Range Likelihood Tree: A Compact and Effective Representation for Visual Exploration of Uncertain Data Sets

Wenbin He [1] Xiaotong Liu [1] Han-Wei Shen [1] Scott M. Collis [2] Jonathan J. Helmus [2]

1) The Ohio State University *

2) Argonne National Laboratory †

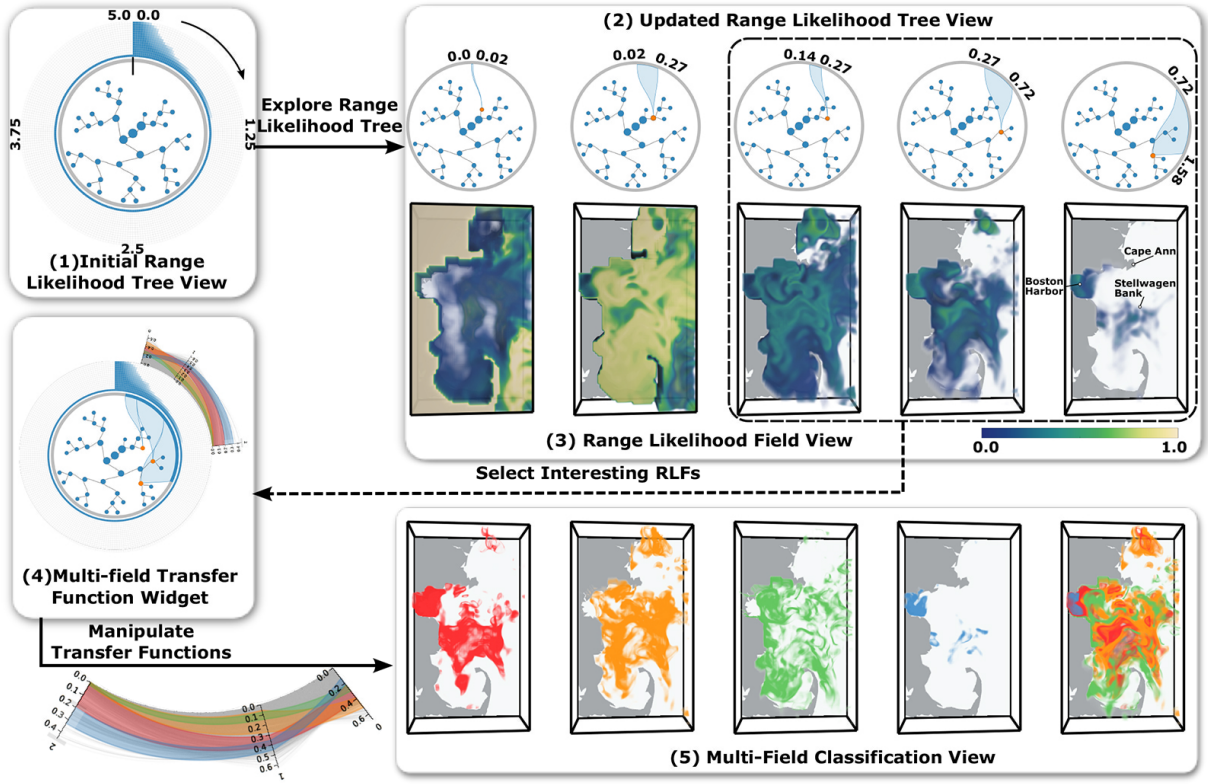


Figure 1: An illustration of our range likelihood tree guided exploration framework using the Massachusetts Bay Sea Trial Ensemble dataset. Starting from the initial range likelihood tree view (1), users can hover on different nodes in the tree to see their corresponding subranges (2), and examine the likelihoods of associated grid points in the range likelihood field view (3); after selecting a few subranges of interest, a multi-field transfer function widget (4) is created, which can be manipulated by the user to explore and classify grid points based on their likelihoods in different subranges in the multi-field classification view (5).

Abstract

Uncertain data visualization plays a fundamental role in many applications such as weather forecast and analysis of fluid flows. Exploring scalar uncertain data modeled as *probability distribution fields* is a challenging task because the underlying features are often more complex, and the data associated with each grid point are high dimensional. In this work, we present a compact and effective representation, called range likelihood tree, to summarize and explore probability distribution fields. The key idea is to decompose and summarize each complex probability distribution over a few representative subranges by cumulative probabilities, and allow users to consider the roles that different subranges play in understanding the probability distributions. In our method, the value domain is first partitioned into subranges, then the distribution at each grid point is transformed according to the cumulative probabilities of the point's distribution in those subranges. Organizing the subranges into a hierarchical structure based on how these cumulative probabilities are spatially distributed in the grid points, the new range likelihood tree representation allows effective classification and identification of features through user query and exploration. We present an exploration framework with multiple interactive views to explore probability distribution fields, and provide guidelines for visual exploration using our framework. We demonstrate the effectiveness and usefulness of our approach in exploratory analysis using several representative uncertain data sets.

*e-mail: {he.495,liu.1952,shen.94}@osu.edu

†e-mail: {scollis,jhelmus}@anl.gov

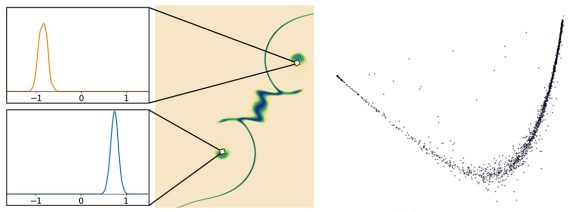
1 Introduction

Understanding uncertainty is one of the major scientific challenges in the era of big data. It has a great influence on many applications such as weather forecast and analysis of fluid flows [6, 13, 15, 20].

As computing power continues to grow, state of the art simulations are able to model uncertainty with increasing complexities. For instance, to help estimate climate changes, scientists perform ensemble simulations with multiple parameterizations and stochastic initial conditions to study different outcomes and analyze the inherent uncertainty in the simulations. Uncertainty quantification and visualization play a fundamental role in validating such simulation outcomes and understanding the underlying scientific phenomena [6, 13, 15]. To visualize uncertainty, one common choice is to model uncertainty at each spatial location as a probability distribution of possible simulation outcomes and form a *probability distribution field* in the entire domain [28, 33, 37, 44]. However, with the increasing complexity of uncertainty (e.g., from a single gaussian to a multi-modal distribution), visual exploration of scalar uncertain data sets is a challenging task because the underlying features are often complex, and the data associated with each grid point are high dimensional. Consequently, conventional scalar field visualization techniques such as isocontour extraction and direct volume rendering are not readily applicable.

The visualization community has made a continuing effort to tackle the challenge of visualization and visual analysis of probability distribution fields [2, 3, 6, 13, 33, 37, 38]. Existing research is mostly focusing on the use of *statistical summaries* and *dissimilarity measures*. Common statistical summaries such as mean and standard deviation are among the most straightforward approaches, but they fail to reveal uncertain behaviors modeled by bimodal or multi-modal distributions [6, 13]. For classification purpose, one can specify a target distribution as a feature of interest, and classify the distribution field based on how dissimilar a distribution is to the target [2, 3, 33] (Figure 2(a)). However, ambiguities may occur when dissimilar distributions have similar distances with respect to the target [2, 3]. In addition, due to the ever increasing complexity of scientific phenomena, it is often difficiult to define distributions of interest in a precise manner. While automatic cluster analysis can work in the absence of precise target definition, meaningful clusters of probability distributions may not be easily obtained from a large number of probability distributions of complex characteristics. The choice of an appropriate dissimilarity measure is also non-trivial and often domain-specific [13]. Furthermore, low-dimensional embedding that respects the dissimilarities among probability distributions, using common projections techniques such as principle component analysis (PCA) and multi-dimensional scaling (MDS), can generate results that are difficult to understand (Figure 2(b)).

