensemble data object may be averaged to obtain the ensemble mean, which is regarded as a representative of the ensemble members. Unfortunately, important information is lost in the averaging process. It is highly beneficial to characterize the uncertainty of an ensemble dataset during the entire analysis process.

In most scientific applications, resulting simulation output is a multivariate field. For instance, the output of a typical Weather Research and Forecasting (WRF) simulation [1] includes more than 100 variables such as temperature, pressure, and wind direction. Therefore, challenges for analyzing an ensemble dataset include both the *uncertainty* with respect to each output variable and the *high dimensionality*.

There have been a large number of uncertainty visualization and analysis [2] methods that go beyond traditional summary statistics [3] and depict ensemble data in more detail. Conventional visualization solutions exploit glyphs [4], [5], [6] and visual variables [7], [8] to encode uncertainties. However, most of them are only designed for 1D or 2D datasets and have limited capabilities to reveal the intrinsic structures in the ensemble dataset. Recent methods can effectively characterize and analyze the uncertainty structures [9], [10] and forms [11], [12] but are designed for data objects with one variable.

To address the second challenge (i.e., high dimensionality), multidimensional projection techniques [13], [14] are widely used to build a low-dimensional layout that respects the distances among data objects in the high-dimensional space. In this low-dimensional layout, closely positioned points indicate similar data objects in the high-dimensional space. However, most conventional projection methods are developed for datasets in which a data object has only one instance. Simply using the ensemble mean as an instance to represent a data object for projection suffers from heavy information loss, because the shape of distribution for each ensemble data object is lost during the averaging process. A conceptual example is shown in Fig. 1 (a-b), where four ensemble data objects with similar ensemble means but different ensemble distributions are lumped together by a conventional multidimensional projection method to the ensemble means. This might cause misleading perceptions from users. For example, users might advocate that these four data objects are almost the same as they are located close to each other.

It remains a challenging task to visually explore the intrinsic structures of a high dimensional ensemble dataset as well as the uncertainty of each individual ensemble data object. The key contribution of this paper is a nov-

- *Haidong Chen, Wei Chen, Honghui Mei, and Jiawei Zhang are with State Key Lab of CAD & CG, Zhejiang University, CHINA, 310058. E-mail: chenhd925@gmail.com, chenwei@cad.zju.edu.cn, xbikvyx@gmail.com, and zjwbeckham@126.com.*
- *Wei Chen is the corresponding author. He is also with the Cyber Innovation Joint Research Center, Zhejiang University.*
- *Song Zhang is with Department of Computer Science and Engineering, Mississippi State University, U.S., 39762. E-mail: szhang@cse.msstate.edu.*
- *Andrew Mercer is with Geosciences Department, Mississippi State University, U.S., 39762. E-mail: mercer@gri.msstate.edu.*
- *Ronghua Liang is with College of Information Engineering, Zhejiang University of Technology, CHINA, 310014. E-mail: rhliang@zjut.edu.cn.*
- *Huamin Qu is with Department of Computer Science and Engineering, Hong Kong University of Science and Technology. E-mail: huamin@cse.ust.hk.*

el uncertainty-aware multidimensional projection approach that generalizes the conventional projection methods to ensemble datasets. Its core is a new dissimilarity measure of the ensemble data objects and an enhanced Laplacian-based projection scheme. Compared with the conventional projection methods that can solely respect the distances among ensemble means, our dissimilarity measure preserves the relationships among the ensemble distributions as well as the ensemble means. By setting the distributional difference weight to zero (see Equation (7)), our approach regresses to a conventional projection scheme. The enhanced Laplacian-based projection scheme achieves a balance between the local and global point layout by imposing global constraints in constructing the Laplacian system.

Fig. 1 (c) illustrates the effectiveness of our uncertainty-aware projection approach, with which four data objects are positioned in the 2D visual plane. In this projection, U_2 and U_3 are located close to each other due to their similar ensemble means and distributions. As U_1 and U_4 follow different distributions from U_2 and U_3, they are positioned far away from U_2 and U_3. Even though all of them have similar ensemble means.

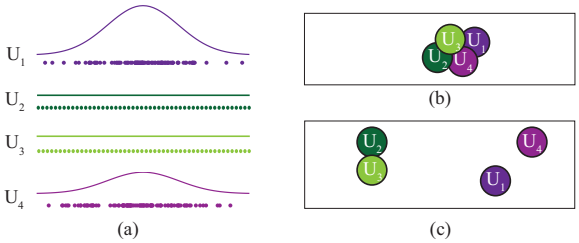


Fig. 1. Projecting an exmample 1D ensemble dataset (a) which consists of four ensemble data objects with similar mean values but different distributions. (b) The result of a conventional multidimensional projection algorithm to the ensemble means. (c) The result of our method.

We note that as with any projection method, our method will not preserve the high-dimensional structures with 100% accuracy. Our goal is to provide users with a projection method to visualize not only the differences between the means of ensemble data objects, but also the differences between the shapes of the ensemble distributions. As demonstrated in Section 6 (see Fig. 6 and Fig. 11), a better separation of data objects with different ensemble distributions can be achieved with our method, but not with previous projection methods using ensemble means.

In addition, we augment the users' abilities to visually study an ensemble dataset with a suite of visual exploration widgets: 1) an uncertainty histogram that allows for selecting the interesting range of uncertainty values in the visualized data; 2) an ensemble bar that embodies key information in an ensemble data object; 3) a parallel coordinates view that shows the details of the ensemble distribution for a particular ensemble data object; and 4) an optional geolocation view that depicts the uncertainty patterns identified in the projection view. Experimental results on a synthetic dataset and numerical weather simulations alike show that ensemble data objects with similar ensemble means but different ensemble distributions can be clearly distinguished with our approach.

In summary, the main contributions of this work include:

- A novel uncertainty-aware multidimensional projection technique for ensemble datasets.
- A multi-view interactive visualization system for effective exploration of intrinsic structures and uncertainties in ensemble datasets.

2 Related Work

Our work relates to multiple topics of visualization including uncertainty visualization, ensemble data visualization, and multidimensional/multivariate data visualization.

Uncertainty spreads throughout the entire data analysis pipeline, including acquisition, transformation, and visualization [15]. Depicting uncertainty can significantly help analysts make better decisions [16]. Thomson et al. [17] proposed a typology for uncertainty visualization in intelligence analysis. Potter et al. [18] presented a comprehensive survey for uncertainty quantification and visualization of scientific data. In the past decade, a large number of uncertainty visualization techniques have been developed. They can be roughly classified into four categories: glyph based, visual variable based, geometry based, and animation based. Glyph based methods encode uncertainties into well-designed glyphs (e.g., the flow radar glyph [4], the circular glyph [19], and the summary plot [5]) and place them into the original data field. Similarly, visual variables such as color [10], [20], brightness [8], [21], blurriness [22], and texture [7], [23] can also be employed to encode uncertainty. The geometry based approaches are a family of techniques that adapt the basic geometry to represent uncertainty, including point [20], line [24], cube [25], and surrounding volume [10], [26]. Coninx et al. [27] and Lundstrom et al. [28] demonstrate that animation can be used to express uncertainty as well. In general, most of these techniques focus on 1D or 2D datasets and cannot be readily extended to multidimensional datasets. Wu et al. [29] introduced the standard error ellipsoid to characterize uncertainty arising in any stage of a visual analytic process for multidimensional datasets. In addition, evaluations of different uncertainty visualization techniques have attracted much attention recently. Deitrick et al. [16] conducted an empirical evaluation to explore the influence of uncertainty visualization on decision making. Sanyal et al. [30] compared four commonly used techniques for 1D and 2D datasets. They identified that the glyph representations are good options generally.

Ensemble data is common in many scientific fields such as meteorology, hydrology, and chemistry. Ensembles effectively represent multiple simulations of a model with varied initial or boundary conditions, yielding slightly different results. They can be regarded as a type of uncertainty data. A convenient visualization method for ensemble data is the small-multiple method [31], [32] associated with linking and brushing operations. Spaghetti plots [33] is another approach to visualize ensemble dataset. Potter et al. [32] introduced a framework, Ensemble-Vis, to visualize and explore ensemble dataset by leveraging multiple coordinated views. Sanyal et al. [6] developed a tool named Noodles to visualize ensemble uncertainty which was modeled with the standard deviation. Unfortunately, Noodles is only applicable for single variable ensemble data. Recent work in this area has progressed from simple visualization to formal analysis

of the results. Thomas et al. [34] presented an interactive system to study off-shore structures in an ensemble ocean forecasting dataset. Gosink et al. [11] proposed a method to characterize different types of predictive uncertainty in ensemble datasets based on the Bayesian model averaging. Mathias et al. [12] developed a Lagrangian framework for ensemble flow field analysis.

The differences between the terms **multidimensional** and **multivariate** are subtle. The term multidimensional refers to independent dimensions, while the term multivariate refers to that of dependent variables [35]. Nevertheless, multivariate data can also be treated as multidimensional, because the relationships among dimensions are typically unknown in advance. Conventional approaches to visualize multidimensional data employ scatter plot matrix, parallel coordinates, star coordinates, etc. Among these approaches, multidimensional projection has gained much attention due to its ability to characterize similarities among data points in a visual space (2D or 3D), and has been proven to be a useful tool in many applications. Daniels et al. [36] employed a LSP-like projection technique for interactive vector field analysis. Chen et al. [37] proposed to embed high-dimensional DTI fibers to a 2D space with MDS for interactive exploration. Joia et al. [14] introduced an advanced projection technique called local affine multidimensional projection (LAMP) to interactively correlate similar data instances. Anand et al. [38] used random projection to find interesting substructures in a high-dimensional dataset. Although these approaches are capable of visualizing data by their low-dimensional layout, they are designed for data without uncertainty. Our approach advances the multidimensional projection scheme by taking uncertainty into account.

