

# EnsembleGraph: Interactive Visual Analysis of Spatiotemporal Behaviors in Ensemble Simulation Data

Qingya Shu <sup>1\*</sup> Hanqi Guo <sup>3†</sup> Jie Liang <sup>1‡</sup> Limei Che <sup>1§</sup> Junfeng Liu <sup>4¶</sup> Xiaoru Yuan <sup>1,2||</sup>

- 1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, P.R. China
- 2) Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University, Beijing, P.R. China
- 3) Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA
- 4) College of Urban and Environmental Sciences, Peking University, Beijing, P.R. China

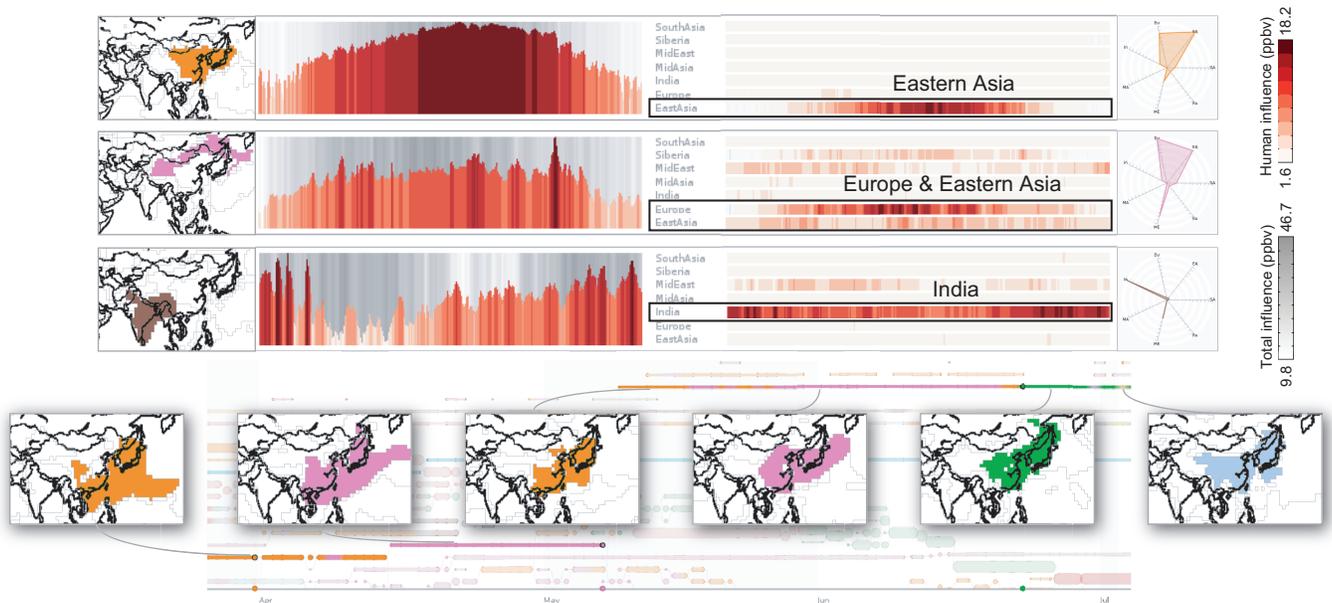


Figure 1: Use case of *EnsembleGraph* to indicate the influences of Eurasian continent emissions on the surface ozone concentration over eastern Asia. According to similar behaviors between ensemble members over space and time, the neighborhood is partitioned into three parts (left thumbnails in the first three rows): eastern China, southwestern China, and northwestern China. Our novel graph-based interface provides an abstraction of the grouped regions. Users can therefore navigate and track regions of interest over space and time. The last row shows tracking partitioned over southeastern China using a graph view and linked spatial view. Users highlight regions for further analysis in the comparison view, where they compare values between individual runs and behavior similarities between ensembles over different subregions (charts in the first three rows).

## ABSTRACT

This paper presents a novel visual analysis tool, *EnsembleGraph*, which aims at helping scientists understand spatiotemporal similarities across runs in time-varying ensemble simulation data. We abstract the input data into a graph, where each node represents a region with similar behaviors across runs and nodes in adjacent time frames are linked if their regions overlap spatially. The visualization of this graph, combined with multiple-linked views showing details, enables users to explore, select, and compare the extracted regions that have similar behaviors. The driving applica-

tion of this paper is the study of regional emission influences over tropospheric ozone, based on the ensemble simulations conducted with different anthropogenic emission absences using MOZART-4. We demonstrate the effectiveness of our method by visualizing the MOZART-4 ensemble simulation data and evaluating the relative regional emission influences on tropospheric ozone concentrations. **Keywords:** ensemble simulation, graph visualization.

## 1 INTRODUCTION

Ensemble simulations have become prevalent in various scientific and engineering domains, such as aerodynamics, climate, and weather research. They are usually used for studying model sensitivities to parameters and initial conditions and for quantifying uncertainties. The visualization of ensemble data sets is a grand challenge, however, because ensemble data are usually multivariate, multivalued, and time-varying and have large data scales.

Our focus in this paper is the behaviors of ensembles—similarities between individual runs in space and time. Currently, scientists typically analyze ensemble data by manual selection and spatiotemporal aggregation. For example, a latitude-longitude box is arbitrarily defined first as the region of interest, and users then

\*qingya.shu@pku.edu.cn

†hguo@anl.gov

‡jie.liang@pku.edu.cn

§limei.che@pku.edu.cn. Now at Baidu Inc.

¶junfeng.liu@pku.edu.cn

||xiaoru.yuan@pku.edu.cn (Corresponding Author)

aggregate values along the temporal dimension (e.g., seasonal or monthly average values). The spatial patterns of different ensemble members can be visualized by contours or pseudo colors. Line charts are plotted to compare the temporal differences between ensemble members. This process has a number of critical issues. First, without an overview, it is difficult to understand the overall patterns of the data set by manual queries back and forth. Second, inappropriately defined regions may lead to information loss in the spatiotemporal aggregation and statistics, because the data properties may be highly inhomogeneous in specific regions. Visualizing such data is challenging but critical, so that scientists can understand their scientific data more effectively.

