An Interactive Framework for Visualization of Weather Forecast Ensembles

Bo Ma and Alireza Entezari, *Senior Member, IEEE*

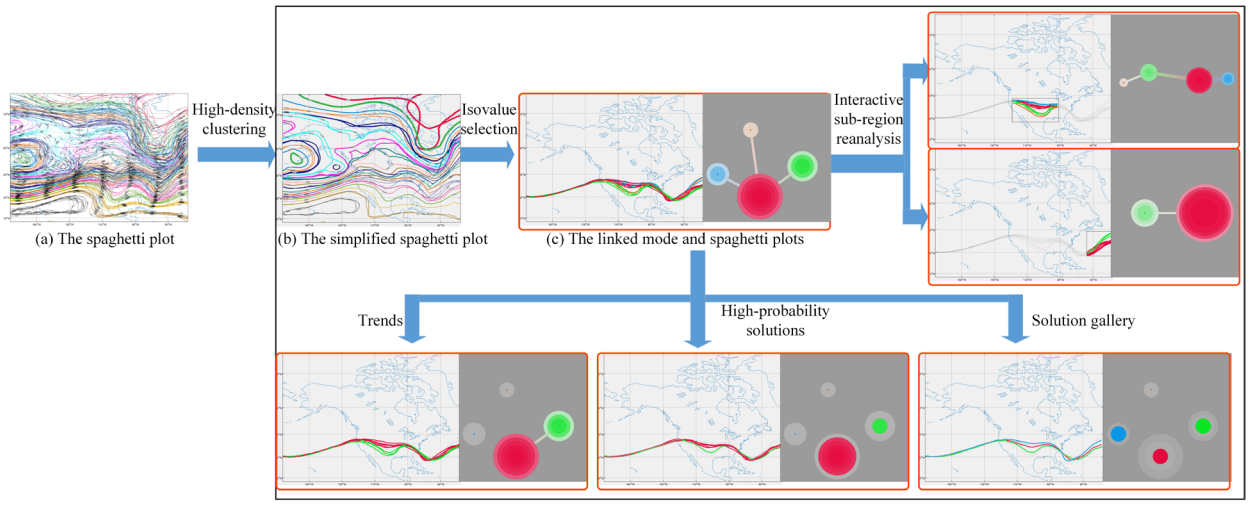


Fig. 1. Interactive exploration of isocontours across multiple isovalues in a weather forecast ensemble. (a) The conventional spaghetti plot suffers from significant visual clutter. Based on our high-density clustering for ensemble isocontours, we present several summary plots and interactions (right box). Users can select isovalues of interest from (b) the simplified spaghetti plot, and the selected isocontours are visualized in (c) the linked mode and spaghetti plots. Users can use the mode plot to select interesting subsets of isocontours (bottom). Users can also interactively select sub-regions for re-analysis and further exploration.

Abstract— Numerical Weather Prediction (NWP) ensembles are commonly used to assess the uncertainty and confidence in weather forecasts. Spaghetti plots are conventional tools for meteorologists to directly examine the uncertainty exhibited by ensembles, where they simultaneously visualize isocontours of all ensemble members. To avoid visual clutter in practical usages, one needs to select a small number of informative isovalues for visual analysis. Moreover, due to the complex topology and variation of ensemble isocontours, it is often a challenging task to interpret the spaghetti plot for even a single isovalue in large ensembles. In this paper, we propose an interactive framework for uncertainty visualization of weather forecast ensembles that significantly improves and expands the utility of spaghetti plots in ensemble analysis. Complementary to state-of-the-art methods, our approach provides a complete framework for visual exploration of ensemble isocontours, including isovalue selection, interactive isocontour variability exploration, and interactive sub-region selection and re-analysis. Our framework is built upon the high-density clustering paradigm, where the mode structure of the density function is represented as a hierarchy of nested subsets of the data. We generalize the high-density clustering for isocontours and propose a bandwidth selection method for estimating the density function of ensemble isocontours. We present novel visualizations based on high-density clustering results, called *the mode plot* and the simplified spaghetti plot. The proposed mode plot visually encodes the structure provided by the high-density clustering result and summarizes the distribution of ensemble isocontours. It also enables the selection of subsets of interesting isocontours, which are interactively highlighted in a linked spaghetti plot for providing spatial context. To provide an interpretable overview of the positional variability of isocontours, our system allows for selection of *informative isovalues* from the simplified spaghetti plot. Due to the spatial variability of ensemble isocontours, the system allows for interactive selection and focus on sub-regions for *local* uncertainty and clustering re-analysis. We examine a number of ensemble datasets to establish the utility of our approach and discuss its advantages over state-of-the-art visual analysis tools for ensemble data.

Index Terms—Spaghetti plots, ensemble visualization, uncertainty visualization, high-density clustering, ensemble forecasting

1 Introduction

Numerical Weather Prediction (NWP) uses computational models to predict future weather states based on current atmospheric measurements. However, the dynamical systems used to model the atmosphere behave chaotically, in that small changes in initial conditions can result in different predictions. Furthermore, the effects of dynamic physical processes of the atmosphere that NWP simulates is model dependent, making weather forecasts vary from model to model. Since the 1990s, ensemble forecasting as a tool has been routinely used to analyze the uncertainty and gauge the confidence in weather forecasts. Each ensemble member is a possible forecast conducted with perturbed initial conditions, different NWP models, different spatial resolutions, or different vertical coordinate systems. Due to the large size and high complexity of ensemble data, visualization plays a central role in assessing the inherent variability and uncertainty in modern weather forecast systems.

• *B. Ma and A. Entezari are with the University of Florida. E-mail: bbo@cise.ufl.edu, entezari@ufl.edu*

Manuscript received 31 Mar. 2018; accepted 1 Aug. 2018.
Date of publication 16 Aug. 2018; date of current version 21 Oct. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2864815

A popular technique for uncertainty visualization of ensemble data is the so-called spaghetti plot, a contour plot of a particular data field, where isocontours of all ensemble members are rendered simultaneously. It is the only conventional visualization technique to directly examine the distribution behavior of all ensemble members in meteorology [41]. Obviously, plotting isocontours of a dense set of isovalues for all ensemble members causes significant overdrawing (visual clutter), making the spaghetti plot challenging to parse. In practice, only isocontours for one or a few selected isovalues are plotted to reduce this visual clutter. By examining the positional variations of the isocontours, forecasters analyze the distribution of ensemble members in regions near the isocontours for the presence of clusters, main trends, and outliers [52, 53]. However, due to the complicated shapes of isocontours and increasing size of ensemble members, spaghetti plots for even a single isovalue can be difficult to interpret.

Recently, there has been a growing body of research in the visualization community on visual analysis and uncertainty quantification in ensemble data [4, 19, 23, 38, 41, 45, 46, 56]. In particular, improving the state-of-the-art of the spaghetti plot has attracted considerable attention. Whitaker *et al.* [56] generalized the concept of statistical data depth to ensemble isocontours to produce a global center-outward ordering. The contour box plot is proposed to visualize statistical quantities of the ordered isocontours, where the mean isocontour, the median isocontour, the 50% confidence band, the 100% confidence band, and the outliers are displayed. Ferstl *et al.* [19, 56] performed clustering analysis on ensemble isocontours and proposed a contour variability plot which aggregates a large number of isocontours into cluster bands. These methods reduce the visual clutter of spaghetti plots by abstracting isocontours into confidence bands which present an overview of the distribution of the isocontours. However, the direct examination of isocontours is no longer possible, and the effectiveness of the produced visual analysis tools is limited to the generation and visualization of static summary plots. For more details, please refer to our supplemental document. Alternatively, Quinan and Meyer [41] enabled interactive highlighting in spaghetti plots which allows the comparison of isocontours across multiple isovalues. However, the per-isovalue interactive highlighting does not allow discerning isocontours with the same isovalue. Sanyal *et al.* [46] integrated uncertainty glyphs in spaghetti plots for exploration of weather forecasting ensembles.

In this paper, we present an interactive framework for uncertainty visualization of ensemble isocontours. Our framework enables a complete exploration of ensemble isocontours, including isovalue selection, interactive isocontour variability exploration, and interactive sub-region selection and re-analysis. We propose a novel visual analysis tool, *the mode plot*, that visually encodes the structure obtained by the high-density clustering paradigm [22] and the distribution behavior of the given ensemble isocontours. Unlike previous approaches that aggregate isocontours into static summaries [19] or directly add additional visual encodings in spaghetti plots [46], we link the mode plot to the spaghetti plot and use interactive highlighting to select subsets of isocontours. The mode plot serves as a scaffold for interactive exploration of spatially coherent subsets of isocontours, while the chosen isocontours are directly highlighted and examined in the spaghetti plot. The linked mode and spaghetti plots provide a focus+context visualization for analyzing the variabilities of ensemble isocontours for a given isovalue. To select informative isovalues from a set of all possible isovalues, we present a simplified spaghetti plot which provides an overview of the variability and distribution of ensemble isocontours across multiple isovalues. Since the ensemble isocontours may exhibit spatial variability in different geographical regions, a global clustering of ensemble isocontours may not be adequate for regional analysis. Our framework allows an interactive selection of interesting sub-regions for clustering re-analysis and uncertainty re-evaluation.

