

Visual Analysis of Transport Similarity in 2D CFD Ensembles

Brad E. Hollister and Alex Pang; Basking School of Engineering; Santa Cruz, CA

Abstract

Currently, there are no all-inclusive methods for visual analysis of ensemble vector fields (EVF) that provide identification of flow trends and general flow similarity over the full extent of transport across ensemble members. Finite-time Variance Analysis (FTVA) provides flow structure information only on particle distributions at the termination of streamline integration. In this paper, we first present a flow structure based on streamline clustering. Second, we discuss a method using streamline clustering to provide information of flow coherence at corresponding spatial regions in the EVF. We consider the regions where bifurcation in flow trends among the EVF members occur. We will also discuss how both methods can be used as a sequential framework for EVF analysis, by using the results of the scalar flow structure to find regions of member flow dissimilarity for further analysis.

1 Introduction

Ensembles of vector field data, as produced via *Computational Fluid Dynamics* (CFD) simulations, are now common within the simulation community, in order to represent the output of a fluid model using distributions of input parameters [29]. The variation in parameter selection can represent uncertainty about boundary conditions, densities or other relevant input.

As a consequence, we are now faced with the challenge of analyzing and visualizing ensemble vector fields (EVF). EVF are made up of individual realizations, each a possible outcome, of the simulation. Flow has traditionally been visualized by advection of mass-less particles, e.g. streamline integration, in a certain vector field. There are many methods to analyze single instance vector fields and quantify their flow.

When extending those methods to ensembles, multiple problems arise. For one, statistical variation likely exists between the members of an EVF. A key visualization problem is first detecting and then displaying that variation. Most importantly, we want to draw attention to significant trends among members. Modes of flow coherence (e.g., trends) should ideally be considered over the full extent of flow: (1) initially, for identification, between the entire paths of each particle's movement sharing a common seed location within the field and then (2) subsequently, within known regions of the field where the modes of variation are clearly evident, as determined from the results of the initial consideration.

Until now, such methods as Finite-time Variance Analysis (FTVA) [27] have been employed to quantify global flow variation in an ensemble. Such methods only investigate variation in the flow through a given seed location at the termination of integration, via the principal components of the covariance matrix computed from the positions of particle deposition. Transport separation, however, may occur anywhere along streamlines with a common seed over the ensemble. FTVA, therefore, overlooks potentially important bifurcation between members.

In this work, we provide the following contributions:

- We utilize proven and efficient streamline clustering methods [1] to characterize, on the scale of the entire field, the flow coherence and bifurcation of the ensemble.
- We quantify via a two-stage streamline clustering method using representative streamlines from their cluster, the degree of flow coherence in regions of known bifurcation across the ensemble members.
- We show how both methods can be used together by first employing the flow structure to identify potential bifurcation and then the exploration of the regions of bifurcation.

2 Related Work

Much work had been done to define and identify global features of flow fields for crisp vector fields. Relevant publications are summarized here.

Lagrangian Coherent Structures (LCS) are a broad class of feature identification for the fluid medium [23]. Perhaps the first notable example is the Finite-time Lyapunov Exponent (FTLE) fields [7] for steady and unsteady vector field visualization.

Generalization of LCS has been discussed in depth [11]. Frameworks for flow field structure definition and visualization have been laid out in [25]. There, the authors discuss pathline predicate definitions relevant for given investigations of flow phenomena.

A variance based FTLE-like method for unsteady uncertain vector fields was first presented in [27]. This method reports the spatial second moment of particle destination, using the principal components of their covariance matrix as a result of initial uncertainty in the vector field. Theisel et al. [21], [22] examined uncertain vector field topology using Gaussian uncertainty. Analysis of streamline separation at infinity using time-discrete Markov Chains was explored in [24], in order to remove the finite-time requirements from [27].

While the papers discussed so far did not utilize EVF, adaption of probabilistic and summary statistics are discussed in the survey paper [20]. Hummel et al. [8] was the first work to apply FTVA from [27] to address EVF visualization. Their paper also used a Minimum Spanning Tree (MST) to detect and visualize trends in particle destinations at finite-time.

With novel numerical schemes to generate ensemble data using non-Gaussian input parameters [29] and [26], techniques to show the subtle variation and modality in output EVF is becoming increasingly needed from the visualization community.

Similar to our work but not appropriate for flow trend detection, are several streamline clustering methods. In [12], the authors extend the point-based clustering algorithm called Density-based Spatial Clustering of Applications with Noise (DBSCAN) to line segments. They applied this method to find representative trajectories in hurricane track data. In [15], the authors use curvature distribution of a field of streamlines to find shape similarity. Neither of these studies are ensemble based, but use crisp vector

fields.

Chen et al. [1] provide an efficient two-stage streamline clustering method based on spatial properties. The first-stage groups streamlines using k-means for feature vectors comprised of the start-point, mid-point, and destination-point of streamlines. Their second-stage finds sub-clusters from the first-stage, based on linear and angular entropy. They summarize flow in regions by finding representative streamlines closest to cluster centroids. Evaluation of fiber clustering methods for diffusion tensor imaging is discussed in [18]. It was from this study that [1] gave an approximate and efficient method.

Guo et al. outline a framework in [6] to provide an interactive assessment of ensemble variation. They call their system eFLAA (ensemble Flow Line Advection and Analysis). They present a novel parallel computation for calculating streamline spatial difference over an ensemble and then visualizing the differences. They compute various features of their ensembles (e.g., carbon dioxide concentration) along streamlines whose variation meets a given threshold.

Mirzargar et al. [17] extend boxplots to curves. They apply their method to quantify and visualize ensemble streamlines and hurricane track data. While they show the band-depth for individual streamlines, they do not delineate bifurcation between member streamlines. Their method is not directly applicable to a dense-field summary of streamline data.

3 Background

We briefly describe the current methods for extracting flow structure from crisp vector fields and EVF. We also discuss information entropy as related to streamline identification and its potential use for EVF statistics.

3.1 Flow Classification

Flow classification is based on material transport in vector fields, and thus provides a global picture of the vector field. The *flow map* Φ is derived from the vector field using integration.

$$\Phi(\mathbf{x}(t); T) = \mathbf{x}(t + T) \quad (1)$$

Equation 1 describes the final location of a particle seeded at \mathbf{x} at time t and advected for an interval T . The field is not required to be time-varying and in such a case, T simply refers to the number of integration steps forward or backward in Φ .