In this work, we propose a visual analysis framework called *EnsembleGraph* (Figure 1), based on behaviors of ensembles. We quantify the behaviors as *behavior vectors*, using metrics that describe similarities between ensemble members in spatiotemporal location (detailed explanation in Section 4). Based on discussions with domain scientists, we design visual analysis tools to support various tasks, include the following:

- **Partitioning the ensemble domain based on behaviors of ensembles.** Clustering of regions with similar behaviors helps find reasonable initial targets. This provides an abstraction and overview of the overall patterns in ensemble data sets.
- **Representing the spatiotemporal distribution of behaviors in ensembles.** A novel interface must be able to expressively show the occurrences of abstracted partitions in space and time. Through this, scientists can quickly and flexibly access data for the target regions.
- **Comparing the different ensemble members.** Comparison is the core mission in ensemble visualization. Our tool should be capable of comparing individual runs in targeted subregions.

To support these tasks, we design the framework with several components. First, we use an automatic ensemble domain partitioning method and do partitioning over the ensemble data, in order to help identify regions with similar relative emission influences. Regarding those grouped partitions as basic units for analysis, we also identify their spatiotemporal relationships in order to give an overview to the whole ensemble data. Second, to support spatiotemporal exploration of all behavior patterns, we map the spatiotemporal extracted regions and their relationships into a graph structure, which provides an intuitive interface for analysis. Third, we provide tools for comparing individual ensemble members, which can be used to validate the findings for the exploration.

The driving application in this paper is to understand the impact of regional emissions on the tropospheric ozone ( $O_3$ ). Tropospheric  $O_3$ , an important greenhouse gas that is harmful to human health and agriculture production. They are formed from chemical reactions of nitrogen oxides, carbon monoxides, and so forth, which are caused mostly by human activities such as industrial and road emissions. These anthropogenic emissions, so called  $O_3$  precursors, are different around the world, because of local industrialization and environmental policies. Those emissions are transported by wind convection, contributing a global atmospheric issue. Thus, tropospheric  $O_3$  is a mixed influence affected by all regional anthropogenic pollutant emissions, yet the mixing weight from each source region is not identical. For example, studies have shown that  $O_3$  concentration over eastern China is affected mostly by domestic pollutant emissions from the industry. Western China, which is less industrialized, has the opposite situation; the  $O_3$  is formed mostly from foreign emissions of upwind neighbors, such as India and Europe [17]. Analyzing and understanding the regional emissions impacts are important for scientists and decision makers to further expedite emission reductions. To this end, scientists

have conducted ensemble simulations under different emission scenarios [17]. The ensemble simulation are based on the Model of Ozone and Related Tracers, version 4 (MOZART-4). It consists of perturbation runs with different emission sources and reference runs (detailed explanation in Section 2.1); With such data sets, scientists would like to investigate the relative importance of different emission sources in the ensemble domain. To support the tasks, we calculate behaviors according to the combination of influences from different source emissions, and we apply our framework for visualization using novel graph-based interface. Two case studies and feedback from domain scientists show the usefulness of our methods.

In summary, the contributions of this paper are as follows:

- Visual analysis framework that helps understand ensemble simulation data based on behaviors of ensembles.
- Novel visual representation for exploring complex ensemble data using a graph visualization method.

We organize the remainder of this paper as follows. In section 2 explain the background and review related work in Section 2. Section 3 gives an overview to our approach. Section 4 describes data processing and graph construction. Section 5 presents visual design and interface. We then demonstrate cases and feedback in Section 6 and Section 7, and conclude the paper in Section 8.

## 2 BACKGROUND

We describe the driving application of this study and then summarize related work on ensemble visualization and graph-based visualization techniques in scientific visualization.

### 2.1 Driving Application

Scientists conduct perturbation experiments with ensemble simulations to evaluate the sensitivities of  $O_3$  concentration to different regional emissions [17]. The simulations are based on the MOZART-4 model. The inputs of the model are from observations and emission inventories, and the outputs are the concentration of a series of chemical species [5]. In this work, we focus on the most important substance— $O_3$ —which dominates the chemical reactions in the model. We use daily surface  $O_3$  concentration in year 2000.

The experiments involve three types of runs in the experiments: a BASE run, a GLOBE run, and perturbation runs. The BASE run is conducted with the actual emissions, and the GLOBE run is conducted by switching off all anthropogenic emissions. The perturbation runs alternatively switch off the emissions from seven different preset regions in the world. As shown in Figure 2, the preset emission source regions are Europe (related to EU run), India (IN run), Middle East (ME run), southeast Asia (SA run), eastern Asia (EA run), mid-Asia (MA run), and Siberia (RU run). Notice that the natural emissions still exist as the background emission, even if the anthropogenic emissions from the source regions are switched off. In our framework, we define the *member behaviors* by the bias of the perturbation runs from BASE and GLOBE runs.

Through the ensemble simulations, one can measure the response of tropospheric  $O_3$  concentration to different anthropogenic emission conditions. Conventionally, scientists choose a range box as the region of interest to study (e.g., rectangular range over eastern China). Next they compare the  $O_3$  concentration of one of the perturbation runs with BASE run and GLOBE run in this area, to evaluate the influence of this source region. For instance, the example in Figure 3 shows the EA run for evaluating emission influence from the eastern Asia region. From maps and statistics, the scientists concluded that the  $O_3$  over eastern China is influenced mostly by emission from the eastern Asia area, indicating very high domestic emission in Eastern China, especially in summer (Figure 3).

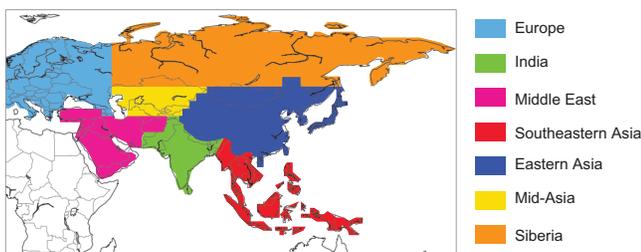


Figure 2: Seven source emission regions. During the ensemble simulation, anthropogenic emissions from one of the seven regions are turned off, in order to calculate relative importance of influence from human activity around that source region.