In this work, we present a compact and effective representation called *Range Likelihood Tree (RLT)*, to summarize and explore probability distribution fields. The key idea is to consider the different roles that *subranges* (subspaces of the value domain) may play in understanding probability distributions, and decompose and summarize each complex probability distribution over a few representative subranges by cumulative probabilities. In our method, the value domain of the entire field is first partitioned into small subranges, and the distribution at each grid point is transformed based on these subranges. For each subrange, a *range likelihood field (RLF)* is formed where the scalar value at each grid point is computed as a cumulative probability of the point's distribution within the subrange. Based on how these cumulative probabilities are distributed spatially in the grid points, RLFs are organized into a hierarchical representation. The use of the new RLT representation allows effective classification and identification of features through user query and exploration. We present an exploration framework with multiple interactive views to explore probability distribution fields through RLFs, and provide guidelines for visual exploration using our framework. We show that RLTs can assist users to progressively gain knowledge by exploring RLFs, examining the corresponding probability distribution field, and classifying the field through interactive transfer function manipulation. We demonstrate the effectiveness and usefulness of our approach in exploratory analysis using several representative uncertain data sets, and verify the visualization results with domain scientists in environment science.



(a) (b)

Figure 2: Visual analysis of probability distribution fields using conventional techniques. (a) Visualizing a distance field constructed from a 2D probability distribution field; regions associated with two distinct distributions can not be readily distinguished from their distances to the target. (b) Visualizing a low-dimensional embedding of a 3D probability distribution field using principal component analysis (PCA); no clear separation of clusters can be found.

2 Related Work

In this section, we briefly review and discuss the previous research that is closely related to our work

Uncertainty Visualization. Brodlie et al. [8] reviewed the state of the art in uncertainty visualization. Glyph-based methods that encode uncertainties by glyphs were explored by Hlawatsch et al. [22] and Wittenbrink et al. [46]. Pfaffelmoser et al. [36] proposed methods to visualize the variability of gradients in 2D uncertain scalar fields. Fout and Ma [18] designed a system that provides verifiable volume rendering via uncertainty analysis. Djurcilov et. al. [17] incorporated uncertainty into volume rendering. Grigoryan et al. [19] proposed a method to visualize surfaces along with uncertainties. Dinesha et al. [16] used high dynamic range (HDR) technology for uncertain volume visualization. Rhodes et al. [39] proposed two techniques for rendering isosurfaces with uncertainty. Besides, animation-based techniques for visualizing uncertainty were explored by Lundstrom et al. [32]. These previous works proposed visualization methods for various models of uncertainty. In this work, we focused on visualizing and exploring scalar uncertain data modeled as probability distribution fields.

Probability Distribution Field Visualization. Luo et at. [33] introduced distributions as a new data type and discussed several techniques for visualizing spatial distribution data. Recently, there are mainly two approaches to analyze and visualize probability distribution fields. One of these approaches is analyzing and visualizing statistical characteristics of distributions. Thompson et al. [44] visualized hixel fields through topology analysis and fuzzy isosurface. Athawale et al. [4, 5] as well as Pöthkow [38] proposed uncertain isocontour extraction methods for probability distribution fields. Pöthkow and Hege [37] exploited nonparametric models for the computation of feature probabilities from uncertain data. Recently, Bensema et al. [6] proposed a technique that classifies ensemble variance based on the modality of the distributions of ensemble predictions. The other approach is analyzing and visualizing dissimilarity measures between distributions. Anderson et al. [2, 3] presented interactive techniques for visualization and analysis of function field data through dissimilarity analysis. Meanwhile, Chen et al. [13] proposed a projection framework that explores uncertainties of a multidimensional ensemble data through uncertainty-aware projection. Unlike previous works, we model a probability distribution field into a compact and effective range-based representation for visual exploration.

Range-based Classification. Value range based methods have been successfully used for feature identification and classification in computer vision and scientific visualization. Otsu et al. [35] presented a method to find a threshold for dividing the intensity value in two subranges. Jolion et al. [24] and Chang et al. [11] further partitioned the intensity value range into multiple subranges. Lundström et al. [30, 31] used value ranges for medical data classification and transfer function design. In our work, we applied range-based classification by transforming a probability distribution to range likelihoods by integrating its cumulative probabilities over a few representative subranges, which are organized into a tree representation that allows simple and effective classification through user query and exploration.

3 System Overview

The main objective of this work is to guide visual exploration of uncertain data sets modeled by probability distributions. Essentially, we transform the problem of distribution-based uncertain data exploration to the problem of exploring a set of subranges that are more intuitive and explanatory. Our system has two major modules: the data transformation module and the interactive visualization module, as shown in Figure 3. Given a probability distribution field, the value domain is first partitioned into subranges, and the distribution at each grid point is transformed according to the cumulative probabilities of the point's distribution in those subranges. For each subrange, a *range likelihood field (RLF)* is formed where the scalar value at each grid point is computed as a cumulative probability of the point's distribution within that subrange. Based on these cumulative probabilities in each RLF, we compute a distance matrix of RLFs, and organize them into a hierarchical representation, called *Range Likelihood Tree (RLT)*, which group similar RLFs into clusters for efficient visual exploration. The resulting RLT from the data transformation module is then used in the visualization module for visual exploration and classification of the underlying probability distribution field.

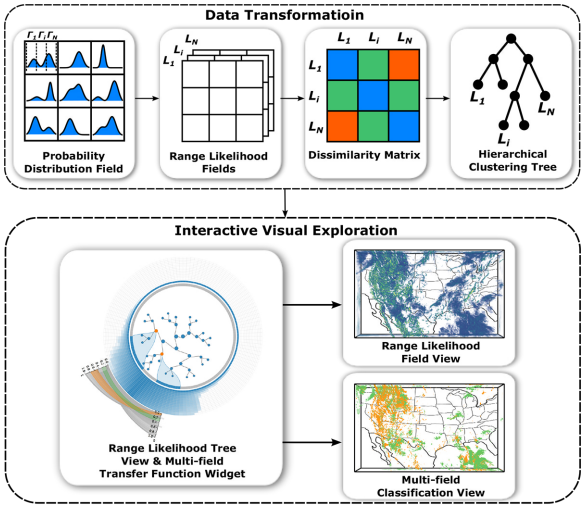


Figure 3: Overview of the analytical workflow.

4 Methods

In this section, we first introduce the concept of range likelihood which transforms a probability distribution into multiple cumulative probabilities over several subranges. Next, we describe how to organize the subranges using a multi-level data model called *Range Likelihood Tree (RLT)*. Finally, we discuss how to classify a probability distribution field using RLT.

4.1 Transforming Distribution into Range Likelihoods

A probability distribution field PF assigns every point p in Euclidean space with a probability density function:

$$PF(p) \sim f_X(p,x) \tag{1}$$

where X is a random variable over the value domain D.