3.2 Finite-time Lyapunov Exponent

Taking the largest eigenvalue of the right-Cauchy Green deformation tensor, Eq. 2, we find the magnitude of the direction of greatest stretching in the flow medium at $\mathbf{x}(t)$. The tensor removes effects of reference frame rotations in $\nabla\Phi$.

$$\lambda_{\max}(\nabla\Phi(\mathbf{x}(t); T)^T \nabla\Phi(\mathbf{x}(t); T)) \quad (2)$$

The finite-time Lyapunov exponent is a logarithmic scaling of the maximum direction (Eq. 3).

$$FTLE(\mathbf{x}(t), T) = \frac{1}{T} \log \sqrt{\lambda_{\max}} \quad (3)$$

FTLE is a scalar field over the vector field domain. Finding its height ridges provides a topological skeleton of the regions in contraction or expansion.

3.3 Ensemble Vector Fields

Ensemble vector fields (EVF) are uncertain vector fields derived from variations between multiple instances (or runs) of an experimental/observation space (i.e., a container or geographical volume for inspection and the related starting conditions, computational model, and fluid characteristics). Repeated runs of the same simulation, with varying simulation input parameters, produce member realizations that taken together can be considered as a distribution of all possible outcomes of the field for a given set of parameters. For the purposes of this study, we limit our definition of an EVF to the definition given in Hummel et al. [8].

In that definition, a time-varying flow field can be described as in Eq. 4, where v is defined over a spatial domain $\Omega \subseteq \mathbb{R}^d$, with dimension d .

$$v : \Omega \times I \rightarrow \mathbb{R}^d \quad (4)$$

The time interval is $I \subseteq \mathbb{R}$. An EVF is a set of m vector fields over the same spatial domain and the ensemble space can be considered to be the intersection of all such vector fields, $\Omega_{EVF} = \Omega_1 \cap \dots \cap \Omega_m$ and $I_{EVF} = I_1 \cap \dots \cap I_m$.

$$EVF : \{1, \dots, m\} \times \Omega_{EVF} \times I_{EVF} \rightarrow \mathbb{R}^d \quad (5)$$

$EVF(i, \dots)$ corresponds to the i -th realization in our ensemble. We can see an example of particle transport in an ensemble (Fig. 1).

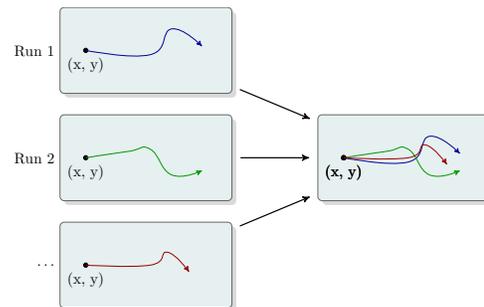


Figure 1: Streamlines seeded at the same positions in all members of the EVF have different transport paths. Seeds in the EVF lead to stronger or weaker path trends. Note that this is similar to FTVA for EVF, but that streamlines may terminate with weak separation but have strong separation anywhere along their trajectories. Here, the green streamline branches from the blue and red streamlines, but all terminate with weak variance.

3.4 Finite-time Variance Analysis

A probabilistic variant of FTLE is called the FTVA, Eq. 6. It takes the covariance matrix of particle positions advected over the ensemble domain from given seed locations. It was first presented by Schneider et al. [27].

$$FTVA(\mathbf{x}(t), T) = \frac{1}{T} \log \sqrt{\lambda_{\max}(\text{Cov}(\mathbf{x}(t); T))} \quad (6)$$

3.5 Streamline Information Entropy

Many works have used information theory [28] applied to streamline geometry [5], [15], [16] for the purposes of selecting streamlines. In this study, we are interested in summarizing

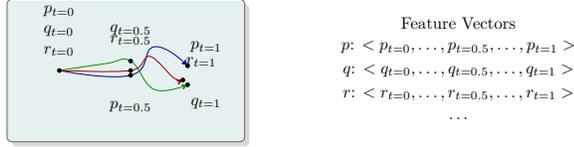


Figure 2: Shown here are three example streamlines all starting at the same location. We use at least the beginning, middle, and end locations. Other points used in the feature vector are evenly spaced over the approximated arc length and registered.

streamlines from the EVF with a common seed. We use this summary in two ways. First, it is used to weight the sampling frequency of points along streamlines (i.e., a higher sampling frequency captures greater streamline variability). Second, we utilize entropy as a reference map to better understand the overall variation in streamline geometry from the EVF.

We use both linear and angular streamline entropy [1]. Equation 7 represents the linear entropy [5], E_L , of single streamline. L_S is its total length and m the number of positions available from the numerical integration. D_j is the length of the j -th segment.

$$E_L = -\frac{1}{\log_2(m+1)} \sum_{j=0}^m \frac{D_j}{L_S} \log_2 \frac{D_j}{L_S} \quad (7)$$

Equation 8 represents the angular entropy [16], with A_j the angle of the line segment j , L_A is the total angular variation along the streamline (e.g. the sum of the absolute values of the A_j), and E_A the total angular entropy for the streamline.

$$E_A = -\frac{1}{\log_2(m)} \sum_{j=0}^{m-1} \frac{A_j}{L_A} \log_2 \frac{A_j}{L_A} \quad (8)$$

Both of these metrics summarize the degree of variation in a streamline over its entire path.

4 Methods

We first describe our method for extracting a cluster-based flow structure from an EVF. Second, we provide an exploratory region-based EVF similarity metric based on the same underlying streamline clustering method.

In section 5, we show how the results from our flow structure can guide a user to probe more deeply into the regions that give rise to global bifurcation in transport among the members.

4.1 Cluster-based Flow Map

For a seed in the simulation domain Ω_{EVF} , we define a feature vector to represent each streamline. We sample position as a spatial feature. The number of features included are at a minimum the initial, middle, and terminal positions of a streamline. Streamline clusters are found for each seed in Ω_{EVF} , where a velocity value has been stored from the simulation. This result is similar to Φ . The *cluster map* Φ_C , is represented as:

$$\Phi_C(\mathbf{x}) = |C_S|, C_S = \{c_1, \dots, c_n\} \quad (9)$$

where \mathbf{x} is the location of the seeded streamlines, C_S is the set of all streamline clusters c_i , i is an integer such that $0 \leq i \leq n$, and n the number of clusters. $|C_S|$ is the cardinality of the finite set C_S . Set c_i contains the similar streamline feature vectors seeded at \mathbf{x} .

We use the mean linear \bar{E}_L , and mean angular entropy \bar{E}_A of a population of streamlines to determine the frequency of sampling. The following steps are performed in computing $\Phi_C(\mathbf{x})$ for each \mathbf{x} :

Step 1 Lookup precomputed \bar{E}_L for \mathbf{x} .