However, such conventional ensemble analysis workflow has some deficiencies. First, the overall pattern is invisible to the users. Because scientists can probe only one small region at one time, they are almost blind to the overall data set, not knowing where to start and having no evaluation for the rationality of targeting regions; second, flexibility of interaction is limited. With this trial-and-error process, scientists have no guarantee of the location and shape of the target region; moreover, data in some locations may be inhomogeneous. To make the access more intuitive and flexible for scientists to explore all potentially interesting features of regional emission influences, we partition the ensemble domain into spatiotemporal subregions, according to the *behaviors* of the ensemble. Then our novel interface provides a visual abstraction and interaction for ensemble analysis. We explain the workflow and the algorithms in the following sections.

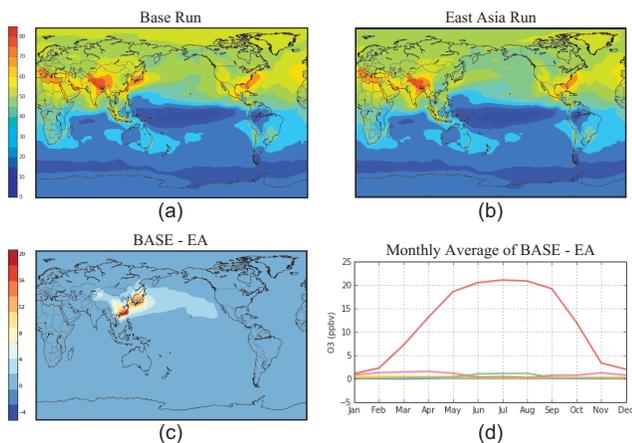


Figure 3: (a): BASE run. (b): EA run — one of the perturbation runs that is simulated for emission absence from the eastern Asia region. (c): Their difference. (d): Monthly  $O_3$  average in an chosen area over eastern China. This indicates that domestic emission influence over eastern China is high, especially in summer.

## 2.2 Related Work

**Ensemble Visualization.** Ensemble simulation data sets are usually multivalued, multivariate, and time-varying. Thus, they are very challenging to visualize [16]. Aiming at showing spatiotemporal information as well as relationships between ensemble members, the related research focuses on uncertainty visualization and comparative visualization.

Previous uncertainty visualization work has used aggregation or distribution to describe data values. Converting the multiple values to statistics (e.g., mean, standard deviation, or peak numbers of distributions) makes common visualization techniques applicable [15, 20, 21]: pseudo coloring, streamlines, pathlines, or isosurfaces. Histograms or parameter-based representations using

Gaussian mixture module also largely reduce the data complexity [18, 36]. Visually embedding or extending uncertainty information into conventional visualization methods helps users identify high-uncertainty regions and outliers: overlaying uncertainty-encoded ribbons and glyphs over spaghetti plots [30], integrating statistics (e.g., skew, kurtosis, and histogram) into the boxplot technique [34], and generalizing functional boxplots to visualize spatial distribution of ensemble contours [25, 38]. Our method abstracts the ensemble domain into partitions with similar data properties (i.e., the behavior of ensemble members) and provides an intuitive interface for exploration. One of the differences is that we focus on the relationships between all ensemble members at each spatiotemporal location, that is, we quantify such properties using *behavior vectors*. Another difference is that we detect and group locations with similar properties in the ensemble domain and regard them as the basic unit for exploration. By doing so, we reduce the complexity of ensemble data and give users a simple portal that lets them quickly identify and compare regions of interest in an ensemble.

Some previous works have also experimented with clustering and classification methods, e.g., clustering ensemble realizations [2], clustering location points according to member distributions [4], or classifying location points by member distributions and ground truths [7]. We also use clustering for locations with similar behaviors. During the clustering procedure, however, our computation of their distance considers individual ensemble members, instead of overall distributions, because counting overall distributions will cause information loss for individual ensemble members.

Comparative visualization is important in ensemble visualization. A recent taxonomy for ensemble data comparison divides existing approaches into location-based comparison and feature-based comparison [27]. Location-based methods conduct data comparison by attributes at fixed locations in an ensemble domain: using a color map to show point-to-point differences between two simulations [26] or statistical aggregation that indicates disagreement between members at a single location [21, 31, 32], line chart and bin chart encoding distribution inside one region [4], similarity matrix for climate simulation models on predefined areas [29], interactive similarity analysis for climate simulation models using multiple criterias [28], and spatiotemporal exploration for off-shore structures in a user-defined area [11, 12]. For ensemble flow field, which is not applicable by traditional scalar field-based methods, Lagrangian-based measurements [10] and transport variances [13] quantify the degree of agreements on the same location in ensemble flow fields. Feature-based methods first extract features from individual runs and then compare them: rendering isosurfaces in a slice-by-slice style [1], constructing isosurface-based comparison for differences between two module runs [26], and displaying isocontours while overlaying uncertainty glyphs [34]. Our method belongs to the location-based type; we compare ensemble members over precalculated regions.

**Graph-Based Methods in Scientific Visualization.** Applying graph visualization on scientific data is a new trend in scientific visualization [37]. Abstracting features from scientific data to graph models can help users gain better navigation and understanding of relationships in the complex data, because graphs in 2D are usually occlusion-free and more intuitive to explore than traditional 3D visualization methods. Many previous works used graph structure to represent large-scale, time-varying volumetric data and the transitions of inside feature transitions over time [3, 9, 14, 39]. FlowGraph [22, 23] and FlowWeb [40] are novel graph-based visualization for flow fields that show the relationships of field lines and data blocks. Sauber et al. [35] use graphs to reveal relationships between variables for multifield data sets. Our work also takes advantage of graphs to abstract complex time-variant ensemble simulation data: we use nodes to represent subregions with similar ensemble behaviors, and we use edges for their spatial overlap in adjacent time

frames. Our graph-based interface helps researchers explore spatiotemporal similarities between ensembles.

