EnsembleGraph: Interactive Visual Analysis of Spatial-Temporal Behavior for Ensemble Simulation Data

Qingya Shu [1] Hanqi Guo [3] Limei Che [1] Xiaoru Yuan [1,2] Junfeng Liu [4] Jie Liang [1,2]

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

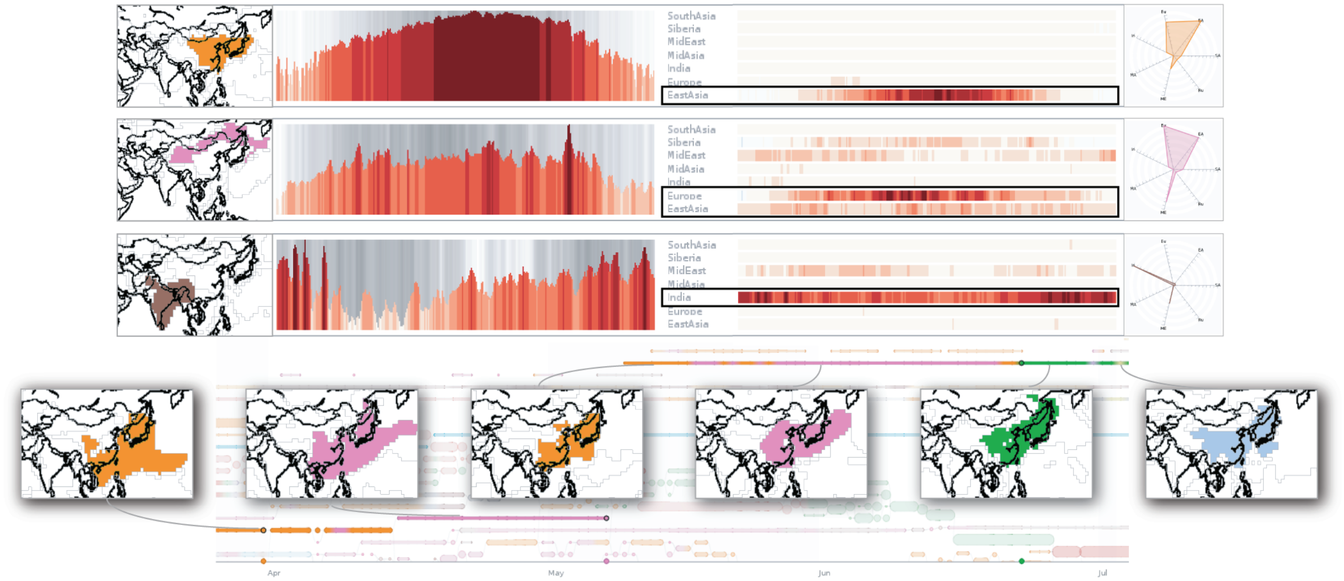


Figure 1: An example case of understanding influence of Eurassian Continent to the surface ozone over area around China. According to the similar behaviors between ensemble members over space and time, the neighbouhood is partitioned into three parts (left thumb nails in the first three rows): Eastern China, Southwest China, and Northwest China. Our novel graph-based interface provides abstraction to the grouped regions. Users can therefore navigate and track regions of interests over space and time. The last row shows tracking partitioned over South Eastern China using graph view and linked spatial view. Users highlight regions for further analysis in detail comparison view, where they compare values between individual runs, and behavior similarities between ensembles over different sub-regions (Charts in the first three rows).

Abstract

We present a novel visualization framework—*EnsembleGraph*—for analyzing ensemble simulation data, in order to help scientists understand behavior similarities between ensemble members over space and time. A graph-based representation is used to visualize individual spatiotemporal regions with similar behaviors, which are extracted by hierarchical clustering algorithms. A user interface with multiple-linked views is provided, which enables users to explore, locate, and compare regions that have similar behaviors between and then users can investigate and analyze the selected regions in detail. The driving application of this paper is the studies on regional emission influences over tropospheric ozone, which is based on ensemble simulations conducted with different anthropogenic emission absences using the MOZART-4 (model of ozone and related tracers, version 4) model. 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. Positive feedbacks from domain experts and two case studies prove efficiency of our method.

Keywords: ensemble simulation, graph visualization

1 Introduction

Ensemble simulations become prevalent in various scientific and engineering domains, such as aerodynamics, climate and weather research, to name a few. They are usually used to studying model sensitivities to parameters and initial conditions, quantifying uncertainties, etc. However, the visualization of ensemble datasets is a grand challenge, because ensemble data are usually multivariate, multivalued, time-varying, and with large data scales.

