In Situ Statistical Distribution-based Data Summarization and Visual Analysis

Soumya Dutta, Subhashis Hazarika, and Han-Wei Shen

As we move towards the exascale computing era, the necessity of effective, scalable, and flexible data reduction techniques is becoming more and more prominent. This is primarily due to the bottleneck stemming from output data size and I/O speed compared to the ever-increasing computing speed as discussed. Therefore, data summarization techniques are needed that can work in the in situ environment, while the data is getting produced, and preserve the important information from the data compactly which will minimize information loss and enable a variety of post hoc analyses. The motivation for developing novel and effective data reduction techniques is discussed in the introductory chapter in detail. In this chapter, statistical distribution-based in situ data summaries are shown to be a pragmatic solution in this respect and is able to preserve important statistical data features. Using only the in situ generated statistical data summaries, which is significantly smaller in size compared to the original raw data, a wide range of data analysis and visualization tasks can be performed such as feature detection, extraction, tracking, query-driven analysis, etc. Besides these, when necessary, the full-resolution data reconstruction is also possible to visualize the data in its entirety with the added advantage of uncertainty quantification. In this part of the chapter, several distribution-based data modeling algorithms are presented along with their in situ performances and demonstrate the usefulness of the distribution data summaries through several application studies.

Soumya Dutta
Los Alamos National Lab, Los Alamos, NM, USA, e-mail: sdutta@lanl.gov

Subhashis Hazarika
Los Alamos National Lab, Los Alamos, NM, USA, e-mail: shazarika@lanl.gov

Han-Wei Shen
The Ohio State University, Columbus, OH, USA, e-mail: shen.94@osu.edu

1 Statistical Distribution Models for Data Summarization

Probability distributions are well known for capturing various statistical properties of data sets. Furthermore, since the distributions are capable of representing a large set of data samples in a compact format, it has been used successfully for modeling scientific data sets and as a result, different types of distribution-based data summaries are proposed as a means of reduced data representation. Before we go into the details of modeling large-scale simulation data using distributions, let us first briefly discuss several statistical distribution representations that have been used in the data science and visualization community for summarizing large-scale data sets. Distribution-based modeling techniques can be classified into two broad categories: (1) Non-parametric distribution models; and (2) Parametric distribution models. Histogram and Kernel Density Estimators (KDE) are popular non-parametric distribution models used extensively in the visualization community, whereas, parametric distributions such as Gaussian distributions, Gaussian Mixture Models (GMM) are also found to be very effective in data analysis. In the following, we briefly introduce the most popular distribution models that are used in various in situ applications and discuss their advantages and disadvantages in the in situ context.

1.1 Non-parametric Distribution Models

Given a set of discrete data samples $\{x_i\}$, a non-parametric distribution in the form of a histogram can be formally defined as:

$$H(s) = \sum_i \delta(x - x_i) \tag{1}$$

where δ is the Dirac delta function defined as:

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The area under a histogram can be normalized, and such histograms are often used as a discrete probability distribution function. Another well known non-parametric distribution model Kernel Density Estimator (KDE) is defined as:

$$f(x) = \frac{1}{nb} \sum_{i=1}^{n} K(\frac{x - x_i}{b}) \tag{3}$$

where $f(x)$ denotes the probability density at x, n is the number of data samples, b (> 0) is the bandwidth, a smoothing parameter, and $K(\cdot)$ is the non-negative kernel function. A range of kernel functions such as uniform, triangular, Gaussian, Epanechnikov kernels can be used for estimating data density.

1.2 Parametric Distribution Models

Compared to the non-parametric distribution models, parametric distribution models offer a more compact distribution representation, since, only the parameters of the models are sufficient to represent the distribution model. The use of parametric Gaussian distributions for data modeling is widely known across various scientific domains. However, the assumption of the normality of data is not always true and can introduce modeling inaccuracies. Gaussian mixture models (GMM) removes this assumption of normality by modeling the data as a convex combination of several Gaussian distributions. The storage footprint for a GMM consists of the parameters of the Gaussian distributions and their weights. Formally, the probability density $p(X)$ of a Gaussian mixture model for a random variable X can be written as:

$$p(X) = \sum_{i=1}^{K} \omega_i * \mathcal{N}(X|\mu_i, \sigma_i) \tag{4}$$

where K is the number of Gaussian components. ω_i, μ_i and, σ_i are the weight, mean, and standard deviation for the i^{th} Gaussian component respectively. Note that the sum of weights in the mixture, $\sum_{i=1}^{K} \omega_i$ is always equal to 1. The computation of parameters for the GMMs is typically done by Expectation-Maximization (EM), which uses an iterative approach to maximize a likelihood function [5]. For an approximate and computationally efficient estimation of parameters of a GMM, an alternative incremental estimation scheme is also available [32, 35] which can satisfy the need of fast-processing in the in situ environment.

1.3 Advantages and Disadvantages of Different Distribution Models in the Context of In Situ Data Reduction

To use distributions as a viable solution for performing in situ data summarization, several constraints need to be discussed. Any in situ data analysis algorithm is expected to be computationally fast so that it does not stall the underlying scientific simulation, and also the additional memory requirement should be as small as possible. In this context, the computation time for the non-parametric model histogram is low as it only requires a scan of data values and counting the frequencies of discretized data values by converting them into bins. However, the storage requirement of histograms is not always small since the frequency of each bin needs to be stored. Furthermore, if summarization is done for multivariate data, then the storage footprint of high-dimensional histograms increases exponentially. The other widely used non-parametric model KDE is computationally more expensive than histograms. Compared to the non-parametric models, the storage requirement of the parametric distributions is always small since only the model parameters are stored. While computation time for estimating parameters for a Gaussian distribution is

low, estimation of model parameters can be expensive at times for high-dimensional parametric distributions such as Gaussian mixture models. Therefore, instead of using the traditional Expectation-Maximization (EM) algorithm all the time for GMM parameter estimation, a faster and approximate incremental mixture model estimation algorithm has been explored for in situ GMM-based data summarization [12]. Another important point is that using only one type of distribution model for all the data may not be the optimal modeling strategy. For example, based on the statistical properties of the data, different distribution models might be suitable at different regions of the data. Therefore, a hybrid distribution-based data summarization would be possible where based on various statistical tests, the most suitable distribution models will be used for summarization [22, 21]. In the following, we briefly discuss various statistical tests that can be done to pick the most suitable distribution model for data summarization.

1.3.1 Various Statistical Tests for Picking the Suitable Distribution Model

Depending on factors like initial data size, type of post hoc analyses targeted and/or in situ computational complexity involved, both parametric and non-parametric models have distinct advantages and disadvantages. The choice of suitable distribution model, therefore, plays an important role in determining the efficiency of distribution-based in situ data summarization strategies. Many standard statistical tests currently exist to decide which distribution model can best represent the underlying data. However, often a single test may not be enough to address all the concerns and trade-offs associated with a real-world in situ scenario. Therefore, users have to carefully design their tests based on their requirements. Depending on the application and scale of operation, the task of selecting a distribution model can be as simple as graphical validation of the shapes of distributions to as complex as solving an optimization function with desired requirements as the function variables. Here, we put forward some of the most commonly used practices prevalent in the field of Statistics and Visualization.

Normality Test: One of the simplest, yet effective statistical test that can be performed is to check for Gaussian/normal behavior in the data distribution. Studies have shown that for the same sample size, Shapiro-Wilk test [33] is the most powerful (i.e, *statistical power*[1]) normality test [31]. It returns a likelihood value, commonly referred to as *pValue*, which lies between 0 and 1. Small *pValues* lead to the rejection of the normality hypothesis, whereas, a value of 1 ascertains normality with high confidence. A *pValue* in the range of $[0.05, 0.1]$ is often considered as a good threshold value to decide normality. For data not satisfying this threshold further evaluations need to be done to find a more suitable distribution model.

Goodness-of-fit Test: Normality tests do not offer a means to evaluate the best-fitted distribution model for the underlying data out of possible candidate models. Kolmogorov-Smirnov (KS) test [36], a type of goodness-of-fit test, is a more generic

[1] *Statistical power* of any test is defined as the probability that it will reject a false null hypothesis.

platform for such comparative validation. It compares the CDF of distribution against the empirical CDF (ECDF) of data. Goodness-of-fit is decided by how close the CDF of a distribution is to the ECDF. If $F(x)$ represents the CDF of the hypothesized distribution and $F_e(x)$ represents the ECDF, then the KS test measure is given as,

$$K = sup_x |F(x) - F_e(x)| \tag{5}$$

Unlike many other statistical tests, the KS test can evaluate the goodness of both parametric and non-parametric distributions at the same time.