### 3 SYSTEM DESIGN AND OVERVIEW

*EnsembleGraph* is intended to help scientists find when and where the ensemble runs are similar. Our motivation starts from close examination of the steps that scientists take in a conventional workflow. We notice that their manual analysis process strongly depends on trial and error, which is time-consuming and easily leads to missing important information in the data. We find that visual analytics can help identify regions with high similarity across ensemble members and help scientists quickly explore the member behavior in targeted regions. Based on discussions with scientists, we list our design goal:

- Enabling subregion detection based on ensemble similarities.** We provide spatial domain partitioning according to similarities across ensemble members. Scientists are interested mostly in finding the subregions that can help them identify patterns between ensemble members inside, such as similar values between individual runs. Predicting their location and shape is difficult, however. Our method quantifies the ensemble similarities on each location and groups locations into subregions, which can serve as guidance for ensemble data exploration.
- Providing overview for subregion selection.** Our visual design includes a graph-based interface to show the evolution of abstracted subregions. The similarities between ensemble members would change over time. Giving an overview of the spatiotemporal distribution of all abstracted subregions helps identify evolving patterns for behavior vectors.
- Facilitating interactive visualization for comparison.** Our interactive interface provides comparison both inside and between subregions. The focus on subregions allows observing actual values by using time series visualization techniques.

*EnsembleGraph* provides three visual components: a spatial view, a temporal view, and a comparison view. The spatial view shows how the ensemble domain is partitioned into subregions according to ensemble similarities. We use a colored map to show the partition results. Users can explore and highlight subregions. The *temporal view* shows the occurrences of all subregions over time. We plot the graph in chronological manner from left to right. Each node in the graph represents a subregion with similar values across individual runs; and links indicate spatial overlapping in adjacent time frames. The layout is designed for easy reading and tracing. The *comparison view* enables users to highlight and compare actual values inside subregions. We enable single- and multi-run viewing modes for comparing ensemble members both inside and between subregions. Our visual interface adopts area chart and pixel-based table-style visualization for displaying time series.

The three components are linked to support spatiotemporal navigation. Users track subregions in the temporal view and relate corresponding partitions in the spatial view. Then the comparison view further provides time series and comparison for focused areas. The last row in Figure 4 illustrates this process.

In order to support the aforementioned navigation functions, the preprocessing steps mainly contains two parts (the first row in Figure 4): the domain partitioning and the region connection. First, we do ensemble domain partitioning based on ensemble behaviors, so as to summarize complex data into abstracted subregions: we quantify the similarities into behavior vectors and group together similar locations into subregions. Second we construct a graph structure for spatiotemporally summarizing to the ensemble domain: we establish connections between regions in neighboring time frames

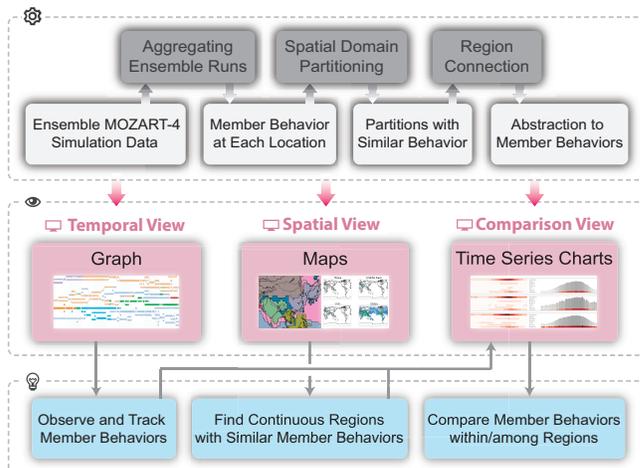


Figure 4: Overview of the *EnsembleGraph* visual analysis framework. The first row is the data preprocessing part: we do partitioning and region tracking for the data domain to provide overall summarization. The second row is our interface including three main components. The temporal view shows regions of similar ensemble behavior over time; the spatial view shows partitioning results as well as spatial patterns of individual runs; and the comparison view visualizes emission influences of highlighted subregions for validation. The third row is the exploration flow.

by detecting their spatial overlaps. We then abstract the ensemble domain into a graph structure for visualization. We explain the preprocessing procedures in the next section.

### 4 DOMAIN PARTITIONING BASED ON ENSEMBLE BEHAVIORS

We provide ensemble domain partitioning based on regional emissions, to help scientists identify so-called influence relationships between regions. Then we construct a graph data structure by their temporal connectivity for further visualization.

#### 4.1 Ensemble Behavior Definition

To establish a metric for representing the emission influences between regions, we quantify the behavior of ensemble by *behavior vectors*. The behavior vector  $\mathbf{v}$  is defined as an  $n$ -dimensional vector for each spatiotemporal location (Figure 5)  $\mathbf{x}$ :

$$\mathbf{v}(\mathbf{x}) = (d_1(\mathbf{x}), d_2(\mathbf{x}), \dots, d_n(\mathbf{x}))^T, \quad (1)$$

where  $n$  is the number of perturbation runs, and  $d_i(\mathbf{x})$  represents the *behavior* of the  $i$ th ensemble member. For the application, our definition of *behavior*  $d_i(\mathbf{x})$  is the influence to the location from the  $i$ th emission source region  $R_i$ . It is calculated as the difference between the BASE run and the  $i$ th run:

$$d_i(\mathbf{x}) = \hat{C}(\mathbf{x}) - C_i(\mathbf{x}), \quad (2)$$

where  $\hat{C}(\mathbf{x})$  and  $C_i(\mathbf{x})$  are the values of the BASE run and the  $i$ th run, respectively. Thus, each spatiotemporal location has a high-dimensional behavior vector. The similarities between behaviors on two locations are defined by the inversion of the Euclidean distance between behavior vectors. Therefore, locations having higher similarity value indicate ensemble members having similar behaviors. In our application, this means that O<sub>3</sub> over two places is influenced by a similar combination of emission sources.

#### 4.2 Spatial Domain Partitioning for Ensemble Data

After quantifying the behavior of ensemble members on each location, we aim at revealing how those behaviors distribute in space,

so as to indicate potentially interesting regions. Our method is to group together the neighboring locations with similar behaviors into subregions and present this as the basic unit for analysis. In our implementation, we first cluster locations with similar behaviors into the same categories, and then merge clustered locations into larger subregions.