3 Uncertainty-aware Multidimensional Projection

Multidimensional projection is emerging as an effective visual exploration technique for high-dimensional datasets. It projects a set of data points in the d-dimensional space into an l-dimensional visual space, typically $l \in \{2, 3\}$. Conventional techniques assume that the value of each data object is deterministic. Many recent scientific simulations produce a collection of values at each data object due to the perturbed initial condition or parameterization. This type of data is called ensemble data.

The conceptual model of our data consists of n ensemble data objects $U = \{U_1, U_2, ..., U_n\}$. Each object has m d-dimensional ensemble members, $U_i = \{U_i^1, U_i^2, ..., U_i^m \mid U_i^t \in \mathbb{R}^d\}$. The goal of our uncertainty-aware projection scheme is to build an l-dimensional representation $V = \{V_1, V_2, ..., V_n \mid V_i \in \mathbb{R}^l\}$ to preserve the relationships among data objects in terms of both the ensemble means and the ensemble distributions.

3.1 Approach Overview

To strike a balance between projection efficiency and accuracy, we employ a two-step multidimensional projection: first, a small set of control points are selected from U based on the ensemble means and projected to a 2D space with the conventional multidimensional scaling method [13]; then all other objects in U are projected to the 2D space with an enhanced Laplacian system that combines the influences from both control points and other data points.

One distinctive feature of our approach is that we take both the ensemble mean differences and the ensemble distribution differences into account to measure the dissimilarity $D(U_i, U_j)$ of any pair of ensemble data objects U_i and U_j. The former is described by the Euclidean distance between the ensemble means $E(\overline{U}_i, \overline{U}_j)$, and the latter is captured by a difference measure between ensemble distributions $J(U_i||U_j)$ based on relative entropy [39]. To reconstruct the continuous ensemble distribution for each data object, a multidimensional Kernel Density Estimation (KDE) method that considers the dimensional correlations is employed.

3.2 Ensemble Data Object and Probability Distribution

For simplicity, we suppose that the data objects are independent. Currently, many uncertainty modelling and visualization methods [9], [10] assume that the ensemble members for a given data object are drawn from a known parametric distribution, for example Gaussian. In practice, the ensemble distribution is often not Gaussian and can vary from data object to data object.

KDE is a non-parametric approach to approximate the underlying continuous distribution by using a sum of kernels centered at each sample. Specifically, the multidimensional KDE [40] for data object U_i is defined as:

$$U_i(x) = \sum_{t=1}^{m} w_t K_{\mathbf{H}} \left(x - U_i^t\right), \tag{1}$$

where $\mathbf{H}$ is a bandwidth matrix, w_t is a weight factor which can be determined by prior knowledge, and $\sum_{t=1}^{m} w_t = 1$. $K_{\mathbf{H}}$ is the multidimensional kernel function satisfying that $K_{\mathbf{H}}(x) \geq 0$ in the entire domain and $\int K_{\mathbf{H}}(x)dx = 1$. As such, the estimated density function can be interpreted as the probability density function.

A range of kernels can be adopted for KDE including uniform, Gaussian, and Epanechnikov kernels. We employ the widely used normal kernel [41], [42], because the kernel type is less important than the bandwidth parameter in terms of influences on the estimation [43] .

In general, there are three choices for $\mathbf{H}$: the scaled identity matrix $\mathbf{H} = h^2\mathbf{I}$, the diagonal matrix $\mathbf{H} = diag(h_1^2, h_2^2, ..., h_d^2)$, and the general symmetric positive definite matrix. The scaled identity matrix implies that the variance in all dimensions are identical. The diagonal matrix implies that each dimension has its own bandwidth and no correlation exists between any two dimensions. In practice, a variety of automatic data-driven bandwidth selection methods can be used to estimate the bandwidth h_k on the k-th dimension, e.g., the Silverman's rule of thumb [43]:

$$h_k = \left(\frac{4{\sigma_k}^5}{3m}\right)^{1/5} \approx 1.06\sigma_k m^{-1/5}, \tag{2}$$

where σ_k is the standard deviation of samples on the k-th dimension, m is the number of samples. More sophisticated approaches like cross-validation can be found in [44]. The general symmetric positive definite matrix encodes the

individual bandwidth for each dimension and the linear relationship between any two dimensions.

Each type of bandwidth matrix has its own advantages and limitations. The scaled identity matrix is prone to under-fitting and has a large fitting error. The general matrix, which requires that $\mathbf{H}$ is symmetric and positive definite, is the most accurate one but needs more parameters and is prone to over-fitting. The diagonal matrix strikes a balance between them.

To take the advantage of simplicity by the diagonal matrix while preserving correlations among dimensions, we choose to perform the KDE process in a space defined by the *principal components transformation* [40]. More specifically, a mean centering process is first applied to each data object before estimation so that the distributional differences, (to be detailed in section 3.3), are not influenced by ensemble means:

$$\widehat{U}_i^t = U_i^t - \overline{U}_i. \tag{3}$$

We then apply the principal components analysis to $\widehat{U}_i = \{\widehat{U}_i^t \mid t = 1, 2, ..., m\}$ that yields a transformation matrix Φ. Φ is composed by the eigenvectors of the covariance matrix of $\widehat{U}_i$. At last, we transform each ensemble member U_i^t into a new set $U_i^* = \{U_i^{t*} \mid t = 1, 2, ..., m\}$ by:

$$U_i^{t*} = \Phi^T \widehat{U}_i^t. \tag{4}$$

Consequently, the probability $U_i(x)$ at location x is computed by $U_i^*(x^*)$ on U_i^* instead, where $x^* = \Phi^T(x - \overline{U}_i)$. Because the bases of the new space are eigenvectors that are orthogonal to each other, the diagonal bandwidth matrix $\mathbf{H}^*$ for estimating $U_i^*(x^*)$ can be quickly estimated with Equation (2) on U_i^*.

3.3 Dissimilarity Estimation

The dissimilarity measure between pairs of data objects plays an essential role in most multidimensional projection techniques. Geometric distances such as the Euclidean distance, the cosine distance, and the geodesic distance have prevailed in the multidimensional projection literature for years. However, these measures cannot be directly applied to ensemble datasets.

A naive way to deal with ensemble datasets is to use the summary statistics like the ensemble means to represent the data objects and calculate the dissimilarities among them. Unfortunately, this simple scheme results in a large amount of information loss. The Anscombe's quartet dataset shows that four data objects with the same means and variances have very different ensemble distributions. Clearly, it is impossible to distinguish them from each other with the ensemble means.

Therefore, it is natural to extend the definition of dissimilarities among ensemble data objects by considering their probability distributions. *Jensen-Shannon divergence* (JSD) [39] or relative entropy is a dissimilarity measure of distributions based on the *Kullback-Leibler divergence*. JSD is symmetric, non-negative, and bounded. It is widely used in the community of uncertain data mining [45]. The distribution difference $J(U_i||U_j)$ between the data object U_i and U_j is defined as:

$$\begin{aligned} J(U_i||U_j) = &\frac{1}{2}\int_{\mathbb{D}} U_i(x)\log\left(\frac{2U_i(x)}{U_i(x)+U_j(x)}\right)dx \\ &+ \frac{1}{2}\int_{\mathbb{D}} U_j(x)\log\left(\frac{2U_j(x)}{U_i(x)+U_j(x)}\right)dx \end{aligned}, \tag{5}$$

where $U_i(x)$ and $U_j(x)$ are the reconstructed probability density functions for U_i and U_j respectively, $\mathbb{D}$ denotes the high-dimensional space where all the ensemble members U_i^t reside. Typically, we set the base of the log function as 2 so that $J(U_i||U_j)$ is bounded by 1. Only when U_i and U_j have the same distributions does $J(U_i||U_j)$ reach 0. By convention, $0 \log 0$ is defined as 0.

In most cases, it is impossible to obtain the analytical solution of Equation (5). Therefore we compute its numerical solution with sampling. In our implementation, the Metropolis-Hastings (MH) [46] algorithm, a Markov Chain Monte Carlo method, is employed, because it can generate a sequence of samples from a function proportional to the target probability density distribution when direct sampling is difficult. As more and more samples are produced, the distribution of samples more closely approximates the target distribution. Consequently, by the law of large numbers, we can rewrite Equation (5) as:

$$\begin{aligned} J(U_i||U_j) \approx &\frac{1}{2S}\sum_{t=1}^{S}\log\left(\frac{2U_i(x_i^t)}{U_i(x_i^t)+U_j(x_i^t)}\right) \\ &+ \frac{1}{2S}\sum_{t=1}^{S}\log\left(\frac{2U_j(x_j^t)}{U_i(x_j^t)+U_j(x_j^t)}\right) \end{aligned} \tag{6}$$

where $\{x_i^t \mid t = 1, 2, ..., S\} \sim U_i(x)$ and $\{x_j^t \mid t = 1, 2, ..., S\} \sim U_j(x)$ are S samples generated by MH.

Therefore, we define the dissimilarity between U_i and U_j as a weighted sum of the Euclidian distance between their ensemble means and the distributional difference between their probability density distributions:

$$D(U_i, U_j) = \alpha E(\overline{U}_i, \overline{U}_j) + \beta J(U_i \parallel U_j), \tag{7}$$

where $\alpha \in [0, 1]$, $\beta \in [0, 1]$ are two parameters adjustable by users, and $\alpha + \beta > 0$. When $\alpha \neq 0$, $\beta = 0$, the distributional differences are missed by our dissimilarity measure. In such a case, our approach will regress to the conventional projection methods, because only the dissimilarities among ensemble means are captured.