Bayesian Information Criterion: Bayesian Information Criterion (BIC) [16] is a commonly used metric for selecting among a finite set of parametric models. It is based on the log-likelihood of a given model on the sample data. It is defined as,

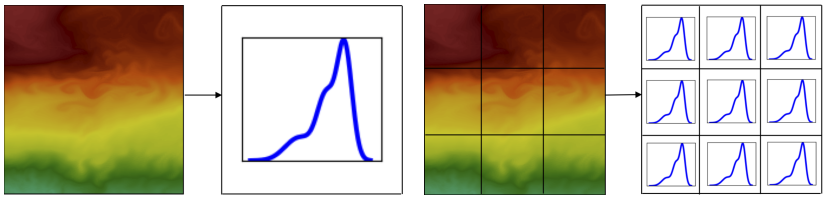
$$BIC = -2L_p + p\log(n) \tag{6}$$

where n is the sample size, L_p is the maximized log-likelihood of the chosen model and p is the number of parameters in the model. A low BIC value indicates a better model. BIC attempts to address the risk of over-fitting by introducing a penalty term $p\log(n)$, which grows with the number of parameters. Therefore, the BIC score is designed to avoid overly complicated models (i.e, with a large number of parameters), which is ideal for distribution-based in situ data summarization approaches. The BIC test is often used for finding out the correct number of Gaussian distributions that would best fit a sample while modeling using a Gaussian mixture model [37, 38].

2 In Situ Distribution-based Data Summarization Techniques

One of the primary advantages of using distribution-based summaries is that the distributions can capture various statistical properties of the data robustly in a compact format. Therefore, in the absence of the full resolution raw data, during the post hoc analysis, a variety of data analysis and visualization tasks can be carried out using such distribution-based data. Furthermore, while the generated results will have uncertainties as the full resolution data is not available, distribution-based data summaries will allow uncertainty quantification for the produced results for conveying uncertainty information to the application scientists.

While modeling scientific data sets using distributions, one can use a global distribution for the whole data domain. For example, a one-dimensional histogram can be used to model a data variable in a data set. In this case, the histogram will be able to answer questions regarding the likelihood of specific values of the scalar field in the data [20, 13, 6], but will not be able to answer questions such as where are those specific values occurring in the domain. This is because the global data distribution is only a coarse statistical summarization of the complete data domain and does not capture any spatial information. Hence, even though significant data reduction using global distribution models can be achieved, still, the applicability



(a) Global distribution-based data model. Data values for all the points is represented by a probability distribution. For each data variable, a separate distribution is created.

(b) Local distribution-based data model. Data values for all the points inside each block is represented by a probability distribution. For each data variable and for each local block, a separate distribution is created.

Fig. 1: Illustration of Local and Global distribution-based data modeling schemes. This image is reprinted from our previous work [11].

and flexibility of such global distribution-based data summaries during the post hoc analysis phase is minimal.

In contrast, to capture the data properties in much finer detail for enabling detailed visual analysis, local region-based distribution modeling techniques have shown great success. In this case, the data domain is first divided into smaller regions/blocks and then a suitable distribution is used to model the data for each region. In this way, even though the storage footprint increases compared to the global model, but, such a local model-based summarization can capture the statistical properties of the data in much more detail compared to the global distributions. In the following, we introduce different schemes of in situ local distribution-based data summarization in detail.

2.1 Local Distribution-based In Situ Data Summarization

As discussed above, the local distribution-based summarization techniques divide the data domain into smaller regions and then use suitable distribution models to reduce the data at each local region. If variables are summarized individually, then univariate distribution models are used. When relationships among multiple variables are required to be captured in the distribution-based data summaries, multiple variables are summarized together using multivariate distribution modeling techniques. Whether univariate modeling is sufficient or multivariate data summaries are needed depends on the specific application tasks. As an in situ data summarization technique, while the univariate distribution-based modeling has its own challenges, multivariate distribution-based modeling techniques are significantly more complex as both the computation cost and storage footprint increases significantly. Therefore, sophisticated distribution modeling schemes are often preferred over standard multivariate distribution-based modeling techniques to address such issues. In the next section, we first discuss various local distribution-based univariate

data summarization techniques and then introduce multivariate data summarization schemes.

2.1.1 Distribution-based Summarization for Univariate Data

Individual data variables can be summarized compactly using univariate distribution-based models such as univariate histograms, Gaussian distributions, GMMs, etc. An important advantage of using the local region-based modeling approach is that it allows modeling of the local statistical properties of the data in detail and therefore, the generated distribution data summaries are flexible and can address a wide range of analysis and visualization tasks in the post hoc exploration phase. Firstly, the data domain is divided into smaller sub-regions (data blocks), and then the desired data variables in each region are summarized using separate univariate distribution models. Finally, all the region-wise distributions are stored into the disk for post hoc analyses.

A straightforward scheme of data domain decomposition used in the literature is regular non-overlapping blocks-wise partitioning. Regular partitioning based data decomposition is computationally less expensive as well as storage efficient. However, since regular partitioning does not consider any data properties, the resulting distribution-based models generated from the data at each partition often show high value variance, and consequently high uncertainty. Furthermore, the distribution summaries only capture the statistical properties of the data values and the spatial organization of such data values inside each block is not preserved in the univariate distribution. Therefore the naive regular partitioning scheme is limited in application in the post hoc analysis phase. To remedy this issue and capture spatial information from the local univariate distribution-based data summaries, two approaches can be taken: (1) By augmenting the spatial distribution information directly to the regular block-wise data summaries; (2) Instead of using regular partitioning, irregular partitioning schemes can be used where spatially contiguous similar data values will be grouped and the distribution-based analysis error will be reduced. Below we briefly present these two approaches.

- **Spatial Distribution-augmented Statistical Data Summarization** An explicit approach of capturing spatial information into the distribution-based data summaries is the direct augmentation of spatial distribution data summaries with the value distribution based data summaries. In this case, the data is still partitioned using regular blocks. Then, for a selected data variable that needs to be summarized, the data values of each partition first summarized using a value-based histogram as shown in Figure 2 (the pink box). Now, to incorporate the spatial information to this value distribution, a spatial Gaussian mixture model (GMM) is estimated for the data points in each histogram bin. For each bin of the value histogram, the data points are identified and then using their spatial locations, a multivariate GMM, termed as spatial GMM, is estimated. Each histogram bin is associated with its unique spatial GMM. While estimating the spatial GMMs, the suitable number of modes for each spatial GMM can be identified by applying

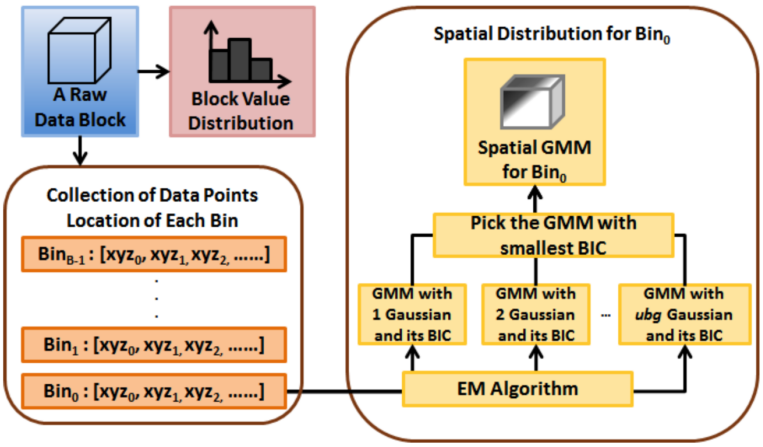
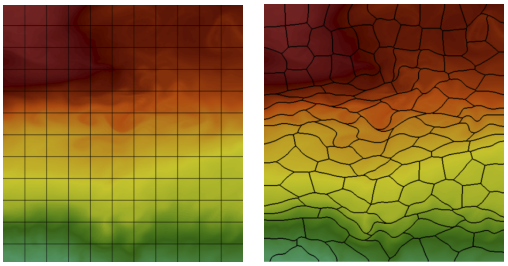


Fig. 2: This diagram shows the steps used to compute the spatial GMM for a raw data block (shown in blue). Besides the computation of the value distribution, the raw data in the block is used to construct the Spatial GMM. First, the locations of the data samples are collected into the corresponding bin interval according to the data value at that location (shown in the bottom left). Then, a Spatial GMM is constructed (shown on the right) for each bin interval using the locations in the interval (illustrated here for Bin_0). This image is reprinted from our previous work [37].

Bayesian Information Criterion (BIC) as has been illustrated in Figure 2. Therefore, for each data block, using this approach, a value histogram and a set of spatial GMMs are stored as the reduced summary data.
Exploration using this spatial GMM augmented distribution-based data summaries is done post hoc. While inferring the data value at a queried location, information from the value histogram and the spatial GMMs are combined using Bayes' rule. Bayes' rule is a popular theorem that is widely used in classification problems. It tells us how to augment the known information with additional evidence from a given condition. In this case, the block value distribution is the known information and the additional evidence are the probabilities from each Spatial GMM at the queried location. More details of this technique can be found in [37].

- **Homogeneity-guided Data Partitioning and Summarization** A second implicit approach for capturing the spatial information in distribution-based modeling and reducing error during post hoc analysis is the use of irregular data-driven partitioning techniques. Naive regular partitioning of data domain does not consider data continuity, and as a result, produces partitions with high data value variance. When distributions are used to reduce such partitions, the resulting distribution models contain high data value variance. Consequently, post hoc analysis using such distribution summaries produces high sampling error leading to increased uncertainty. Therefore, to reduce the data value variation in the partitions, a su-



(a) Regular block-wise partitioning.

(b) SLIC-based partitioning.