Our framework is based on the high-density clustering paradigm from a new and expanding branch of statistics – topological data analysis (TDA) [55]. The high-density clustering also falls broadly in the class of hierarchical clustering. However, unlike most cases whose cluster hierarchies are generated based on various heuristics or computational convenience, the high-density clustering studies the mode hierarchy of the density estimate for a given dataset. It directly links the clustering task to the fundamental problem of non-parametric density estimation. We propose a bandwidth selection method which maximizes the number of significant modes of the density estimate. As a result, main trends and outliers of ensemble isocontours can efficiently be identified and separated by the high-density clustering. The proposed bandwidth selection method provides two intuitive tuning parameters, which also allow users to customize the clustering result. Our proposed mode plot visually encodes the structure of nested high-density clusters and provides a great deal of information that goes beyond the clustering task. In particular, it frees the user from having to assert prior knowledge on the right number of clusters, provides a useful summary of the entire ensembles, and allows the focus and selection of interesting subsets of isocontours at various granularities. Our implementation enables efficient clusterings and interactive explorations of ensemble isocontours.

The contributions of this paper include:

- High-density clustering for ensemble isocontours: We introduce and generalize the high-density clustering algorithm for studying the uncertainty and distribution behavior of ensemble isocontours. We propose a bandwidth selection method for estimating the density function of ensemble isocontours.
- Interactive exploration of ensemble isocontours: We present the mode plot to visually encode high-density clustering results and provide an effective summary of the distribution of ensemble isocontours. The mode plot serves as a scaffold for selection of subsets of isocontours, which are directly visualized in the linked spaghetti plot via interactive highlighting.
- Isovalue selection and sub-region re-analysis: We propose a simplified spaghetti plot that presents a highly interpretable overview of isocontour variabilities across multiple isovalues, from which one can select isovalues of interest. Our framework enables an interactive selection of sub-regions for clustering re-analysis which allows for assessing the *local* variability of ensemble isocontours.

2 Related Work

Uncertainty visualization has attracted considerable attention in the visualization community in the past two decades. For reviews and taxonomies of uncertainty visualization, we refer to some excellent surveys [6, 7, 39]. Ensemble visualization is a special category of uncertainty visualization, where uncertainty information is given by a set of possible outcomes of the data. Visualization of NWP ensembles is a workhorse application for ensemble visualization, and a recent survey on visualization in meteorology is given by Rautenhaus *et al.* [42]. Obermaier *et al.* [33] classifies the ensemble visualization as local-based visualization or feature-based visualization. Location-based methods examine the statistical properties at fixed locations of the ensemble, while feature-based methods extract meaningful features (e.g., isocontours for scalar fields or pathlines for flow fields) from individual ensemble members and compare them across all ensembles. Our approach belongs to feature-based methods, and we review literature that is most relevant to our work.

Many approaches have been proposed to incorporate statistical models into the process of isocontour (or isosurface) detection and extraction. Pöthkow *et al.* [36] modeled the uncertain scalar data as random fields and introduced the level crossing probability to measure the spatial distribution of uncertain isocontours. The authors further extended this idea to consider the correlation between random variables [38] and modeling of random fields with non-parametric models [37]. Pfaffelmoser *et al.* [35] proposed an algorithm to efficiently compute the isosurface crossing probabilities through a correlated random variable field. Athawale and Entezari [3] presented a closed-form solution for the level crossing probabilities for data uncertainty quantified by the uniform distribution. They further proposed an isosurface extraction algorithm [4] for uncertain scalar fields.

Feature-based methods infer the uncertainty of the data field based on the variability and spatial distribution of extracted isocontours.

Whitaker *et al.* [56] introduced the concept of statistical data depth and its generalization for measuring the location centeredness of ensemble isocontours. This enables a global center-outward ordering from which a number of isocontour-relevant statistical quantities are derived. These statistical quantities are further visually encoded in a new visualization called the contour boxplot which abstracts and summarizes the conventional spaghetti plot. Ferstl *et al.* [19] employed principal component analysis (PCA) to transform ensemble isocontours to a Euclidean space based on a popular representation of the isocontour: the signed distance transform. Clusters in the transformed space are detected by fitting a multivariate normal distribution (or ellipse) to each cluster. A so-called contour variability plot is proposed which shows the median isocontour and the confidence band for each cluster. These methods aggregate and summarize isocontours in static confidence envelopes which obscure geometric information which is readily seen in spaghetti plots, e.g., the distances between ensemble isocontours within the confidence envelope. Recently, there has been significant interest in uncertainty visualization of isosurfaces of 3D ensemble data [14, 15, 23, 43]. In our work, we summarize the distribution of ensemble isocontours in the mode plot and interactively select subsets of interesting isocontours highlighted in the spaghetti plot.

Our work is also related to the interactive exploration of the uncertainty of ensemble data. Recently, a number of visualization systems have been proposed for interactive visual analysis of ensemble simulations, including Ensemble-Vis [40], Noodles [46], Met.3D [43], WeaVER [41], and Ovis [24]. Demir *et al.* [13] proposed multi-charts to encode statistical information of 3D ensemble data, which are linked to the spatial view via volume rendering. Kehrer *et al.* [25] employed brushing and linking techniques for exploration of ensemble data based on statistical moments. Shu *et al.* [49] proposed an interactive visual analysis tool for studying the spatiotemporal similarities in time-varying ensembles. The brushing and linking technique connects multiple displays of the data and plays an essential role in the interactive visual analysis. In our framework, we link the mode plot to the spaghetti plot, where the selection of data in the mode plot is reflected in the spatial context via interactive highlighting.

Clustering, a fundamental approach to discover grouping patterns, plays an essential role in reducing the complexity of ensemble data. Thomas *et al.* [51] used the idea of clustering contours to detected symmetric regions in scalar fields. Ma *et al.* [29, 30] clustered volumetric features for transfer function design and volume compression quality assessment. Oeltze *et al.* [34] compared different techniques for streamline clustering. Obermaier *et al.* [32] performed clustering analysis for interactive visualizations of trends in time-varying ensembles. Hao *et al.* [21] used cluster trees to explore the relationships between ensemble members. Biswas *et al.* [5] compared the clustering structures of ensemble data across multiple resolutions to investigate the model's sensitivity to variations in input parameters. Want *et al.* [54] performed agglomerative clustering to ensemble members and encoded the result in dendrograms to aid the exploration of parameter correlations. Ferstl *et al.* [18] compared a number of clustering techniques in characterizing the uncertainty and variability of streamline ensembles. A hierarchical clustering of ensemble isocontours across multiple forecasting times was proposed to analyze the temporal growth of uncertainty [20]. In particular, Kumpf *et al.* [27] recognized that clustering results of meteorological ensembles can be sensitive to small changes in selected geographic regions. We use the proposed mode plot to efficiently encode the hierarchical high-density clustering result which provides a useful 2D summary of the distribution of ensemble isocontours. We approach the sensitivity of clustering to the selection of geographic regions via interactive sub-region selection and clustering re-analysis.

3 High-density clustering for ensemble isocontours

In this section, we describe the high-density clustering algorithm and its generalization for ensemble isocontours of a particular isovalue. We use a popular and powerful representation for isocontours (or isosurfaces) called the signed distance transform, which has been successfully used for solving various visualization problems [8, 19]. Given an ensemble

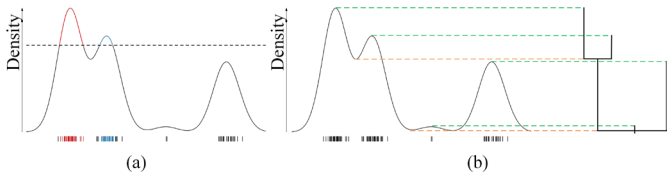


Fig. 2. (a) An illustration of high-density clusters. There are two high-density clusters (highlighted as red and blue) in the upper level set. (b) An illustration of the dendrogram for encoding high-density clusters. The dendrogram records the creations and merges of modes.

of n scalar fields $\mathbf{s}_1, ..., \mathbf{s}_n \in \mathbb{R}^m$ (defined on the same grid structure, e.g., longitude-latitude grid, with m grid points) and an isovalue of interest, we compute the signed distance transforms $\mathbf{d}_1, ..., \mathbf{d}_n \in \mathbb{R}^m$ for the corresponding isocontours. Each grid point of $\mathbf{d}_i$ records its (signed) closest distance to the isocontour of $\mathbf{s}_i$, where the sign indicates the side of the isocontour on which the grid point is seated. Each $\mathbf{d}_i$ (representing the isocontour of member i) can be treated as a point in an m-dimensional vector space [19], which directly enables the applications of various algorithms, including high-density clustering. There is a number of ways to compute the signed distance transform, and we discuss our implementation in Section 6. Note that the span of $\mathbf{d}_1, ..., \mathbf{d}_n$ is a vector space whose dimension is bounded by the minimum of m and n ($n << m$). Without loss of any information, we use principal component analysis (PCA) to reduce the dimensionality of each $\mathbf{d}_i$ to $n-1$, i.e., $\mathbf{d}_1, ..., \mathbf{d}_n \in \mathbb{R}^{n-1}$.

In the following, we describe the high-density clustering algorithm for ensemble isocontours represented as their Euclidean embeddings $\mathbf{d}_1, ..., \mathbf{d}_n \in \mathbb{R}^{n-1}$. The high-density clustering algorithm essentially studies the mode structure of the density estimate for a given (multidimensional) data. The only parameter involved in the clustering process is the smooth bandwidth for the kernel density estimator. We propose a bandwidth selection method that maximizes the number of significant modes of the density estimate, thus allowing the identification of different main trends and outliers from the clustering result.