In summary, the proposed dissimilarity measure is endowed with the following properties:

- **Symmetry** ($D(U_i, U_j) = D(U_j, U_i)$) The dissimilarities between any pair of data objects are symmetric. Therefore it can be seamlessly incorporated into many state-of-the-art multidimensional projection techniques.
- **Uncertainty-awareness** The dissimilarity considers the differences in ensemble distributions therefore differentiating data objects indistinguishable by summary statistics.

3.4 The Enhanced Laplacian-based Projection

Our scheme of projecting ensemble data objects to the visual space is inspired by the Least Squares Projection (LSP)

technique [47]. Generally, it is a two-step local technique. In the first step, a subset of data objects are projected to the visual space. In the second step, the rest of the data objects are interpolated according to the *K*-nearest neighborhood graph. However, this method inherits the drawbacks of local methods. It may bias the data objects projected in the second step in favor of control points projected in the first step [48] (see Fig. 3 (a)). Thus, instead of directly using the Laplacian system originally proposed in [47], we propose to add global constraints into the equations. In other words, our method associates each data object with two small sets: a near set and a random set, instead of only one *K*-nearest set. Both sets are used for constructing the neighborhood graph for interpolation. The random set plays the role of the global constraints. This simple scheme strikes a balance between the locality maintained by the near set and the global layout preserved by the random set.

A large number of samples from the ensemble distributions may be needed to obtain an accurate estimation of the relative entropy (Equation (6)), especially when the ensemble distribution is complicated. To avoid pairwise distributional difference calculations, we select control points and construct the neighborhood graph only based on the ensemble means $\overline{U} = \{\overline{U}_1, \overline{U}_2, ..., \overline{U}_n\}$. Control points $C = \{C_i \mid C_i \in U, i = 1, 2, ..., K\}$ are selected with a K-center algorithm [49]. If we do not have any priori knowledge about the dataset, the number of control points is set as $K = \sqrt{n}$ [50]. The near set N_i for each ensemble object U_i is then defined as the ensemble objects whose means are the *K*-nearest neighbors of $\overline{U}_i$. Apparently, N_i might not be the true *K*-nearest neighbors of U_i because only the ensemble mean information is utilized. However, in practice the *K*-nearest neighbors in N_i can be guaranteed by enlarging the near set. In such a case, all other objects in N_i that are not the real *K*-nearest neighbors of U_i can be deemed as objects in the random set R_i. Fig. 2 shows an example to build the neighborhood graph for the ensemble data object U_1.

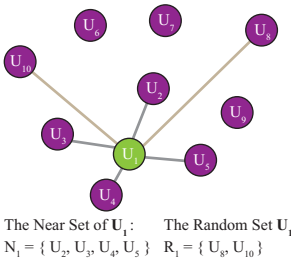


Fig. 2. The neighborhood graph for U_1 with a near set of size 4 and a random set of size 2.

In our implementation, an iterative majorization algorithm called Scaling by Majorizing a Convex Function (SMACOF) [13] is employed to build a low-dimensional representation $\{V_i^c \mid V_i^c \in \mathbb{R}^2, i = 1, 2, ..., K\}$ for control points. V_i^c is the 2D projection of the control point C_i.

Technically, the Laplacian-based projection scheme rests on the theory of convex combination, that is, the low dimensional representation for each high-dimensional data object can be regarded as a linear combination of its neighborhoods in the visual space. Mathematically, let $V_i \in \mathbb{R}^2$ be the projection of ensemble data object U_i, according to the convex combination theory, V_i can be written as:

$$V_i = \sum_{U_j \in \{N_i \cup R_i\}} \tau_{ij} V_j, \tag{8}$$

where $\tau_{ij} > 0$, $\sum \tau_{ij} = 1$. Typically, the inverse of dissimilarity $D_{inv}(U_i, U_j) = 1/D(U_i, U_j)$ between the ensemble data object U_i and U_j is used to define the weight:

$$\tau_{ij} = \frac{D_{inv}(U_i, U_j)}{\sum_{U_j \in \{N_i \cup R_i\}} D_{inv}(U_i, U_j)}. \tag{9}$$

By reorganizing Equation (8) for all ensemble data objects into a matrix representation, a sparse linear system constrained by the control points can be derived [47] which can then be solved in a least squares sense. The solution of this system is the low-dimensional representation of the ensemble dataset U.

Fig. 3 demonstrates the effect of using a random set. A synthetic dataset consisting of 500 data objects in 5 clusters is employed. Each ensemble data object has 80 members. Fig. 3 (a) displays the projection result without the random set. In this case, our method regresses to the Least Squares Projection technique. Fig. 3 (b) show the result with the size of the random set tested at 8. It is reasonable that a large random set preserves more global relationships among points, leading to low stress. This is empirically verified by the plot of standard normalized stress over the size of the random set (see Fig. 3 (d)). For comparison, the projection result of MDS implemented with SMACOF [13], which optimizes the global relationships among points, is shown in Fig. 3 (c). MDS provides the smallest stress but requires the pairwise dissimilarity matrix that it is usually not feasible in projecting large dataset. Fig. 3 (e) shows the time of our method with different sizes of the random set. From this plot, we can see that the consuming time is closely linear to the sum of $|N_i|$ and $|R_i|$, $|N_i| = 20$ in this example. This is because the distributional difference estimation is much more time-consuming compared to building and solving the linear system. In summary, the addition of a random set provides a balance between the fast but inaccurate local methods (e.g., Laplacian-based projection methods) and the accurate but slow global methods (e.g., MDS).

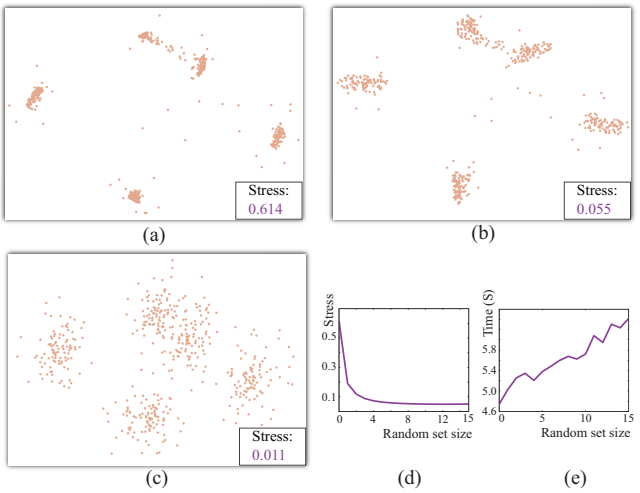


Fig. 3. The influence of different sizes of the random set in our approach. (a) $|R_i| = 0$, which is the LSP technique. (b) Our method with $|R_i| = 8$. (c) MDS projection. (d) The stress plot. (e) The time plot.

4 Uncertainty Quantification and Visualization for Individual Ensemble Data Object

The uncertainty-aware 2D projection of the multidimensional ensemble data objects helps users understand the relationships among data objects in a simple 2D layout. It is equally important to allow users to further examine the uncertainty of each ensemble data object. This section describes our solutions for quantifying and visualizing uncertainty of the i-th ensemble data object U_i.

4.1 Uncertainty Quantification

Quantifying uncertainty plays an important role in the pipeline of uncertainty visualization [15]. It offers the variables for visual encoding. In this section, we describe two measures for quantifying the overall uncertainty of an ensemble data object and the detailed deviation for each ensemble member. We also discuss the limitations and application scenarios of each measure.

In the 1D case, the standard deviation is a widely used metric to quantify uncertainty of a random variable. Inspired by this rule, our first measure models the *overall uncertainty* O_i of the ensemble data object U_i as a sum of the standard deviations in all dimensions:

$$O_i = \sum_{k=1}^{d} \sigma_i^k, \tag{10}$$

here σ_i^k represents the standard deviation on the k-th dimension. The deviation δ_i^t of the t-th ensemble member of U_i is defined as its Euclidean distance to the ensemble mean:

$$\delta_i^t = \| U_i^t - \overline{U}_i \| . \tag{11}$$

The first measure is simple and can be used for most applications. However, it does not take the correlations among dimensions into account. The second measure employs the covariance matrix to quantify uncertainty for each ensemble data object. The covariance matrix can be considered as a generalization of variance to multiple dimensions. It measures the dispersion of an ensemble data object with respect to the ensemble mean. A geometrical representation of a covariance matrix is the hyper ellipsoid [29], whose axes correspond to the eigenvectors and square roots of eigenvalues of the covariance matrix. Similar to [29], we use the volume of the hyper ellipsoid to represent the overall uncertainty of an ensemble data object:

$$O_i = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \prod_{k=1}^{d} \sqrt{\lambda_k}, \tag{12}$$

where $\Gamma(\cdot)$ indicates the Gamma function, λ_k represents the eigenvalues of the covariance matrix $\sum_i$. Once established, the deviation of an ensemble member is defined as the Mahalanobis distance to the ensemble mean:

$$\delta_i^t = \sqrt{\left(U_i^t - \overline{U}_i\right)^T \sum_i^{-1} \left(U_i^t - \overline{U}_i\right)}. \tag{13}$$

The second measure requires more computation than the first one due to a matrix inversion. In addition, a sufficient number of ensemble members are demanded in the second measure to correctly represent correlations among dimensions. Thus, the second measure is not a good option when the number of ensemble members is much smaller than the dimensionality of the dataset.

4.2 From Quantification to Visualization

Following the basic principles identified by Sanyal et al. [30] that the glyph size and color are two effective variables to display uncertainty in a 2D space, we design a color bar based representation called *ensemble bar* to depict the overall uncertainty and the distribution of ensemble members. Our goal is to provide users a quick preview of uncertainty patterns in an ensemble data object rather than complete details, which will be displayed in the parallel coordinates view. The height of the ensemble bar encodes the overall uncertainty: a higher bar implies an ensemble data object of larger overall uncertainty. The pattern shown in the bar depicts the distribution of all ensemble member deviations. To help users easily explore the distribution differences of different ensemble data objects, all ensemble bars have an identical width.