Fig. 3: Different types of data partitioning schemes. This image is reprinted from our previous work [11].

pervoxel generation algorithm SLIC (Simple Linear Iterative Clustering) [2] is used. SLIC is a computationally efficient variant of the local K-means clustering algorithm and produces spatially connected data clusters, which are homogeneous in nature. Each SLIC cluster/supervoxel is treated as a local data partition and is summarized using an appropriate distribution model. Since the data are partitioned into near homogeneous regions, using normality tests often it is found that a single Gaussian distribution is sufficient to capture the statistical data properties of each partition. When a single Gaussian is not sufficient, a GMM can be used for summarization. A detailed description of this hybrid distribution-based in situ data summarization scheme can be found in [15].

Compared to the traditional K-means clustering, SLIC adopts a local neighborhood-based approach, where similar data points within a local neighborhood are grouped into one cluster. During the optimization stage, from each cluster center, distances only to the points in the predefined neighborhood are compared. This reduces the total number of distance computations significantly by limiting search in a local neighborhood. As a result, the algorithm performance is boosted significantly. Furthermore, SLIC uses a weighted distance measure that provides contributions from both the spatial locality of the data points and their scalar value similarities. The distance measure can be defined as:

$$D(i,j) = \beta \cdot ||c_i - p_j||_2 + (1-\beta) \cdot |val_i - val_j| \tag{7}$$

Here, c_i is the location of the cluster center i and p_j is the location of point j. val_i and val_j are the data values at i^{th} cluster center and j^{th} data point respectively. The mixing weight β is configured based on the importance of spatial vs value components, such that $0 <= \beta <= 1$, and $\beta + (1-\beta) = 1$. Smaller values of β will give higher weightage on the difference of data values than their spatial locations. Due to these properties, SLIC partitions the data domain into smaller sub-regions where each partition contains points which are: (a) spatially as contiguous as possible; (b) homogeneous in value domain. In Figure 3b, we

show an illustrative example of the SLIC algorithm applied on a 2D data. As can be seen, SLIC partitions similar valued data points along non-axis aligned boundaries compared to the regular partitioning scheme shown in Figure 3a.

2.1.2 Distribution-based Summarization for Multivariate Data

Many times, scientific simulations are designed to measure multiple physical variables/attributes (like pressure, temperature, precipitation, etc.) at the same time. These variables are used to perform various multivariate analyses to gain in-depth insights into the underlying physical phenomenon. Therefore, instead of modeling individual variables as independent univariate distributions, it is often necessary to model them together as multivariate distributions in order to preserve the variable inter-dependence. However, the benefits of distribution-based data summarization are not always readily applicable when using standard multivariate distributions. Unlike their univariate counterparts, it becomes increasingly difficult to work with the corresponding standard multivariate distribution representations when the number of variables (i.e, dimensionality) increases. In this section, we discuss the challenges associated with multivariate distribution-based data summarization and pragmatic solutions to address them.

- **Multivariate Histogram.** Compared to univariate histograms (Section 1.1), computing and storing multivariate histograms is a non-trivial task. The storage footprint of a multivariate histogram can increase exponentially with the number of variables and the desired level of discretization (i.e, number of bins). This makes them ineffective for the purpose of in situ data reduction. Sparse representations of the multivariate histograms can be constructed to bring down the exponential storage cost [29]. Based on the sparseness of the multivariate histogram, the large multi-dimensional array can be transformed into a much smaller size. This transformation, encoded with dictionary-based data structures, can be used to map the transformed multi-dimensional array back to the original array. To further reduced the storage overhead, the multivariate histogram can be stored as a sequence of the index and frequency pairs where the indices are represented as bitstrings computed from a space-filling curve traversal of the multi-dimensional array. However, such sparse representations are sensitive to how the data is distributed and the number of histogram bins used. Therefore, despite cutting down the exponential storage cost of multivariate histogram representations, they are not always effective for data reduction when compared with the original size of the raw data.
- **Multivariate GMM.** As discussed in Section 1.2, because of their compact representation and good modeling accuracy, univariate GMMs are frequently used for distribution-based data summarization. Their multivariate counterparts can also be represented by Equation 4 above, with multivariate Gaussian kernels instead of univariate Gaussians. However, the estimation of multivariate GMM using Expectation-Maximization is computationally expensive compared to the univariate GMMs. The computation time and model complexity increase rapidly

with the number of variables. As a result, the in situ estimation of multivariate GMMs will only add to the overall simulation execution time for large-scale simulations. This can overshadow the advantages of data reduction and I/O bottleneck alleviation for distribution-based data summarization.

- **Copula-based Multivariate Distribution Modeling.** Given the challenges associated with standard multivariate distribution models, it is important to take a fresh look at modeling multivariate distributions for in situ analysis. One such way is to use copula functions [22]. Copula functions offer a statistically robust mechanism to decouple the process of multivariate distribution estimation into two independent tasks: *univariate distribution estimation* and *dependency modeling*. As a result, the exponential cost of storage and/or distribution estimation time can be reduced significantly because we can independently model the individual variables using arbitrary univariate distribution types, while the copula function captures the dependency among them separately.
  A copula function is a multivariate distribution function, whose univariate marginals are standard uniform distributions. In terms of cumulative density functions (CDF), $C : [0,1]^d \rightarrow [0,1]$ represents a d-dimensional copula (i.e., d-dimensional multivariate CDF) with uniform marginals. Sklar's theorem [34] stated that every joint CDF in $\mathbb{R}^d$ implicitly consists of a d-dimensional copula function. If F is the joint CDF and $F_1, F_2, ...F_d$ are the marginal CDF's for a set of d real valued random variables, $X_1, X_2, ...X_d$ respectively, then Sklar's theorem can be formally represented as;

$$\begin{aligned} F(x_1, x_2 ... x_d) &= C(F_1(x_1), F_2(x_2), ...F_d(x_d)) \\ &= C(u_1, u_2, ...u_d) \quad (\text{using } F_i(x_i) = u_i \ \sim U[0,1]) \end{aligned} \tag{8}$$

  where, the joint CDF F is defined as the probability of the random variable X_i taking values less than or equal to x_i. Therefore, to model any multivariate distribution using a copula-based strategy, we need the following two sets of information.

  1. Univariate marginal distributions of the individual variables (i.e, F_i's).
  2. A copula function that captures the dependency among the variables (i.e, C).

  Copula-based multivariate distribution modeling techniques generally approximate the function $C(.)$ using standard copula functions [30]. One such popular copula function is the Gaussian copula, which is derived from a standard multivariate normal distribution. For the purpose of data reduction, the Gaussian copula is ideal because it requires the storage of only the correlation matrix, which can be efficiently computed in an in situ environment. Using this flexible multivariate distribution modeling approach, for each local region, we can store the multivariate data summaries (comprising of univariate distributions and a copula function) to achieve multivariate relationship-aware in situ data reduction. Figure 4 provides a schematic overview of a copula-based in situ multivariate data summarization workflow. The summaries can be utilized to carry out various multivariate post hoc analyses.

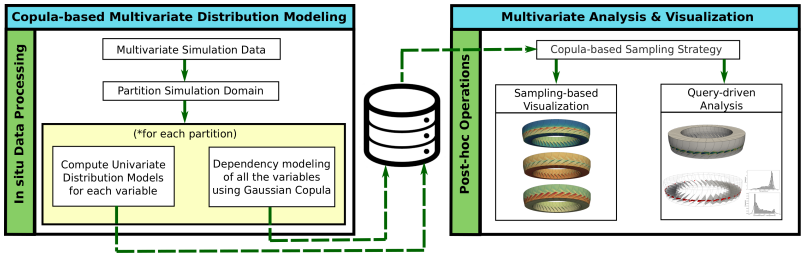
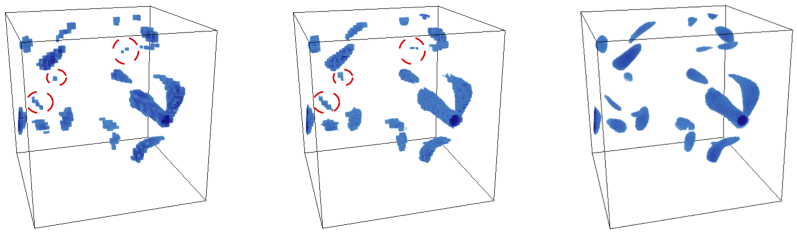


Fig. 4: Overview of a copula-based in situ multivariate data summarization workflow. Multivariate data summaries are created in situ using Gaussian copula functions. The summaries can be utilized later to perform different multivariate post hoc analysis and visualization tasks. This image is reprinted from our previous work [22].



(a) Distribution similarity-based identified feature using regular block partitioning.

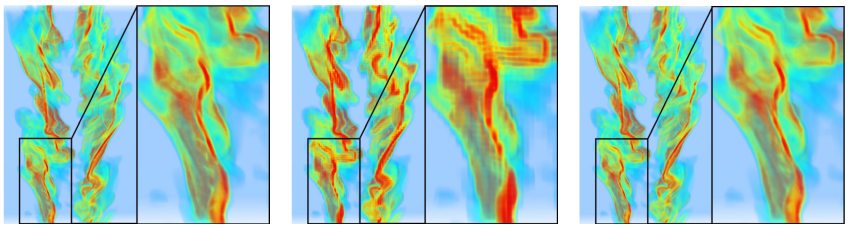
(b) Distribution similarity-based identified feature using K-d tree based partitioning.

(c) Distribution similarity-based identified feature using SLIC-based partitioning.

Fig. 5: Distribution data-driven probabilistic feature search using SLIC-based data summaries in Vortex data set. This image is reprinted from our previous work [15].