3.1 High-density Clustering

Given a set of isocontours of an isovalue $\mathbf{d}_1, \ldots, \mathbf{d}_n \in \mathbb{R}^{n-1}$, the kernel density estimator (KDE) at an arbitrary point $\mathbf{d} \in \mathbb{R}^{n-1}$ is

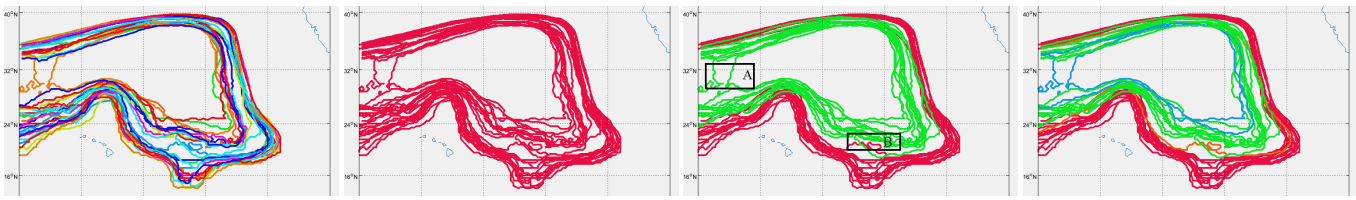
$$\hat{f}_h(\mathbf{d}) = \frac{1}{nh^{n-1}} \sum_{i=1}^{n} K\left(\frac{\mathbf{d}-\mathbf{d}_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{d}-\mathbf{d}_i) \qquad (1)$$

where K is a smooth function called the kernel function, a non-negative function that is symmetric around 0 and integrates to 1, e.g., Gaussian kernel $K(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|\mathbf{x}\|_2^2}{2}}$. $K_h(\mathbf{x}) = \frac{1}{h^p} K(\frac{\mathbf{x}}{h})$ is a scaled version of K which is also a kernel function, and p is the dimension of $\mathbf{x}$. h is called the bandwidth of the density estimator which governs the level of smoothness. $\frac{1}{n}$ is the normalization factor to ensure the integration of $\hat{f}_h$ equals 1. Note that $\hat{f}_h$ is a non-parametric function because it depends on the data. For example, Fig. 2 (a) plots the density function estimated from a synthetic 1D point cloud. The point cloud, plotted in the bottom, is composed of 120 samples drawn from a Gaussian mixture of three normal distributions and 2 outliers. The KDE results in a multi-modal distribution which expresses its capability for estimating arbitrary shape distributions.

For any threshold value $\varepsilon \geq 0$, the *upper level set* of $\hat{f}$ is

$$U_\varepsilon = \{\mathbf{d} | \hat{f}(\mathbf{d}) \geq \varepsilon\} \qquad (2)$$

The *high-density clusters* at density level ε are the connected components of the upper level set U_ε. Fig. 2 (a) demonstrates the high-density clusters for a particular ε (indicated as a horizontal dashed line), where the upper level set consists of two connected components (highlighted as red and blue). The data points that belong to the high-density clusters are also highlighted accordingly at the bottom. Note that these high-density clusters generally do not partition the entire dataset, instead only data points belonging to the underlying U_ε are considered in the



(a) (b) (c) (d)

Fig. 3. An illustration of bandwidth selection methods using mode clustering, where the bandwidth h is computed from (a) the rule of thumb [50] (h = 33.74), our method with $[\sigma_{sig}, \sigma_{outlier}]$= (b) $[15, 0]$ (h = 658), (c) $[10, 0]$ (h = 333), and (d) $[10, 2]$ (h = 324).

clustering. As ε decreases, the region covered by U_ε expands and the density requirement for being a "high-density cluster" is relaxed. As such, new connected components may be born, and existing clusters can merge into a single connected component. Considering the high-density clusters for all ε, a complete clustering structure is captured by a hierarchy of nested subsets of the data.

The evolution of high-density clusters and their order of inclusions is typically denoted as a dendrogram (or level set tree), as shown on the right of Fig. 2 (b). Decreasing ε from $\hat{f}_{max}$ to 0, whenever we reach a mode (green dash lines), there is birth of a new connected component in U_ε as well as a leaf branch in the dendrogram. On the other hand, the connections of modes (orange dashed lines) result in the merges of the connected components in U_ε as well as the branches in the dendrogram. Each branch represents the high-density clusters with the same topology, i.e., they contain the same set of bumps (or modes). Each leaf branch corresponds to a single mode where the high-density clusters have the simplest topology – a bump. The high-density clustering characterizes the mode hierarchy and the shape of the density estimate $\hat{f}$.

High-density clustering has a direct connection to persistent homology in TDA [55]. Persistent homology is a new branch of statistics devoted to discovering the shape of the data based on its connectivity structures: connected components (0th dimensional topological features), cycles (1st dimensional topological features), or high dimensional voids. High-density clustering is a special case of persistent homology where the evolution of the connected components of the density estimate at multiple density levels are studied.

3.2 Bandwidth Selection

The high-density clustering paradigm directly links the clustering task to the fundamental problem of non-parametric kernel density estimation. The sole parameter involved in the entire clustering process is the smoothing bandwidth h, which exhibits a strong influence on the result. From an exploratory point of view, all choices of bandwidths lead to useful density estimators [47], thus useful clusterings. While large bandwidths provide a picture of the global structure of the data, small bandwidths reveal local cluster structures. However, in our case, interactive bandwidth exploration brings in additional interactions, thus impeding the uncertainty exploration of ensemble isocontours. Therefore, we need to select a bandwidth that allows the identification and separation of main trends, as well as outliers, from the clustering result.

Commonly-used methods for bandwidth selection include the rule of thumb [50], least squares cross-validation [44], or plug-in method [48]. Roughly speaking, all these methods are aimed at obtaining a density estimator that minimizes the L2 loss function (or the mean integrated squared error). These methods are not recommended in TDA [9, 55], because the accuracy of the density estimator (with respect to L2 risk) is not tightly coupled with the topology (or shape) of the data. One can estimate the density function poorly but can still obtain the correct evolutions of topological features (or mode structure of the distribution) of the data. Chazal *et al.* [9] suggests a bandwidth selection method that maximizes the number of statistically significant topological features, including connected components, circles, and high-dimensional features. In the case of high-density clustering, each mode defines a (0th dimensional) topological feature, i.e., a connected component. The "significance" of a local mode is determined by its persistence (i.e., merge time minus birth time) in high-density filtrations of U_ε, as shown in Fig. 2 (b). Persistence confidence bands are constructed using bootstrapping, and a mode is treated as significant if it falls within the confidence bands. A subset of bandwidths is chosen from all possible bandwidths that maximize the number of significant features (and total significant persistence). However, this approach is also computationally prohibitive for high-dimensional data, and the computational overhead comes both from the evaluation of high-dimensional density estimator and the bootstrapping.

Our idea for bandwidth selection also involves maximizing the number of significant modes. Instead of the persistence, we propose to define the significance of a mode based on the size of its mode cluster. A mode cluster is defined as the basins of attraction of the mode, i.e., the sets of data points whose gradient ascents stop at the same mode. The size of a mode cluster is the number of ensemble members in the underlying cluster. Let $c_1(h), c_2(h), \ldots, c_k(h)$ denote the sizes of the mode clusters at scale h (one can compute these quantities efficiently using the mean-shift algorithm [11]). Formally, we define the number of significant modes $N(h)$ as

$$N(h) = \#\left\{i | c_i(h) \geq \sigma_{sig}\right\} \tag{3}$$

where σ_{sig} is a threshold parameter that defines the smallest size of a mode cluster to be qualified as significant. $N(h)$ is small when h is small since the data is under-smoothed, resulting in most $c_i(h)$'s being smaller than σ_{sig}. On the other hand, for a large h, $N(h)$ is also small because the data is over-smoothed and produces very few modes. The optimization of $N(h)$ balances the tradeoff between the number of modes and the size of the mode clusters. Since a slight increase/decrease of h may only slightly increase/decrease the sizes of the mode clusters but not change $N(h)$, the optimal solution forms a set of similar bandwidths. To select a single bandwidth, we introduce another parameter, $\sigma_{outlier}$, which specifies the upper bound of the number of outlier modes: $O(h) = \#\left\{i | c_i(h) < \sigma_{sig}\right\}$. From the optimal solution set, we choose an h that maximizes $O(h)$ subject to $O(h) \leq \sigma_{outlier}$. A small $\sigma_{outlier}$ picks a relatively large smoothing bandwidth that merges outlier modes to significant modes, while a large $\sigma_{outlier}$ selects a small bandwidth to allow the separation of outliers. For the case of $O(h) > \sigma_{outlier}$ for all candidate bandwidths, it is not possible to merge the required number of outlier modes to significant modes given the optimization of $N(h)$, due to the high dissimilarity between the outliers and the main trends. In this case, we choose the largest h from the solution set to limit the number of outliers. The default settings for σ_{sig} and $\sigma_{outlier}$ are 30% of the number of ensemble members and 2 respectively. However, users can conveniently change these parameters according to their own needs as demonstrated in Section 5.