We provide multiple options for the clustering algorithm, including  $k$ -means [19], hierarchical clustering [8], and DBSCAN [6]. Users can select different algorithms and tune parameters to improve the results with our tool. For simplicity, we use  $k$ -means for algorithm description in this paper, which is one of the most used methods for vector quantization. The clustering result keeps points in the same groups close to each other and points in different groups distinct from each other. The  $k$ -means algorithm starts with some randomly selected centroids and then iteratively changes the centroids for each cluster, until the clustering results become stable. To keep the results stable, we avoid random initialization with the unsupervised preclustering algorithm Canopy [24]. It quickly covers all data points with several circles with the same size (so called canopies) in the high-dimensional space. The centroids of canopies are then used as input for  $k$ -means, thus avoiding the random seeding problem. After the clustering, we obtain the cluster labels for each location in the domain.

We repeat the clustering for all time steps. To improve the performance of the clustering for the sequential time steps, we use the previous labeling results as the initial input for the  $k$ -means clustering in the next time steps. This approach is based on the assumption that the ensemble values do not significantly change in subsequent time steps.

After the clustering, we group locations with the same clustering labels, in order to detect all continuous regions with similar values across runs. We thus obtain partitioned results for ensemble domains, and they keep the important features in the data. Subregions are used as basic units in the graph-based visualization and analysis, which represent data properties of all locations inside.

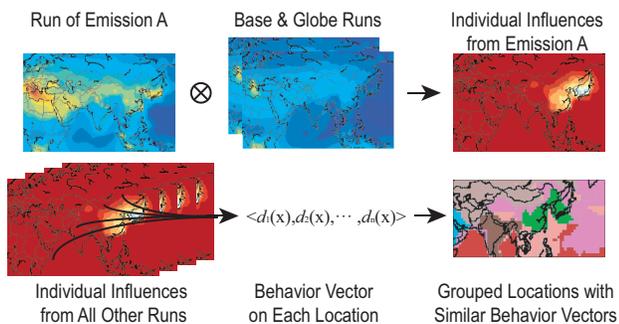


Figure 5: Partitioning ensemble domain according to the behavior vectors on each location. First, we calculate the behavior vector on each location according to the `BASE` run and the perturbation runs to quantify behavior of ensemble on each location using a vector. Then we classify all behavior vectors by clustering methods and use this result for ensemble domain partitioning.

### 4.3 Partitioned Subregions Connection

We correspond subregions in subsequent time steps, in order to capture the evolution of similar ensemble behaviors over time. Feature-tracking algorithms have been well studied over decades; we choose an effective way to track neighboring features based on their spatial overlap [33].

Two parameters in the overlap-based subregion tracking: the subregion size threshold  $\gamma_r$  and the overlap size threshold  $\gamma_o$ . Regions with small sizes less than  $\gamma_r$  are skipped so as to avoid uninteresting, noisy regions. The overlap size threshold  $\gamma_o$  is for region con-

nectivity. Smaller  $\gamma_o$  usually leads to larger connectivities between subregions. Users can interactively change the parameters to obtain different granularities of subregions for analysis. When users change the parameters, our tool updates the visualization results in the background, in order to reduce the delay in the user interface. More details are discussed in the next section.

Next, we construct the edges connecting the subregions that are regarded as the same feature. The split and merge events are also recorded in the graph for further interactive visual analysis.

## 5 VISUALIZATION AND INTERACTION DESIGN

We construct and visualize the graph data structure of extracted subregions that have similar ensemble behaviors based on the design goals. In addition to the graph visualization that gives the overview, we provide linked views for interactive exploration and comparison of the selected subregions.

The visualization tool consists of three main components: the temporal view, the spatial view, and the comparison view. In addition to the three main components, we provide a control panel for users to select clustering algorithms and their parameters.

Our prototype system is implemented in C++ with OpenGL and Qt libraries. We separately run the client on a lightweight machine and the server on a powerful workstation with larger memory and I/O performance. This design enables large data handling. We would like to further extend this system to cluster and supercomputing environments.

### 5.1 Temporal View

The temporal view visualizes the abstracted subregions by using a graph in a chronological manner, in order to provide an overview of the ensemble data. It maps the time-varying partitioning results into a 2D plane, which is intuitive for interactive exploration. For example, the bottom part of Figure 6 shows a visualization result of an ensemble flow simulation evolution.

Our visualization design involves three principles for the graph layout. First, graph nodes must be aligned to their times of occurrence from the left to the right. Thus the layout is in a “streaming” style, analogue to the storyline visualization. Second, edge crossings need to be reduced as much as possible for better readability. Third, the recurring subregions should be aligned horizontally as straight as possible, so that users can easily trace them in the visualization. In our implementation, we first use the “dot” algorithm in the GraphViz library to generate an initial layout, which places nodes from the left to the right. Then we visually wrap the straightened nonbranch paths, to emphasize connected subregions for easy tracking. In the visualization, the sizes of nodes are proportional to the sizes of counterpart regions, and the colors of nodes encode various variables selected by users.

The temporal view also links to other views in the system. The selected subregions are also shown in the spatial view and the comparison view for detailed analysis. In addition, we provide an optional star glyph visualization in this view for probing. Users can highlight subregions by clicking on the interested nodes. The averaged ensemble values of the selected subregion are visualized by a star glyph. Each highlighted node correspond to one polygon in the star glyph, and the coordinates on the axis are the averaged property from individual ensemble members. Figure 1 shows some example star glyphs on the right side.

### 5.2 Spatial View

The spatial view visualizes the spatial distribution and the input ensemble data of the subregions. We provide a data mode and a partition mode in this view. The data mode shows data values for all ensemble members. Users choose ensemble members and then navigate in a map. By dragging the slider bar, users can change the

current time step for display. The partition mode shows the distribution of subregions with solid colors and allows users to highlight locations of interests. Users can focus on a subregion by double-clicking in the map view and submit the selected subregion to the server side to query for retrieving data values. The focusing action will be broadcast to the temporal view and the comparison view, to update the current exploration status.

### 5.3 Comparison View

The comparison view enables two types of comparisons: an ensemble member comparison on selected subregions, and a subregion comparison on the selected ensemble run. Users can also examine detailed information in the selected subregions for further analysis. Both types of comparison were indicated by the domain scientists in our study to be important.