3 Post hoc Visual Analyses Using Distribution-based Data Summaries

One of the primary requirements of any in situ data summarization technique is to be flexible during post-hoc analysis so that a variety of visualization and analysis tasks can be performed using it. Since data analysis algorithms are often constrained by storage and computation cost in the in situ environment, a majority of the exploration tasks are still preferred to be done post hoc by the application scientists where they can refine the analysis results interactively, change search criteria as new information is learned, and visualize the data on demand. In this section, we discuss how the various types of aforementioned in situ distribution-based data summaries can be used to enable a wide range of analyses tasks in the post hoc exploration phase.



(a) True similarity field generated using raw data for comparison.

(b) Similarity field generated using Block GMM based data summarization.

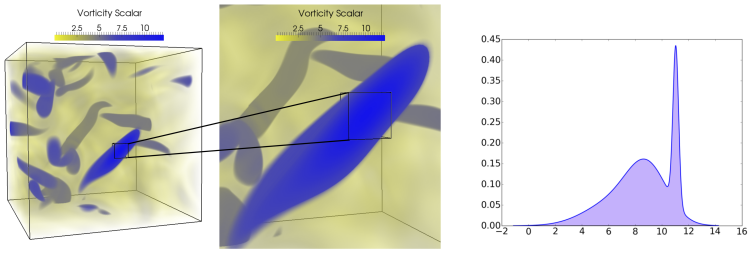
(c) Similarity field generated using spatial GMM augmented data summaries.

Fig. 6: Distribution-based feature similarity field visualization using spatial GMM based data summaries in Turbulent Combustion data set. This image is reprinted from our previous work [37].

3.1 Stochastic Feature Analysis

Analysis and visualization of various scientific features in the simulation data sets is one of the primary tasks that application scientists perform routinely. The distribution-based data summaries can be used to carry out this task robustly. By representing the user specified target features in the form of a distribution, such feature can be searched in the distribution-based data summaries and the features can be extracted and visualized. Feature extraction can be done by matching the target feature distribution to the in situ generated distributions of the local regions and all the regions with a high similarity can be explored interactively. In Figure 5 an example of distribution-based feature extraction is shown. This example uses the homogeneity-guided SLIC-based data partitioning scheme and the data for SLIC partitions is summarized using univariate GMM-based modeling. As can be seen in Figure 5c, the SLIC-based data summaries are able to model the data accurately and hence the extracted features does not have discontinuity and artifacts which are visible from the results generated using the naive regular partitioning scheme (Figure 5a), and also in the K-d tree based partitioning scheme (Figure 5b). More results and a comprehensive quantitative study of this technique can be found in [15].

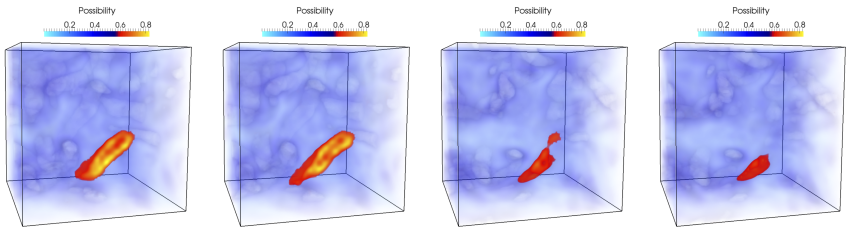
Another example of post hoc feature exploration using the spatial distribution augmented data summaries is shown in Figure 6. Given a target feature, a new feature similarity field is generated where the high valued regions are highlighted as the regions of interest. The true similarity field is shown in Figure 6a, which is generated using the ground truth raw data for comparison purposes. Figure 6b shows the similarity field generated using regular block-wise partitioning and GMMs are used as the distribution model. Finally, Figure 6c depicts the feature similarity field resulted from spatial GMM augmented data summaries. As can be seen that the spatial GMM based data summaries produce the most accurate feature similarity field with minimal artifact [37].



(a) Feature selection in Vortex data set with a zoomed in view.

(b) Estimated GMM of the user interested feature.

Fig. 7: Selected feature in Vortex data set, a zoomed in view and the GMM of the selected region. This image is reprinted from our previous work [14].



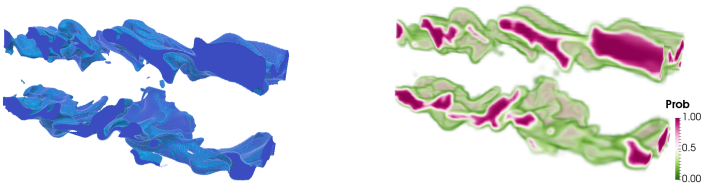
(a) T=3. (b) T=6. (c) T=14. (d) T=15.

Fig. 8: Extraction and tracking using Vortex data set. Tracked feature for 4 selected time steps are displayed. This image is reprinted from our previous work [14].

3.2 Feature Extraction and Tracking

Feature tracking in scientific data sets is an important task. Application scientists are often interested in extracting and tracking the temporal evolution of scientific features (a region of interest) such as vortex cores, hurricane eye, eddies in ocean etc., to learn about the temporal development of various physical phenomenon in detail. The proposed distribution-based data summaries can be used to track such scientific features robustly over time. In this study, a regular block-wise distribution modeling is used where parametric distribution Gaussian mixture model is used to model the data for each local block. Since features in scientific simulations are often hard to be defined by a precise descriptor, a value-based distribution is used to represent the target feature. Finally, using stochastic similarity measures and extracted motion information from the distribution-based data summaries, the feature is extracted and tracked over time robustly. More details of this distribution-driven feature tracking algorithm be found in [14].

In Figure 7, target feature selection in the form of a GMM is displayed where user can highlight a region of interest using an interactive box filter. The feature shown



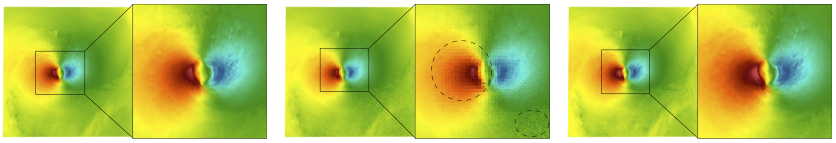
(a) Ground truth (b) Probability field

Fig. 9: Multivariate query-driven analysis of Combustion data set for the query $0.3 < mixfrac < 0.7$ and $y\_oh > 0.0006$. This image is reprinted from our previous work [22].

here is a vortex core in a pseudo-spectral simulation of coherence vortex structures. The tracking results of this feature is provided in Figure 8 where the tracked vortex feature is shown for four different time steps. Note that, even though the shape of the feature changes over time, still the tracking algorithm is able to extract and track the feature robustly in future time steps.

3.3 Multivariate Query-driven Analysis and Visualization

Query-driven analysis techniques are highly effective for analyzing and visualizing large scale data. By selecting a subset of the data domain that meets a user-defined criteria, analysis activities can be focused only on the selected region instead of considering the entire domain. This makes the work-flow of scientists more manageable and effective. These type of selective analyses are particularly common with multivariate data to trim down the variable subspace. Many query-driven strategies rely on computing local data statistics to execute the queries efficiently. Therefore, the use of statistical distributions as local data summaries inherently facilitates such query-driven strategies. With distributions as the underlying data representation, we can report queried region as a probability field. A high probability value at a region signifies that the query of interest has high likelihood at that region. Figure 9 shows the query results on the Combustion data set for the bi-variate query $0.3 < mixfrac < 0.7$ and $y\_oh > 0.0006$. Figure 9a shows the ground truth deterministic region of the original raw data, while Figure 9b shows the corresponding probability field satisfying the given query, i.e, $P(0.3 < mixfrac < 0.7 \text{ AND } y\_oh > 0.0006)$.



(a) Raw data (Ground truth). A zoomed view is shown on the right for better visual comparison.

(b) Reconstruction using regular block partitioning scheme. A zoomed view is shown on the right.

(c) Reconstruction using proposed SLIC-based scheme. A zoomed view is shown on the right.

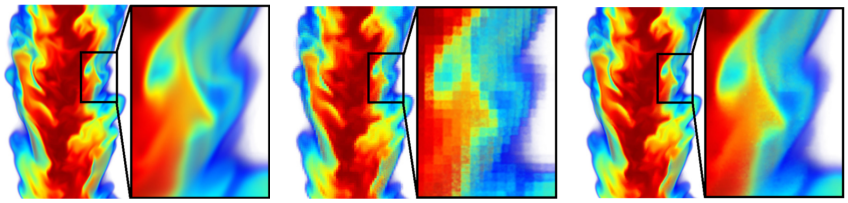
Fig. 10: Visual comparison of U-velocity of Hurricane Isabel data. The reconstructed fields are generated using Monte Carlo sampling of distribution-based summarized data. This image is reprinted from our previous work [15].

3.4 Distribution Sampling-based Data Reconstruction

Often the application scientists want to visualize the data in its entirety to explore certain data features in detail. To enable visualization the full resolution data, the distribution-based data summaries can be used to reconstruct the full resolution data. To reconstruct the data, statistical sampling techniques [17] are used to sample data values from the distribution-based data summaries for reconstructing the data set. In the following, we present reconstruction results created from various types of distribution-based data summaries for both univariate and multivariate data.