Fig. 3 illustrates the proposed bandwidth selection method for ensemble isocontours using the NCEP's SREF (Short-Range Ensemble Forecast) dataset. There are 26 members in this dataset, and we transform the ensemble isocontours to Euclidean space in $\mathbb{R}^{25}$. We did not manage exhaustive comparisons with existing methods due to their various computational overheads for our high dimensional data. For example, the popular R package 'ks' [17] provides rich bandwidth selection methods; however, it supports at most 6-dimensional data. Therefore, we compare our approach with the rule of thumb method. For each selected bandwidth, we examine its performance for mode clustering using the mean-shift algorithm. Fig. 3 shows the results of mode clustering with different bandwidths, where isocontours of different mode clusters are uniquely colored. Fig. 3 (a) shows that the rule of thumb method selected a very small bandwidth that significantly

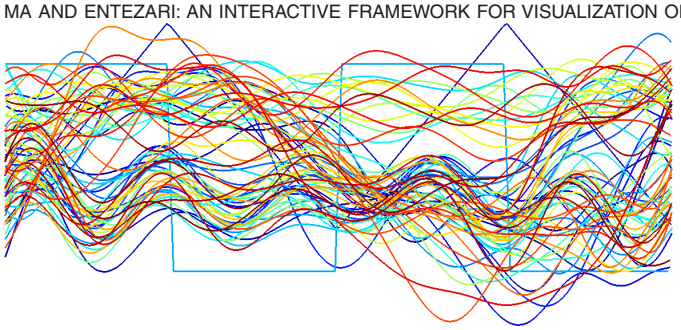


Fig. 4. The spaghetti plot for the synthetic data. The synthetic ensemble contains four trends and two outliers.

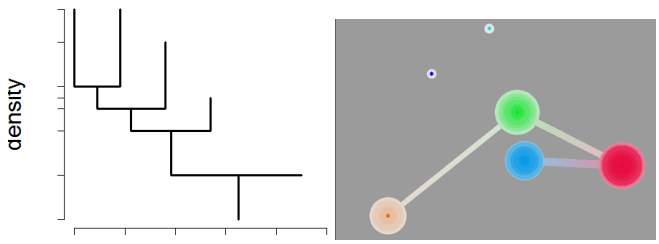


Fig. 5. The comparison of (a) the dendrogram and (b) the proposed mode plot for the synthetic data. In (b), each mode is represented as a concentric circle glyph and edges indicate the connections of modes.

under-smoothed our high-dimensional data, resulting in a small local mode for each isocontour. We suspect such phenomena may due to the limited efficacy of the signed distance transforms for representing isocontours. Using our method with $\sigma_{sig} = 15$ and $\sigma_{outlier} = 0$, Fig. 3 (b) shows that all isocontours are merged into a single cluster, because σ_{sig} is too large and no trend contains more than 15 isocontours. Reducing σ_{sig} to 10, we can see the two strong modes in (c), there seem to be some outliers for both trends (i.e., isocontours that pass through region A and B for the green and red clusters respectively). By setting $\sigma_{outlier} = 2$, we can further separate these outliers, as shown in (d).

4 Visualization

The dendrogram described in Section 3.1 is a standard tool to visualize the arrangement of hierarchical clusters. In this section, we present a new visualization, called *the mode plot*, to encode high-density clusters which is particularly effective for understanding the distribution behavior and variability of ensemble isocontours. The proposed mode plot is further linked to the spaghetti plot for interactive explorations of interesting subsets of isocontours, including trends, high probability solutions, and solution galleries. We also propose a so-called simplified spaghetti plot to provide an interpretable overview of ensemble isocontours across multiple isovalues, from which one can select informative isovalues for detailed analysis (using linked mode and spaghetti plots). Furthermore, our framework enables interactive selections of geographical sub-regions for clustering re-analysis and local explorations of ensemble isocontours.

To demonstrate the effectiveness of our approach, we synthesized an ensemble data, consisting of 72 members. All members are isocontours of implicit distance fields (361×199) of some perturbed functions. There are four trends synthesized from four sinusoidal functions (with size 15, 15, 20, and 20) and 2 outliers: a triangle wave and a square wave. The spaghetti plot of the synthetic data is shown in Fig. 4, where the overlapping of isocontours makes it difficult to draw any useful conclusions. We perform the high-density clustering with a bandwidth selected from our proposed method with $\sigma_{sig} = 15$ and $\sigma_{outlier} = 2$.

4.1 Visual Encoding of High-density Clusters

The dendrogram, shown in Fig. 2 (b) or Fig. 5 (a), primarily displays topological information of the density estimate, i.e., the creations and merges of mode branches. However, it does not sufficiently encode the connectivity between different modes which is important to understand the density estimate. For example, when two modes (say A and B) are connected in U_ε, their corresponding branches are merged into a new branch. With the expansion of U_ε, the two modes can be connected to another mode C, where the newly-created branch is merged with the branch of C. However, the connectivity among the three modes cannot be inferred from the dendrogram (e.g., is C connected to A or B?). Moreover, each branch represents a set of high-density clusters with the same topology, but it does not describe how the underlying cluster evolves and grows.

To overcome the limitations of dendrograms, we present the mode plot to encode high-density clusters for efficient explorations of ensemble isocontours. Our mode plot is an extension of the visualization proposed by Chen *et al.* [10], where we propose new concentric-circle glyphs to encode mode hierarchies and connectivities for the distribution of ensemble isocontours. We also introduce a number of interactions for mode plots, allowing convenient selections of desired isocontours. The basic idea for constructing a mode plot is straightforward: create a concentric circle for each mode to encode the evolution of the respective mode cluster, and connect the concentric circles when their corresponding modes are connected in U_ε. Fig. 5 (b) shows the proposed mode plot for the synthetic data, which is generated with the following steps:

Mode clustering and mode projection: We find local modes $l_1, \ldots, l_k$ and mode clusters of the kernel density estimate with our proposed bandwidth selection method (mean-shift algorithm can be used (Section 6)). The local modes are further projected into $\mathbb{R}^2$, denoted as $l_1^*, \ldots, l_k^*$, using multidimensional scaling (MDS).

Visual encoding of mode cluster evolution: We first plot the projected modes as uniquely colored hexagrams (shown in the centers of the concentric circles). Then, we uniformly sample 20 density levels $\varepsilon_1, \ldots, \varepsilon_{20}$ in the range $[1/20, 1] \times \hat{f}_{max}$. For each density level, ε, starting from ε_1 to ε_{20}, we compute the sizes of the intersections between the mode clusters and the current high-density cluster U_ε, denoted as $n_1(\varepsilon), \ldots, n_k(\varepsilon)$. If $n_i(\varepsilon) > 0$, we draw a circle around l_i^* whose radius is proportional to $n_i(\varepsilon)$. Each circle is further filled with the same color as its corresponding mode (i.e., the center hexagram), where the color saturation is modified based on the current density level ε and a sigmoid function: $s(\varepsilon) = 1/(1 + e^{-9(\varepsilon/\hat{f}_{max} - 0.65)})$. The non-linear monotonic increasing sigmoid function maps high density levels to high color saturations, which also increases the visual contrast of circles at high, medium, and low density levels.

Visual encoding of mode connectivity: For each ε, if two modes are connected in U_ε, we draw an edge between the two modes that has a width inversely proportional to ε and a gradient color varying along the connected circles. Note that all modes will eventually merge at an arbitrary low-density level, whereas we ignore such trivial connections as they do not convey much information.

The mode plot in Fig. 5 (b) clearly shows the mode structure of the density estimate, while it is difficult to perceive the same information from the dendrogram in (a). The relative positions of the concentric circle centers describe the similarity between the corresponding modes. Although modes can be readily discerned in the mode plot, modes are assigned unique colors to enable the separation and classification of isocontours in the linked spaghetti plot (Section 4.2). The maximum radius of each concentric circle reflects the size of the respective mode cluster. As such, outlier modes are represented as small isolated circles, e.g., the tiny dark blue and cyan circles. For each concentric circle, a large high-saturation core indicates the mode is seated in a high-density region, while modes born at low-density levels are encoded with small low-saturation cores. The radial changes of color saturation reflect the distribution of isocontours within mode clusters, a high contrast core and exterior implies the presence of inner-cluster outliers (e.g., Fig. 11 (d) red) whereas a low contrast concentric circle indicates isocontours have similar proximities to each other (e.g., Fig. 9 (g) red and green). A wide high-saturation line between two concentric circles means the corresponding modes are connected at a high-density level.

Both the inner-cluster variability and intra-cluster similarity influ-

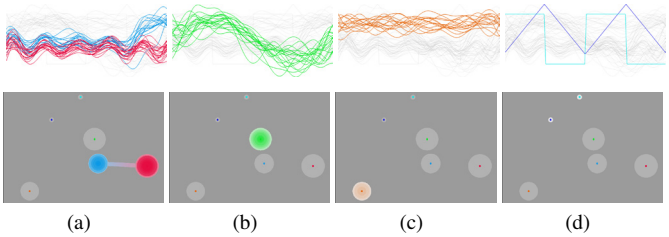


Fig. 6. The selections of different combinations of modes for the examination of various possible solutions.