Let $\delta_i = \{\delta_i^1, \delta_i^2, ..., \delta_i^m\}$ be the deviations of the m ensemble members from the ensemble mean $\overline{U}_i$. The height of the ensemble bar is given by:

$$H_i = (1 - O_i)H_{min} + O_i H_{max}, \tag{14}$$

where H_{min} and H_{max} represent the minimum and maximum heights of all ensemble bars, respectively. $O_i \in [0, 1]$ is the normalized overall uncertainty of U_i.

In order to convey the distribution of ensemble member deviations in the bar, we first let that the left and right edges of the bar correspond to the minimum deviation $min(\delta_i)$ and the maximum deviation $max(\delta_i)$, respectively. Then, we equally discretize the entire bar into a set of bins and count the number of ensemble members in each bin. We draw each bin as a colored rectangle. The color map is set as a linearly varied sequential color map (e.g., from white to green in this paper).

The number of bins has a great impact on the pattern shown in an ensemble bar. In practice, various useful guidelines and rules of thumb can be employed. In our implementation, we take the square root of the number of ensemble members as the number of bins. This scheme is simple and has been widely used in many applications, such as Excel. Users are allowed to interactively adjust the number of bins to discretize the ensemble bar.

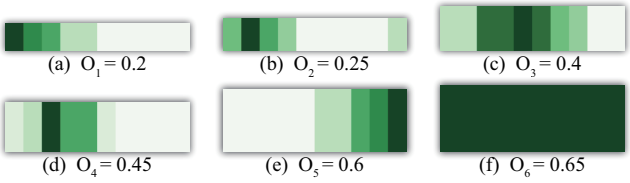


Fig. 4. Various types of ensemble bars show different types of ensemble data objects, each of which consists of 100 members. The overall uncertainty increases from (a) to (f). In each case, a bar in dark green indicates the peak of the distribution of ensemble member deviations. The ensemble member deviations approximately follow: (a-b) the positively skewed distribution; (c-d) the normal distribution; (e) the negatively skewed distribution; (f) the uniform distribution.

In general, the ensemble bar allows users to compare overall uncertainties of different ensemble data objects

through the heights of the ensemble bars (see Fig. 4). In addition, the patterns of bins enable users to identify different types of distributions. Taking the uniform distribution as an example, an approximately monochromatic ensemble bar is produced with our method (see Fig. 4 (f)). This indicates that all bins in the ensemble bar almost have the same number of ensemble members. Similarly, for positively skewed distribution of which more ensemble members are close to the ensemble mean, an ensemble bar with the dark-green bin on the left is generated (see Fig. 4 (a-b)).

5 Visual Exploration and Interactions

This section describes an integrated system with a suite of visualization and interaction tools for visually exploring multidimensional ensemble datasets.

5.1 The Exploration Workflow

The exploration process starts by an uncertainty histogram, on which users can select target uncertainty intervals (e.g., highly uncertain data objects). Then, users can interactively investigate the low-dimensional layout of the ensemble data objects in the 2D projection view equipped with a set of interactions such as zoom in/out, and pointer/lasso selection. Once points of interest are specified, the geospatial locations associated with each data object will be highlighted in the Geo-Location view. Meanwhile, the ensemble bar view shows all selected data objects. Users can further drill down to a specific ensemble bar to explore the high-level distribution and the overall uncertainty of a data object. Afterwards, users can select bins of interest in the ensemble bar to examine the selected ensemble members in an animated continuous parallel coordinates view. The parallel coordinates view also allows users to study the distribution of ensemble members on each dimension. To investigate the source of uncertainty (parameter perturbations in this paper), a radial plot is provided to display the parameter configurations that generate these ensemble members.

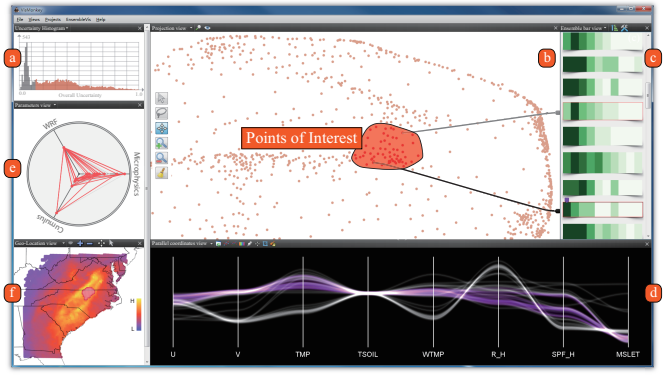


Fig. 5. The main interface of our visual exploration system. (a) The uncertainty histogram. (b) The 2D projection view. (c) The ensemble bar view. (d) The continuous parallel coordinates view. (e) The parameters view. (f) The Geo-Location view.

5.2 Exploration Tools

The main user interface is composed of a set of linked widgets and views to enhance interactivity. Fig. 5 shows an overview of our system.

Uncertainty histogram. At the top left corner, an uncertainty histogram widget allows users to select data points with overall uncertainty in a desired range.

Projection view. This view displays the intrinsic structures of an ensemble dataset in the 2D plane. The distances among 2D data points encode the similarities of the corresponding ensemble data objects with regard to both ensemble means and ensemble distributions. Users can either select a point or a group of points for further analysis. A set of operations are provided in this view to assist exploration including: zooming, panning, hovering, pointer selection, lasso selection, and multi-selection with the control key pressed.

Ensemble bar view. To facilitate comparison of overall uncertainties for different data objects, the ensemble bar view is displayed at the right side of the main interface. The overall uncertainties are perceived by the heights of the bars, and the ensemble distributions are depicted by different types of patterns in the ensemble bars. Following interactions are provided in this view:

- Choosing a color map and the bin number;
- Dragging bars together for detailed comparison;
- Reordering bars according to the distributional patterns;
- Selecting bins of interest to further examine ensemble members in the parallel coordinates view.

Parallel coordinates view. A continuous representation is employed to convey uncertainty at each dimension. Bright areas on each dimension indicate high certainty. Users can detect outliers with line dyeing, brushing, local zooming, and axis reordering. To ensure a smooth interaction, animated transitions are implemented in this view.

Parameters view. Studying the effect of parameters in an ensemble simulation (that is, ensemble members are generated by different parameterizations) is an important task. For this purpose, we employ a radial plot where each parameter corresponds to one of the equiangular axes. Initially, all parameter configurations are shown as a context. Once users select a set of ensemble members, the corresponding parameter configurations are highlighted in this view.

Geo-Location view. For geolocation-related applications, a geo-location view is employed. Uncertainty features identified in the projection view may be easily understood and verified by the domain experts in this view.

6 Experiments and Discussion

We conducted three experiments to demonstrate the effectiveness and usefulness of our uncertainty-aware projection method and the exploration system. The system was implemented with the standard C++ and Qt 5.0. The experiments were performed on a PC equipped with an Intel Core 2 Duo 3.0 GHz CPU, 4GB host memory and an NVIDIA Quadro 4000 video card with 1.5 GB video memory.

6.1 The Synthetic Dataset

We generated a synthetic 5D dataset consisting of 300 ensemble data objects to demonstrate the effectiveness and the influence of dissimilarity weights. We first randomly chose

three anchor points that constitute an equilateral triangle as cluster centers in a 5D space. Then, we generated 100 ensemble data objects in each cluster. Each data object had 250 ensemble members following either a uniform or a normal distribution. The ensemble mean of a data object in a cluster was set as the cluster center with a small random bias, and the correlation matrix was set as a fixed scaled identity matrix. This process generated a dataset with three clusters. The Euclidean distances between any two cluster centers are equal. The geometrical differences of data objects in the same cluster are subtle but the distributional differences are significant. The equilateral triangle structure in this dataset is designed to demonstrate capability of our method to preserve the high-dimensional features in a visual space.

Fig. 6 shows the effect of the dissimilarity weights α and β (see Equation (7)) in our uncertainty-aware multidimensional projection method. Without considering the distributional differences ($\beta = 0$), all data objects with similar ensemble means are projected close to each other (see Fig. 6 (a)), even though data objects within a cluster have distinct distributions. This is consistent with the result of the conventional multidimensional projection methods designed for certain datasets. When the distributional differences are incorporated ($\beta > 0$), data objects within a cluster are gradually separated (see Fig. 6 (b-c)). For demonstration, we color the uniformly distributed data objects with green and normally distributed data objects with purple. From Fig. 6 (b-c), we can observe that there are three clusters. In each cluster, we can further identify two sub-clusters, each of which corresponds to a particular type of distribution. However, these structures cannot be observed in Fig. 6 (a), as data objects within a cluster become distinguishable only when ensemble distributions are considered.

In general, our method not only preserves the global structure of the three clusters but also distinguishes data objects between the two types of ensemble distributions within each cluster. The degree of separation of the data objects within a cluster depends on the relative value of α and β. When the ensemble mean is not considered ($\alpha = 0$), all data objects are separated into two clusters solely based on the two types of ensemble distributions (see Fig. 6 (d)).

6.2 The NBA Players' Statistics Dataset

We applied our method to a dataset consisting of 933 NBA players' career statistics[1] since 1981. In this dataset, we treat each player as an ensemble data object, and the statistics in a season as an ensemble member. Each ensemble member is a 16-dimensional vector that describes the statistics in a season including: Games Played, Minutes Played, Field Goals, Field Goals Attempt, 3-Point Field Goals, 3-Point Field Goal Attempted, Free Throws, Free Throw Attempted, Offensive Rebounds, Defensive Rebounds, Assists, Steals, Blocks, Turnovers, Personal Fouls, and Points. Each ensemble data object has at least 5 ensemble members. Because the number of the ensemble members for most ensemble data objects is smaller than the dimensionality 16, we chose the first uncertainty quantification measure to evaluate

1. The statistics of a player who has career over 5 seasons are collected from a professional basketball statistics website (http://www.basketball-reference.com).