Step 2 Lookup precomputed \bar{E}_A for \mathbf{x} .

Step 3 Calculate the number of streamline sample points, $\propto (\bar{E}_L + \bar{E}_A)$.

Step 4 For each streamline, assign a feature vector.

Step 5 Perform DBSCAN on all streamline feature vectors.

Step 6 Record the number of clusters found in Φ_C .

The number of regularly sampled features is proportional to the mean linear and angular entropy (see step 3 above). We linearly interpolate the number of samples between a minimum and a maximum positive integer and take the floor of the result. The upper-limit on the number of samples is dependent on the data or user constraints. The α for interpolation is equal to the ratio of the average of the linear and angular entropy (at the seed) to the absolute value of the difference between the maximum and minimum total entropy (linear and angular entropy combined) from the data set.

Because we desire to detect bundles of streamlines that may start out together, diverge, and finally converge over the ensemble members, we need to sample spatial features that are registered between the streamlines. Note that our method of clustering is inspired by [1]. They found sub-clusters based on entropy from initially grouping streamlines sampled at three spatial locations each. We use streamline entropy to determine sample frequency for streamlines at a seed. In Fig. 2, the blue and red streamlines are spatially similar. However, if the minimum three points are used for the feature vector, all streamlines in the example would be found in a single cluster.

4.2 Spatial Feature Registration

Streamline registration of spatial features is accomplished via an approximation of arc length. t is a real number on the interval $[0, 1]$ and is considered a fraction of the total arc length of a curve (streamline). The arc length L of curve S is defined as in Eq. 10 on the interval $[a, b]$. $ds^2 = dx^2 + dy^2$, for the infinitesimal line segment ds .

$$L_S = \int_a^b ds = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx \quad (10)$$

S is the streamline from which we have a set of points derived from numerical integration in the vector field. l can be considered an ordered list of those points and can be accessed by index i . For our finite approximation, when n is the number of points from integration, we have:

$$L_S = \sum_{i=0}^{n-1} \text{dist}(l(i), l(i+1)) \quad (11)$$

where dist is the Euclidean distance between two points. Parameter t then, is the fraction of L_S we wish to consider for comparison between a registered set of streamlines.

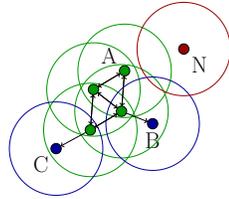


Figure 3: Illustration of DBSCAN cluster analysis requiring minimum points constituting a cluster. Points around A are core points. Points B and C are not core points, but are density-connected via the cluster of A (and thus belong to this cluster). Point N is Noise, since it is neither a core point nor reachable from a core point. DBSCAN also requires a maximum distance parameter ϵ that determines density-connected points [4].

4.3 Cluster Parameter Selection

Hummel et al. used a MST for terminal point trend clustering [8]. That study reported using a fraction of the average length of streamlines for the minimum distance between clusters.

We apply DBSCAN to assign cluster labels to member streamlines. Refer to Fig. 3 for an example. DBSCAN takes two parameters: ϵ , the maximum distance between features in a cluster, and $minPts$, the minimum number of data points in a cluster. The value for ϵ can be chosen by using a k-distance graph, plotting the distance to the $k = minPts$ nearest neighbor. Good values of ϵ are where this plot shows a strong bend. If ϵ is chosen too small, a large part of the data will not be clustered. Whereas for a too high value of ϵ , clusters will merge and the majority of objects will be in the same cluster [4].

We, however, take an approach similar to [8], setting the minimum distance between clusters to be related to their spatial domain. We use five percent of the diagonal distance across the full simulation domain as ϵ . For p-values, most authors refer to statistically significant as $P < 0.05$ [19]. Thus, five percent presents itself as a good “rule-of-thumb” for the fraction of the domain. We do not use the length of the streamlines themselves because we apply our clustering to multiple points along the streamlines. ϵ needs to be a function of the spatial domain size instead.

Ester et al. recommends $minPts \geq D + 1$, where D is the dimension of the data set [4]. Karami et al. provide adaptive strategies for parameter selection but at significant computational overhead [10]. In our study, $minPts$ is set to five percent of the training data set size (e.g. the number of streamlines for a seed).

4.4 Region-based EVF Flow Similarity

In Fig. 4, EVF exhibit regional flow coherency when representative flow lines for the region can themselves be clustered.

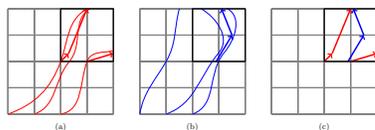


Figure 4: Schematic for observing regional clustering across ensemble members. (a) and (b) represent separate realizations with the upper quadrant (heavy outline) considered. (c) EVF union of members (a) and (b). Arrows are representative flow for the region.

We summarize the possible combinations of coherence over

the ensemble members in Fig. 5. The *lower-left quadrant*: coherent flow in individual members and among members. *Lower-right*: incoherent flow in members but coherent among members. *Upper-left*: coherent flow in members but incoherent among members, and in the *upper-right*, incoherent flow in individual members and among members.

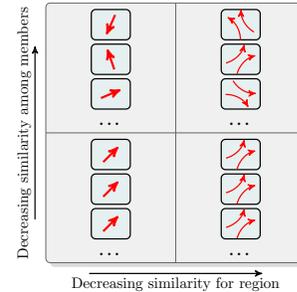


Figure 5: Matrix showing primary combinations of EVF flow similarity. Each box shows hypothetical representative flow (arrows) for a given region in a member of the vector field.

We utilize the following steps to summarize flow in a region from the EVF. After steps 1 through 4 are complete, ensemble flow coherence is visualized in a region using representative streamlines.

- Step 1** Define a spatial region ($\subset \Omega_{EVF}$) for inspection.
- Step 2** Gather precomputed streamline segments spanning the region.
- Step 3** For each member, cluster streamline segments.
- Step 4** Assign a representative streamline per cluster by using the streamline closest to the cluster centroid via Euclidean distance.

5 Experiments

5.1 Implementation

Our results were obtained from code written in Python, utilizing the SciPy package, Sci-kit Learn [9], and HDF5 [2] via H5py [3].

Table 1: Timings for flow maps and FTVA pre-computation for the data sets in this study. Number of members reflects the members used in the computations and not necessarily the total available members. In cases where less members are used than available, those members used were randomly chosen from the available set. Compute times are dependent on number of ensemble members and field resolutions.

	resolution	members	time steps	flow map	FTVA
Lock	128x128	20	1100	30375.94s	206.69s
Ocean	53x90	30	1100	9285.76s	54.09s
Stir	152x152	15	1100	32126.12s	335.73s

The PC system used an Intel Core i7-3930k with 32 GB of RAM. All Python scripts were run as single-threaded processes. Tables 1 and 2 show compute times for algorithms used in this study. Time spend on file I/O is excluded. We omit timings for regional analysis, since compute times vary widely based on dimensions of the selected area.