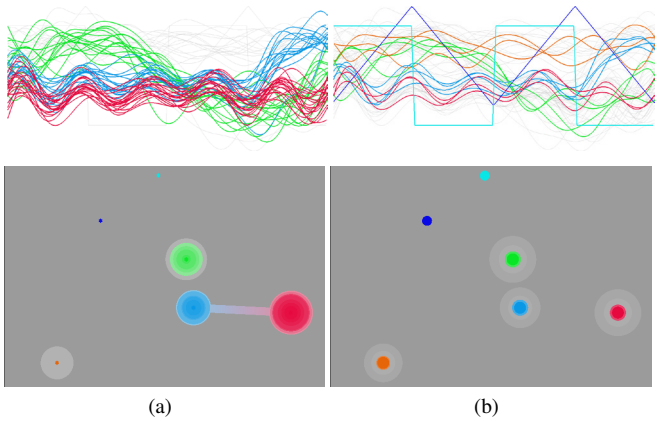


Fig. 7. (a) High-density filtration with $\varepsilon = 0.6 * \hat{f}_{max}$ for the extraction of high-probability solutions. (b) Mode percentile filtration with 20th percentile for the generation of the solution gallery.

ence the significance of modes as well as their connectivities. For example, as shown in Fig. 5 (b), the four large concentric circles correspond to four trends (at density regions from high to low): red, blue, green, and brown. Comparing the blue and green concentric circles, although the green is larger than the blue in maximum radius, it has a lower saturation core. Since both the green and blue concentric circles have similar distances to other trend modes (i.e., red and brown), the intra-cluster similarity is less likely to be the factor for such phenomena. Therefore, by simply inspecting the mode plot, we conclude that the lower saturation green is due to its higher inner-cluster variability compared to blue, which can be further verified via interactions (Section 4.2). On the other hand, the low-saturation brown mode and its large distances to other modes distinguish it from the high probability solutions.

4.2 Interactive Exploration of Ensemble Isocontours

We link the mode plot with the spaghetti plot and use interactive highlighting to explore the uncertainty and variability of ensemble isocontours. The mode plot serves as an attribute space summary of the distribution of ensemble isocontours, as well as a classification widget for interactive explorations. In the spatial context, interactive highlighting in the spaghetti plot enables direct visualization of the positional variations of the selected isocontours. Our interactive framework provides a focus+context visualization that enables both an overview of the clustering structures and the selection of isocontours of interest. We propose the following interactions for selecting trends, high-probability solutions, and solution galleries:

- **Mode selection:** The mode plot provides an overview of the distribution of ensemble isocontours, from which we enable the selection of modes of interest and allow users to focus on different solutions (either trends or outliers). Fig. 6 demonstrates various mode selections at density level $\varepsilon = 0$, where the entire concentric circles for the selected modes are highlighted in the mode plot, and the corresponding isocontours are highlighted in the spaghetti plot. This operation allows users to the focus on a subset of solutions, e.g., (a) main-trend solutions, (b) an important solution with less confidence, (c) a less-likely solution, (d) outliers. The interaction can be a pre-filter and used in conjunction with other interactions.

- **High-density filtration:** We allow the selection of isocontours in the high-density clusters at various density levels, which helps to identify the high-probability solutions (i.e., selection of isocontours with top density values) in a global manner. Fig. 7 (a) shows the high-density filtration at density level $\varepsilon = 0.6 * \hat{f}_{max}$. In the mode plot, the selected high-density clusters are represented as the partially colored concentric circles, and the connections indicate the corresponding modes belong to same high-density clusters. In the spaghetti plots, the selected isocontours are highlighted, and all other isocontours are greyed out. The selected high-probability isocontours can be used as predictions in weather forecasts, where it enables flexible control of the number of high-confidence isocontours.

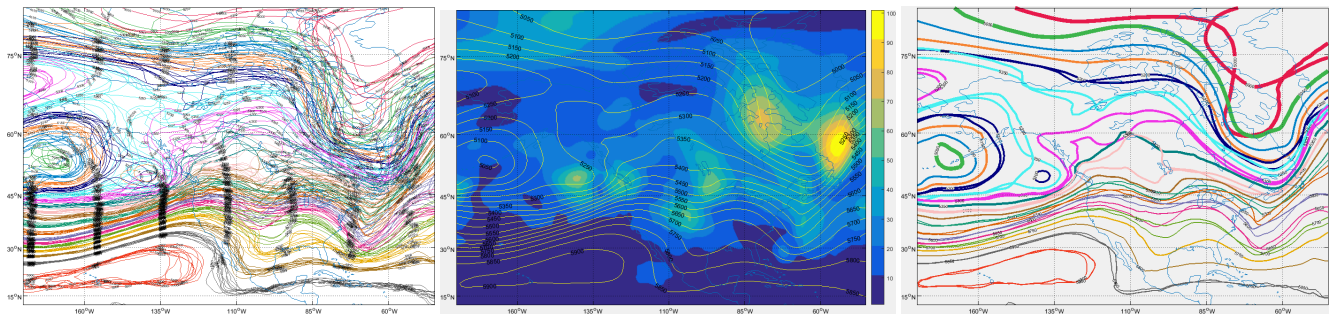
- **Mode percentile filtration:** It is often useful to generate a gallery of all possible solutions to sense the variability of ensemble isocontours [53]. For each mode, we rank the isocontours within each mode cluster based on their density values, where we can specify a percentile to select high-rank isocontours (e.g., 10th percentile selects the 10% highest ranked isocontours). By default, we have four percentile levels 10th, 25th, 50th, and 95th correspond to four solution galleries, where the concentric circle glyphs in mode plots are modified to four percentile levels, and the radius of circles indicates the sizes of the selected isocontours at different levels. A gallery of all possible solutions (including outliers) for the 20th percentile is shown in Fig. 7 (b).

4.3 Isovalue Selection

Current literature [4, 19, 23, 38, 41, 46, 56] assumes that the isovalues of interest are given, which is valid in scenarios in which threshold values of some fields are of interest. For example, a threshold of temperature, wind speed, and relative humidity is critical for fire forecasting. However, in operational weather forecasting, the uncertainty of ensemble isocontours is frequently used to assess the uncertainty associated with the weather forecast. Since ensemble isocontours of different isovalues exhibit different variability behaviors, it is essential to select interesting isovalues that provide the most useful information [53].

The brute force approach involves selecting isovalues from the spaghetti plot of a full set of isovalues, as shown in Fig. 8 (a). Even though the isocontours of different isovalues are assigned unique colors, overlaying a large number of isocontours brings significant visual clutter, making it difficult to interpret. To overcome this problem, the conventional method for isovalue selection in weather forecasting uses the mean-spread plot, as shown in Fig. 8 (b), where the mean field and the spread field (i.e., the standard deviation of ensemble members) are mapped to contours and colors. Since the goal is to select isovalues that carry the most uncertainty information, isovalues whose mean isocontours pass through multiple high uncertainty (or spread) regions are chosen for detailed analysis. However, this approach suffers from two drawbacks: 1. The ensemble isocontours are represented as a single mean isocontour, which could be misleading when the ensemble isocontours form clusters of diverging shapes. 2. The spread could also be misleading for regions with large gradients, since small displacements of members can lead to a large spread due to the large gradient magnitude [52]. For example, as shown in (b), the large spreads in the eastern regions are mostly due to the large gradient (indicated by the close isolines), while the western areas have relative low spreads, even though the isocontours exhibit multiple variations as shown in (a).

We propose to select isovalues of interest from a simplified spaghetti plot which renders a set of representative isocontours for multiple isovalues. For each isovalue, we choose the rank 1 isocontours for all significant modes as representatives. Specifically, we define significant modes using the same criteria in our bandwidth selection method in Section 3.2, i.e., a significant mode has a mode cluster size $\geq \sigma_{sig}$. We further encode the variability of isocontours within each mode cluster to the contour width of its representatives, where a large width



(a) (b) (c)

Fig. 8. The comparison of (a) the spaghetti plot, (b) the mean-spread plot, and (c) the simplified spaghetti plot for isovalue selection. The isocontours are from the Geopotential height field at 5-dam intervals of GEFS/R dataset.

indicates a high inner-mode variability. The simplified spaghetti plot shown in Fig. 8 (c) significantly reduces the visual clutter observed in (a). We can select informative isovalues by directly examining the positional variabilities of representative isocontours and their contour widths. For example, from (c) we can observe multiple possible trough locations in the U.S. south while this information is hard to obtain from (a). Moreover, the thin contours of these troughs indicate the low uncertainty associated with the representative isocontours. On the other hand, the thick representative isocontours in the north-east region indicate the relatively high variability within their mode clusters.