We design the comparison view in a list style. Each list item relates to one selected subregion. Inside each item is a thumbnail map and a detailed visualization result. Thumbnails on the left side give a preview of the location of corresponding subregions. The same color scheme is used as the ones in the temporal view and the spatial view. The ensemble members are shown on the right side of the view, using area charts and pixel-based visualizations.

We provides three viewing modes for visualizing behaviors of ensemble members for each highlighted subregion: (1) **the natural-anthropogenic mode** compares natural influence and anthropogenic influence; (2) **the domestic-foreign mode** compares emission influences from one area with ones from other areas; and (3) **the individual influence mode** compares individual influences through a pixel-based visualization. **The natural-anthropogenic mode** takes the *BASE* run as the background chart and plots the anthropogenic influence (defined by the difference between the *BASE* run and the *GLOBE* run) as the foreground. This lets users gain knowledge about overall  $O_3$  concentration and the influence fraction from human activities. Users can also switch the foreground chart to the *GLOBE* run, to focus on the natural emission influences. **The domestic-foreign mode** visualizes the domain-foreign ratio of emission influences, which is calculated by dividing domain emission influences (differences between the *BASE* run and perturbation runs) by foreign emission influences (differences between individual runs and the *GLOBE* run). We use a red-to-blue color map for this ratio as the foreground and a white-to-gray color map for total anthropogenic influences in the background. **The individual influence mode** visualizes temporal distribution of influences from all source regions. In the pixel-based table-style visual representation, each row is one chart for the corresponding source region, while each column represents the corresponding time step. Red means positive values, and blue means negative values.

## 6 RESULTS

This section presents two case studies with *EnsembleGraph*: identifying source region emission influences of tropospheric  $O_3$  over China in Case I and investigating spatial patterns in the Southern Hemisphere in Case II. Both case studies use daily output from nine simulation runs from MOZART-4, in the year 2000 (366 timesteps in total). We use surface  $O_3$  concentration, with spatial resolution at  $192 \times 96$ .

### 6.1 Case I: Analyzing Source Region Emission Influences on $O_3$ over China

Scientists would like to analyze how  $O_3$  over China is influenced by anthropogenic emissions from the Eurasian continent. Specifically, they want to find out which places are more influenced by domestic emissions and which places are influenced by foreign regions, and how influences change over time. *EnsembleGraph* partitions the area by relative importance between regional anthropogenic emission influences. The partition map in the spatial view shows that

the area over China is covered mostly by three partitions. The first one (the first row in Figure 1) covers eastern China area and neighboring regions such as Japan and the Korean peninsula. The second one (the second row in Figure 1) covers southwestern China, being connected to India. The third one (the third row in Figure 1) is northwest China, connecting to middle Asia and Siberia. The partitioning results agree with the geographical terrain: southwest China is the Qinghai-Tibet plateau, and northwest China is separated from eastern China by mountains.

From the visualization results we can observe and compare actual values inside these subregions. Inspired by the analysis tasks of domain scientists, we use the single-run mode (area charts in Figure 1) to display the differences between the *BASE* run and the *GLOBE* run, in order to show the total anthropogenic emission influences in each region. We find that in eastern China (the first row), influences happen mostly in summer and almost disappear in winter. In northwest China (the second row), most anthropogenic influences appear in spring and summer. On the southwest direction (the third row) the anthropogenic influences last almost the whole year. The multi-run mode (pixel-based table in Figure 1) allows comparison for temporal distribution of each regional emission. The eastern China area is influenced almost only by east Asia (the first row). In northwest China (the second row), Europe and east Asia are the dominant emission source regions in spring and summer. Emission influences over southwest China (the third row) are totally different: the area largely affected by India throughout the whole year, and occasionally affected by the middle East during spring. These findings are similar to the previous work [17]. In their study, the researchers selected two box areas over Xinjiang province ( $40^\circ\text{N}$ - $45^\circ\text{N}$ ,  $84^\circ\text{E}$ - $90^\circ\text{E}$ ) and the Qinghai-Tibet Plateau ( $29^\circ\text{N}$ - $34^\circ\text{N}$ ,  $86^\circ\text{E}$ - $92^\circ\text{E}$ ). Eastern China suffers a higher domestic pollutant influence of  $O_3$  concentration because of its industrial prosperity. Western China, which is less industrialized, is influenced mostly by foreign emissions of upwind neighbors, such as India.

### 6.2 Case II: Observing Source Region Emission Influences on $O_3$ over the Southern Hemisphere

This section shows how scientists use *EnsembleGraph* to study spatial patterns over the Southern Hemisphere by different regional emission patterns. First, by exploring the temporal view using panning and zooming, scientists can discover an obvious node chain that stays throughout the whole year. It seems related to a very large region indeed. Partition map in the spatial view shows that it belongs to the largest segment covering the whole Southern Hemisphere. It explains that most parts of the Southern Hemisphere have similar ensemble behaviors. By double clicking on the Southern Hemisphere partition and submitting this selected region to the server side, scientists can filter out all other unrelated regions, leaving only a trunk with several branches in the graph, which indicates that the Southern Hemisphere can be as considered as a whole region or separated into two or three regions according to the different member behaviors. The counterpart subregions are shown in the partitioning map in Figure 6 (thumbnails in the left side).

In the comparison view, we can find temporal member behavior differences: Most southern and most northern areas of these three subregions have higher  $O_3$  concentration in January and December. Through the natural-anthropogenic comparison mode, however, we found that  $O_3$  over the most southern subregion is caused by natural emissions, when  $O_3$  over the latter subregion is affected by anthropogenic emissions (see Figure 6). The most northern strip, although it appears to have lower total  $O_3$  concentration, suffers from relatively higher anthropogenic emission influences during the whole year. The individual influence comparison mode shows the relative importance: for the most northern subregion, southern Asia emissions have been the dominant source throughout the year, fol-

lowed by eastern Asia and the Middle East during the summer in the Northern Hemisphere. The other two subregions in the south are similarly more affected by southern Asia, eastern Asia and Siberia during the summer.



Figure 6: Southern Hemisphere divided into three regions according to ensemble behaviors.