Since each grid point is associated with a probability distribution, exploring all distributions in a probability distribution field with different shapes and modalities is difficult. To obtain an effective overview of a complex and heterogeneous probability distribution field, a probability distribution can be summarized using simpler cumulative probabilities over several subranges. For example, probability distributions in Figure 4 can be represented by two subranges $[0,0.5)$ and $[0.5,1]$ with cumulative probabilities as $[0.7,0.3]$, $[0.5,0.5]$, and $[0.3,0.7]$, respectively. Formally, given N subranges $\{\Gamma_1,\Gamma_2,...,\Gamma_N\}$ that partition the value domain D of a probability distribution field, the distribution at each grid point is transformed into these subranges. For each subrange Γ_i, a *range likelihood field (RLF)* L_X is formed where the scalar value at each grid point p is computed as a cumulative probability of the point's distribution $f_X(p,x)$ within the respective subrange:

$$L_X(p;\Gamma_i) = \int_{\Gamma_i} f_X(p;x)dx, i = \{1,...,N\}. \tag{2}$$

We denote the transformation from a probability distribution into a range likelihood field as *range likelihood transformation*.

Figure 4: Three probability distributions from different spatial locations in an uncertain scalar field. They are summarized by two subranges (in two distinct colors) with the probabilities that each distribution falls within the subranges as $[0.7,0.3]$, $[0.5,0.5]$, and $[0.3,0.7]$, respectively.

Essentially, a RLF is a univariate scalar field, which can be visualized through isocontour extraction or direct volume rendering. Through range likelihood transformation of the distribution at every grid point p, a probability distribution field can be represented by multiple RLFs (forming a *multi-field*), while each RLF corresponds to the cumulative probabilities of a subrange that may highlight a salient feature from the distribution field. As shown in Figure 5, a probability distribution field (visualized as several sampled scalar fields in Figure 5(a)) is transformed into several representative RLFs of different ranges (Figure 5(b)). Such range likelihood transformation of a distribution field has several key benefits:

1. Each RLF indicates a set of grid points of non-zero likelihoods in its corresponding subrange, which may represent a particular feature. The corresponding set of grid points in the distribution field can be quickly retrieved given a RLF.

2. Associating visualization results with subranges allows users to consider the roles that different subranges play in understanding the probability distributions and to explore and relate a feature to the domain knowledge.

3. Given multiple RLFs, every grid point is associated with multiple likelihoods in different subranges, which can be flexibly combined to classify the distribution field for feature identification.

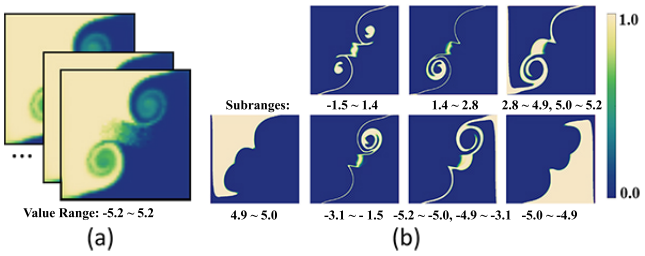


Figure 5: Transforming distribution into range likelihoods. (a) Visualization of scalar fields sampled from a 2D distribution field. (b) Visualization of range likelihood fields (RLFs) transformed from the distribution field.

4.2 A Multi-level Data Model for Range Likelihood Fields

Since a value domain D can be partitioned into subranges in a number of ways, an appropriate partitioning is critical to the effectiveness of range likelihood transformation. A simple and popular approach is to uniformly partition the value domain into N subranges of equal sizes (e.g., histograms with equal sizes of bins). However, a more effective partitioning should take the input data into consideration, as features may not be uniformly distributed in the value domain. As a result, when the domain is not partitioned fine enough (N is too small), distinct features may not be separated; on the other hand, when the domain is overly partitioned (N is too large), features may be broken into small fragments, and storage overhead may become too large as each subrange corresponds to a range likelihood field.

To provide an effective partition of the value domain, we take a divide-and-conquer approach: (1) the value domain is first partitioned into N subranges, and the distribution at each grid point is transformed into its subranges; the initial value of N is set as 256 to balance the accuracy and storage overhead; (2) a distance matrix is computed between every pair of the N RLFs based on a similarity measure; (3) the N RLFs are organized into a binary tree using hierarchical clustering; and (4) a compact tree representation is produced after merging similar RLF clusters. We call this data representation *Range Likelihood Tree (RLT)*. Below we describe the key steps to construct such a representation.

Measuring similarity of RLFs Since each grid point has a likelihood in a given RLF, we consider two RLFs *similar* if their associated grid points of non-zero likelihoods are spatially similar (illustrated in Figure 6). Hence we consider RLFs as distributions with *spatial location* being the random variable for comparison:

$$\hat{L_X}(p;\Gamma_i) = \frac{L_X(p;\Gamma_i)}{\int_R L_X(p;\Gamma_i)dp}, \tag{3}$$

which scales the likelihood at a spatial location p for subrange Γ_i by the total sum of the likelihoods over the entire spatial domain R. To compute the distance between two distributions, we employ a statistical measure *JensenShannon divergence (JSD)* [10]. We refer to detailed experimental validation of JSD in comparison with alternative measures in Section 7.

Hierarchical Clustering of RLFs Based on the similarity measure, we compute a distance matrix for every pair of RLFs. Agglomerative hierarchical clustering is then applied to the distance

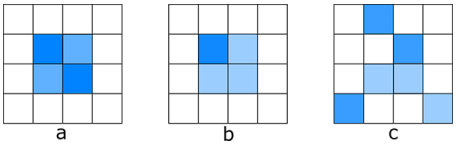


Figure 6: An illustration of three range likelihood fields (RLFs), highlighting grid points with non-zero likelihoods. Based on spatial locality of grid points, RLF-a is similar to RLF-b, while RLF-c is different from RLF-a and RLF-b.

matrix to create an unbalanced, binary clustering tree of RLFs in a bottom-up manner. Starting with each RLF in a separate cluster, pairs of resulting clusters are merged successively until all RLFs are contained in one large cluster. Distance between clusters is determined based on average distance of all pairs of RLFs between clusters. Using agglomerative hierarchical clustering, the number of clusters is not required beforehand and the clustering tree can be split easily into the desired number of clusters. In this way, RLFs of similar features can be clustered into one subtree, while RLFs of distinct features will be organized into different branches.

Pruning of Hierarchical Clustering Tree As a result of hierarchical clustering, a hierarchical clustering tree of RLFs is constructed, which is called *Range Likelihood Tree (RLT)*. Each node of the tree contains a cluster of RLFs and RLF clusters at a higher level are more distinct than those at a deeper level. To keep a compact representation of the underlying probability distribution field with representative RLFs, the hierarchical-clustering tree is further pruned by merging similar RLF clusters. Starting from the root, for each node, we evaluate the dissimilarity between its child nodes. If the dissimilarity is below a predefined threshold D_t, nodes in the subtree are reduced into one node and the RLFs in the leaf nodes $\{L_1, L_2, ..., L_n\}$ of this subtree are merged into one RLF L based on:

$$L(p) = \sum_{i=1}^{n} L_i(p) \tag{4}$$

If the dissimilarity is above the threshold D_t, then we keep this node and apply the same operation on its child nodes. The pruning of a RLT also reduces the storage cost because fewer RLFs needs to be kept to represent the underlying probability distribution field.

4.3 Range Likelihood Based Probabilistic Classification

By selecting multiple RLFs from the RLT, every grid position p is now associated with a *feature vector* of likelihoods $\vec{l}$, where each element in this vector represents the cumulative probability of the distribution at position p over a subrange. The problem of probability distribution field classification is then transformed to the problem of multi-field classification. To enable visual classification of multi-field data, we start with a set of user selected feature vectors and cluster them into several clusters. Each cluster is represented as a mean vector and a covariance matrix. Then these clusters are used to generate a multi-field transfer function, which is later used to assign the color and opacity to each feature vector hence the corresponding voxel of the data.