4.4 Sub-region Selection and Clustering Re-analysis

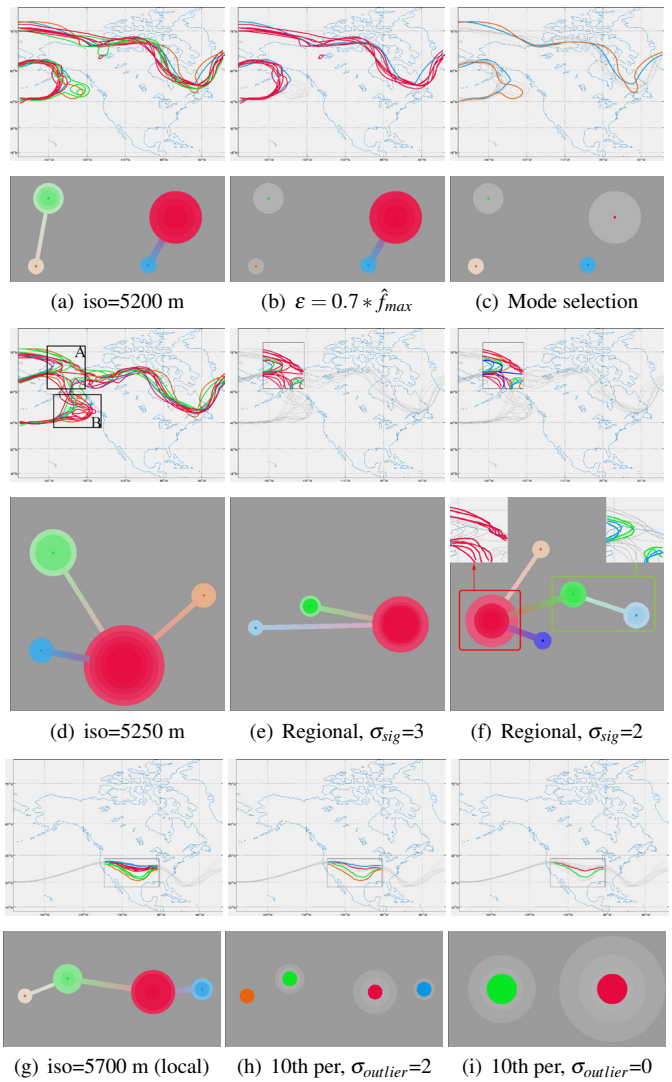
We have been discussing the uncertainty and variability analysis for ensemble isocontours in a given geographical region. Forecasters are also often interested in the positional variations of ensemble isocontours at different sub-regions [52, 53], in which case clustering analysis has to be repeated, because it is sensitive to the selection of regions [27]. Although the importance of this requirement has been recognized [18, 27], it has not been approached by existing methods. Our interactive framework allows selection of any rectangular region via brushing in the spaghetti plot and efficiently recomputes the high-density clustering for the selected region. For example, the right images of Fig. 1 show the clustering re-analysis for two selected regions, where the variability of the (middle) trough and the (eastern) ridge are better characterized in the local sense.

5 Results

We have conducted a number of experiments on ensemble forecasting datasets. Due to space limitations, we present some of our results; more experiments can be found in our supplemental video.

Our first experiment uses the historical weather forecast dataset from 2nd-generation Global Ensemble Forecast System Reforecast (GEFS/R). The dataset is publicly available from the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory [1]. The dataset consists of 11 forecast members on a regular longitude-latitude grid, and the horizontal resolution is 55 km (0-8 days) and 70 km (8-16 days). Although 11 elevation layers are potentially available to enable characterization of the dataset in 3D, forecasters almost always examine the 2D slices of the dataset at different elevation levels to interpret weather predictions [41]. We use the isocontours of the Geopotential height field at 500-hPa, initialized at 0000 UTC on 19 November 2001 and valid at 1200 UTC on 22 November 2001 in North America. This classic example has been used to study the uncertainty of ensemble weather forecasts [52, 53]. We set $\sigma_{sig} = 3$ (30% of the number of members) and $\sigma_{outlier} = 2$ for high-density clusterings.

Fig. 8 (a) shows the spaghetti plot of the ensemble isocontours at 5-dam intervals. The simplified spaghetti plot, shown in (c), provides an overview of the representative isocontours and their associated uncertainties. We select three isovalues (m): 5200, 5250, and 5700 that exhibit different variabilities across multiple regions for detailed analysis. In Fig. 9, the first column shows overviews of the high-density clustering results, where all modes are selected at the lowest density level in the mode plots and isocontours are highlighted accordingly in



(a) iso=5200 m (b) $\varepsilon = 0.7 * \hat{f}_{max}$ (c) Mode selection

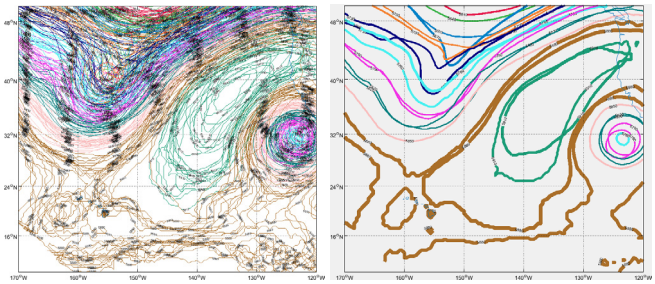
(d) iso=5250 m (e) Regional, σ_{sig}=3 (f) Regional, σ_{sig}=2

(g) iso=5700 m (local) (h) 10th per, $\sigma_{outlier}$=2 (i) 10th per, $\sigma_{outlier}$=0

Fig. 9. The demonstration of our framework for exploration of ensemble isocontours for isovalue (m) 5200 (a-c), 5250 (d-f), and 5700 (local) (g-i) of the GEFS/R dataset. The left column shows the summaries of high-density clusterings, and the right columns show interactions.

the spaghetti plots. The second and third columns show the interactive explorations of ensemble isocontours.

In Fig. 9 (a), the mode plot shows two trends: red and green, and comparing the size and saturation of the corresponding glyphs one can conclude that the red trend is stronger. The high saturation blue glyph, as well as its connection (with a relatively wide line) to the red glyph, indicates its similarity to the red trend, i.e, it is an inner-cluster outlier.



(a) (b)

Fig. 10. The comparison of (a) the spaghetti plot and (b) the simplified spaghetti plot of the Geopotential height field at 3-dam intervals of SREF dataset.

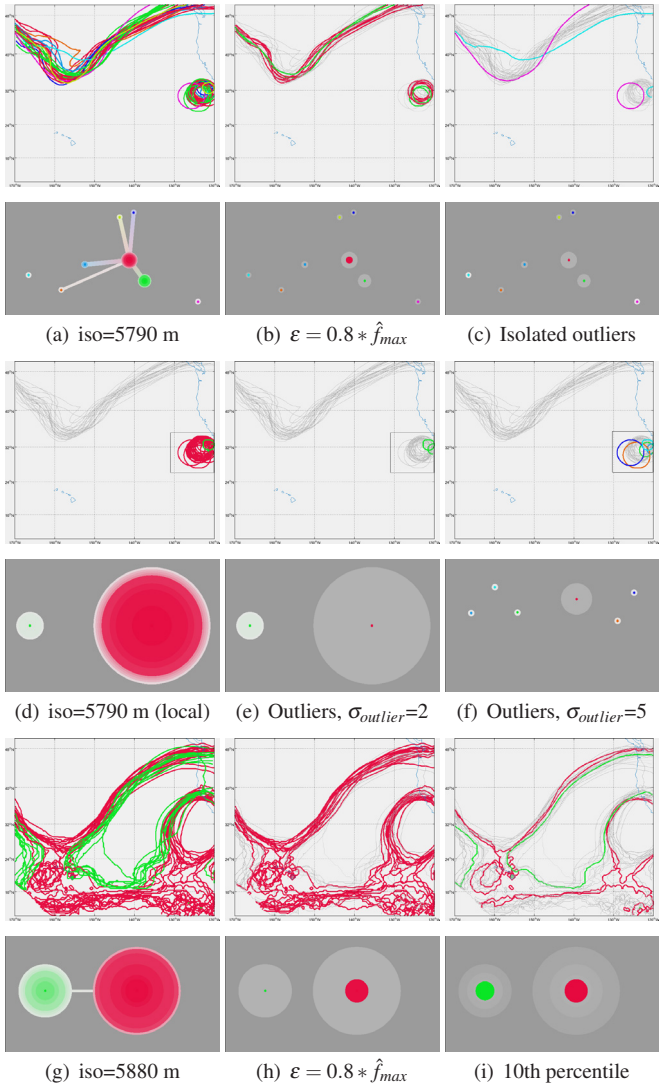
The brown glyph is also connected to the green trend; however, its relatively large distances to other glyphs and low saturation imply its distinctness. These initial conclusions can be further verified from the highlighted spaghetti plot (top), as well as the interactive selections of high-probability solutions and different modes, as shown in (b) and (c).

The mode plot in Fig. 9 (d) summarizes the distribution of ensemble isocontours with complicated variations over multiple regions. From the highlighted spaghetti plot, we notice that the two trends (green and red) mostly differ in region B while the variations and topologies in region A are not apparent. Therefore, we select region A for clustering re-analysis (took 0.32s), and the result is shown in (e). However, the dominant red trend is still composed of isocontours with various shapes and topologies. Because of the high variability of these isocontours, it is difficult to form significant clusters that have more than σ_{sig} (i.e., 3) members. To further structure the solution, we relax σ_{sig} to 2 (re-clustering took 0.55s), and the resulting mode plot in (f) shows one strong trend (red), one trend consists of two modes (connected green and blue), and two outliers that are connected to the red trend. The selections of the two trends are shown in the insets of the mode plot, where we observe isocontours with two different topologies. Additional interactions can be found in our supplemental video.

Fig. 9 (g) - (i) show the regional analysis for Fig. 1 (c). In Fig. 1 (c), the mode plot shows the dominance of the red trend and the distribution of different solutions, while it is difficult to obtain the same information from the spaghetti plot. Due to the overlapping and complex shapes of isocontours, it is difficult to compare different clusters in (even color-coded) spaghetti plots directly (e.g., the blue isocontours are occluded by the red isocontours in Fig. 1 (c)). The same drawback occurs in confidence envelope-based plots [19, 56], as we will discuss in Section 6. The mode plot in (g) summarizes the local analysis, where we see four possible trough locations with different strengths in the U.S. south. Compared to the global clustering (shown in Fig. 1 (c)), the local analysis better captures the location variations of the troughs. The 10th percentile solution gallery for all modes is shown in (h), where we observe the differences and similarities between the four troughs. To reduce outliers, we set $\sigma_{outlier} = 0$ for re-clustering (took 0.21s), and the representatives for the two main trends are shown in (i).