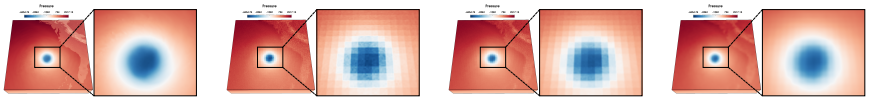
Figure 10 shows the reconstruction result for the U-velocity field of Hurricane Isabel data set. In this example, GMM-based data summaries were generated from the in situ SLIC-based partitioning scheme [15]. For comparison, in Figure 10b, we have shown the reconstruction result when regular block-wise partitioning scheme is used. As can be seen, the reconstructed result produced from the SLIC-based data summaries (Figure 10c) match closest to the ground truth shown in Figure 10a. The result of regular block-wise partitioning contains artifacts and discontinuities (as highlighted by black dotted regions), which are corrected in reconstruction obtained from SLIC-partitioning based data summaries.

Visualization of the reconstructed full resolution data using the spatial distribution-augmented data summaries [37] has been presented in Figure 11. In this example, mixture fraction field of turbulent combustion data is used. The rendering of the ground truth data is depicted in Figure 11a. For demonstrating the efficacy of the spatial distribution-based data summaries, in Figure 11b, we have provided the reconstruction result generated from regular block-wise GMM-driven data summaries, which do not use any spatial distribution information. Finally, Figure 11c shows the result produced from spatial distribution-augmented data summaries, which obtains a smooth reconstruction of the data. It is evident that the augmentation of the spatial distribution information makes the reconstruction more accurate and removes the block boundary irregularities, which are visible in the reconstruction result created from block-wise GMM-guided data summaries (Figure 11b).

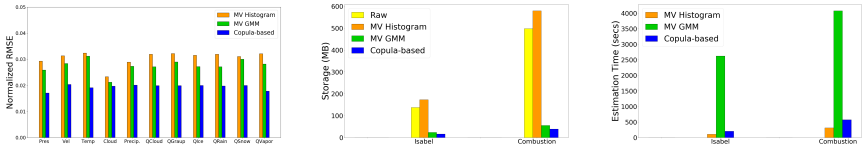


(a) Rendering from raw data. (b) Reconstructed data rendering using block-wise GMM based summary data. (c) Reconstructed data rendering using spatial GMM based summary data.

Fig. 11: Visual comparison of volume rendering in combustion data set. The samples are drawn from the PDFs, which are calculated at all grid points of the raw data, using Monte Carlo sampling. This image is reprinted from our previous work [37].



(a) Original raw pressure field (b) Pressure field reconstructed from multivariate histogram. (c) Pressure field reconstructed from multivariate GMM. (d) Pressure field reconstructed from copula-based model.



(e) Reconstruction errors for all 11 variables in Isabel data set. (f) Storage footprint of different multivariate models (g) Estimation times of different multivariate models

Fig. 12: Qualitative and quantitative results for multivariate sampling-based scalar field reconstruction of Hurricane Isabel data set. This image is reprinted from our previous work [22].

For multivariate data, it is important to reconstruct the scalar fields of the different variables at the same time. This can be achieved only when the variable relationships are factored in during distribution modeling. Figure 12 shows the multivariate reconstruction results for the Hurricane Isabel data set with 11 physical variables, using multivariate histograms (sparse), multivariate GMM, and copula-based multivariate modeling strategy. Figures 12a-d shows the qualitative reconstruction results of only the Pressure variable. Whereas, Figure 12e shows the quantitative results of normalized root mean squared error (RMSE) for all the 11 variables for the three different multivariate distribution modeling approaches. As can be seen, the flexible copula-based multivariate distribution modeling approach performs better than standard multivariate distribution model. The storage overhead and estimation times for the 3 different multivariate models are reported in Figure 12f and 12g

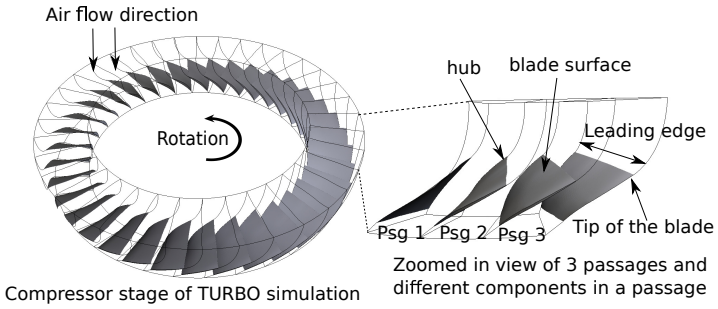


Fig. 13: A diagram of the compressor stage of TURBO simulation and a zoomed in view of it on the right. Different components of a blade is shown. This image is reprinted from our previous work [12].

respectively, for both the Isabel as well as the Combustion data sets. The results highlight the fact that multivariate histograms have higher storage footprint, while, multivariate GMMs have large estimation time cost associated as compared to the flexible copula-based method.

4 Demonstration of An In Situ Distribution-guided End-to-End Application Study

In this section, we describe an end-to-end real-life example of an application study using in situ generated distribution-based data summaries for solving a practical domain-specific problem. In this application study, we explore rotating stall phenomenon in a transonic jet engine simulation data sets. The data is generated from a large-scale computational fluid dynamics (CFD) simulation code, TURBO [10, 9]. TURBO is a Navier-Stokes based, time-accurate simulation code, which was developed at NASA. TURBO simulation models the flow of air through a rotor in the engine turbine compressor stage. The model of the rotor of the compressor stage is shown in Figure 13. The rotor consists of 36 blades and so there are 36 blade passages. A zoomed-in view of the rotor is shown on the right where the tip, the hub, and the leading edge of the blade is highlighted. It has been shown previously that the data generated from TURBO can capture the stall phenomenon in great detail. However, the volume of data generated from TURBO is very large, and therefore, in situ data summarization is critical for enabling timely exploration of the simulation data with high temporal fidelity at an interactive rate.

One of the primary goals of this study was to develop techniques that can detect the rotating stall as early as possible such that the experts can employ stall preventing measures. Rotating stall, if fully developed, can potentially damage the turbine compressor blades. Therefore, early detection of the stall is critical. Furthermore, the reasons behind the inception of rotating stall in transonic engines are still not

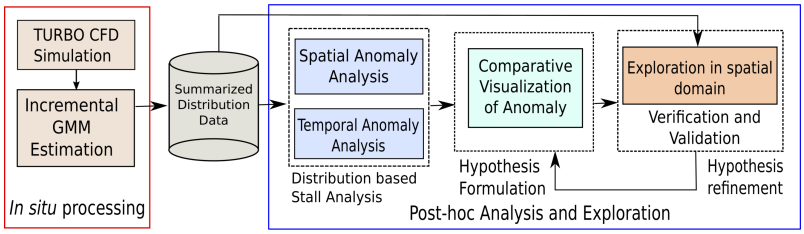


Fig. 14: A schematic of the in situ distribution anomaly-guided stall analysis. This image is reprinted from our previous work [12].

fully understood and hence is an open research problem. Besides identifying the precursors of the rotating stall, the experts also want to understand the role of different variables during the inception of the stall. In the following discussion, we present two application studies for analyzing and visualizing the rotating stall phenomenon using both univariate and multivariate in situ distribution-based data summaries and demonstrate their effectiveness.

4.1 Univariate Distribution Anomaly-guided Stall Analysis

Since the rotating stall is referred to as an instability in the flow data, it can be characterized as an abnormality in the simulation data. In an ideal condition, the simulation is expected to be axisymmetric, and hence, variables such as Pressure, Entropy are expected to have identical values around the compressor stage. Any region where Pressure or Entropy values deviate from its expected behavior can be identified as abnormalities and therefore a region containing potential stall. To capture such regions and interactively analyze rotating stall post hoc, the large-scale simulation data was first summarized in situ using block-wise GMM-based data summaries. To compute the GMM-models efficiently, in this work, instead of using the traditional Expectation-Maximization (EM) algorithm, an approximate incremental mixture model estimation technique was used. More details of this incremental modeling can be found in [12]. The summarization was performed for Pressure and Entropy variables and the distribution-based summaries were stored into disks. Then using the reduced GMM-based distribution data, post hoc stall analysis was carried out. In Figure 14, a schematic of the complete end-to-end analysis workflow is presented. As we can see, the data was summarized in situ and then in the post hoc analysis phase, the data summaries were used to detect regions that showed spatial and temporal distribution anomalies. Through interactive visualization, the domain experts verified their hypothesis and explored the stall features efficiently.