Table 2: Timings for pre-computation of clustering for terminal points (term.) and multiple streamline samples (3 pts., 13 pts., and variable pts. between 3 and 13) for the data sets in this study. Included is the total calculation time of the linear and angular entropy pre-computations. Compute times are dependent on number of ensemble members and field resolutions. Identical resolution and number of members used for these timings are shown in table 1.

	term.	3 pts.	13 pts.	var. pts.	entropy
Lock	389.23s	22523.08s	23900.47s	23086.69s	20359.93s
Ocean	97.32s	4590.12s	4704.98s	3581.92s	11842.21s
Stir	602.34s	31761.07s	32555.85s	14291.28s	28710.68s

5.2 Data Sets

Lock-exchange The initial conditions are heavy fluid on one side and light fluid on the other, separated by a barrier (the lock) [29]. At initial time, that barrier is removed, and the flow is allowed to evolve. See Fig. 6a. Initial uncertainty originates from not knowing the position of the interface between the two fluids. In other words, the volumes of heavy and light fluid on each side is not exactly known, and the initial barrier slides left and right accordingly. At the start of the simulation, the probability distribution of the position of the barrier is Gaussian. Therefore, after infinite time, it is expected that the barrier is characterized by a similar Gaussian distribution, but with the light fluid on top of the heavy one, and with the variance of distribution stretched if the size of the whole lock domain is not square. However, the probability distributions of the interface or the dominant dynamics in between this start and infinite time are not assumed Gaussian. The lock-exchange data has the following parameters: 128 x 128 grid with velocity measurements, 1000 realizations.

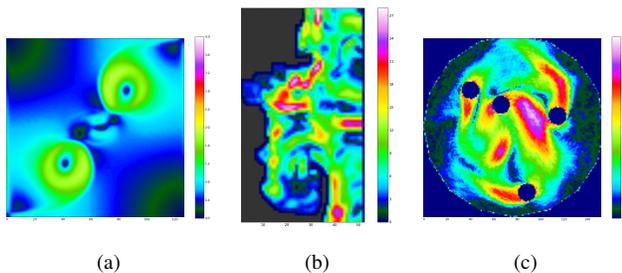


Figure 6: Single member velocity magnitude fields from, (a) Lock-exchange data, (b) Ocean data, and (c) Industrial Stirring data.

Ocean This data set covers a region of the Massachusetts Bay on the east coast of the United States of America [13, 14]. See Fig. 6b. The Massachusetts Bay volume in the study was divided into 53 x 90 grid with 16 depths. The depths at these 53 x 90 grid points vary significantly: depths as shallow as 90 meters and as deep as 196 meters were recorded. The important visualization concern for this data set is understanding where ocean current streamlines seeded at the same location split into distinct paths in different realizations. For example, streamlines may deviate geometrically between their common seed positions and their individual termination position in a set of streamlines from multiple realizations, but still have similarly located terminal positions. See Fig. 7.

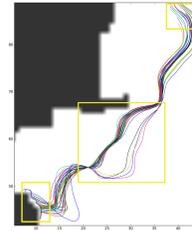


Figure 7: Member streamline bifurcation between members in ocean data set. Seed location is at the red cross marker. Streamlines separate along their trajectories forming two distinct clusters as seen in the central border selected region in yellow. However, the distribution of their terminal positions alone (FTVA) do not account for these separate bundles, especially as seen in the spread of the terminal positions in the upper-right and lower-left of the Fig. (additional yellow boxes).

Industrial Stirring The stirring data set is a set of 15 two-dimensional flow fields resulting from the simulation of mixing in a stirring apparatus [8]. See Fig. 6c. The device consists of two counter-rotating pairs of mixing rods that stir a medium in a cylindrical tank. The ensemble was generated by slightly varying the viscosity of the fluid to investigate mixing quality of the device for a range of different fluids. The primary question for this data set regards the effectiveness of the stirring process. An ensemble visualization is expected to be able to identify regions where the mixing quality is high or low throughout the ensemble.

5.3 Results and Analysis

This study does not use individual member variances (FTLE) in the consideration of FTVA [8], but compares our new visualizations to FTVA only. Using FTLE generalizes the application of FTVA to sensitivity between otherwise identical simulation runs (where variations due to numerical error and other noise-based variation is potentially present). Perhaps a more informative metric on FTVA, and streamline clustering in general, is streamline entropy, as discussed in section 4. Thus, our visualizations refer to both average linear and angular entropy maps, as well as FTVA maps, for interpretation of streamline clustering and sampling frequency for individual streamlines.

Lock-exchange The first data set to be evaluated is the lock-exchange simulation ensemble. The most interesting aspects of the flow occur at the mixing interface between the two fluids. Figure 8a and Fig. 8b, show FTVA applied for forward and backward integration, respectively. Figure 8c and Fig. 8d display the terminal position clusters. Our methods of clustering entire streamlines are shown in Fig. 8f (three sample points for all streamlines) and Fig. 8g (ten additional sample points along each streamline). Similar flow patterns are seen using both methods, although our method captures aspects of both terminal end point distributions with either streamline sampling frequency. As we increase the sampling rate used in Fig. 8f to the one used in Fig. 8g, there are areas where cluster counts increase and are not seen using terminal positions alone. These clusters arise due to variations captured by using more samples and thus detect trends of overall streamline geometry.