Our second experiment uses NCEP's SREF (Short-Range Ensemble Forecast), which contains 26 member runs: 13 Nonhydrostatic Multiscale Model on the B-grid (NMMB) members and 13 Weather Research and Forecasting (WRF) members. The simulation runs four times per day and includes forecasts at 3-hour intervals expanding to 87 hours in duration. We use the SREF version that runs on NCEP's 243 grid, which is a longitude-latitude grid over the Eastern North Pacific. The data is publicly available from NOAA [2]. We use the isocontours of the Geopotential height field at 500-hPa, initialized at 0000 UTC on 9 September 2017 and valid at 0000 UTC on 11 September 2017. We set $\sigma_{sig} = 8$ (30% of the number of members) and $\sigma_{outlier} = 2$ for high-density clusterings.

Fig. 10 compares the spaghetti plot and the simplified spaghetti plot of ensemble isocontours at 3-dam intervals. We do not select isovalues with diverging trends, instead we select two isovalues (m) to test the performance of our approach: 5790 (representative isocontours with subtle differences) and 5880 (complex topology with high inner-cluster



(a) iso=5790 m (b) $\varepsilon = 0.8 * \hat{f}_{max}$ (c) Isolated outliers

(d) iso=5790 m (local) (e) Outliers, $\sigma_{outlier}$=2 (f) Outliers, $\sigma_{outlier}$=5

(g) iso=5880 m (h) $\varepsilon = 0.8 * \hat{f}_{max}$ (i) 10th percentile

Fig. 11. The demonstration of our framework for exploration of ensemble isocontours for isovalue (m) 5790 (a-c), 5790 (local) (d-f), and 5880 (g-i) of the SREF dataset. The left column shows the summaries of high-density clusterings, and the right columns show interactions.

variations). The mode plot of Fig. 11 (a) shows two relatively strong trends (red and green) together with a number of outliers, where four outliers are connected to the red mode and two outliers are isolated. Again, this information cannot be obtained from the cluttered spaghetti plot shown at the top. The high-density filtration at $\varepsilon = 0.8 * \hat{f}_{max}$ is shown in (b), where the high-density isocontours from both the red and green trends are selected as a prediction. The isolated outliers are shown in (c). Since the isocontours are mainly composed of two disconnected components in two geographical regions, we select the component in the eastern part for regional analysis (re-clustering took 0.20s). The clustering result in (d) shows one main trend (red) and one outliers (picked out in (e)). The low saturation exterior of the red glyph indicates that there are additional outliers in the underlying trend. Therefore, we increased $\sigma_{outlier}$ to 5 and performed clustering re-analysis (took 0.66s) to separate these outliers from the main trend. The newly extracted outliers are shown in (f). In (g), although the red isocontours exhibit higher variability compared to the green ones (as shown in the spaghetti plot), the red trend is actually stronger (indicated by the mode plot) due to its large size. The high probability solutions with $\varepsilon = 0.8 * \hat{f}_{max}$ are presented in (e), which further shows the dominance of the red trend. Fig. 11 (f) shows the 10th percentile representative isocontours, where we see a cleaner plot to examine the positional differences of isocontours from the two trends.

6 Discussion

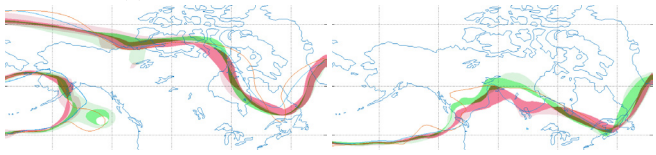
We compute the signed distance transforms of isocontours from scalar fields using an efficient image-based method. Our method is inspired by the Marching Cubes algorithm [28], where cells that are crossed by an isosurface are used to extract the triangle meshes of the isosurface. For 2D scalar fields, a cell in a Cartesian grid is a square between four neighboring grid points. A cell is crossed by an isocontour if the isovalue falls between the minimum and maximum of the scalar values of its four grid points. Therefore, we can form a binary image of cells, where each pixel represents a cell and is assigned value 1 if the underlying isocontour crosses the cell. Thus, we use the signed distance transform of the binary image as an approximation. The accuracy of using the binary image to approximate the geometry of an isocontour depends on the resolution of the data. Fortunately, one can always increase the resolution of the scalar field by subdividing the cells of the original grid into smaller cells and interpolating data over the finer grid [12]. As shown in Fig. 12, the finer the subdivision, the more accurately the binary image represents the isocontour. Note that the subdivision also increases the resolution of the signed distance transforms. However, it does not bring extra overhead to our clustering process, as the dimension of the isocontours is bounded by the number of ensemble members.

Although the high-density clusters are defined based on the density estimate, our interest is the evolution of population clusters, not the density estimate itself. Thus, we can avoid evaluation of the density estimate, which is very expensive for high-dimensional data. Our implementation is based on the relationship between the (hierarchical) high-density clustering and the (flat) mode clustering, as described in Section 4.1. Intuitively, one can treat the high-density clustering as a hierarchical version of the mode clustering. First, we use the mean-shift algorithm, together with our proposed bandwidth selection method, to find the modes and mode clusters. The high-density clusters at level ε are composed of the intersections of the mode clusters and U_ε. Then, for any pairs of modes, we need to determine whether they are connected in U_ε. Several methods have been proposed for solving this problem [26, 31]. In particular, an efficient heuristic – modes are not connected if there is a low-density region (that is not in U_ε) between the modes – is suggested in [10]. For any two mode clusters that intersect U_ε, a density estimation is performed along the line segment formed by the closest points between the two clusters. Then, the density values for samples along the line are examined, and the two modes are not connected if there is a sample whose density value is smaller than ε. However, this approximation is insufficient for high-dimensional data, as a single line covers only very limited areas. In our implementation, we extend this idea and examine the regions that are covered by the line segments between every pair of data points from the two clusters. We use an efficient C-based library [16] for the implementation of the mean-shift algorithm. The average pre-processing time (over the set of isovalues) for bandwidth selection and overall clustering (on a machine with Intel Core i7 3.6 GHz CPU, 16 GB RAM) is 0.0072s and 0.33s for the GEFF/R dataset and 0.14s and 0.94s for the SREF dataset.

Our method naturally provides a density ranking for ensemble isocontours, which can be potentially used to generate confidence envelopes as in [19, 56]. For example, we can use mode clusters to form bands. For each mode cluster, we create a 50th percentile band enclosed by a 100th percentile band similar to [56]. The confidence bands for isovalue (m) 5250 and 5350 of the GEFF/R dataset are shown in Fig. 13. From the confidence bands, we obtain both qualitative and quantitative interpretations of shapes of isocontours and the associated variability. However, this method suffers from a similar problem as the spaghetti plots; that is, it is difficult to compare isocontours of different trends directly. The clusters of isocontours are abstracted into confidence bands, and it is difficult to obtain information like the number of isocontours within each band and the positional differences between the isocontours. For example, in Fig. 13 (a), it is not possible to compare the significance of the two trends as in the mode plot of Fig. 9 (a). Therefore, we chose not to integrate the confidence band method into our framework. Instead, we prefer a direct visualization of isocontours (using interactive highlighting) with the guidance of the proposed mode plot. However, we envision that confidence envelope methods could be useful for visualizations of 3D ensembles, as overlaying isosurfaces with transparent layers often do not provide an effective visualization due to occlusions. In the future, we plan to extend our framework for 3D ensemble datasets and investigate the confidence envelope technique for the visualization of 3D isosurfaces.

(a) (b) (c)

Fig. 12. The renderings of the binary images of cells for isovalue (m) 5350 of the GEFS/R dataset, where each cell is divided (a) 0 times, (b) 1 times, and (c) 3 times.



(a) (b)

Fig. 13. The illustration of our density-based confidence bands for isovalue (m) (a) 5200 and (b) 5350 of the GEFS/R dataset.

One of the limitations of our approach is the possible occlusion problem in the mode plot. For example, when there are many outlier clusters that connect to strong modes, it is possible that the connections (or edges) may intersect each other or even occlude other (tiny) modes. Although this happens rarely in our experiments, we plan to further improve the mode plot by modifying the objective function of MDS to have a more intelligent arrangement of modes. The simplified spaghetti plot is still a contour plot that overlays multiple isocontours and therefore inherits the limitation of conventional spaghetti plots; that is, it does not scale well. As the data load increases, the simplified spaghetti plot eventually becomes cluttered and not useful. Therefore, we plan to improve our isovalue selection method using our abstract summary product – the mode plot. For example, one possible solution is to stack the mode plots of consecutive isovalues to generate a 3D scalar field. Then, one might infer the distribution of ensemble isocontours across a dense set of isovalues (i.e., each isovalue corresponds to a slice) from the visualization of the 3D scalar field using volume rendering. With a similar idea, we also plan to extend the mode plot for the exploration of the temporal growth of the uncertainty in ensemble forecasts. Finally, we note that the proposed clustering approach can also be applied to the underlying fields of ensemble data as these fields can be treated as vectors. However, statistical analysis of the underlying fields is different from the analysis of ensemble isocontours of these fields [56]. We plan to investigate this problem and develop visualizations to further understand the discrepancies and connections between the underlying fields and their ensemble isoconoturs for uncertainty analysis.