To cluster user-selected feature vectors which can be treated as points in multi-dimensional space, we use a Gaussian mixture model (GMM) to find clusters as well as their mean vectors and covariance matrices. Since the number of clusters could change for different user-selected feature vectors, instead of using the expectation maximization (EM) algorithm which requires to set number of clusters before clustering, we employ the Minimum Volume Ellipsoid Estimator (MVE) method [24] to automatically cluster feature vectors into an optimal number of clusters. It partitions the multi-field space

into candidate clusters, and iteratively performs the Kolmogorov-Smirnov test to find the best fitting clusters, while each cluster is fit into a Gaussian function. Finally, the user-selected feature vectors are fit by a GMM and the membership of a feature vector for one Gaussian cluster is evaluated by:

$$G(\vec{l},\vec{\mu},\Sigma) = e^{-\frac{1}{2}(\vec{l}-\vec{\mu})'\Sigma^{-1}(\vec{l}-\vec{\mu})} \quad (5)$$

where $\vec{l}$ is a feature vector, $\vec{\mu}$ is the mean vector that the Gaussian is centered over, and Σ is the covariance matrix of the Gaussian.

The fitted GMM is then used to guide the visual classification of the multi-field data by treating each Gaussian component of the GMM as an Ellipsoid Gaussian transfer function [21, 25, 45], which maps a given feature vector $\vec{l}$ with an opacity α:

$$\alpha(\vec{l}) = \alpha_{max} G(\vec{l},\vec{\mu},\Sigma) \quad (6)$$

where $\vec{\mu}$ and Σ is a mean vector and a covariance matrix of a Gaussian component of the fitted GMM. α_{max} is a constant to scale the Gaussian transfer function. The final color C and the opacity α are evaluated by combining multiple Gaussian transfer functions together:

$$C = \frac{\sum \alpha_i C_i}{\sum \alpha_i} \quad and \quad \alpha = \sum \alpha_i \quad (7)$$

where C_i and α_i are the color and opacity generated by the i-th Gaussian transfer function.

5 Range Likelihood Tree Guided Exploration Framework

Modeling probability distribution fields as range likelihood trees, we transform the problem of distribution-based uncertain data exploration to the problem of exploring the cumulative probabilities in a set of value ranges that are meaningful to users. We develop a RLT guided exploration framework with interactive visualization techniques for exploring probability distribution fields. In this section, we describe the design considerations and choices of our interface, and provide guidelines for visual exploration using our framework.

5.1 User Interface

The user interface consists of three components: the range likelihood tree (RLT) view, the multi-field transfer function widget, and the spatial view. We provide design details for each component as follows.

5.1.1 Range Likelihood Tree View

The RLT view is to facilitate users in understanding the underlying probability distribution field through visual exploration of ranges at different hierarchies. To achieve an efficient and tidy arrangement of RLT while illustrating the distributions regarding selected ranges, we employ a composite radial visualization (Figure 7), which consists of two-layered radial visualizations: a RLT visualization in the center, surrounded by a density visualization of distributions over the entire value domain.

RLT Layout. To provide a compact visualization of RLT, a radial tree layout is chosen in this work. Compared with a linear tree layout, the radial tree layout is more compact and with shorter distance between each node and its corresponding value range. The Reingold-Tilford drawing algorithm [9] is employed to create the radial tree layout. In our design, the root of a RLT is placed at the center of the visualization, and the distance of a node from the root encodes the level of the node (as shown in Figure 7).

Distribution Density Map. Since a probability distribution field often contains a large number of distributions, a simple plotting of all distribution as curves will likely lead to problems with overplotting and visual clutter. To provide an effective overview of distributions

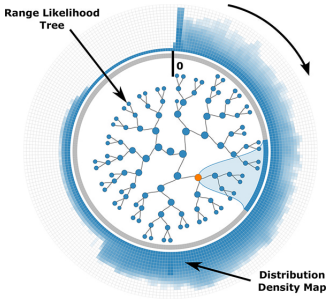


Figure 7: The composite radial visualization of a range likelihood tree with distribution density map.

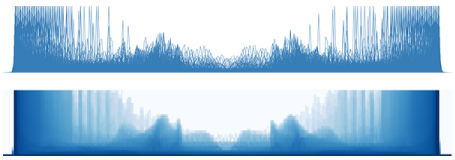


Figure 8: Visualizing distributions of the material ensemble dataset using superimposed curves (top) and curve density estimation (bottom).

in the underlying probability distribution field, we employ a curve density estimation technique [26] to create a density-based visualization of distributions, as shown in Figure 8. The density map in the RLF view is drawn using a ring metaphor that surrounds a RLT to form a composite radial visualization, where values start at the 12 o'clock position (marked with a line bar in Figure 7) and increase in a clockwise order.

Visual Linking. When the user selects a node of interest, it is desired that the underlying range is simultaneously highlighted for examining the distributions within the selected range. Explicit visual linking in which curves are rendered between related elements has been shown expressive and effective [43]. This also motivated our choice of a radial visualization, where links are typically drawn as arcs connecting related elements [29]. Our composite radial visualization supports such visual linking using an easy interaction — when the user hovers over a node of interest, visual links are immediately drawn as arcs to connect the node with the range in the distribution density map (as shown in Figure 7).

5.1.2 Multi-field Transfer Function Widget

Through visual exploration of RLFs in the RLT view and the RLF view, users can progressively understand the underlying probability distribution field when focusing on one RLF at a certain level in a RLT at a time. In addition, multiple RLFs can be selected to enable further analysis, as uncertain features may exist in multiple distinct ranges. As a result, multiple RLFs form a *multidimensional range likelihood field (MRLF)* where every distribution at a grid point in the underlying probability distribution field is transformed into a likelihood vector across the multi-field. To visualize these likelihood vectors, we employ the parallel coordinates plot (PCP) [23], a popular choice in multivariate transfer function design that enables users to manipulate transfer functions through traditional brushing such as axis brushing [7]. In our approach, a *coordinates axis* represents the scale of likelihoods for one selected RLF, and a polyline that connects multiple coordinates axes represents a multi-field likelihood vector.

Inspired by flexible linked axes [14], which extend the conventional parallel coordinates design (Figure 9 (top)) by flexibly positioning coordinates axes (Figure 9 (bottom)), we orient the coordinates axes so that they are aligned with the distribution density

map (in the RLT view) to form an integrated radial visualization (Figure 1(4)). In this way, a coordinates axis can be placed near the corresponding subrange in the RLT visualization, which preserves the user's mental map during exploration.

As coordinates axes are positioned in the polar coordinate system, we encode likelihood vectors as polycurves across the coordinates axes. To overcome the overplotting problem of drawing polycurves of all likelihood vectors, we employ adaptive sampling to draw a set of likelihood vectors that is statistically similar to the overall population. We start with a set of likelihood vectors which are sampled from the field randomly and keep adding more samples until their distribution is similar to the overall distribution. Overlaying the sampled polycurves, transfer functions are highlighted using ribbon metaphors across coordinates axes in an illustrative manner [34].

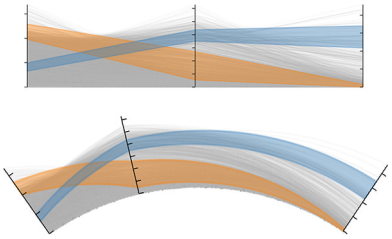


Figure 9: Multi-field transfer function widget using parallel coordinates axes (top) and flexible coordinates axes (bottom).

5.1.3 Spatial View

The spatial view has two sub-views: the range likelihood field (RLF) view and the multi-field classification view.