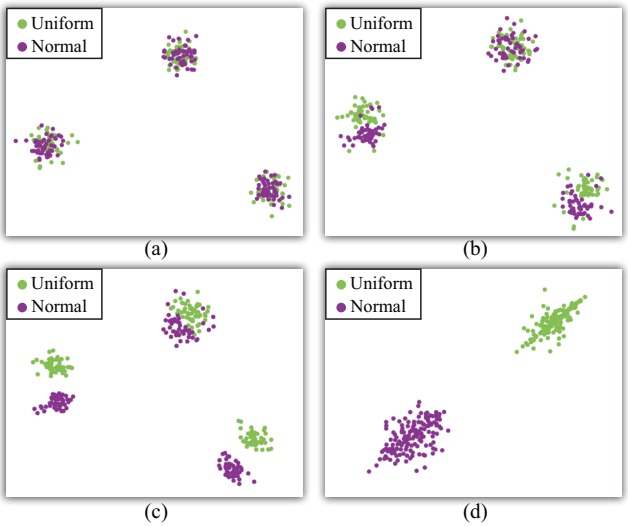


Fig. 6. The projections produced with different dissimilarity weights on the ensemble mean difference and the ensemble distribution difference in the dissimilarity measure. (a) $\alpha = 1.0$, $\beta = 0.0$. (b) $\alpha = 0.5$, $\beta = 1.0$. (c) $\alpha = 0.25$, $\beta = 1.0$. (d) $\alpha = 0.0$, $\beta = 1.0$.

the overall uncertainty (Equation (10)) and the ensemble member deviations from mean (Equation (11)). The overall uncertainty in this case reflects the fluctuation of a player's performance over years, or a player's inconsistency.

Fig. 7 shows the results of the uncertainty-aware projection with different parameters. The color encodes the position of a player. We can see a triangular structure in Fig. 7 (a-b). The points in the upper part mainly represent players who play center (**C**) and players who play both center and forward (**C-F**). The points in the lower part primarily represent players who play guard (**G**) and players who play both forward and guard (**F-G**). In addition, it is easy to find that yellow points are distributed in both the upper and the lower part. This is because players of forward (**F**) usually are versatile. For demonstration, we annotated a set of players with their names. The results in Fig. 7 (a-b) show that points on the left side are role players and the points on the right side are key players. More specifically, the points in the upper right corner represent excellent players of **C** or **C-F**, and the points in the lower right corner represent excellent players of **G** or **F-G**.

Fig. 7 (c) shows the result of projection using only distributional differences. This result indicates that the seasonal fluctuation patterns of the players in **C** and **C-F** positions are quite different from those of players of **F-G** and **G** positions. Furthermore, if we only use the geometrical differences among players to model their dissimilarities, the consistency of a player will be hidden. For example, without considering the distributional differences, the points in a region indicated by the magenta circle in Fig. 7 (a) are lumped together. This implies that their ensemble means are similar. However, Fig. 7 (c) reveals that their ensemble distributions are distinct. This explains why points inside the magenta circle in Fig. 7 (b) are much more dispersed. And this dispersion may help a general manager identify consistent players from a pool of players who are similar in mean performance.

To further inspect the difference between the results

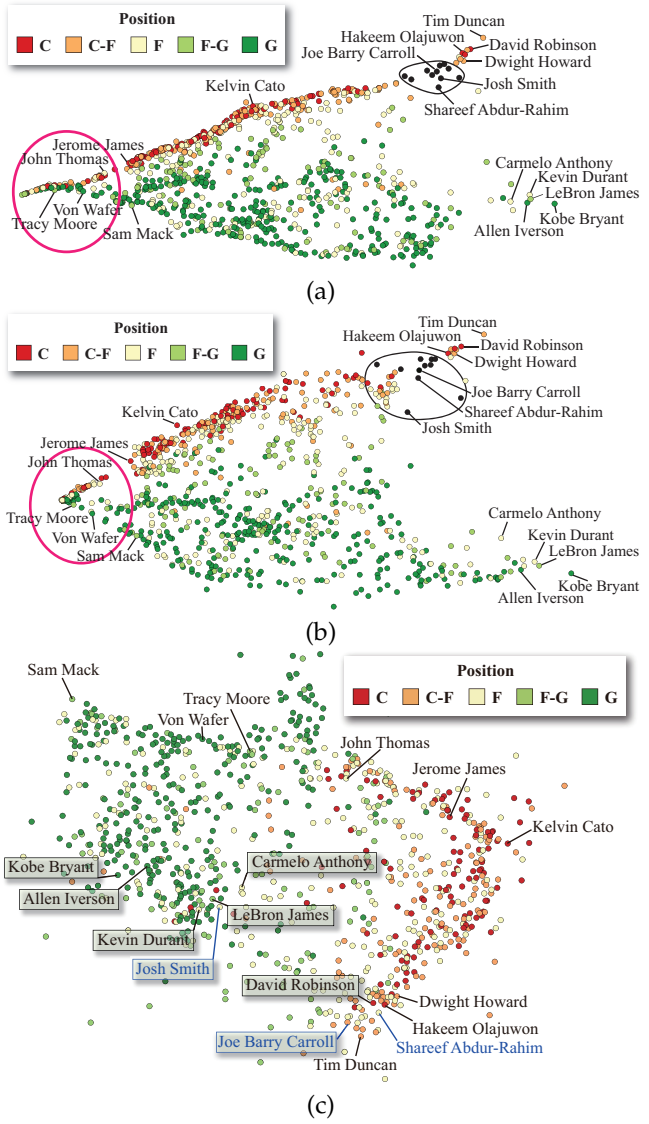


Fig. 7. Results for the NBA players' statistics dataset. (a) The result that considers only the geometrical differences, $\alpha = 1.0$, $\beta = 0.0$. (b) The result of our method with $\alpha = 0.55$, $\beta = 0.45$. (c) The result that considers only the distributional difference, $\alpha = 0.0$, $\beta = 1.0$. In this example, $|R_i| = 8$ and $|N_i| = 2$.

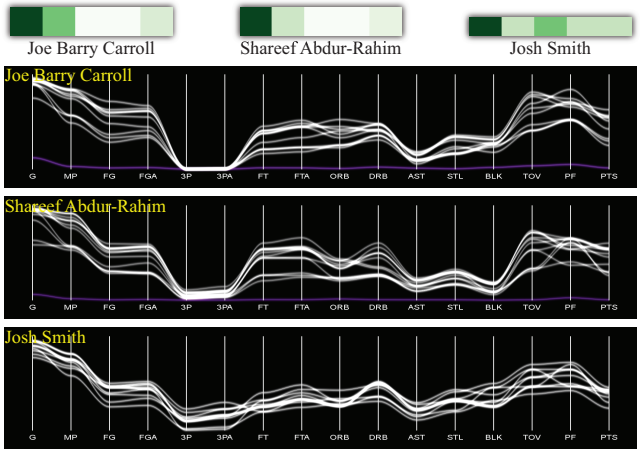


Fig. 8. The ensemble bars and parallel coordinates representation for the three selected players.

with and without considering distributional differences, we highlight a group of points in a region indicated by the black circle in Fig. 7 (a). The same group of points are also highlighted in black in Fig. 7 (b) for comparison. Because Fig. 7 (a) only considers the geometrical difference among ensemble means, the highlighted points close to each other have subtle differences among ensemble means. Take Josh Smith and Joe Barry Carroll as an example, they are located close to each other in Fig. 7 (a). However, they are dispersed when taking the distributional difference into account in Fig. 7 (b). This is because their distributional difference is significant, which can be clearly verified in Fig. 7 (c) (notice the annotations in blue). The ensemble bars and the parallel coordinates views shown in Fig. 8 further verify this observation. On the other hand, Joe Barry Carroll and Shareef Abdur-Rahim are positioned closer to each other when considering the distributional difference. This is because their ensemble distributions are quite similar to each other, which is evident in their ensemble bars and the parallel coordinates views in Fig. 8.

From the heights of the ensemble bars in Fig. 8, we can conclude that the consistency of Josh Smith is higher than those of Joe Barry Carroll and Shareef Abdur-Rahim. Further examining the details, we can see that the high overall uncertainties of Joe Barry Carroll and Shareef Abdur-Rahim are partly caused by some flat and close-to-zero outlier ensemble members (highlighted in purple). This was caused by seasons in which they may have injuries.

6.3 The Numerical Weather Simulation Dataset

The dataset used in this experiment is an ensemble WRF simulation of the 1993 superstorm at 6:00 PM, Mar. 13th, 1993. We select a region with latitude from 30°N to 40°N, and longitude from 85°W to 75°W. The dataset consists of 7,050 geospatial grid points. Each grid point contains 8 variables, including wind speed in the zonal (E-W) and meridional (N-S) directions respectively, temperature, soil temperature, water temperature, relative humidity, specific humidity, and mean sea level pressure. Forty runs of the WRF simulation with different cumulative schemes and microphysics schemes generated the ensemble members. In other words, our simulation produced an 8-dimensional ensemble dataset with 7,050 ensemble data objects, each of which has 40 ensemble members. As the number of ensemble members is much larger than the dimensionality of the dataset, the second uncertainty quantification measure was employed in this case.

Fig. 9 presents our results. To help users easily associate the geospatial locations in the projection view, we use color to encode the geospatial locations of data points. Specifically, we synthesize a 2D texture that covers the entire simulation region, i.e., latitude from 30°N to 40°N, and longitude from 85°W to 75°W. For simplicity, a linear gradient fill mode from violet to orange along the diagonal for this texture is adopted. In this way, the points sampled from the sea area will be encoded with color close to violet and the points sampled from the land area will be encoded with color close to orange (see Fig. 9). For clarity, the map is drawn as a background. From the projection result, we can easily find that points in the right part are closer to the sea area and points in the left part are closer to the land area.