The GMM-based data summaries were first used to estimate the spatial and temporal region-wise anomalies in the data set. To detect such regions, block-wise GMM-based distribution-based summaries were compared over space and time for

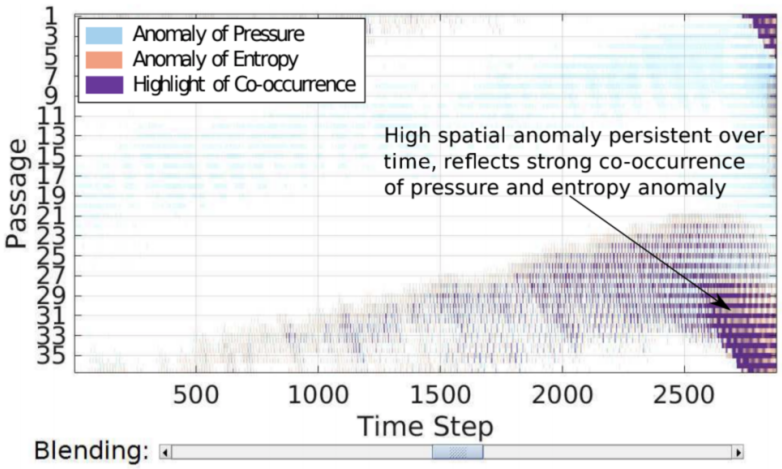
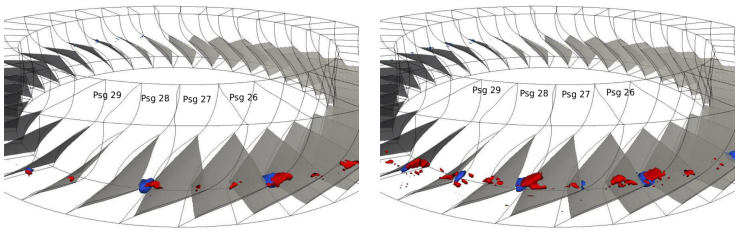


Fig. 15: In situ generated distribution-based spatial anomaly pattern study. The image shows spatial anomaly of Pressure and Entropy variables where the stalled regions are highlighted in blended purple color. This image is reprinted from our previous work [12].



(a) Spatial anomalies at time step 2200. (b) Spatial anomalies at time step 2540.

Fig. 16: Visualization of detected spatial anomaly regions of Pressure (in blue surfaces) and Entropy (in red surfaces). The regions are detected near the blade tip regions of several rotor passages. These regions act as blockage to the regular airflow and create flow instability which eventually leads to stall. This image is reprinted from our previous work [12].

each blade passage. Finally, the detected regions that indicated spatial anomaly were plotted as shown in Figure 15. It can be seen that the abnormalities develop gradually over time and when the stall happens, such abnormal regions become pronounced (indicated by the dark purple region in the plot). Visualization of such detected spatially anomalous regions in the data domain is shown in Figure 16. As can be seen, the detected regions are near the tip of the blades as expected and the anomalies are observed for both Pressure (the blue-colored regions) and Entropy variable (the red-colored regions). The compressor blade passages containing such abnormalities

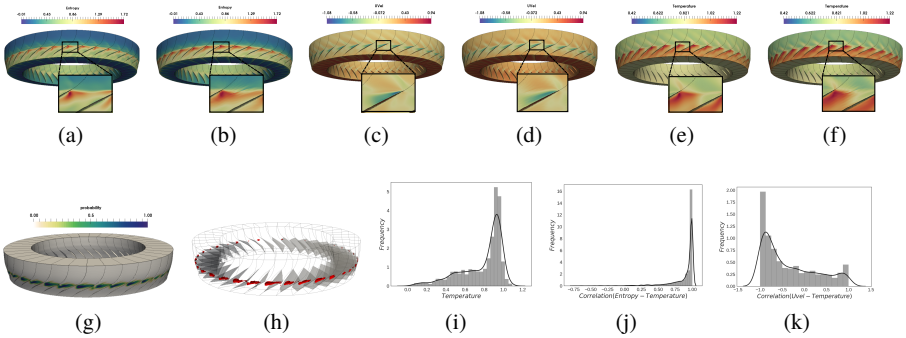


Fig. 17: Post hoc analysis results of the jet turbine data set. (a) Raw Entropy field. (b) Sampled scalar field of Entropy. (c) Raw U-velocity field. (d) Sampled scalar field of U-velocity. (e) Raw Temperature field. (f) Sampled scalar field of Temperature. (g) Probabilistic multivariate query result for $P(Entropy > 0.8 \text{ AND } Uvel < -0.05)$ (h) Isosurface for probability value of 0.5. (i) Distribution of Temperature values in the queried region i.e., $P(Temp|Entropy > 0.8 \text{ AND } Uvel < -0.05)$. (j) Distribution of correlation coefficients between Entropy and Temperature for the queried region. (k) Distribution of correlation coefficients between U-velocity and Temperature for the queried region. This image is reprinted from our previous work [22].

are identified as stalled regions. A similar analysis was also done using the temporal anomaly plots. Using both spatial and temporal anomaly-based analysis, the domain expert was able to confirm the effectiveness of distribution-based data summaries in detecting rotating stall. For a detailed discussion on this topic, interested readers are referred to [12].

4.2 Multivariate Distribution Query-driven Stall Exploration

Scientists were also interested in understanding the importance of the variables Entropy, U-velocity, and Temperature towards the formation of stall-like features in the turbine passages. This requires that the distribution-based data summaries capture the multivariate relationship among the variables for post hoc analyses. Copula-based multivariate distribution modeling strategy, as discussed in Section 2.1.2, was employed to create multivariate data summaries for local partitions. To model the individual variables, a Gaussian distribution was used for partitions with a high normality test score, and a GMM (with 3 modes) was used otherwise. To retain the spatial context of data within a partition, spatial variables (x, y, and z dimensions) were also modeled using uniform distributions. The dependency among all these variables (i.e., 6, 3 physical + 3 spatial) was modeled using Gaussian copula functions.

Table 1: Percentage timing of in situ GMM-based univariate summarization with half and full annulus runs. This table is reused from our previous work [12].

Configuration	2 revs.		4 revs.	
	Simulation	In situ	Simulation	In situ
Half annl. (164 cores)	97.3%	2.7%	97.5%	2.5%
Full annl. (328 cores)	97.63%	2.37%	97.42%	2.58%

The in situ multivariate data summaries were later used to perform various multivariate post hoc analyses and visualizations. Figure 17a,c,e show the original scalar fields for Entropy, U-velocity, and Temperature respectively. The corresponding sampled scalar fields, reconstructed from the data summaries are shown in Figure 17b,d,f respectively. Scientists knew that Entropy values greater than 0.8 and negative U-velocities correspond to potentially unstable flow structures, which can lead to stalls. To focus the study on regions with such multivariate properties, a multivariate query $Entropy > 0.8$ and $Uvel < -0.05$ was made to select the region. The corresponding probability field is shown in Figure 17g, whereas, Figure 17h shows the isosurfaces of probability value 0.5 across the blade structures. Figure 17i shows the distribution of Temperature values in this queried region (i.e., $P(Temp|Entropy > 0.8 \text{ AND } Uvel < -0.05)$). The peak in the distribution suggests that Temperature values around 0.9 can be related to potential engine stall. Figure 17j and k show how Temperature is correlated with Entropy and U-velocity respectively, in the selected region. There is a strong positive correlation with Entropy and a significant negative correlation with U-velocity. Such exploratory analysis activity can help scientists to gain more insights into the multivariate relationships in their simulation. For more detailed discussions, the interested readers are referred to [22].

From the above analyses, we can observe that the various distribution-based techniques led to a detailed understanding of the rotating stall inception problem and also how different variables can be used to detect stall quickly before it becomes destructive. In the future, the early detection capabilities developed can be used to implement some stall preventing measures. One potential measure is to install sensors at the appropriate places that will be measuring abnormalities using the proposed techniques for early detection of the event in the turbine stage so that when abnormalities are detected, these sensors would recommend the users to act and prevent engine destruction. Also, the knowledge learned from these analyses could lead to a better turbine stage design that will make the engine safer.

4.3 Storage and Performance Evaluation

The performance studies presented here with TURBO simulation were done using a cluster, Oakley [7, 8], at the Ohio Supercomputer Center. Oakley contains 694 nodes

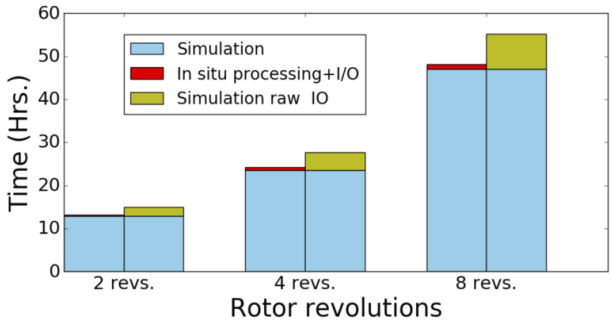


Fig. 18: In situ timing comparison for univariate data modeling using TURBO simulation with and without raw output. Pressure and Entropy variables were summarized using regular block-wise partitioned data and GMMs were used as the distribution model. With the *in situ* pathway, the raw I/O time can be saved. This image is reprinted from our previous work [12].

with Intel Xeon x5650 CPUs (12cores per node) and 48 GB of memory per node. A parallel high-performance and shared disk space Lustre was used for I/O.

4.3.1 Performance Evaluation of Univariate Data Modeling:

One single revolution of the complete rotor, i.e. the full annulus model, in TURBO simulation generates 5.04 TB raw data. To perform the studies presented above, 8 revolutions were run which generated a total of 40.32 TB data and a total of 28800 time steps were run. Since the in situ call was made at every 10^{th} time step, it required processing of 4.032 TBs of data. However, for this experimentation, only the rotor was considered and the data for the two stators were not stored. The raw data for the rotor part in PLOT3d format is 690 MB per time step, and hence, storing all the raw data for 8 revolutions at every 10^{th} time step would require 993.6 GB storage. In this study, for summarizing the data using local distribution-based models, the spatial domain was partitioned using non-overlapping regular blocks of size 5^3. The output of the in situ summarized data for two variables in VTK format took only 51.8 GB for 8 revolutions resulting in approximately 95% data reduction.