When consulting the linear entropy map (Fig. 8d), the pat-

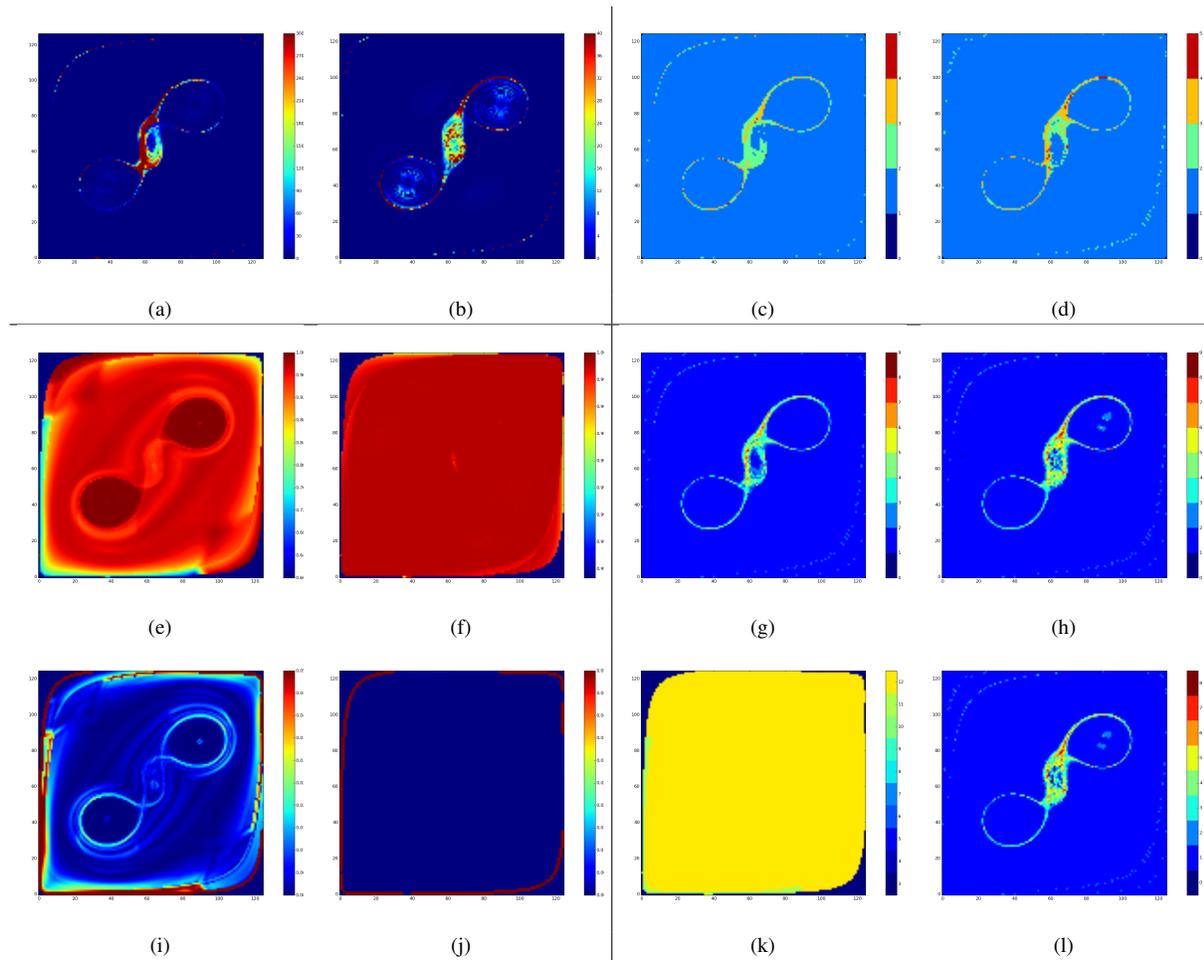


Figure 8: Comparison of transport visual summaries for the lock-exchange data set. Methods from [8] are along first row separated by the horizontal line. The vertical line separates entropy maps on the left and cluster results on the right half of the Fig. (a) FTVA for forward integrated streamlines. (b) FTVA for backward integrated streamlines. (c) Number of trend clusters from terminal positions in forward integration. (d) Number of trend clusters from terminal positions in backward integration. (e) Map of average linear streamline entropies for ensemble. (f) Map of average angular streamline entropies for ensemble. (g) Streamline clusters sampled at three points per streamline. (h) Streamline clusters sampled at ten additional points per streamline. (i) Gradient magnitude for linear entropy map. (j) Gradient magnitude for angular entropy map. (k) Sample map, i.e. the map of the points sampled on each streamline for their corresponding seed location. (l) Cluster map for streamlines sampled variably based on entropy. (Note: color bars for sample and cluster maps contains discrete colors labeled from top to bottom in increasing order.)

tern of the flow field where both variance and distinct flow trends emerge is summarized for the lock-exchange data. Small field differences occurring in angular entropy are seen in Fig. 8e, providing a nearly constant field except near the transitions in entropy in the upper-left and lower-right corners of the domain. This indicates that the number of clusters and variances we see in the flow occur primarily from variation along streamline lengths, i.e. their linear entropy. However, when both linear and angular entropy inform the sampling frequency for streamlines, the overall higher magnitude of angular entropy (in this example) dominates the influence on sampling frequency for streamlines in the interior of the domain. The cluster map using variable sampling between a minimum of three samples and a maximum of thirteen samples (Fig. 8k), shows a result consistent with the uniform sampling in Fig. 8h. (The number samples for streamlines at each seed location is shown in Fig. 8j.) Finally, we see from the magnitude of the gradient of the linear entropy map (Fig. 8h), that a larger number of clusters are found for streamlines with seed locations near the gradient ridges. Where linear entropy changes over the domain, we see streamline geometry variance over the ensemble members (and thus streamline trends).

We now investigate regions from the lock-exchange simulation domain. We show this for two separate regions using the method outlined in section 4.4. We can see regional clustering in Fig. 9, for a region exhibiting incoherent flow patterns within each member of the region. This is similar to the lower-left quadrant of Fig. 5. The flow is simplified using representative streamlines for the region. If we track the streamlines as entering from the bottom of the selected region, some of the representative streamlines flow more from top to bottom that veering to the right or left. Thus, we consider three distinct flow trends from the members of the ensemble for this region.

In contrast, Fig. 10 displays more regional coherence of the type shown in the lower-left quadrant of Fig. 5. The region has a single representative streamline per member. The summary streamlines also show little variation as shown in Fig. 10d. There is coherence both within the region per member, and between members, for a strong overall coherence in the ensemble.

This similarity is different than that shown in Fig. 8. In the full-field analysis, we do not know where in the field trends occur, only that they do for particular seeds. When applying streamline clustering for a region, we show the EVF giving rise to trends seen mapped to seed locations as in Fig. 8. However, this insight is limited to the region itself and the trends produced by flow through the region may be mapped to more than one seed, either within the region itself or outside it. We will next show two more data sets, using our method applied to the entire field as we did for the lock-exchange in Fig. 8.

Ocean Figure 11 is analogous to Fig. 8, but shows results for the ocean data. The primary variance occurs in the central region of the simulation domain for both integration directions. This is somewhat intuitive, since streamlines seeded there have the potential to cover a larger area and thus their terminal positions to differ over greater distances. The trend/clustering analysis for terminal points is shown in Fig. 11c and Fig. 11d, for forward and backward integration respectively.