The Range Likelihood Field View is used to visualize a RLF of interest. After the user selects one node in the RLT, a RLF is constructed on the fly based on equation (4) and the RLF view is updated interactively with direct volume rendering. We adopt a perception-aware color map [41] to color the RLF in a blue-green-yellow manner with a linear opacity scale. This view is linked with the RLT view — when the user hovers over a node of interest, the RLF view is simultaneously updated to visualize the RLF within the selected subrange.

The Multi-Field Classification View mainly shows the probabilistic classification results of multiple RLFs. The user first brushes on the coordinates axes of PCP to select a set of multi-field likelihood vectors. Then Gaussian transfer functions are automatically generated based on the user selected likelihood vectors based on equation (6) and (7). The classification results of the MRLF is rendered in this view based on these Gaussian transfer functions.

5.2 Exploration Guidelines

After transforming a given probability distribution field into a RLT, the RLT serves as the exploration space where users can search for features of interest and perform further analysis such as probabilistic classification. We follow the visual information seeking mantra "Overview first, zoom and filter, then details-on-demand" by Ben Shneiderman [42] to guide the exploration process:

Overview First. The RLT view (Figure 1(1)) provides an overview of the multi-level structure of RLFs as well as the distributions in the underlying probability distribution field. By selecting a node in the RLT, its corresponding RLF is constructed (based on equation (4)) to obtain an overview of how likely the uncertain values at each grid point falls within this selected range.

Zoom and Filter. To understand how much the distributions fall within a RLF and where they are spatially located, users can hover a node in the RLT view (Figure 1(2)) and examine the likelihoods of associated grid points in the RLF view (Figure 1(3)). When no prior knowledge is available about the uncertain data set, it is natural to start from the root and follow a binary-split path. As the user zooms into a deeper level of the hierarchy, the RLF in focus becomes smaller and more distributions will be filtered out.

Details on Demand. When users select a set of representative RLFs of interest, a multi-field transfer function widget (Figure 1(4)) is created. users can manipulate the transfer function widget to explore and classify grid points based on their likelihoods in different subranges in the multi-field classification view (Figure 1 (5)).

6 Results

We demonstrate the effectiveness of our methods through experiments on three uncertain data sets in different applications: Massachusetts Bay Sea Trial (MBST-98), D-FTLE of temporal downsampled Hurricane Isabel, and High-Resolution Rapid Refresh (HRRR) ensemble dataset. The datasets are listed in Table 1. All the experiments were performed on a desktop computer with an Intel(R) Core(TM) i7-4790K CPU 4.0GHz processor, 16GB memory, and an NVIDIA GTX 970 GPU. In particular, we verified the visualization results (in Section 6.2 and Section 6.3) with two domain scientists, who are experts in environment science with over ten years of experience.

6.1 Massachusetts Bay Sea Trial Ensemble Dataset

The Massachusetts Bay Sea Trial (MBST-98) was an interdisciplinary forecast simulation [27, 40] based on the Littoral Ocean Observing and Predicting System (LOOPS). The MBST-98 took place in Massachusetts Bay from 17 August to 5 October 1998 with 600 ensemble runs. The scientific focus of this simulation was phytoplankton and zooplankton patchiness which were encoded in variables *chlorophyll-a concentration* and *zooplankton concentration*. In order to investigate the chlorophyll-a concentration, we selected the CHL variable at the time step representing September 2, 1998 and perform kernel density estimation based on all 600 ensembles at every grid point to model the uncertainty as probability density functions.

From the different components provided by our system, users can better explore the distribution field. Figure 1(1) presents an overview of the multi-level structure of the RLFs as well as the distributions in the underlying probability distribution field. Based on the distribution density map, we can see that the underlying probability distributions have high probabilities when chlorophyll-a concentration is low and have decreasing probability when chlorophyll-a concentration increases. By selecting ranges in the RLT view, their corresponding RLFs are shown in the RLF view, which help users comprehend the cumulative probabilities at every grid point in the

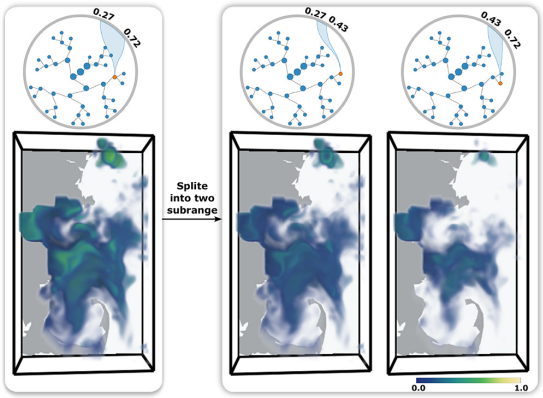


Figure 10: Exploring ranges at different levels in the RLT of the Massachusetts Bay Sea Trial Ensemble dataset.

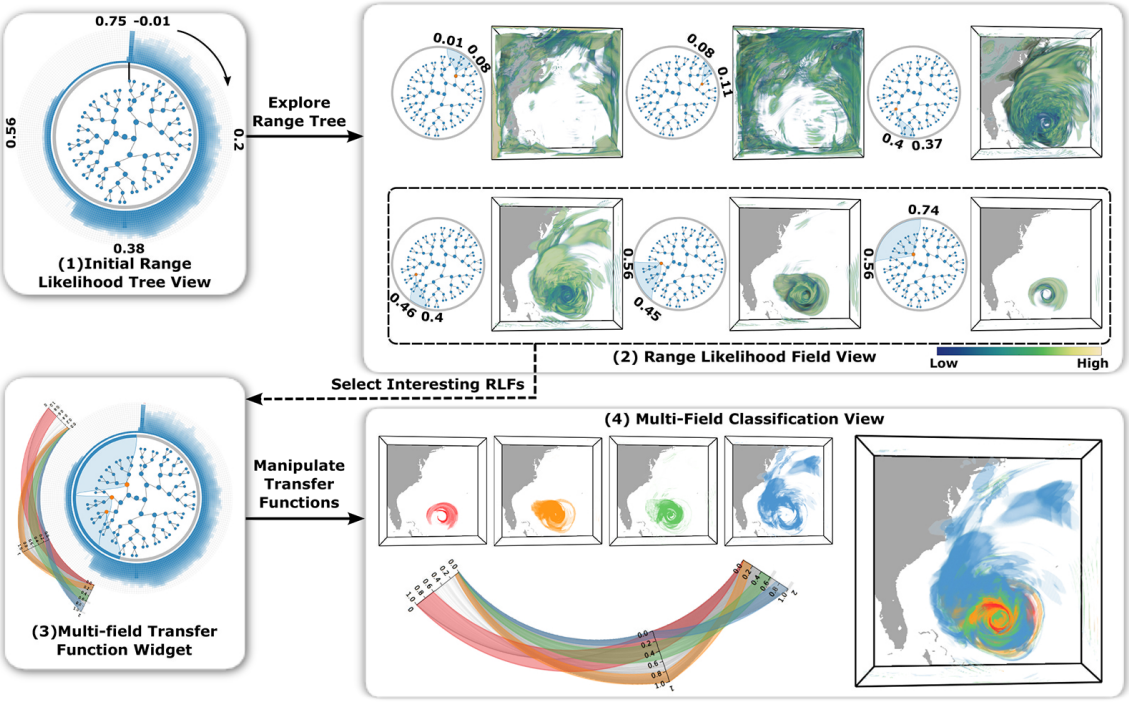


Figure 11: Experiments on the Temporally Down-Sampled Hurricane Isabel dataset. See Section 6.2 for details about the exploration process.