In Figure 18, timing performance of the in situ processing is presented. It can be seen that the in situ summarization time is significantly smaller compared to the simulation time. Furthermore, the raw data I/O shown here can be completely removed if the in situ pathway is taken. From Table 1, it is also observed that in situ distribution-based summarization for two variables only take about 2.5% additional time. While estimating the in situ processing overhead, both the half-annulus model (consists of 18 passages) and the full-annulus model (the complete rotor with 36 passages) were tested. It is observed that the in situ data summarization time is consistent for both these runs. Hence, by performing in situ processing, we have

demonstrated a scalable rotating stall analysis to help the expert achieve a better understanding of the phenomenon. In the following, we present performance results for the multivariate data summarization case.

4.3.2 Performance Evaluation of Multivariate Data Modeling:

For the copula-based in situ multivariate data summarization case study, the simulation domain was partitioned regularly into non-overlapping partitions of size 5^3. Two full revolutions of the turbine data comprised of 7200 time steps. In situ data summaries for local partitions were created every 10^{th} time step, thereby storing 720 time steps. Since the size of the raw data produced at each time step was 690 MB, therefore, two full revolutions of the simulation accumulated 496.8 GB of data. Compared to this size, storing multivariate data summaries took only 19.6 GB which resulted in 96% data reduction. Moreover, writing the raw data to the storage disk took around 13% of the simulation execution time, whereas, the combined time to estimate the multivariate summaries and writing them out to disk took 15.4% of the overall simulation time. Besides reducing the storage footprint, the data summaries offer significantly faster post hoc analysis time. Performing multivariate queries on a regular workstation machine took on average 64.6 seconds, whereas, reconstruction of the full scalar fields from the data summaries took only 178.3 seconds on average.

5 Discussion and Guidelines for Practitioners

5.1 Discussion

The above sections demonstrate the efficacy and usefulness of various in situ distribution-based data summarization techniques for performing flexible and explorable data analysis and visualization. It has been shown that when the data is reduced in the form of distributions then features in such summary data can be searched efficiently by defining the feature as a distribution. This is primarily beneficial for scientific features which are hard to be defined precisely due to the complexity of the feature [14]. In such cases, a statistical distribution-based feature descriptor is found to be effective. Another advantage is that the distribution-based data representations can be directly used in these cases and a full reconstruction is not necessary for finding or tracking features. This also helps in accelerating the post hoc analysis. However, a reconstruction of the full resolution data is also possible from the distribution-based data summaries which can be used to explore scientific features using traditional techniques when precise feature descriptors are available.

Another observation for the distribution-based data analysis techniques is that typically the distribution estimation is done in situ and feature analysis and visualization are conducted in the post hoc analysis phase. This strategy is adopted by keeping in mind that a majority of the visual analysis tasks require interaction with

the data, forming new hypotheses, and then refining and verifying those hypotheses as results are studied. These process engages application scientists in the exploration loop and often can take a considerable amount of time. Therefore, these kinds of exploratory analyses are not suitable for an in situ environment when the simulation is running because doing so could slow down the simulation significantly which is undesirable. However, we acknowledge that if the application scientists know precisely about the data features that they are interested in, then extraction and visualization of such features directly in the in situ environment might be a viable solution. In such cases, visualization artifacts such as images of the features can be stored for post hoc analysis.

It is to be noted that, besides applying traditional statistical methods for estimating distributions as discussed above, there is a recent surge in the use of deep learning-based models to estimate data distribution in the field of machine learning. Two such prospective methods are the Generative Adversarial Networks (GANs) [18] and Variational Autoencoders (VAEs) [26]. Such deep learning-based models adopt unique optimization strategies to model very high-dimensional distributions of a wide range of objects. They convert a purely statistical problem of distribution estimation into an optimization problem (i.e, find the parameter values that minimize some objective function). However, to model the distribution perfectly, deep learning methods need multiple iterations over the data, which can be infeasible in situ solution. Recent efforts into the application of such methods in the field of scientific visualization [4, 19, 25] have been to mostly perform post hoc analyses. Bringing in the advantages of such powerful models to an in situ environment is an exciting research prospect in the near future.

5.2 Guidelines for Practitioners

In this section, we briefly provide some guidelines for the users and practitioners about how the appropriate distribution models can be selected and how some of these techniques can be implemented in the simulation. Given a particular task, the first choice is to decide whether univariate distribution models are sufficient or multivariate models will be needed. If multivariate models are necessary, then we recommend using the statistical Copula based approach. This technique is suitable when several variables are needed to be summarized and can tackle the curse of dimensionality problem that arises often while estimating high-dimensional distributions. If univariate distribution-based models are sufficient, then we found that GMM-based summarization performed best. In Section 1.3, we have provided several statistical tests that can be used to select the appropriate number of Gaussian models when estimating a GMM. Note that, using the traditional Expectation-Maximization algorithm for estimating GMM parameters can be costly at times, and hence, if performance is critical, an alternate incremental Gaussian mixture model estimation strategy is suggested in [12]. The use of an incremental algorithm will trade some estimation accuracy but will result in faster parameter estimation.

A majority of the above techniques advocate for the local region-based distribution models. In such modeling, since the data blocks are non-overlapping, the distribution estimation for each data block is independent and hence no additional data communication is required. So, it is straightforward to compute such models. First, the data in each processing node needs a partitioning and then an appropriate distribution model can be used. Users and practitioners are encouraged to consider using EDDA [1], the open-source distribution-based analysis library, which came out of the research done at the Ohio State University and implements building blocks of several of the distribution estimation techniques that have been discussed in this chapter. The library is under development and so some of the advanced techniques might not be readily available. However, we believe this library will be a useful starting point for the practitioners who are interested in using distributions in their analyses.

If the users are interested in conducting the feature analysis in the in situ environment, then additional data communication among computing nodes will be needed. Since a data feature could span across multiple computing nodes, a strategy needs to be developed which will send data distributions to the neighboring processing nodes so that the complete feature can be extracted and analyzed. Sending data distributions to the neighboring blocks is expected to be a cheap operation since the size of distribution parameters is significantly smaller compared to the raw data. Note that this will require new research to come up with a desired and scalable solution. However, we believe that with the present advances made in the distribution-based analysis domain, as discussed throughout this chapter, the strategy of estimating distributions in situ and performing feature analysis post hoc have resulted in promising results and a variety of complex and important visual-analysis tasks were satisfactorily performed.

6 Additional Research Possibilities and Future Scopes

Above sections present various in situ distribution-based data modeling techniques for both univariate and multivariate scalar data sets. The applicability of such distribution-based data summaries for solving various domain specific problems is also demonstrated in Section **??**. In order to study the usefulness of the above distribution-based data summaries, in the context of a broader set of visualization tasks, we plan to conduct a comprehensive evaluation where comparison among various data reduction techniques such as distribution-based summaries, data compression techniques, and sampling-based reduction approaches will be considered. Besides analyzing scalar data sets, distribution-based data summaries can also be used for analyzing and visualizing vector field data sets, ensemble data sets, and also particle-based data. Several studies have already been done for summarizing vector fields using distribution-based representations [23, 27]. To generate streamlines from such distribution-based vector data summaries, He et al. adopted a Bayesian framework using particle filtering technique [23]. In another work, to trace the parti-

cles accurately using distribution-based vector fields, Li et al. used winding angle of particles trajectories for correctly predicting the particle path using a Bayesian approach [27]. A recent work demonstrated usefulness of distribution-based techniques for in situ particle data reduction [28].

Among other future possibilities, applications of distribution-based data summaries have also been tested for summarizing and analyzing large ensemble data sets. In one approach, Wang et al. [38] first captured the relationship between high and low resolution ensemble data members. Then for future runs of the simulation using different parameter combinations, the data was summarized in situ using GMM-based data models. During post hoc analysis, the high-resolution data was reconstructed from the GMM-based down-sampled data summaries using the prior knowledge to improve the reconstruction quality. The in situ study was conducted using Nyx cosmology simulation [3]. For more details about this technique, please refer to [38]. Besides statistical super resolution, distribution-based representations of ensemble data can also be used for studying data features which are characterized as a range of data values. Study of such features were done by He et al. [24] using range likelihood trees.

7 Conclusion

In this chapter we have described various methods of in situ distribution-based data summarization techniques, which on one hand can achieve significant data reduction, and on the other hand can also be used as a flexible data product for post hoc visual analysis. We discussed in details the advantages and disadvantages of using different parameter and non-parametric distribution models for data summarization from the perspective of their in situ feasibility. Using a real world large-scale CFD simulation, we discussed the challenges and possible solutions for distribution-based data modeling for both univariate and multivariate cases. Additionally, several important post hoc data analysis and visualization tasks have been briefly discussed which highlight the effectiveness of the in situ generated distribution-based data summaries in solving a wide range of visualization and data analysis problems.