Our focus in this paper is the behaviors of ensembles—the similarities between individual runs in space and time. Currently, daily routine for scientists to analyze such data is merely based on manually selection and spatial temporal aggregation. First, a latitude-longitude box is arbitrarily defined as the target region to start with, and then they aggregate values along temporal dimension, e.g. seasonal or monthly average values. Second, they visualize and investigate spatial patterns of different ensemble members by plotting contour lines or pseudo colored maps. Line charts are plotted to compare the temporal differences between ensemble members. Nevertheless, there are a number of critical issues involved in the routine: First, without an overview, it is difficult to understand the overall patterns of the dataset 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 could be highly inhomogeneous in specific regions; To solve the complexity of spatial-temporal data, visualizing such kind of data is quite challenging but also rewarding, so that scientists can understand their scientific data more effectively.

In this work, we propose a visual analysis framework based on behaviors of ensembles. We quantify the behaviors as *behavior vectors*, using metrics that describe similarities between ensemble members in spatialtemporal location(detail explanation in Section 4). According to close discussion with domain scientists, we design visual analysis tools to support various tasks. the visual analysis tasks for such kind of data can be based on . Specifically, the tasks are

- T1: Partitioning ensemble domain based on ensemble behaviors.
- T2: Investigating the spatiotemporal distribution of behavior patterns.
- T3: Comparing the different behavior patterns.

In order to support the tasks, we design the framework with several components. First, we proposed an automatic ensemble domain partitioning method, We data partitioning over the ensemble dataset, in order to extract regions with similar behaviors. We then define *behavior patterns* as basic units for spatiotemporal aggregations of locations that have similar behaviors. Second, to support spatial temporal exploration of all behavior patterns, we use a graph-based user interface to give an overview of behavior patterns over space and time. Third, we provide tools to compare the behavior patterns, which can be used to validate the findings for the exploration.

The driving application in this paper is the impacts of regional emissions on the tropospheric ozone (O_3). Scientists conduct ensemble simulations under different emission scenarios, in order to understand and evaluate the regional impacts on ozone formation [14]. Tropospheric ozone is known as an important greenhouse gas, and it is also harmful to human health and agriculture production. They are formed from chemical reactions of nitrogen oxides, carbon monoxides, etc., which are mostly caused by human activities such as industrial and road emissions. These anthropogenic emissions, so called ozone precursors, are different around the world, due to local industrializations and environmental policies. Influence from those emissions is transported by wind convection, bringing a global atmospheric issue. Thus, tropospheric ozone is a mixed influence affected by all regional anthropogenic pollutant emission, yet the mixing weight from each source regions are not identical. For example, previous studies have shown that, ozone concentration over East China is mostly affected by domestic pollutant emissions due to the industrial prosperity. Meanwhile, Western China, which is less industrialized, has the opposite situation. The ozone is mostly formed from foreign emissions of upwind neighbors, such as India and Europe [14]. Analyzing and understanding the regional emissions impacts are important for scientists and decision makers to further expedite emission reductions. The application data is ensemble simulation based on Model of Ozone and Related Tracers, version 4 (MOZART-4). It consists of perturbation runs with different emission sources, and reference runs (detail explanation in Section 2.1); With such dataset, scientists would like to investigate the relative importance of different emission sources to regions in the ensemble domain. In order to support these driven tasks, we calculate behaviors according to the combination of influences from difference source emissions, and apply our frame work for visualization using novel graph-based interface. Two case studies and feedback from domain scientists shows usefulness of our methods.

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

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

We organize the remainder of this paper as follows. We first explain background of our driven application in Section 2.1, and review related work in Section 2.2, Section 3 gives overview to our approach. Section 4 describes data processing and graph construction. Section 5 is about visual design and interface. We then demonstrate cases and feedbacks in Section 6, and finally make conclusion in Section 7.

2 Background

We introduce our ensemble datasets and driving application, and summarize closely related work on ensemble data visualization, which has been more and more focused in recent years. In addition, we also review graph-based visualization techniques for scientific data sets.

2.1 Driving application

Scientists conduct perturbation experiment for evaluating sensitivities of ozone concentration to different regional emissions [14]. The simulation is based on MOZART-4. The input data of the model is from observation and the emission inventories, and the outputs are the concentration of a series of chemical species. In this work we focus on the ozone, the most important substance in the model and which dominants the chemical reactions.