selected ranges. Figure 1(2) and 1(3) show five chlorophyll-a concentration value subranges as well as their corresponding RLFs, which highlight distinct features. We can see that the land regions are highlighted with range likelihoods aroud 1.0 with chlorophyll-a concentration near 0.0 mg/m^3 (milligram per cubic meter), and the ocean regions are highlighted with different range likelihoods when chlorophyll-a concentration is greater than $0.0mg/m^3$. For the subrange $0.27 \sim 0.72mg/m^3$, chlorophyll-a concentration has a higher likelihood at northeast of Cape Ann compared with other regions, and for the subrange $0.72 \sim 1.58mg/m^3$ chlorophyll-a concentration has a higher likelihood near Boston Harbor. Besides these highlighted regions, there are a large number of transparent regions in these two RLFs, which represents the likelihoods in those regions are zero. This is consistent with the fact that nutrients advected over Stellwagen Bank (due to tidal mixing) as well as along the coastline (due to wind driven upwelling and episodic wind mixing) [40]. We also observed that in some regions such as Stellwagen Bank, chlorophyll-a concentration has a non-zero likelihood in the selected value ranges, which indicates that chlorophyll-a concentration at those regions could fall within any of those ranges. With the RLT, the user can explore the probability distribution field with different levels of detail. For example, in Figure 10 the value range $0.27 \sim 0.72mg/m^3$ is further divided into two subranges. We observed that the feature at northeast of Cape Ann splits into two features.

By selecting multiple RLFs with respect to the different chlorophyll-a concentration value ranges, the user can view the correlation of the multi-field likelihoods in the transfer function widget (Figure 1(4)) and manipulate the transfer function for visual classification of the distributions in the multi-field classification view (Figure 1(5)). With this multi-field classification, insightful visualization results can be obtained. For instance, in Figure 1(4), four Gaussian transfer functions are constructed and in Figure 1(5) the classification results are shown. The blue part of the result is mainly at the region of the Boston Harbor, where the cumulative probability in range $0.72 \sim 1.58mg/m^3$ is around 0.27 and the cumulative probability in range $0.14 \sim 0.27mg/m^3$ is around 0.17; The red feature surrounding the blue feature at the Boston Harbor has a higher cumulative probability in range $0.14 \sim 0.27mg/m^3$ but lower cumulative probability in range $0.72 \sim 1.58mg/m^3$. Many other features can also be identified from the visualization results.

6.2 Temporally Down-Sampled Hurricane Isabel Dataset

Finite-time Lyapunov exponent (FTLE) is becoming a popular method in fluid dynamics for transport behavior analysis of unsteady flows. The higher FTLE value a region has, the more divergent flows are in the region. Guo et al. [20] extended the concept of FTLE to D-FTLE, which computes a probability density function of FTLE values for every grid point. In this experiment, we study unsteady flows using D-FTLE on the Hurricane Isabel dataset. The uncertainty is introduced through temporal down-sampling using the method proposed by Chen et al. [12]. The Hurricane Isabel dataset is an atmospheric simulation data created by the Weather Research and Forecast model (WRF) with 48 time steps (hourly average), which models a strong hurricane in the West Atlantic region in September 2003. We down-sampled the Hurricane Isabel dataset by aggregating every 12 time steps into one. Following the method in [12], for every 12 time steps, we fit a quadratic Bezier curve and store the quadratic Bezier curve parameters and the interpolation error distributions at every grid point. The variables U, V, and W which correspond to the wind vector directions are used in this experiment. A D-FTLE distribution field is then generated starting from time step 0 and advecting 6 hours.

An overview of the multi-level structure of the RLFs and a density map of the underlying distributions are shown in Figure 11(1). We can see that most distribution of the FTLE has low probability when the FTLE value is high, suggesting that high FTLE values only exist in a small part of the spatial domain. Selecting ranges of FTLE values from low to high, their corresponding RLFs are visualized in the RLF view (Figure 11(2)). We found that high FTLE values have higher likelihoods in the hurricane eyewall, elevated in the sur-

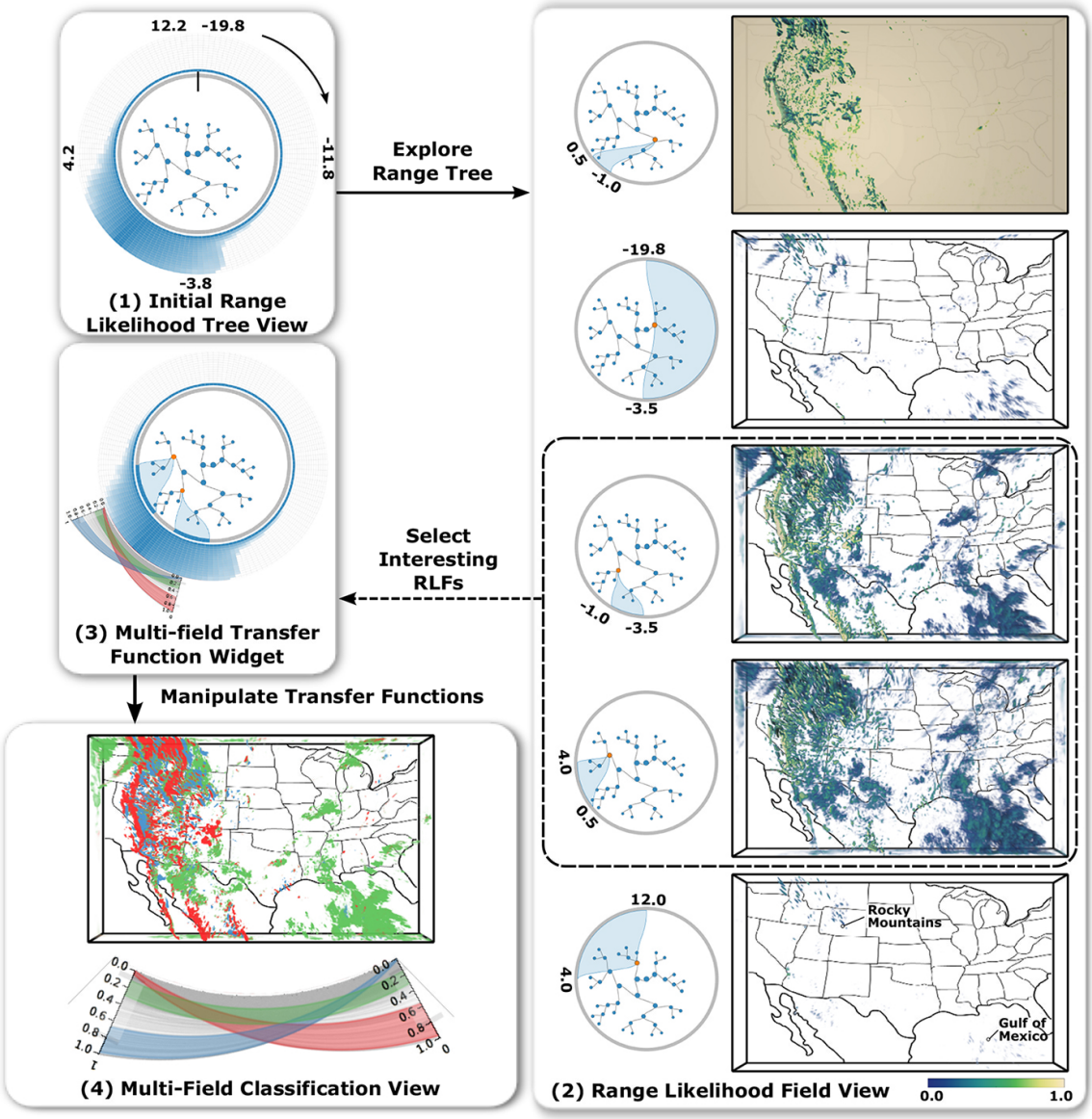


Figure 12: Experiments on the Ensemble HRRR Simulation dataset. See Section 6.3 for details about the exploration process.

rounding rainbands and can seperate the spiral bands, especially the extended northerly band. Meteorologists with whom we discussed the visualization results confirmed that the structure of the hurricane eyewall and rainbands can be highlighted by selecting RLFs corresponding to high FTLE values. In particular, as the FTLE value increases, the RLF represents the structure closer to the hurricane eye.