### 6.3 Domain Scientists Review

We have worked with domain scientists to validate our findings and evaluate our tool. For the first case, they confirmed our findings, namely that eastern China is significantly influenced by its domain emission, while northwestern and southwestern China are isolated from those emissions because of mountains and plateaus, and are more affected by the upwind areas. For the second case, the partitioning results can be interpreted as follows. The middle part is located in the westerly of Southern Hemisphere, which has almost no land as obstacles and thus experiences strong wind convection. That leads to similar emission influences around this latitude and also isolates the air above the Antarctic continent, which is the southernmost region in the partitioning results. To confirm the influence mechanism from the Eurasian continent to the Southern Hemisphere, we need more simulation data to apply to our framework. We have also received positive feedback on the tool. The scientists showed particular interests in the partitioning results, which help them efficiently locate subregions with similar ensemble behaviors. Without such a visual analysis tool, it is time-consuming to define a subregion to study the ensemble member similarities. The tool also provides a flexible user interface to visualize the ensemble value distributions and compare the ensemble members. Scientists can locate the features that they have found with their traditional workflows. Our tool has the potential to help them find even more interesting features in their future experiments by adding more customizable operators.

## 7 GENERAL APPLICATION CASE: ENSEMBLE LOCK-EXCHANGE SIMULATION

Our next experiment involves an ensemble flow simulation to study the model sensitivities to different perturbations. The simulation is an experiment of the lock-exchange problem: a light fluid and a heavy fluid are separated by a barrier at the start condition; at the first time step, the barrier is removed, to allow the two fluids mix. Scientists slightly change the initial density difference between the two fluids to evaluate the sensitivity of flow mixing progress. In our application we use the output fluid density and focus on density similarities between ensemble members of different initial perturbations. We use 100 runs of 100 time steps, with a spatial resolution of  $128 \times 128$  for analysis.

Figure 7 shows the visualization results of this data set. We use the offset fluid density values from the mean values to quantify the similarity of ensemble members: at each location, the behavior vector consists of the offset density values of all ensemble members, and the ensemble member similarity between two locations is defined as their Euclidean distance. Higher similarities between two

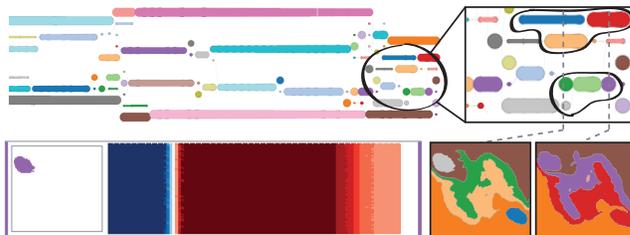


Figure 7: Ensemble flow simulation process. First row: a detail part at the end of the graph. Second row left: time series inside one ellipse region; Second row right: spatial view at the two time steps.

locations mean similar output from the same run. *EnsembleGraph* therefore groups such locations and describes the evolution of such subregions during this ensemble flow mixing process. We observe some relatively larger sized node paths, relating to the two separated liquids that keep stable at first and then slowly start to rotate and mix together in the simulation process. We also observe that the graph becomes messy at the end of the simulation: many small splitting regions and crossing edges appear. The first row shows a detailed part at the end of the graph, and two time frames: One gray high-speed ellipse region deviates from the large region and merge at the other side with the purple region. On the symmetric side, the dark blue ellipse region morphs into the larger red part. The second row shows time series inside one ellipse region. Each row is one run; red means positive values, and blue means negative values, comparing to the average. We can see that the value inside the highlighted subregion is first below average, then suddenly rises to a very high level, and finally slowly falls to a normal level. This behavior means that two rotating liquids with different densities pass across the highlighted subregions during the experiments, and the overall density gradually falls to average after mixing.

## 8 CONCLUSIONS AND FUTURE WORK

We present a visual analysis framework *EnsembleGraph* for ensemble simulation data analysis. The goal of this work is to enable interactive exploration of similarity patterns in spatiotemporal ensemble domains. Our approach involves partitioning the data domain and then constructing a graph data structure to represent subregions with similar ensemble values for visualization. The graph-based user interface design with multiple linked views enables the efficient spatiotemporal explanation of the data. Two application cases are demonstrated: regional emission influences and ensemble lock-exchange flow simulation. With emission simulation data, *EnsembleGraph* enables scientists to evaluate and compare regional anthropogenic emission influences on global tropospheric  $O_3$ .

We plan to develop more domain-specific algorithms and a customized interface for ensemble domain partitioning, in order to support more applications that incorporate ensemble simulations. Extending the framework to support large scale-data is also our future work: The clustering algorithm for each time step is easily extendable to a parallel environment or Map/Reduce framework. We can also compress large-scale data by grouping similar time steps and downsampling the spatial resolution. We envision that our new visual analysis framework will be adopted for visual analysis of a wider range of spatiotemporal data in different domains.

## ACKNOWLEDGMENTS

This work is supported by NSFC No. 61170204, NSFC Key Project No. 61232012, and the Strategic Priority Research Program - Climate Change: Carbon Budget and Relevant Issues of the Chinese Academy of Sciences Grant No. XDA05040205. This material is also partially based upon work supported by the U.S. Department of Energy, Office of Science, under contract number DE-AC02-06CH11357.