The experiments include three types of runs, namely the base run, the globe run and the perturbation runs. The base run is conducted with real input data, and the globe run is conducted by switching off all anthropogenic emissions. The perturbation runs alternatively switch off simulate emission source from different regions in the world. Note that natural emissions still exist even if emission source regions is switched off. As shown in Figure. 2, seven emission source regions are defined, including: Europe, India, Mid-East, Southeast-Asia, East-Asia, Mid-Asia, and Siberia. In our frame work, we define member behaviors by the bias of the perturbation runs from base and globe runs.

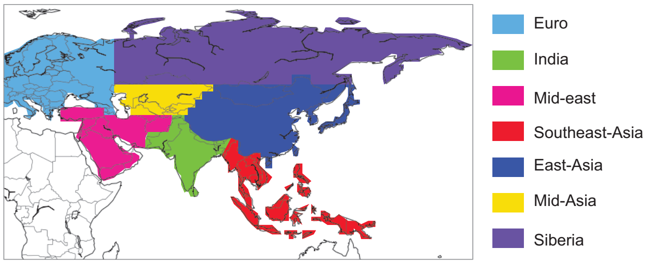


Figure 2: Seven source emission regions. During the ensemble simulation, emissions from one region are turned of, in order to calculate the relative importance of its regional emission influence on ozone.

Through the ensemble simulation, the response of tropospheric ozone concentration to different anthropogenic emission conditions can be measured. Conditionally, scientists choose a range box as region of interests to study, e.g., rectangular range over Eastern

China; Next they compare the ozone concentration of “East Asia” run with base run and global run in this area. Then they conclude from charts of maps and aggregated values, e.g., the ozone over Eastern China is mostly influenced by emission from East Asia area, indicating very high domestic emission in Eastern China (Figure 3). To make it more flexible for exploring all potential interesting features of regional influences, we compute the ensemble member behaviors and partition data domain using this measure. We deliberate work flow and the algorithms in the following sections.

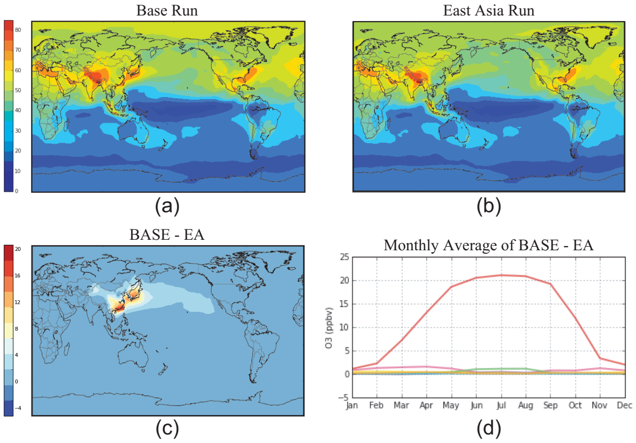


Figure 3: (a): The base run; (b): The run “East Asia”; (c): Their difference; (d): Monthly ozone average in an chosen area over Eastern China. This indicates that domestic emission influence over Eastern China is very high.

2.2 Related Work

Ensemble Visualization. Ensemble simulation data sets is usually multivalued, multivariate and time-varying, thus, it is very challenging to visualize [13]. Ensemble visualization inherited from uncertain visualization at the beginning, because each location has multiple values. Many studies took great efforts on fitting existing visualization techniques to support these complex data sets. One way is utilizing operators to convert multiple values to scalar for visualization, e.g., mean, standard deviation, or peak numbers of distributions [12,17]. These manipulations therefore made common visualization techniques applicable [16]: pseudo coloring, streamlines, pathlines or isosurfaces, etc. Another way is to visually embed uncertainty information into conventional visualization methods. For example, by overlaying uncertainty-encoded ribbons and glyphs over spaghetti plots, *Noodles* helped users to identify high uncertainty regions and outliers [27]. Sanyal et al. [31] extended boxplot by integrating important statistics (e.g., skew, kurtosis and histogram) to create visual signatures for data distribution. This hybrid *summary plot* largely reduced massive information while also highlighted salient features of ensemble data sets. Other works use distribution to describe data property on each location. Parameter-based or histogram-based approaches largely reduce the total ensemble data size, while at the same time keep the meaning of data and its visualization efficiency. For example, Liu et al. [15] used Gaussian mixture model to fit data distribution on each location, *Hixel* [33] stored histograms on each pixel as a novel data representation.