To further investigate the distributions around the hurricane eye, we select three RLFs with respect to three FTLE value ranges $0.56 \sim 0.74$, $0.46 \sim 0.56$, and $0.4 \sim 0.46$. The probabilistic classification results of the underlying probability distribution of FTLE values are shown in Figure 11(3) and Figure 11(4). The red part of the result is within the hurricane eye, with high cumulative probability in the highest FTLE value range and low cumulative probability in the lowest FTLE value range. The outside feature with blue color has low cumulative probability in the highest FTLE value range and high cumulative probability in the lowest FTLE value range. Feedback from the scientists confirms that these classifications well separate the outflow cloud shield (blue region) from the hurricane eye and eyewall.

6.3 Ensemble HRRR Simulation Dataset

The High-Resolution Rapid Refresh (HRRR) [1] is an National Oceanic and Atmospheric Administration (NOAA) real-time atmospheric model based on the Weather Research and Forecasting (WRF) model. The HRRR is hourly updated and available to the public through the Unidata's THREDDS Data Server2. We use an ensemble of the HRRR simulation output at time step representing 22:00:00 UTC, August 29, 2015, which consists of more than 20 variables and starts from 10 different hours. Based on the domain experts' focus of interest, we select the variable vertical velocity for analysis. We perform kernel density estimation at every grid point to model the uncertainty as probability density functions.

The RLT, RLFs, and the probabilistic classification results are shown in Figure 12. In Figure 12(2), we selected several representative value ranges and visualize their corresponding RLFs. It can be

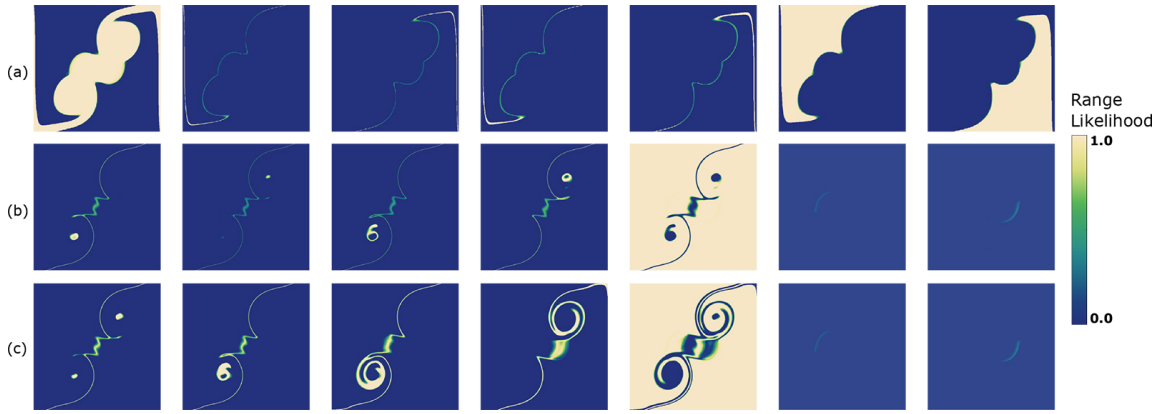


Figure 13: Comparison of different clustering approaches using the 2D material ensemble data set. Each image represents a RLF with respect to one cluster. (a): Hierarchical clustering based on the Euclidean distances between range likelihood fields. (b): Hierarchical clustering applied on likelihood distributions (scaled range likelihood fields) using the Euclidean distance. (c): K-means clustering using the Euclidean distances between likelihood distributions.

seen that vertical flows have high likelihood to occur near the Rocky Mountains region due to the topography and median likelihood to occur near the Gulf of Mexico compared with other regions. To further analyze the upward and downward motion of the wind, we select two RLFs with respect to the upward flow and downward flow for classification.

The domain scientists helped us explain the classification results (in Figure 12(4)): near the Rocky Mountains the wind directions are with either upward motion (color in red) or downward motion (color in blue) due to orographic lift which occurs when an air mass is moving over rising terrain and the foehn winds which occurs in the downwind side of a mountain. The red and blue regions are interweaving since winds are moving across mountains. Near the Gulf of Mexico the wind directions have around 0.2 likelihood in either direction.

7 Discussion and Future Work

The range likelihood tree representation brings a novel perspective to visual summarization and exploration of probability distribution fields. Unlike previous statistical summary and dissimilarity based methods, our method summarizes complex probability distributions with the range likelihood fields over a few representative subranges by considering the different roles that different subranges may play in understanding probability distributions. Our data model enables effective classification through user query and exploration.

In our approach, clustering of RLFs was done in the preprocessing stage prior to the exploration process. Table 1 reports the computational performance of the clustering algorithm. While we used JensenShannon divergence (JSD) and hierarchical clustering in our approach, we also experimented with alternative dissimilarity measures (e.g., Euclidean distance) and clustering methods (e.g., k-means clustering). Figure 13 shows the clustering results with alternative clustering approaches and dissimilarity measures using the 2D material ensemble dataset. Compared with these alternative methods, the clustering results of JSD and hierarchical clustering (Figure 5(b)) highlights more distinct regions with respect to different features. We also experimented with different numbers of initial subrange partitioning, and found the resulting leaf nodes in the pruned trees were roughly the same with minor variations in all four datasets. Therefore, we used $N = 256$ as the initial number of partitions in our experiments.

We showed our system to the domain scientists to assess the potential of our work applied to weather research. The domain experts were able to understand the concept of range likelihood fields, and found the range likelihood tree effective as a compact representation of a probability distribution field. The composite visualization of the range likelihood tree and the distribution density map are intuitive for them to understand, and the transfer function widget is easy to use. Overall, the positive feedback we received from the domain scientists encouraged us to further apply the exploration framework to solve more real-world scientific problems in our future work. In particular, we plan to combine our approach with statistical summary of distributions such as entropy to analyze overall uncertainty at spacial locations. We would also like to investigate multivariate distributions and high-dimensional value ranges in analyzing multivariate uncertain data. We would also like to extend our work to time-varying probability distribution field visualization and exploration.

Table 1: Average runtime (in seconds) for the clustering algorithm.

Data Set	Resolution	Runs	Partitions	Size	Time
MBST-98	$106 \times 180 \times 32$	600	256	0.58GB	151.71s
D-FTLE (Isabel)	$250 \times 250 \times 50$	200	256	2.98GB	852.28s
HRRR	$449 \times 264 \times 40$	10	128	2.26GB	346.56s

8 Conclusion

In this work, we present a compact and effective representation, called range likelihood tree (RLT), to summarize and explore probability distribution fields. The RLT representation decomposes and summarizes complex probability distributions with the range likelihood fields over a few representative subranges, which allows effective classification and identification of features through user query and exploration. We present an exploration framework with multiple interactive views to explore probability distribution fields through range likelihood fields, and provide guidelines for visual exploration using our framework. We demonstrated the effectiveness and usefulness of our approach in exploratory analysis using several representative uncertain data sets, and verified the visualization results with domain scientists in environment science.