Our system can help users explore the dataset from multiple perspectives and levels-of-detail. We present several exploration scenarios below.

Visual cluster analysis. Identifying highly correlated data objects is an important data exploration task to gain insight into an ensemble dataset. This is because highly correlated data objects (i.e., the clusters) constitute the major structure of a dataset.

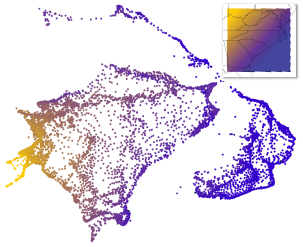


Fig. 9. The result of our projection method with $\alpha = 0.4$, $\beta = 0.6$, $|R_i| = 24$ and $|N_i| = 4$. The color of a point is indexed from a 2D linear map (from violet to orange) with the geo-coordinates.

Fig. 10 shows the result of our uncertainty-aware projection method colored by the overall uncertainty. In this result, a color set from purple (low uncertainty) to yellow (high uncertainty) is employed. Because the overall uncertainty varies significantly, we employ a log transformation and scale the overall uncertainty to [0,1].

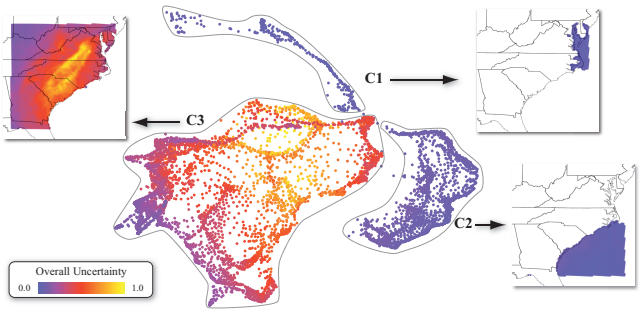


Fig. 10. Cluster identification and analysis based on our uncertainty-aware projection method. Three preliminary clusters are clearly shown with our method. The geospatial locations encoded with overall uncertainties are displayed on the map. In this example, $\alpha = 0.3$, $\beta = 0.7$, $|R_i| = 24$, and $|N_i| = 4$.

From the resulting 2D manifold, we can easily find that the entire dataset can be roughly grouped into three clusters **C1**, **C2**, and **C3**. With the lasso tool, we can further select the points in these clusters and show the geospatial locations associated with each data object in the Geo-Location view. The results show that **C1** contains data objects located in the Chesapeake bay and Delaware bay area, **C2** contains data objects located in the sea area, and **C3** contains data objects located over the land. This separation of data points based on land types is particularly interesting given that the geospatial location is not part of the data used in projection. From the results in the Geo-Location view, we can further find that data objects over the sea have lower uncertainties than those over the land.

For comparison, we also projected this dataset by only considering the geometrical differences and the distributional differences respectively. The near set, the random set, the neighborhood graph, and the control points used in each projection were identical, while α and β were different. Fig. 11 (a) shows the result that only takes the geometrical differences into account, i.e., $\alpha = 1.0$ and $\beta = 0.0$. Fig. 11 (c) displays the result that only considers the distributional differences, i.e., $\alpha = 0.0$ and $\beta = 1.0$. Fig. 11 (b) presents the result that considers both geometrical and distributional differences with $\alpha = 0.3$ and $\beta = 0.7$.

From Fig. 11 (a), we can see that the major structure shown in this result is similar to that in the result produced by our uncertainty-aware projection method in Fig. 11 (b). However, the difference between the data objects over the sea and the land is captured much more clearly by taking both the geometrical and distributional differences into account. Noticing the red points in **C1** highlighted in Fig. 11 (a), many of them are lumped with points in **C3**. This is because the geometrical differences of ensemble means among these points are subtle. On the other hand, from Fig. 11 (c) we know that the distributional differences between data objects in **C1** and **C3** are significant because **C1** and **C3** are clearly separated in this projection.

Furthermore, more points outside the 3 main clusters (e.g., points in the orange circle) are produced by the method without considering distributional differences in Fig. 11 (a). To investigate the reason why our uncertainty-aware method avoids these outliers, three groups (**G1**, **G2**, and **G3**) of points are highlighted in different colors. For inspection, the points of **G1**, **G2**, and **G3** are also highlighted in Fig. 11 (b) and Fig. 11 (c). We calculate the average pairwise geometrical difference between two groups by:

$$\overline{E}(P,Q) = \frac{\sum_{U_i \in P, U_j \in Q} E(\overline{U}_i, \overline{U}_j)}{|P|\,|Q|}. \tag{15}$$

Similarly, the average pairwise distributional difference between two groups is calculated by:

$$\overline{J}(P,Q) = \frac{\sum_{U_i \in P, U_j \in Q} J(U_i, U_j)}{|P|\,|Q|}. \tag{16}$$

In this case, $\overline{E}(G1,G2) = 0.29$, $\overline{E}(G1,G3) = 0.27$, meaning that the average geometrical dissimilarity of **G1** to **G2** and **G3** are almost the same. This faithfully explains why points in **G1** are located between **G2** and **G3** in Fig. 11 (a). However, $\overline{J}(G1,G2) = 0.82$, $\overline{J}(G1,G3) = 0.44$, meaning that the ensemble distributions of points in **G1** have higher dissimilarities to those points in **G2**. The resulting projection in Fig. 11 (c) visually verifies this observation. From the result presented in Fig. 11 (b), we conclude that our uncertainty-aware method strikes a balance between methods that only consider geometrical differences or distributional differences.

Another interesting observation is that there are two points located between **G1** and **G2** in Fig. 11 (b) (the region indicated by the black circle). We highlight their geospatial locations in the Geo-Location view. From the magnified map, we can see that they are sampled from the Florida shore of the Gulf of Mexico. For demonstration, we also highlight these two points both in Fig. 11 (a) and Fig. 11 (c) (the black points in the black circles). From Fig. 11 (a), we can infer that the ensemble means of these two points are similar to points over the land. From Fig. 11 (c), we can infer that their ensemble distributions are much more similar to points over the sea. These observations further prove the necessity of modeling the dissimilarities among ensembles with both geometrical and distributional differences.

Ensemble distribution investigation. Our system can also be employed to investigate the patterns of uncertainty

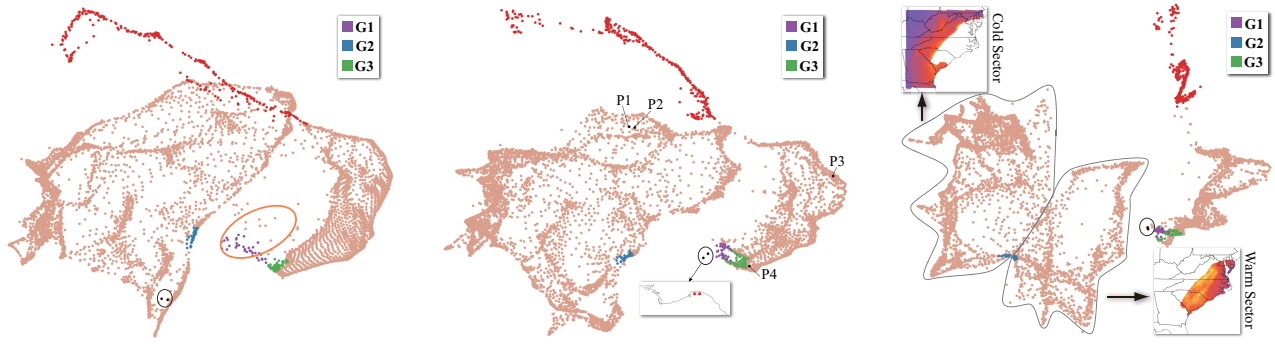


Fig. 11. (a) The result that considers only geometrical differences, $\alpha = 1.0$, $\beta = 0.0$, i.e. the projection method applied to the ensemble means. (b) The result of our uncertainty-aware projection, $\alpha = 0.3$, $\beta = 0.7$. (c) The result that considers only the distributional differences, $\alpha = 0.0$ and $\beta = 1.0$. In this example, $|R_i| = 24$ and $|N_i| = 4$.

in different areas and study the reason for the uncertainty of a single data object. For simplicity, we selected a set of representative points (**P1**, **P2**, **P3**, and **P4** in Fig. 11 (b)) in different areas for analysis. Fig. 12 shows the ensemble bars and parallel coordinate representations for the selected points. From the heights of the ensemble bars, we can see that **P1** or **P2** has a higher overall uncertainty than those of **P3** and **P4**. From the patterns depicted in the ensemble bars of **P1** and **P2**, we can infer that their ensemble member deviations follow an approximately uniform distribution. By examining their parallel coordinates representations, we can visually verify this observation. From the ensemble bars of **P3** and **P4**, we may conclude that their ensemble member deviations approximately follow a positively skewed distribution which is further confirmed by their parallel coordinates representations. Another interesting observation about the ensemble bars of **P3** and **P4** is that the color of the last bin varies greatly (see the magnified bins on the right side). This suggests that **P4** has more outlier members. By selecting the last bin in the ensemble bars of **P3** and **P4**, more ensemble members are highlighted in the parallel coordinates representation of **P4**.