Our streamline clustering method provides a much higher sensitivity for visualizing trends in the streamlines than conventional FTVA. The number of clusters increase from Fig. 11f to 11g

at the higher sampling frequency. This is due to detecting more variation on the streamlines and seeing a higher resolution of the trends. Figure 11a through Fig. 11d fail to detect most of the flow behavior that occurs near the upper coastal region and the flow trends present there, i.e. flow bundles that separate along the intermediate positions of the streamlines but have similar positions at their terminal positions. See Fig. 7 for an example of this.

In Fig. 11d and 11e, it can be seen that streamlines seeded centrally have higher average streamline entropy (linear and angular). Again, we see more clusters near the ridges in the gradient magnitudes of Fig. 11h and Fig. 11i. Near the center of the domain, the number of the clusters drops to zero in Fig. 11g (the higher streamline sampling frequency). The lack of trends for these seeds is not seen in Fig. 11c and Fig. 11d (and in the lower sampling rate of our method in Fig. 11f). Our method uncovers the highly variable and chaotic flow mapped to this seeding region. This behavior is also shown when adaptive sampling is applied in Fig. 11k.

Industrial Stirring Figure 12 applies the same method to the industrial mixing simulation ensemble. As discussed in [8], the design of the stirring machinery shows needed improvement due to the low variance in much of the domain via FTVA. This is corroborated and repeated here in Fig. 12a and Fig. 12b. The trend analysis from [8] additionally shows much of the domain possessing at least two clusters of terminal particle positions for both the forward and backward integration.

Our method shown in Fig. 12d through Fig. 12k, sharply contrasts parts of the previous analysis from [8]. We find even in regions of high variance, little evidence of good transport. As can be seen in Fig. 12d and Fig. 12e, there are irregular domain regions showing very low average linear and angular streamline entropy. (This is most evident in the ovoid structure to the far-right middle section of the domain.) Interestingly, the region along the lower-left of the cylindrical tank possesses high average entropy, but little to no clusters (see Fig. 12f, Fig. 12g and Fig. 12k). This would appear to contradict the regions with low entropy and also no clusters, except for the fact that we had already observed that regions with a high number of trends generally occur at the ridges of the gradient magnitudes of the entropy maps. We see that this region with high entropy in the lower-left of the domain also exhibits low gradient magnitude (not a region containing a ridge) and thus agrees with the earlier assessment.

There is little difference between the average linear and angular entropy maps for this data set (Fig. 12d and Fig. 12e) and this signature may be useful for classifying such overall behavior. In regions of the flow that both have low average entropy and low levels of cluster count, we would want to improve the overall transport. This analysis may suggest that a potential geometrical or material design might be implemented to prevent lack of agitation at the fluid and paddle points of contact, since this behavior is consistent across the ensemble where fluids of varying parameters of viscosity were used in the simulation.

6 Conclusion

In this paper, we first presented a flow structure based on streamline clustering over their spatial extent. Using the mean linear and angular streamline entropy maps, we showed that where variations in entropy is greatest, there is in general a correspondingly high number of clusters for those streamlines.

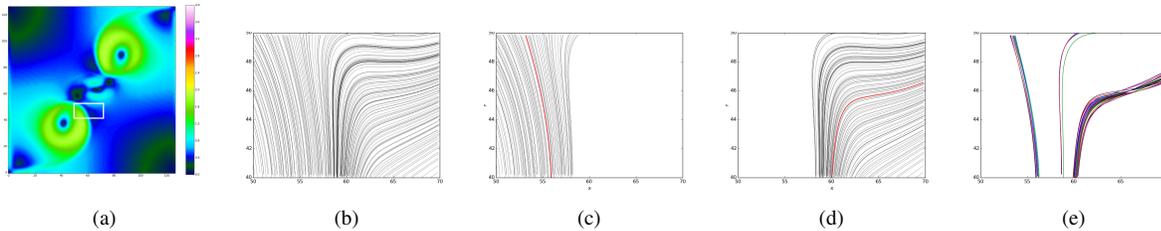


Figure 9: Streamline clusters for an incoherent flow region in lock-exchange data set. (a) Region location (shown by white box selected rectangle) from lock-exchange velocity magnitude field. (b) All streamlines from a single member. (c) First cluster from (a) with representative streamline. (d) Second cluster from (b) with representative streamline. Representative streamlines are highlighted in red. (e) Plot of representative streamlines for 20 members, each a random color.

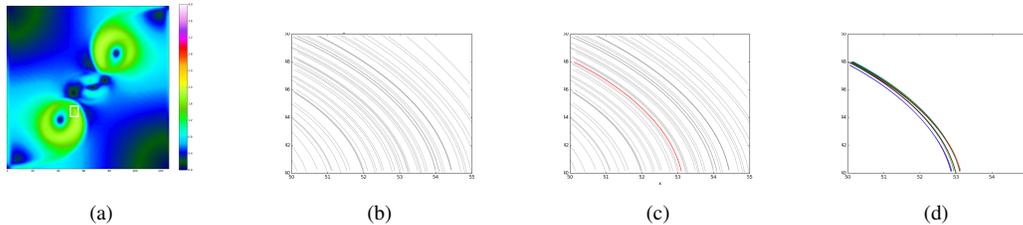


Figure 10: Streamline clusters for a coherent flow region. (a) Region location (shown by white box selected rectangle) from lock-exchange velocity magnitude field. (b) All streamlines from single member. (c) Single cluster with representative from (b). Representative streamlines are highlighted in red. (d) Plot of representative streamlines for 20 members, each a random color.

Preliminary results revealed that related methods of trajectory similarity/clustering did not capture the behavior of spatial bifurcation or flow bundling as we had anticipated. For example, TRACLUS [12] is a direct extension of DBSCAN to line-segment data. TRACLUS tends to cluster trajectories without regard to individual path integrity, and often finds patterns in partitioned segments of the initial streamlines instead.

We followed our analysis of flow structure by investigating flow coherence at regions of bifurcation in a 2D EVF. Finally, we discussed how both methods can be used in a sequential framework for EVF analysis. The methods presented here are not limited to steady-flow. For the purpose of clarity in this initial study, we chose to focus on a single time-step in the simulation.

Future work will employ better adaptive strategies for streamline sampling frequency and incorporate multiple similarity metrics. Additionally, new methods of region analysis over the entire simulation domain may prove useful via algorithmic versus manual inspection.