## REFERENCES

- [1] O. S. Alabi, X. Wu, J. M. Harter, M. Phadke, L. Pinto, H. Petersen, S. Bass, M. Keifer, S. Zhong, C. Healey, et al. Comparative visualization of ensembles using ensemble surface slicing. In *VDA'12: Proc. Visualization and Data Analysis*, pages 82–94, 2012.
- [2] U. D. Bordoloi, D. L. Kao, and H.-W. Shen. Visualization techniques for spatial probability density function data. *Data Science Journal*, 3:153–162, 2004.
- [3] P.-T. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. Interactive exploration and analysis of large-scale simulations using topology-based data segmentation. *IEEE Trans. Vis. Comput. Graph.*, 17(9):1307–1324, 2011.
- [4] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3D ensemble visualization. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2694–2703, 2014.
- [5] L. K. Emmons, S. Walters, P. G. Hess, J.-F. Lamarque, G. G. Pfister, D. Fillmore, C. Granier, A. Guenther, D. Kinnison, T. Laepple, J. Orlando, X. Tie, G. Tyndall, C. Wiedinmyer, S. L. Baughcum, and S. Kloster. Description and evaluation of the model for ozone and related chemical tracers, version 4 (MOZART-4). *Geoscientific Model Development*, 3(1):43–67, 2010.
- [6] 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-96: Proc. International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.
- [7] L. Gosink, K. Bensema, T. Pulsipher, H. Obermaier, M. Henry, H. Childs, and K. I. Joy. Characterizing and visualizing predictive uncertainty in numerical ensembles through Bayesian model averaging. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2703–2712, 2013.
- [8] J. C. Gower and G. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64, 1969.
- [9] Y. Gu and C. Wang. TransGraph: Hierarchical exploration of transition relationships in time-varying volumetric data. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2015–2024, 2011.
- [10] H. Guo, X. Yuan, J. Huang, and X. Zhu. Coupled ensemble flow line advection and analysis. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2733–2742, 2013.
- [11] T. Höllt, A. Magdy, G. Chen, G. Gopalakrishnan, I. Hoteit, C. Hansen, and M. Hadwiger. Visual analysis of uncertainties in ocean forecasts for planning and operation of off-shore structures. In *Proc. IEEE Pacific Visualization Symposium 2013*, pages 185–192, 2013.
- [12] T. Höllt, A. Magdy, P. Zhan, G. Chen, G. Gopalakrishnan, I. Hoteit, C. Hansen, and M. Hadwiger. Ovis: A framework for visual analysis of ocean forecast ensembles. *IEEE Trans. Vis. Comput. Graph.*, 20(8):1114–1126, 2014.
- [13] M. Hummel, H. Obermaier, C. Garth, and K. I. Joy. Comparative visual analysis of Lagrangian transport in CFD ensembles. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2743–2752, 2013.
- [14] H. Jänicke and G. Scheuermann. Visual analysis of flow features using information theory. *IEEE Comput. Graph. Appl.*, 30(1):40–49, 2010.
- [15] D. T. Kao, A. Luo, J. L. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *IV'02: Proc. International Conference on Information Visualisation*, pages 219–226, 2002.
- [16] J. Kehler and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.*, 19(3):495–513, 2013.
- [17] X. Li, J. Liu, D. L. Mauzerall, L. K. Emmons, S. Walters, L. W. Horowitz, and S. Tao. Effects of trans-Eurasian transport of air pollutants on surface ozone concentrations over Western China. *Journal of Geophysical Research: Atmospheres*, 119(21):12338–12354, 2014.
- [18] S. Liu, J. A. Levine, P.-T. Bremer, and V. Pascucci. Gaussian mixture model based volume visualization. In *LDAV'12: Proc. IEEE Symposium on Large Data Analysis and Visualization*, pages 73–77, 2012.
- [19] S. P. Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.
- [20] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *VisSym'03: Proc. Symp. Data Visualization*, pages 29–38, 2003.
- [21] A. Luo, A. Pang, and D. Kao. Visualizing spatial multivalued data. *IEEE Comput. Graph. Appl.*, 25(3):69–79, 2005.
- [22] J. Ma, C. Wang, C. Shene, and J. Jiang. A graph-based interface for visual analytics of 3D streamlines and pathlines. *IEEE Trans. Vis. Comput. Graph.*, 20(8):1127–1140, 2014.
- [23] J. Ma, C. Wang, and C.-K. Shene. FlowGraph: A compound hierarchical graph for flow field exploration. *Proc. IEEE Pacific Visualization Symposium 2013*, pages 233–240, 2013.
- [24] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *SIGKDD 00': Proc. International Conference on Knowledge Discovery and Data Mining*, pages 169–178. ACM, 2000.
- [25] M. Mirzargar, R. T. Whitaker, and R. M. Kirby. Curve Boxplot: Generalization of boxplot for ensembles of curves. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2654–2663, 2014.
- [26] T. Nocke, M. Flechsig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *Proc. Simulation Conference 2007*, pages 703–711, 2007.
- [27] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *IEEE Comput. Graph. Appl.*, 34(3):8–11, 2014.
- [28] J. Poco, A. Dasgupta, Y. Wei, and W. Hargrove. Visual reconciliation of alternative similarity spaces in climate modeling. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1923–1932, 2014.
- [29] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. SimilarityExplorer: A visual inter-comparison tool for multifaceted climate data. *Comput. Graph. Forum*, 33(3):341–350, 2014.
- [30] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. *Comput. Graph. Forum*, 29(3):823–832, 2010.
- [31] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johnson. Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated ViSUS-CDAT systems. *Journal of Physics: Conference Series*, 180(1):1–5, 2009.
- [32] K. Potter, A. T. Wilson, P.-T. Bremer, D. N. Williams, C. M. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-Vis: A framework for the statistical visualization of ensemble data. In *ICDM'09: Proc. IEEE International Conference on Data Mining Workshops*, pages 233–240, 2009.
- [33] R. Samtaney, D. Silver, N. Zabusky, and J. Cao. Visualizing features and tracking their evolution. *Computer*, 27(7):20–27, 1994.
- [34] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1421–1430, 2010.
- [35] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Trans. Vis. Comput. Graph.*, 12(5):917–924, 2006.
- [36] D. Thompson, J. A. Levine, J. C. Bennett, P.-T. Bremer, A. Gyulassy, and P. P. Pébay. Analysis of large-scale scalar data using hixels. In *LDAV'11: Proc. IEEE Symposium on Large Data Analysis and Visualization*, pages 23–30, 2011.
- [37] C. Wang. A survey of graph-based representations and techniques for scientific visualization. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARS*, pages 41–60. The Eurographics Association, 2015.
- [38] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour Boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2713–2722, 2013.
- [39] W. Widanagamaachchi, C. Christensen, V. Pascucci, and P.-T. Bremer. Interactive exploration of large-scale time-varying data using dynamic tracking graphs. In *LDAV'12: Proc. IEEE Symposium on Large Data Analysis and Visualization*, pages 9–17, 2012.
- [40] L. Xu and H.-W. Shen. Flow Web: a graph based user interface for 3D flow field exploration. In *Proc. IS&T/SPIE Visualization and Data Analysis 2010*, pages 1–12, 2010.