In order to reduce massive information while at the same time emphasize feature characteristics in ensemble data, some previous works utilized clustering in their visualization pipeline. Bordoloi et al. [2] have conducted clustering algorithm for both locations and ensemble realizations: Grouping regions with similar distributions can present overall distribution information, meanwhile grouping realizations with similar data outcomes can help understanding relationships between members and simulation conditions. Another example is automatic clustering locations with similar distributions in *Multi-Chart* by Demir et al. [4]. By doing so, it detected regions with similar member distributions. Our work also used clustering. However, the difference from the aforementioned methods is that during clustering procedure we consider distance from each members, instead of overall distributions. To achieve our goal, we aim to compare relative importance of each perturbation runs. Only counting overall distributions will cause information loss of the individual ensemble members.

Many works provide visual analytics for ensemble data sets. Nocke et al. proposed various visual representations to different types of visual comparison tasks for climate simulations [21]. For example, slice-based volume visualization for inner simulation comparison, pseudo colored isosurfaces for inter simulation comparison, and graphical table and pixel based visualization for global comparison. They integrated them in their *SimEnvVis* framework. A recent taxonomy for ensemble data comparison divides existing approaches as location-oriented comparison and feature-oriented approaches [22]. The first type, location-oriented method, conducts data comparison by attributes at fixed locations in ensemble domain. An example is to visually compare statistical measures on each location (e.g. means and variances) to indicate disagreements between members [17, 29]. *Multi-Charts* [4] maps 3D field to 1D using Hilbert space filling curves, and then enables comparison for region distributions using line char and bin chart techniques. Gosink et al. [5] do classification for data on each location point, according to member distribution and ground truths. For ensemble flow field, which is not applicable by traditional scalar field based methods, Lagrangian-based measurements are proposed to evaluate differences between field lines starting at same locations [7]. Individual and joint transport variances are also introduced to visualize agreements or disagreements in ensemble flow fields [10]. The second type of ensemble data comparison is feature-based approach. These methods first extract features from individual runs, and then compare those extracted features. One example is to render their isosurfaces in a slice-by-slice style to ease visual comparison [1]. Sanyal et al. [31] used spaghetti plot to simultaneously display multiple isocontours, and they further visualized the ensemble uncertainty by drawing special-designed glyphs and ribbons encoded by uncertainty metrics. Contour boxplot [35] and curve boxplot [20] generalize functional boxplot to visualize spatial distribution of contours. They are excellent examples for visualizing quantitative properties for contour ensembles. In this paper we calculate the ensemble behaviors at each location for comparison. So our visual analysis framework tends to belong to the location-oriented approaches. However one of our differences is that during the visual analysis workflow, the basic unit for spatiotemporal exploration is the grouped behavior patterns.

Increasing amount of visual analysis systems support comparison and analysis approach for ensemble data sets through visualization interface. Well designed interactions or powerful data mining techniques are used to gain a good understanding to ensemble data. For example, *Ensemble-Vis* [29] and *ViSUS/CDAT* [28] system combined multiple linked views for visual analysis to ensemble data sets, in order to take the advantages of each. Piringer et al. provided an interactive system [24] which combines multiple previews with detailed windows for hundreds of 2D function ensemble outcomes. *SimilarityExplorer* [26] enables inter-comparison and similarity analysis for model structure and model outputs. Another work from Poco et al. [25] lets users interactively group climate simulation models using multiple similarity criterias. *Ovis* [8,9] by Hölt et al. is a visual analysis framework for ocean simulation data. It provides interactively spatial temporal exploration for off-shore structures, and analysis for features of interests. Our work provides visual analysis for ensemble simulation data by MOZART-4 model as our driven applciation. The driven application aims at understanding regional emissions in MOZART-4 simulation. We support exploration for all potential interesting regions in the data domain, based on similarities of behavior between ensemble members.

Graph-based Methods in Scientific Visualization. Graph visualization applied in scientific data analysis is a new trend [34]. By abstracting features from scientific data to graph models, it could help users to gain better navigation and understanding over the complex data. It is usually occlusion-free and more intuitive to explore graphs in 2D than traditional 3D visualization methods. For example, TransGraph [6] maps time-varying volumetric data into 2D plane using graph visualization techniques. Sauber et al. [32] use graph to visualize the relationships between variables of multi-field datasets. Bremer et al. [3] and Widanagamaachchi et al. [36] use graph methods to show features evolving in large scale time-varying simulation datasets. Similar techniques are also used in flow field data, e.g. FlowGraph [18, 19] and FlowWeb [37] show relationships of field lines and data blocks, which provide flexible data navigation and query interface. Janicke and Scheuermann [11] used graph representation to visualize features in time-varying volume data and their transitions. Our work develops a novel approach, which creates a new graph-based method to explore and understand spatial temporal data,

3 Workflow

In this section we describe our user exploration work flow to our visual analysis framework, and also the pipeline to our system.