Acknowledgement

We sincerely acknowledge the contributions from Ko-Chih Wang, Wenbin He, Cheng Li, Chun-Ming Chen, Kewei Lu, and Tzu-Hsuan Wei. This project was supported in part by the US Department of Energy Los Alamos National Laboratory contract 47145, UT-Battelle LLC contract 4000159447, NSF grants IIS-1250752, IIS-1065025, and US Department of Energy grants DE-SC0007444, DE-DC0012495. We would also like to thank Prof. Jen-Ping Chen from the Department of Mechanical and Aerospace Engineering, Ohio State University for providing access

to the TURBO simulation and offering invaluable domain feedback for the in situ application studies. The in situ experiments used computing resources at the Ohio Supercomputer Center [7]. The Hurricane Isabel data set has kindly been provided by Wei Wang, Cindy Bruyere, Bill Kuo, and others at NCAR. Tim Scheitlin at NCAR converted the data into the Brick-of-Float format described above. The Turbulent Combustion data set is made available by Dr. Jacqueline Chen at Sandia National Laboratories through the US Department of Energy's SciDAC Institute for Ultrascale Visualization. This research was released under LA-UR-20-20838.

References

1. Edda - extreme-scale distribution-based data analysis library. https://sites.google.com/site/gravityvisdb/edda
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012). DOI 10.1109/TPAMI.2012.120
3. Almgren, A.S., Bell, J.B., Lijewski, M.J., Lukić, Z., Andel, E.V.: Nyx: A MASSIVELY PARALLEL AMR CODE FOR COMPUTATIONAL COSMOLOGY. The Astrophysical Journal **765**(1), 39 (2013). DOI 10.1088/0004-637x/765/1/39. URL https://doi.org/10.1088%2F0004-637x%2F765%2F1%2F39
4. Berger, M., Li, J., Levine, J.A.: A generative model for volume rendering. IEEE Transactions on Visualization and Computer Graphics pp. 1–1 (2018). DOI 10.1109/TVCG.2018.2816059
5. Bilmes, J.: A gentle tutorial on the em algorithm including gaussian mixtures and baum-welch. Tech. rep., International Computer Science Institute (1997)
6. Biswas, A., Dutta, S., Shen, H., Woodring, J.: An information-aware framework for exploring multivariate data sets. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2683–2692 (2013). DOI 10.1109/TVCG.2013.133
7. Center, O.S.: Ohio supercomputer center (1987). URL http://osc.edu/ark:/19495/f5s1ph73
8. Center, O.S.: Oakley supercomputer. http://osc.edu/ark:/19495/hpc0cvqn (2012)
9. Chen, J.P., Hathaway, M.D., Herrick, G.P.: Prestall behavior of a transonic axial compressor stage via time-accurate numerical simulation. Journal of Turbomachinery **130**(4), 041014 (2008). DOI 10.1115/1.2812968. URL http://turbomachinery.asmedigitalcollection.asme.org/article.aspx?articleid=1467850
10. Chen, J.P., Webster, R., Hathaway, M., Herrick, G., Skoch, G.: Numerical simulation of stall and stall control in axial and radial compressors. In: 44th AIAA Aerospace Sciences Meeting and Exhibit. American Institute of Aeronautics and Astronautics (2006). DOI 10.2514/6.2006-418. URL http://arc.aiaa.org/doi/abs/10.2514/6.2006-418
11. Dutta, S.: In situ summarization and visual exploration of large-scale simulation data sets. Ph.D. thesis, The Ohio State University: http://rave.ohiolink.edu/etdc/view?acc_num . . . (2018)
12. Dutta, S., Chen, C., Heinlein, G., Shen, H., Chen, J.: In situ distribution guided analysis and visualization of transonic jet engine simulations. IEEE Transactions on Visualization and Computer Graphics **23**(1), 811–820 (2017). DOI 10.1109/TVCG.2016.2598604
13. Dutta, S., Liu, X., Biswas, A., Shen, H.W., Chen, J.P.: Pointwise information guided visual analysis of time-varying multi-fields. In: SIGGRAPH Asia 2017 Symposium on Visualization, SA '17. Association for Computing Machinery, New York, NY, USA (2017). DOI 10.1145/3139295.3139298. URL https://doi.org/10.1145/3139295.3139298
14. Dutta, S., Shen, H.: Distribution driven extraction and tracking of features for time-varying data analysis. IEEE Transactions on Visualization and Computer Graphics **22**(1), 837–846 (2016). DOI 10.1109/TVCG.2015.2467436

15. Dutta, S., Woodring, J., Shen, H., Chen, J., Ahrens, J.: Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets. In: 2017 IEEE Pacific Visualization Symposium (PacificVis), pp. 111–120 (2017). DOI 10.1109/PACIFICVIS.2017.8031585
16. Findley, D.F.: Counterexamples to parsimony and bic. Annals of the Institute of Statistical Mathematics **43**(3), 505–514 (1991)
17. Gentle, J.E.: Random Number Generation and Monte Carlo Methods. Springer-Verlag New York (2007). DOI 10.1007/b97336
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014). URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
19. Han, J., Tao, J., Wang, C.: Flownet: A deep learning framework for clustering and selection of streamlines and stream surfaces. IEEE Transactions on Visualization and Computer Graphics pp. 1–1 (2018). DOI 10.1109/TVCG.2018.2880207
20. Hazarika, S., Biswas, A., Dutta, S., Shen, H.W.: Information guided exploration of scalar values and isocontours in ensemble datasets. Entropy **20**(7) (2018)
21. Hazarika, S., Biswas, A., Shen, H.W.: Uncertainty visualization using copula-based analysis in mixed distribution models. IEEE Transactions on Visualization and Computer Graphics **24**(1), 934–943 (2018). DOI 10.1109/TVCG.2017.2744099
22. Hazarika, S., Dutta, S., Shen, H., Chen, J.: Codda: A flexible copula-based distribution driven analysis framework for large-scale multivariate data. IEEE Transactions on Visualization and Computer Graphics **25**(1), 1214–1224 (2019). DOI 10.1109/TVCG.2018.2864801
23. He, W., Chen, C., Liu, X., Shen, H.: A bayesian approach for probabilistic streamline computation in uncertain flows. In: 2016 IEEE Pacific Visualization Symposium (PacificVis), pp. 214–218 (2016). DOI 10.1109/PACIFICVIS.2016.7465273
24. He, W., Liu, X., Shen, H., Collis, S.M., Helmus, J.J.: Range likelihood tree: A compact and effective representation for visual exploration of uncertain data sets. In: 2017 IEEE Pacific Visualization Symposium (PacificVis), pp. 151–160 (2017). DOI 10.1109/PACIFICVIS.2017.8031589
25. He, W., Wang, J., Guo, H., Wang, K., Shen, H., Raj, M., Nashed, Y.S.G., Peterka, T.: Insitunet: Deep image synthesis for parameter space exploration of ensemble simulations. IEEE Transactions on Visualization and Computer Graphics **26**(1), 23–33 (2020)
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). URL http://arxiv.org/abs/1312.6114. Cite arxiv:1312.6114
27. Li, C., Shen, H.W.: Winding angle assisted particle tracing in distribution-based vector field. In: SIGGRAPH Asia 2017 Symposium on Visualization. Association for Computing Machinery, New York, NY, USA (2017). DOI 10.1145/3139295.3139297. URL https://doi.org/10.1145/3139295.3139297
28. Li, G., Xu, J., Zhang, T., Shan, G., Shen, H., Wang, K., Liao, S., Lu, Z.: Distribution-based particle data reduction for in-situ analysis and visualization of large-scale n-body cosmological simulations. In: 2020 IEEE Pacific Visualization Symposium (PacificVis), pp. 171–180 (2020)
29. Lu, K., Shen, H.W.: A compact multivariate histogram representation for query-driven visualization. In: Proceedings of the 2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV), LDAV '15, pp. 49–56 (2015)
30. Rank, J.: Copulas: From Theory to Application in Finance. Bloomberg Financial. Wiley (2007). URL https://books.google.com/books?id=133HkvhOHC8C
31. Razali, N.M., Wah, Y.B.: Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics **2**(1), 21–33 (2011)
32. Schindler, K., Wang, H.: Smooth foreground-background segmentation for video processing. In: Proceedings of the 7th Asian Conference on Computer Vision - Volume Part II, ACCV'06, pp. 581–590. Springer-Verlag, Berlin, Heidelberg (2006). DOI 10.1007/11612704_58. URL http://dx.doi.org/10.1007/11612704_58

33. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**, 591–611 (1965)
34. Sklar, M.: Fonctions de Répartition À N Dimensions Et Leurs Marges. Université Paris 8 (1959)
35. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2, pp. –252 Vol. 2 (1999). DOI 10.1109/CVPR.1999.784637
36. Stephens, M.A.: Edf statistics for goodness of fit and some comparisons. Journal of the American Statistical Association **69**(347), 730–737 (1974). DOI 10.1080/01621459.1974.10480196
37. Wang, K., Kewei Lu, Wei, T., Shareef, N., Shen, H.: Statistical visualization and analysis of large data using a value-based spatial distribution. In: 2017 IEEE Pacific Visualization Symposium (PacificVis), pp. 161–170 (2017). DOI 10.1109/PACIFICVIS.2017.8031590
38. Wang, K., Xu, J., Woodring, J., Shen, H.: Statistical super resolution for data analysis and visualization of large scale cosmological simulations. In: 2019 IEEE Pacific Visualization Symposium (PacificVis), pp. 303–312 (2019). DOI 10.1109/PacificVis.2019.00043