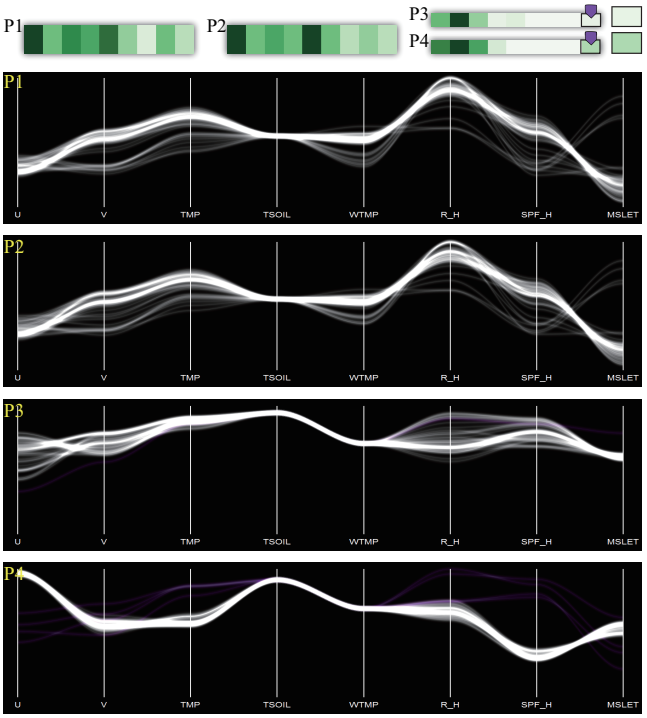


Fig. 12. The ensemble bars and parallel coordinates representations for the four selected data objects. The selected bins are magnified on the right side. All selected ensemble members are highlighted in purple.

6.4 Discussion

Parameter configuration. Fig. 3 (d-e) reveal that the near set $|N_i|$ and the random set size $|R_i|$ have a great impact on projection efficiency and accuracy. A large near set and random set require more computation time and resource, while a high projection accuracy (i.e., low stress) can be obtained. Because the dissimilarity estimation process relates to the number of ensemble members and the dimensionality of the dataset, it is intractable to find an optimal size for the near set and the random set. In our implementation, a good balance between efficiency and accuracy is achieved by setting $|N_i|$ close to $\sqrt{n}/3$ and $|R_i|$ no larger than 4.

The experiment in section 6.1 demonstrates the influence of the dissimilarity weights α and β. Empirically, if users need to put more emphasis on geometrical differences, a large α should be assigned. On the other hand, if users want to highlight the distributional differences among data objects, a large β needs to be assigned. Without any prior knowledge, α and β are both set to 0.5 for a preliminary study of an ensemble data set.

Projection performance. The overall projection complexity of our method is determined by three steps: building the neighborhood graph, estimating dissimilarities, and solving the sparse linear system.

Building the neighborhood graph involves defining a near set and a random set for each data object. A standard procedure to find the nearest neighbors is prohibitive $O(n^2)$. By reorganizing the input data objects into $\sqrt{n}$ clusters [47], the complexity can be reduced to $O(n^{\frac{3}{2}})$.

A core component of our method is to estimate the distributional differences among data objects in the neighborhood graph. The complexity is $O(n(|N_i| + |R_i|)mdS)$. Here $|N_i|$ is the size of the near set, $|R_i|$ is the size of the random set, m is the number of ensemble members, d is the dimension of the dataset, and S is the number of samples. In our

implementation $|N_i| = \sqrt{n}/3$, both $|R_i|$ and S are constants, thus the complexity becomes $O(n^{\frac{3}{2}}md)$.

As the derived linear system is sparse, symmetric, and positive definite, the conjugated gradient method is employed. In this case, the complexity is $O(n\sqrt{k})$ [47], where k is the condition number of the matrix $L^T L$. (L represents the coefficient matrix of the linear system.)

In general, the complexity of our method is $O(max(n^{\frac{3}{2}}md, n\sqrt{k}))$. In practice, dissimilarity estimation dominates the projection time in our experiments. For instance, several hours are required to compute the dissimilarities in the numerical weather simulation dataset on our computer. A parallel implementation of the entire system would alleviate this problem. The computation consumption can be further decreased by employing a fast KDE algorithm, e.g., the KD-tree based KDE. This is an avenue for future work. In addition, from Equation (7), we know that changing the value of α and/or β for exploration does not require the re-estimation of the geometrical and distributional differences. Accordingly, we can obtain a new projection in seconds by solving a new linear system.

Strengths and limitations. By considering both geometrical and distributional differences, data objects with similar summary statistics can be correctly distinguished with our approach. Our enhanced Laplacian-based projection scheme strikes a balance between the computation time and accuracy, which can be easily verified in Fig. 3 (d-e). Currently, both the probability reconstruction and the dissimilarity estimation processes in our approach are limited to numerical data objects. In the future, we will seek to adapt our approach for categorical multivariate ensemble data visualization and exploration.

7 Evaluation

We evaluated the exploration system with a meteorologist collaborator to analyze the ensemble simulation of the 1993 superstorm. We examined the ensemble data at the 44th hour of the simulation out of the entire 49-hour simulation. We intentionally chose a time point towards the end of the simulation when there are more diversities among the ensemble distributions. According to the findings in Section 6.3, the dissimilarity weights were specified as $\alpha = 0.3, \beta = 0.7$ so that data objects from different land types were clearly separated. The meteorologist can interactively adjust these weights to compare projection patterns. The observations and feedback are summarized in this section.

7.1 The Uncertainty-aware Projection Method

The meteorologist observed that, without considering the distributional differences among data objects, the Chesapeake bay/Delaware bay and the surrounding land are overlapped in the projection view (see Fig. 11 (a)). When considering both geometrical and distributional differences, these two types of land covers are clearly separated (see Fig. 11 (b,c)). The meteorologist hypothesized that this separation is caused by the more precise ensemble simulations over the sea than over the land. This hypothesis came from the past experience with WRF simulations and was validated by examining the parallel coordinates of data objects over the land and the sea.

The meteorologist also observed that the warm sector and cold sector (see Fig. 11 (c)) are much better separated in the projection view when considering distributional differences rather than considering geometrical differences only. This is because the uncertainty between the ensemble members significantly increased in the warm sector located closer to the center of storm at that time point. The reason for larger uncertainty is the uncertain timing of the cold front associated with the storm. The different ensemble members timed the cold frontal passage differently, which affects when each point in the warm sector leaves the warm sector and enters the cold sector.

Based on these observations, the meteorologist agreed that projecting the data objects according to both geometrical and distributional information provides a novel and useful perspective on the data.

7.2 The Uncertainty Visualization

The meteorologist stated that it is a routine but an effective means to use color for showing uncertainty on a geospatial map. The color map provided a useful overview of the uncertainty of the data objects in regions of high or low uncertainty. The meteorologist further confirmed that the high uncertainty region (i.e. the yellow region) in Fig. 10 coincides with the path of the super storm center.

Besides color mapping, the ensemble bar representation provides a supplementary visualization for the ensemble uncertainty. The meteorologist asserted that this representation was intuitive and helpful for him to study the consensus of ensemble members and compare different data objects. The meteorologist concluded that the ensemble bars can be used to characterize and identify different types of uncertainty. For example, the ensemble bars with dark green color to the left indicate positive skew distributions where most ensemble member agree each other with few outliers (e.g., in the sea surface). And ensemble bars with dark green color in the middle or right represent little to no agreement among ensemble members (e.g., in the warm sector).

7.3 The Linked Multi-view Interface

The meteorologist found the linked interaction mode helpful. A cluster of points in the projection view can be easily selected and their ensemble distributions can then be reviewed in the ensemble bar view immediately. The selection of bins in an ensemble bar allows a group of ensemble members to be quickly selected and reviewed in the parallel coordinates view and the parameter view. In particular, the last several bins in an ensemble bar usually contain the outlier ensemble members that require further analysis.

The parallel coordinates view helps the meteorologist examine the distribution of individual variables in a data object. The meteorologist prefers the continuous parallel coordinates plot for a more intuitive presentation of the distribution. The ability to reorder the axes in the parallel coordinates view helps organize the similarly distributed variables close to each other for comparison.

The meteorologist pointed out that the soil temperature variable was not simulated, but inherited from the input

NARR dataset, therefore having no uncertainty in the ensembles. The meteorologist expressed the desire to interactively select a group of variables for analysis. Currently, this feature is not supported in our system due to the time-consuming dissimilarity computation. However, a parallel computing system can help achieve this goal in the future.

8 Conclusion

We present a visualization and exploration system for multidimensional ensemble dataset whose kernel is a novel uncertainty-aware multidimensional projection method. This new method considers not only the ensemble means but also the ensemble distributions. Experiments on the artificial dataset demonstrate the ability of the uncertainty-aware projection to distinguish between ensemble data objects with similar means but different ensemble distributions. Results on both real-word and simulation datasets verify that 1) differences in ensemble distributions are an important part and crucial for proper ensemble analysis, and 2) our uncertainty-aware projection can distinguish ensemble distributions in the real-world data.

Acknowledgments

This work was partly supported by National High Technology Research and Development Program of China (2012AA12090), Major Program of National Natural Science Foundation of China (61232012), National Natural Science Foundation of China (61422211), Zhejiang Provincial Natural Science Foundation of China (LR13F020001), National Science Foundation (NSF1117871), Pacific Northwest National Laboratory under U.S. Department of Energy Contract (DE-AC05-76RL01830), National Science Foundation of China (61379076), Program for New Century Excellent Talents in University of China (NCET-12-1087), Zhejiang Provincial Natural Science Foundation under Grant No. LR14F020002, and the NUS-ZJU SeSama center.

References

[1] J. Michalakes, J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock, and W. Wang, "The weather research and forecast model: software architecture and performance," in *Proceedings of the 11th ECMWF Workshop on the Use of High Performance Computing In Meteorology*, 2004, pp. 156–168.

[2] C. Johnson and A. Sanderson, "A next step: Visualizing errors and uncertainty," *IEEE Computer Graphics and Applications*, vol. 23, no. 5, pp. 6–10, 2003.