7 Conclusion

In this paper, we propose an interactive framework for uncertainty exploration of ensemble isocontours. We generalize the high-density clustering for ensemble isocontours and propose a novel bandwidth selection method for estimating the density function of ensemble isocontours. We propose the mode plot, together with the linked spaghetti plot, to interactively select and explore interesting subsets of isocontours. To select informative isovalues, we propose the simplified spaghetti plot to provide a highly interpretable overview of the variability of ensemble isocontours for multiple isovalues. Our framework also allows for interactive selection of sub-regions for local uncertainty re-analysis. Our framework offers a focus+context visualization and enables a complete exploration of ensemble isocontours.

Acknowledgments

The authors thank Florian Ferstl [19] for providing the source code for contour variability plots. This work was supported by the US National Science Foundation (NSF IIS-1617101), and the Office of Naval Research (N00014-14-1-0762).

REFERENCES

[1] ESRL/PSD GEFS Reforecast Version 2. https://www.esrl.noaa.gov/psd/forecasts/reforecast2/.

[2] National Oceanic and Atmospheric Administration. http://nomads.ncep.noaa.gov/, 2004.

[3] T. Athawale and A. Entezari. Uncertainty quantification in linear interpolation for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2723–2732, Dec 2013. doi: 10.1109/TVCG.2013.208

[4] T. Athawale, E. Sakhaee, and A. Entezari. Isosurface visualization of data with nonparametric models for uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):777–786, Jan 2016. doi: 10.1109/TVCG.2015.2467958

[5] A. Biswas, G. Lin, X. Liu, and H. W. Shen. Visualization of time-varying weather ensembles across multiple resolutions. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):841–850, Jan 2017. doi: 10.1109/TVCG.2016.2598869

[6] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. *Overview and State-of-the-Art of Uncertainty Visualization*, pp. 3–27. Springer London, 2014. doi: 10.1007/978-1-4471-6497-5_1

[7] K. Brodlie, R. AllendesOsorio, and A. Lopes. *A Review of Uncertainty in Data Visualization*, pp. 81–109. Springer London, 2012. doi: 10.1007/978-1-4471-2804-5_6

[8] S. Bruckner and T. Möller. Isosurface similarity maps. *Computer Graphics Forum*, 29(3):773–782, 2010. doi: 10.1111/j.1467-8659.2009.01689.x

[9] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

[10] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, pp. 1–13, 2017.

[11] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, Aug 1995. doi: 10.1109/34.400568

[12] H. E. Cline, W. E. Lorensen, S. Ludke, C. R. Crawford, and B. C. Teeter. Two algorithms for the three-dimensional reconstruction of tomograms. *Medical physics*, 15(3):320–327, 1988.

[13] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3d ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2694–2703, Dec 2014. doi: 10.1109/TVCG.2014.2346448

[14] I. Demir, M. Jarema, and R. Westermann. Visualizing the central tendency of ensembles of shapes. In *SIGGRAPH ASIA 2016 Symposium on Visualization*, SA '16, pp. 3:1–3:8. ACM, 2016. doi: 10.1145/3002151.3002165

[15] I. Demir, J. Kehrer, and R. Westermann. Screen-space silhouettes for visualizing ensembles of 3d isosurfaces. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 204–208, April 2016. doi: 10.1109/PACIFICVIS.2016.7465271

[16] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox.

[17] T. Duong et al. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software*, 21(7):1–16, 2007.

[18] F. Ferstl, K. Bürger, and R. Westermann. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):767–776, 2016.

[19] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann. Visual analysis of spatial variability and global correlations in ensembles of iso-contours. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization*, EuroVis '16, pp. 221–230, 2016. doi: 10.1111/cgf.12898

[20] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann. Time-hierarchical clustering and visualization of weather forecast ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):831–840, Jan 2017. doi: 10.1109/TVCG.2016.2598868

[21] L. Hao, C. G. Healey, and S. A. Bass. Effective visualization of temporal ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):787–796, Jan 2016. doi: 10.1109/TVCG.2015.2468093

[22] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., 99th ed., 1975.

[23] S. Hazarika, S. Dutta, and H. W. Shen. Visualizing the variations of ensemble of isosurfaces. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 209–213, April 2016. doi: 10.1109/PACIFICVIS.2016.7465272

[24] T. Höllt, A. Magdy, P. Zhan, G. Chen, G. Gopalakrishnan, I. Hoteit, C. D. Hansen, and M. Hadwiger. Ovis: A framework for visual analysisof ocean forecast ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1114–1126, Aug 2014. doi: 10.1109/TVCG.2014.2307892

[25] J. Kehrer, P. Filzmoser, and H. Hauser. Brushing moments in interactive visual analysis. *Computer Graphics Forum*, 29(3):813–822, 2010. doi: 10.1111/j.1467-8659.2009.01697.x

[26] S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. *arXiv preprint arXiv:1105.0540*, 2011.

[27] A. Kumpf, B. Tost, M. Baumgart, M. Riemer, R. Westermann, and M. Rautenhaus. Visualizing confidence in cluster-based ensemble weather forecast analyses. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):109–119, Jan 2018. doi: 10.1109/TVCG.2017.2745178

[28] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, pp. 163–169, 1987.

[29] B. Ma and A. Entezari. Volumetric feature-based classification and visibility analysis for transfer function design. *IEEE Transactions on Visualization and Computer Graphics*, 2018 (in press). doi: 10.1109/TVCG.2017.2776935

[30] B. Ma, S. K. Suter, and A. Entezari. Quality assessment of volume compression approaches using isovalue clustering. *Computers & Graphics*, 63:18–27, 2017.

[31] G. Menardi and A. Azzalini. An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, 24(5):753–767, 2014.

[32] H. Obermaier, K. Bensema, and K. I. Joy. Visual trends analysis in time-varying ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(10):2331–2342, Oct 2016. doi: 10.1109/TVCG.2015.2507592

[33] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *IEEE Computer Graphics and Applications*, 34(3):8–11, May 2014. doi: 10.1109/MCG.2014.52

[34] S. Oeltze, D. J. Lehmann, A. Kuhn, G. Janiga, H. Theisel, and B. Preim. Blood flow clustering and applications invirtual stenting of intracranial aneurysms. *IEEE Transactions on Visualization and Computer Graphics*, 20(5):686–701, May 2014. doi: 10.1109/TVCG.2013.2297914

[35] T. Pfaffelmoser, M. Reitinger, and R. Westermann. Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. *Computer Graphics Forum*, 30(3):951–960, 2011. doi: 10.1111/j.1467-8659.2011.01944.x

[36] K. Pöthkow and H. C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, Oct 2011. doi: 10.1109/TVCG.2010.247

[37] K. Pöthkow and H.-C. Hege. Nonparametric models for uncertainty visualization. *Computer Graphics Forum*, 32(3pt2):131–140, 2013. doi: 10.1111/cgf.12100

[38] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic Marching Cubes. *Computer Graphics Forum*, 2011. doi: 10.1111/j.1467-8659.2011.01942.x

[39] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In A. M. Dienstfrey and R. F. Boisvert, eds., *Uncertainty Quantification in Scientific Computing*, pp. 226–249. Springer Berlin Heidelberg, 2012.

[40] K. Potter, A. Wilson, P. T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 233–240, Dec 2009. doi: 10.1109/ICDMW.2009.55

[41] P. S. Quinan and M. Meyer. Visually comparing weather features in forecasts. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):389–398, Jan 2016. doi: 10.1109/TVCG.2015.2467754

[42] M. Rautenhaus, M. Böttinger, S. Siemen, R. Hoffman, R. M. Kirby, M. Mirzargar, N. Röber, and R. Westermann. Visualization in meteorology—a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[43] M. Rautenhaus, M. Kern, A. Schäfler, and R. Westermann. Three-

dimensional visualization of ensemble weather forecasts part 1: The visualization tool met.3d (version 1.0). *Geoscientific Model Development*, 8(7):2329–2353, 2015. doi: 10.5194/gmd-8-2329-2015

[44] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pp. 65–78, 1982.

[45] E. Sakhaee and A. Entezari. A statistical direct volume rendering framework for visualization of uncertain data. *IEEE Transactions on Visualization and Computer Graphics*, 23(12):2509–2520, Dec 2017. doi: 10.1109/TVCG.2016.2637333

[46] 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*, 16(6):1421–1430, Nov. 2010. doi: 10.1109/TVCG.2010.181

[47] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[48] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

[49] Q. Shu, H. Guo, J. Liang, L. Che, J. Liu, and X. Yuan. Ensemblegraph: Interactive visual analysis of spatiotemporal behaviors in ensemble simulation data. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 56–63, April 2016. doi: 10.1109/PACIFICVIS.2016.7465251

[50] B. W. Silverman. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[51] D. M. Thomas and V. Natarajan. Multiscale symmetry detection in scalar fields by clustering contours. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2427–2436, Dec 2014. doi: 10.1109/TVCG.2014.2346332

[52] UCAR/COMET. Ensemble Forecasting Explained. `https://www.meted.ucar.edu/training_module.php?id=156#.WqVkZ5MbPVo`, 2004.

[53] UCAR/COMET. Introduction to Ensemble Prediction. `https://www.meted.ucar.edu/training_module.php?id=170#.WqViqpMbPVo`, 2005.

[54] J. Wang, X. Liu, H. W. Shen, and G. Lin. Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):81–90, Jan 2017. doi: 10.1109/TVCG.2016.2598830

[55] L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1), 2018. doi: 10.1146/annurev-statistics-031017-100045

[56] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2713–2722, Dec 2013. doi: 10.1109/TVCG.2013.143