Acknowledgments

This work was supported in part by NSF grants IIS-1250752, IIS-1065025, and US Department of Energy grants DE-SC0007444, DE-DC0012495, program manager Lucy Nowell.

REFERENCES

[1] C. Alexander, S. S. Weygandt, D. C. D. S. Benjamin, T. G. Smirnova, E. P. James, M. H. P. Hofmann, J. Olson, and J. M. Brown. The highresolution rapid refresh: Recent model and data assimilation development towards an operational implementation in 2014. *In Proceedings 26th Conference on Weather Analysis and Forecasting*, 2014.

[2] J. C. Anderson, L. Gosink, M. A. Duchaineau, and K. Joy. Feature identification and extraction in function fields. In *EuroVis 2007*, 2007.

[3] J. C. Anderson, L. Gosink, M. A. Duchaineau, and K. Joy. Interactive visualization of function fields by range-space segmentation. *Computer Graphics Forum*, 28(3):727–734, June 2009.

[4] T. Athawale and A. Entezari. Uncertainty quantification in linear interpolation for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2723–2732, 2013.

[5] T. Athawale, E. Sakhaee, and A. Entezari. Isosurface visualization of data with nonparametric models for uncertainty. *IEEE Trans. Vis. Comput. Graph.*, 22(1):777–786, 2016.

[6] K. Bensema, L. Gosink, H. Obermaier, and K. Joy. Modality-driven classification and visualization of ensemble variance. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2015.

[7] J. Blaas, C. P. Botha, and F. H. Post. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 2008.

[8] K. Brodlie, R. A. Osorio, and A. Lopes. A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. 2012.

[9] C. Buchheim, M. Jünger, and S. Leipert. Improving walkers algorithm to run in linear time. In *Graph Drawing*. Springer, 2002.

[10] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 2007.

[11] J.-H. Chang, K.-C. Fan, and Y.-L. Chang. Multi-modal gray-level histogram modeling and decomposition, 2002.

[12] C. Chen, A. Biswas, and H. Shen. Uncertainty modeling and error reduction for pathline computation in time-varying flow fields. In *2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, China, April 14-17, 2015*, pages 215–222, 2015.

[13] H. Chen, S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, and H. Qu. Uncertainty-aware multidimensional ensemble data visualization and exploration. *Visualization and Computer Graphics, IEEE Transactions on*, 21(9):1072–1086, 2015.

[14] J. H. Claessen and J. J. Van Wijk. Flexible linked axes for multivariate data visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2310–2316, 2011.

[15] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3d ensemble visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2694–2703, 2014.

[16] V. Dinesha, N. Adabala, and V. Natarajan. Uncertainty visualization using hdr volume rendering. *The Visual Computer*, 28, 2012.

[17] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang. Visualizing scalar volumetric data with uncertainty. *Computers and Graphics*, 2002.

[18] N. Fout and K.-L. Ma. Fuzzy volume rendering. *IEEE Transactions onVisualization and Computer Graphics*, 18(12):2335–2344, 2012.

[19] G. Grigoryan and P. Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 10(5):564–573, Sept. 2004.

[20] H. Guo, W. He, T. Peterka, H.-W. Shen, S. M. Collis, and J. J. Helmus. Finite-time lyapunov exponents and lagrangian coherent structures in uncertain unsteady flows. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[21] H. Guo, H. Xiao, and X. Yuan. Multi-dimensional transfer function design based on flexible dimension projection embedded in parallel coordinates. In *PacificVis*. IEEE Computer Society, 2011.

[22] 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*, 2011.

[23] A. Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.

[24] J.-M. Jolion, P. Meer, and S. Bataouche. Robust clustering with applications in computer vision. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):791–802, 1991.

[25] J. Kniss, S. Premoze, M. Ikits, A. Lefohn, C. Hansen, and E. Praun. Gaussian transfer functions for multi-field volume visualization. In *Proceedings of IEEE Visualization 2003*, pages 497–504, 2003.

[26] O. D. Lampe and H. Hauser. Curve density estimates. In *Computer Graphics Forum*, volume 30, pages 633–642, 2011.

[27] P. F. J. Lermusiaux. Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *J. Comput. Phys.*, 217(1):176–199, Sept. 2006.

[28] S. Liu, J. A. Levine, P. Bremer, and V. Pascucci. Gaussian mixture model based volume visualization. In *Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on*, pages 73–77. IEEE, 2012.

[29] X. Liu and H.-W. Shen. Association analysis for visual exploration of multivariate scientific data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):955–964, 2016.

[30] C. Lundström, P. Ljung, and A. Ynnerman. Extending and Simplifying Transfer Function Design in Medical Volume Rendering Using Local Histograms. *IEEE VGTC Symposium on Visualization*, 2005.

[31] C. Lundström, P. Ljung, and A. Ynnerman. Local histograms for design of transfer functions in direct volume rendering. *IEEE TVCG*, 12(6):1570–1579, 2006.

[32] C. Lundstrm, P. Ljung, A. Persson, and A. Ynnerman. Uncertainty visualization in medical volume rendering using probabilistic animation. *IEEE Trans. Vis. Comput. Graph.*, 13(6), 2007.

[33] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *VisSym*, 2003.

[34] K. T. McDonnell and K. Mueller. Illustrative parallel coordinates. In *Computer Graphics Forum*, volume 27, pages 1031–1038, 2008.

[35] N. Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[36] T. Pfaffelmoser, M. Mihai, and R. Westermann. Visualizing the variability of gradients in uncertain 2d scalar fields. *IEEE transactions on visualization and computer graphics*, 2013.

[37] K. Pöthkow and H.-C. Hege. Nonparametric models for uncertainty visualization. In *Computer Graphics Forum*, volume 32, pages 131–140. Wiley Online Library, 2013.

[38] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic marching cubes. In *Computer Graphics Forum*.

[39] P. J. Rhodes, R. S. Laramee, R. D. Bergeron, and T. M. Sparr. Uncertainty Visualization Methods in Isosurface Rendering. In *Eurographics 2003 - Short Presentations*. Eurographics Association, 2003.

[40] A. R. Robinson. Real-time forecasting of the multidisciplinary coastal ocean with the littoral ocean observing and predicting system (loops). *The Third Conference on Coastal Atmospheric and Oceanic Prediction Processes*, 1999.

[41] F. Samsel, M. Petersen, T. Geld, G. Abram, J. Wendelberger, and J. Ahrens. Colormaps that improve perception of high-resolution ocean data. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015.

[42] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[43] M. Steinberger, M. Waldner, M. Streit, A. Lex, and D. Schmalstieg. Context-preserving visual links. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2249–2258, 2011.

[44] D. Thompson, J. A. Levine, J. C. Bennett, P.-T. Bremer, A. Gyulassy, V. Pascucci, and P. P. Pébay. Analysis of large-scale scalar data using hixels. In *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on*, pages 23–30. IEEE, 2011.

[45] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi. Efficient volume exploration using the gaussian mixture model. *IEEE Trans. Vis. Comput. Graph.*, 17(11):1560–1573, 2011.

[46] C. Wittenbrink, E. Saxon, J. J. Furman, A. Pang, and S. Lodha. Glyphs for Visualizing Uncertainty in Environmental Vector Fields. In *IEEE Transactions on Visualization and Computer Graphics*, volume 2410, pages 266–279, 1995.