[3] K. Potter, S. Gerber, and E. Anderson, "Visualization of uncertainty without a mean," *Computer Graphics and Applications*, vol. 33, no. 1, pp. 75–79, 2013.

[4] M. Hlawatsch, P. Leube, W. Nowak, and D. Weiskopf, "Flow radar glyphs-static visualization of unsteady flow with uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 1949–1958, 2011.

[5] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson, "Visualizing summary statistics and uncertainty," *Computer Graphics Forum*, vol. 29, no. 3, pp. 823–832, 2010.

[6] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead, "Noodles: A tool for visualization of numerical weather model ensemble uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1421–1430, 2010.

[7] R. Botchen, D. Weiskopf, and T. Ertl, "Texture-based visualization of uncertainty in flow fields," in *Proceedings of Visualization, 2005. VIS 05. IEEE*, 2005, pp. 647–654.

[8] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang, "Visualizing scalar volumetric data with uncertainty," *Computers & Graphics*, vol. 26, no. 2, pp. 239–248, 2002.

[9] K. Pöthkow, B. Weber, and H. Hege, "Probabilistic marching cubes," *Computer Graphics Forum*, vol. 30, no. 3, pp. 931–940, 2011.

[10] K. Pöthkow and H. Hege, "Positional uncertainty of isocontours: Condition analysis and probabilistic measures," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 10, pp. 1393–1406, 2011.

[11] L. Gosink, K. Bensema, T. Pulsipher, H. Obermaier, M. Henry, H. Childs, and K. Joy, "Characterizing and visualizing predictive uncertainty in numerical ensembles through bayesian model averaging," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2703–2712, 2013.

[12] H. Mathias, O. Harald, G. Christoph, and I. Kenneth, "Comparative visual analysis of lagrangian transport in cfd ensembles," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, p. In press, 2013.

[13] I. Borg and P. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[14] P. Joia, F. Paulovich, D. Coimbra, J. Cuminato, and L. Nonato, "Local affine multidimensional projection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2563–2571, 2011.

[15] A. Pang, C. Wittenbrink, and S. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

[16] S. Deitrick and R. Edsall, "The influence of uncertainty visualization on decision making: An empirical evaluation," *Progress in Spatial Data Handling*, pp. 719–738, 2006.

[17] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel, "A typology for visualizing uncertainty," in *Proceedings of SPIE*, vol. 5669, 2005, pp. 146–157.

[18] K. Potter, P. Rosen, and C. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," *Uncertainty Quantification in Scientific Computing*, pp. 226–249, 2012.

[19] T. Pfaffelmoser, M. Mihai, and R. Westermann, "Visualizing the variability of gradients in uncertain 2d scalar fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 11, pp. 1948–1961, 2013.

[20] G. Grigoryan and P. Rheingans, "Point-based probabilistic surfaces to show surface uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 5, pp. 564–573, 2004.

[21] V. Dinesha, N. Adabala, and V. Natarajan, "Uncertainty visualization using hdr volume rendering," *The Visual Computer*, vol. 28, no. 3, pp. 265–278, 2012.

[22] C. H. Lee and A. Varshney, "Representing thermal vibrations and uncertainty in molecular surfaces," in *SPIE Conference on Visualization and Data Analysis*, vol. 4665, 2002, pp. 80–90.

[23] P. Rhodes, R. Laramee, R. Bergeron, and T. Sparr, "Uncertainty visualization methods in isosurface rendering," in *Eurographics*, vol. 2003, 2003, pp. 83–88.

[24] B. Zehner, N. Watanabe, and O. Kolditz, "Visualization of gridded scalar data with uncertainty in geosciences," *Computers & Geosciences*, vol. 36, no. 10, pp. 1268–1275, 2010.

[25] G. Schmidt, S. Chen, A. Bryden, M. Livingston, L. Rosenblum, and B. Osborn, "Multidimensional visual representations for underwater environmental uncertainty," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp. 56–65, 2004.

[26] T. Pfaffelmoser, M. Reitinger, and R. Westermann, "Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields," *Computer Graphics Forum*, vol. 30, no. 3, pp. 951–960, 2011.

[27] A. Coninx, G. Bonneau, J. Droulez, and G. Thibault, "Visualization of uncertain scalar data fields using color scales and perceptually adapted noise," in *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, 2011, pp. 59–66.

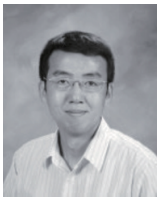
[28] C. Lundstrom, P. Ljung, A. Persson, and A. Ynnerman, "Uncertainty visualization in medical volume rendering using probabilistic animation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1648–1655, 2007.

[29] Y. Wu, G. Yuan, and K. Ma, "Visualizing flow of uncertainty through analytical processes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2526–2535, 2012.

[30] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead, "A user study to compare four uncertainty visualization methods

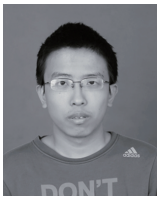
for 1d and 2d datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1209–1218, 2009.
[31] A. Wilson and K. Potter, "Toward visual analysis of ensemble data sets," in *Proceedings of the Workshop on Ultrascale Visualization*, 2009, pp. 48–53.
[32] K. Potter, A. Wilson, P. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johnson, "Ensemble-vis: A framework for the statistical visualization of ensemble data," in *IEEE International Conference on Data Mining Workshops*, 2009, pp. 233–240.
[33] P. Diggle, P. Heagerty, K. Liang, and S. Zeger, *Analysis of longitudinal data*. Oxford University Press, 2002.
[34] T. Hollt, A. Magdy, G. Chen, G. Gopalakrishnan, I. Hoteit, C. Hansen, and M. Hadwiger, "Visual analysis of uncertainties in ocean forecasts for planning and operation of off-shore structures," in *Proceedings IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2013, pp. 185–192.
[35] P. Wong and R. Bergeron, "30 years of multidimensional multivariate visualization," *Scientific Visualization, Overviews, Methodologies, and Techniques*, pp. 3–33, 1997.
[36] J. Daniels, E. Anderson, L. G. Nonato, and C. Silva, "Interactive vector field feature identification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1560–1568, 2010.
[37] W. Chen, Z. Ding, S. Zhang, A. MacKay-Brandt, S. Correia, H. Qu, J. Crow, D. Tate, Z. Yan, and Q. Peng, "A novel interface for interactive exploration of dti fibers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1433–1440, 2009.
[38] A. Anand, L. Wilkinson, and T. Dang, "Visual pattern discovery using random projections," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 43–52.
[39] S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.
[40] D. Scott and S. Sain, "Multi-dimensional density estimation," *Handbook of Statistics*, vol. 23, pp. 229–263, 2004.
[41] M. Minnotte and D. Scott, "The mode tree: A tool for visualization of nonparametric density features," *Journal of Computational and Graphical Statistics*, vol. 2, no. 1, pp. 51–68, 1993.
[42] O. Lampe and H. Hauser, "Interactive visualization of streaming data with kernel density estimation," in *IEEE Pacific Visualization Symposium*, 2011, pp. 171–178.
[43] B. Silverman, *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1986.
[44] D. Scott, *Multivariate density estimation*. Wiley, 1992.
[45] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 751–763, 2011.
[46] C. Andrieu, N. De Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1, pp. 5–43, 2003.
[47] F. Paulovich, L. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 3, pp. 564–575, 2008.
[48] S. Ingram, T. Munzner, and M. Olano, "Glimmer: Multilevel mds on the gpu," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 2, pp. 249–261, 2009.
[49] T. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985.
[50] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

Haidong Chen received the BS degree from Northeast Normal University in 2009 and the master degree from Zhejiang University in 2011. He is currently pursuing the PhD degree in Zhejiang University. His research interests include uncertainty visualization and high-dimensional data visualization.

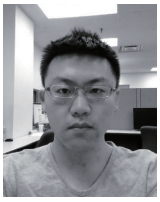


Song Zhang received the BS degree in computer science from Nankai University in 1996 and the PhD degree in computer science from Brown University in 2006. He is an associate professor in the Department of Computer Science and Engineering, Mississippi State University. His research interests include scientific visualization, data Analysis, medical imaging, and computer graphics.

Wei Chen received the PhD degree in Fraunhofer Institute for Graphics, Darmstadt, Germany, in July 2002. He is a professor in State Key Lab of CAD & CG at Zhejiang University, P.R. China. From July 2006 to September 2008, he was a visiting scholar at Purdue University. He has performed research in computer graphics and visualization and published more than 60 peer-reviewed journal and conference papers in the last five years. His current research interests include visualization and visual analytics.



Honghui Mei is currently pursuing the PhD degree in Zhejiang University. His research interests include visualization and visual analytics.



Jiawei Zhang is a Ph.D. student in Electrical and Computer Engineering at Purdue University. His research interests include Visual Analytics and Information Visualization.

Andrew Mercer received the PhD in meteorology from the University of Oklahoma in 2008. He is a meteorologist whose previous research has involved using artificial intelligence and advanced statistics to model weather phenomena. His current research interests lie in the integration of artificial intelligence into all facets of the geosciences.

Ronghua Liang received the PhD degree in computer science from Zhejiang University in 2003. He is currently a Professor of Computer Science and the Vice Dean of College of Information Engineering, Zhejiang University of Technology, China. His research interests include Information Visualization, Computer Vision, and Medical Visualization.

Huamin Qu obtained a BS in Mathematics from Xi'an Jiaotong University, China, an MS and a PhD (2004) in Computer Science from the Stony Brook University. He is an associate professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His main research interests are in visualization and computer graphics. He has co-authored more than 60 refereed papers including 20 papers in the IEEE TVCG. He is on the steering committee of the IEEE Pacific Visualization Conferences and is an associate editor of IEEE TVCG.