References

- [1] C.-K. Chen, S. Yan, H. Yu, N. Max, and K.-L. Ma. An illustrative visualization framework for 3d vector fields. In *Computer Graphics Forum*, volume 30, pages 1941–1951. Wiley Online Library, 2011.
- [2] C. M. Chilan, M. Yang, A. Cheng, and L. Arber. Parallel i/o performance study with hdf5, a scientific data package. *TeraGrid 2006: Advancing Scientific Discovery*, 2006.
- [3] A. Collette. *Python and HDF5*. ” O’Reilly Media, Inc.”, 2013.
- [4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [5] S. Furuya and T. Itoh. A streamline selection technique for integrated scalar and vector visualization. *Vis ’08: IEEE Visualization Poster Session*, 2(4), 2008.
- [6] H. Guo, X. Yuan, J. Huang, and X. Zhu. Coupled ensemble flow line advection and analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2733–2742, 2013.
- [7] G. Haller. Distinguished material surfaces and coherent structures in three-dimensional fluid flows. *Physica D: Non-linear Phenomena*, 149(4):248–277, 2001.
- [8] M. Hummel, H. Obermaier, C. Garth, and K. I. Joy. Comparative visual analysis of lagrangian transport in cfd ensembles. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2743–2752, 2013.
- [9] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.
- [10] A. Karami and R. Johansson. Choosing dbscan parameters automatically using differential evolution. *International Journal of Computer Applications*, 91(7):1–11, 2014.
- [11] J. Kasten, I. Hotz, and H.-C. Hege. On the elusive concept of lagrangian coherent structures. In *Topological Methods in Data Analysis and Visualization II*, pages 207–220. Springer, 2012.
- [12] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.
- [13] P. F. Lermusiaux. Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *Journal of Computational Physics*, 217(1):176–199, 2006.
- [14] A. Love, D. L. Kao, and A. Pang. Visualizing spatial multivalued data. *IEEE Computer Graphics and Applications*, 25(3):69–79, May/June 2005.
- [15] K. Lu, A. Chaudhuri, T.-Y. Lee, H.-W. Shen, and P. C. Wong. Exploring vector fields with distribution-based

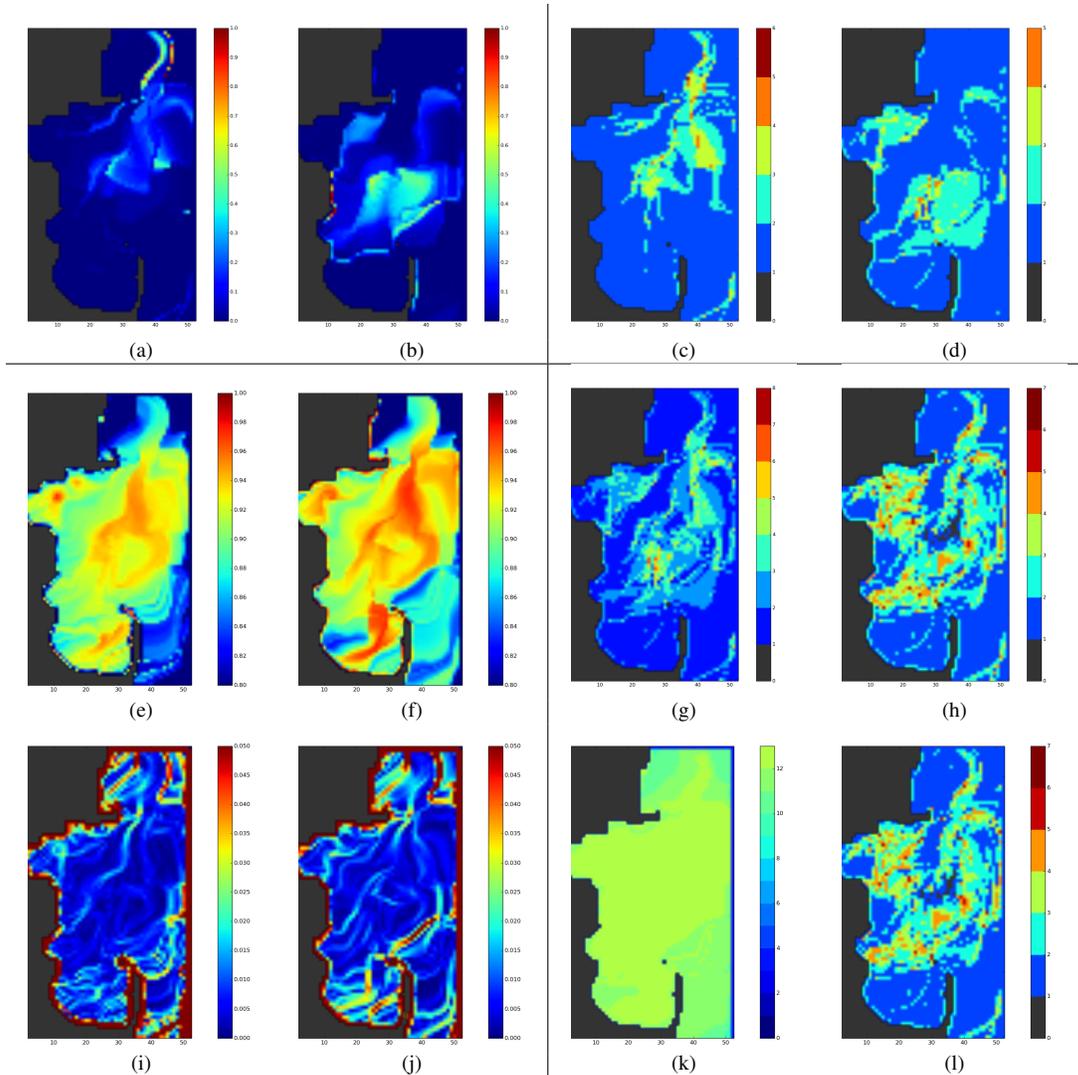


Figure 11: Comparison of transport visual summaries for the Massachusetts Bay data set at surface level. Methods from [8] are along first row separated by the horizontal line. The vertical line separates entropy maps on the left and cluster results on the right half of the Fig. (a) FTVA for forward integrated streamlines. (b) FTVA for backward integrated streamlines. (c) Number of trend clusters from terminal positions in forward integration. (d) Number of trend clusters from terminal positions in backward integration. (e) Map of average linear streamline entropies for ensemble. (f) Map of average angular streamline entropies for ensemble. (g) Streamline clusters sampled at three points per streamline. (h) Streamline clusters sampled at ten additional points per streamline. (i) Gradient magnitude for linear entropy map. (j) Gradient magnitude for angular entropy map. (k) Sample map, i.e. the map of the points sampled on each streamline for their corresponding seed location. (l) Cluster map for streamlines sampled variably based on entropy. (Note: color bars for sample and cluster maps contains discrete colors labeled from top to bottom in increasing order.)

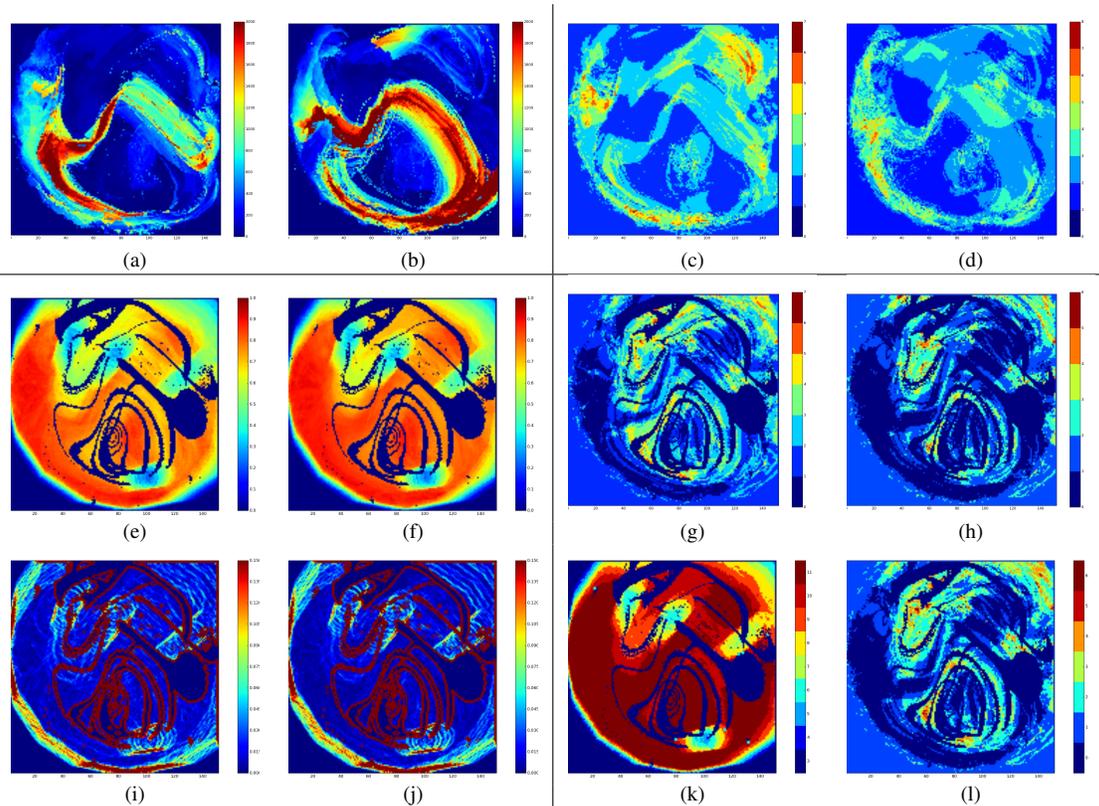


Figure 12: Comparison of transport visual summaries for the industrial stirring data set. Methods from [8] are along first row separated by the horizontal line. The vertical line separates entropy maps on the left and cluster results on the right half of the Fig. (a) FTVA for forward integrated streamlines. (b) FTVA for backward integrated streamlines. (c) Number of trend clusters from terminal positions in forward integration. (d) Number of trend clusters from terminal positions in backward integration. (e) Map of average linear streamline entropies for ensemble. (f) Map of average angular streamline entropies for ensemble. (g) Streamline clusters sampled at three points per streamline. (h) Streamline clusters sampled at ten additional points per streamline. (i) Gradient magnitude for linear entropy map. (j) Gradient magnitude for angular entropy map. (k) Sample map, i.e. the map of the points sampled on each streamline for their corresponding seed location. (l) Cluster map for streamlines sampled variably based on entropy. (Note: color bars for sample and cluster maps contains discrete colors labeled from top to bottom in increasing order. Also notice that all fields shown in this Fig. are slightly truncated in their upper right corner from Fig. 6c. We use the intersection of the simulation region for all members in the ensemble.)

- streamline analysis. In *PacificVis*, pages 257–264, 2013.
- [16] S. Marchesin, C.-K. Chen, C. Ho, and K.-L. Ma. View-dependent streamlines for 3d vector fields. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1578–1586, 2010.
- [17] M. Mirzargar, R. Whitaker, and R. Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. 2014.
- [18] B. Moberts, A. Vilanova, and J. J. van Wijk. Evaluation of fiber clustering methods for diffusion tensor imaging. In *Visualization, 2005. VIS 05. IEEE*, pages 65–72. IEEE, 2005.
- [19] R. Nuzzo. Statistical errors. *Nature*, 506(13):150–152, 2014.
- [20] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *Computer Graphics and Applications, IEEE*, 34(3):8–11, 2014.
- [21] M. Otto, T. Germer, H.-C. Hege, and H. Theisel. Uncertain 2d vector field topology. In *Computer Graphics Forum*, volume 29, pages 347–356. Wiley Online Library, 2010.
- [22] M. Otto, T. Germer, and H. Theisel. Uncertain topology of 3d vector fields. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, pages 67–74. IEEE, 2011.
- [23] T. Peacock and G. Haller. Lagrangian coherent structures: The hidden skeleton of fluid flows. *Physics today*, 66(2):41–47, 2013.
- [24] W. Reich and G. Scheuermann. Analysis of streamline separation at infinity using time-discrete markov chains. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2140–2148, 2012.
- [25] T. Salzbrunn, C. Garth, G. Scheuermann, and J. Meyer. Pathline predicates and unsteady flow structures. *The Visual Computer*, 24(12):1039–1051, 2008.
- [26] T. P. Sapsis and P. F. Lermusiaux. Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D: Nonlinear Phenomena*, 238(23):2347–2360, 2009.
- [27] D. Schneider, J. Fuhrmann, W. Reich, and G. Scheuermann. A variance based fite-like method for unsteady uncertain vector fields. In *Topological Methods in Data Analysis and Visualization II*, pages 255–268. Springer, 2012.
- [28] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [29] M. Uecker mann, P. F. Lermusiaux, and T. Sapsis. Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *Journal of Computational Physics*, 233:272–294, 2013.

7 Author Biography

Brad Hollister received his PhD at the University of California at Santa Cruz in 2015. He is currently a Postdoctoral Fellow at the University of Utah's Scientific Computing and Imaging Institute.

Alex Pang is a Professor of Computer Science at UC Santa Cruz. He received his PhD in Computer Science from UCLA in 1990, and his BS in Industrial Engineering from University of the Philippines with magna cum laude in 1981.