During the exploration work flow, we provided user an overview to partitioned ensemble domain. It consists three parts: the spatial view, the temporal view, and the detail comparison view. The spatial view shows how ensemble domain is partitioned into subregions with similar ensemble behaviors. The temporal view shows the occurrences of each sub-regions over time. The detail comparison view lets user highlight and compare partitions.

In the spatial view, we use colored map to show partition results. User observes spatial distributions of grouped sub-regions, and chooses the interesting ones for analyzing. This lets user get initial overview for ensemble behaviors, which cannot be achieved by manually probing rectangular areas.

In temporal view, our interface which uses graph visualization techniques gives abstraction to all sub-regions over space and time. We choose graph as visual representation metaphor because its powerful data visualization ability in 2D plane. Each node in graph represents a spatiotemporal behavior pattern in ensemble domain. Node color encodes the behavior pattern. The graph is plotted in a streaming style from left to right, so that user could easily track behavior patterns over time.

We enable linked interactions between the two visual components to support spatial temporal navigation for ensemble data. User tracks regions in the temporal view, and relates counterpart partitions in spatial view. After interaction, partitions that are not highlighted will fade out, so that the regions of interests are emphasized. The detail comparison view then compares relative importance from source emission regions on these highlighted sub-regions, including the total anthropogenic influences, the natural influences, and the individual influences.

In general, a complete user navigation includes, getting overview using abstraction in temporal view, choosing sub-regions of interests by relating them in spatial, and further analysis through comparison view. The third row in figure 4 illustrates this process.

Pipeline of our system contains two parts: the domain partitioning, and the region connection. Figure 4 illustrates the pipeline. We partition ensemble field based on calculated ensemble behaviors, in order to summarize all behavior patterns in ensemble domain. In each location, we calculate the behavior vector, by aggregating all perturbation runs and the base run. Then we assign all behavior vectors into several groups by k-means clustering algorithm, and group regions together that are assigned with similar behavior vectors. By doing so, we partition the ensemble field into continuous regions. Such spatial partitioning method is executed for each single time steps.

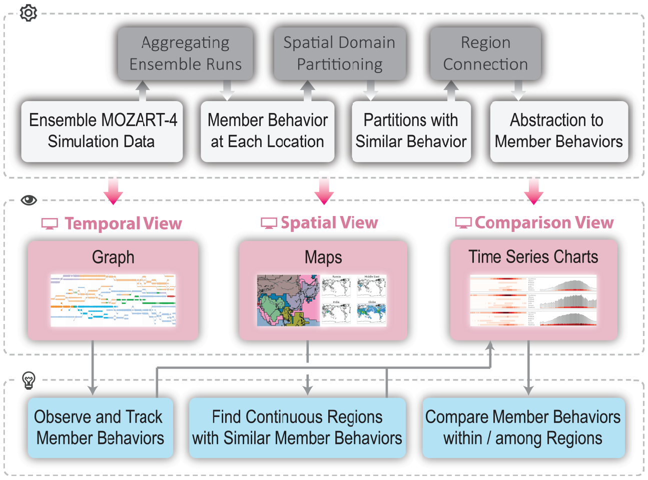

Aggregating Ensemble Runs
Spatial Domain Partitioning
Region Connection
Ensemble MOZART-4 Simulation Data
Member Behavior at Each Location
Partitions with Similar Behavior
Abstraction to Member Behaviors
Temporal View
Spatial View
Comparison View
Graph
Maps
Time Series Charts
Observe and Track Member Behaviors
Find Continuous Regions with Similar Member Behaviors
Compare Member Behaviors within / among Regions

Figure 4: Overview to our visual analysis framework. The first row is data preprocessing part: we do partitioning and region tracking for 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; The detail comparison view visualizes emission influences of highlighted sub-regions for validation. The third row is the exploration flow.

We connect the continuous regions over time to obtain spatial temporal summarization to all behavior patterns. For doing so, we establish connection between regions in neighboring time steps, by detecting their spatial overlapping. Regions that appear in close location are connected together as related regions. This helps to track the behavior pattern change over time. The summarized data structure is submitted to the front-end interface, which is discussed in the next section.

4 Domain Partitioning based on Ensemble Behaviors

In our visual analysis framework, we first calculate the ensemble behavior for every spatiotemporal location, and then group subregions by this behavior, over space and time. A graph data structure is then constructed by the temporal connectivity between them for further visualization.

4.1 Ensemble Behavior Definition

We quantify the ensemble behavior by *behavior vectors*, The behavior vector $\mathbf{v}$ is defined as a n-dimensional vector for each spatiotemporal location $\mathbf{x}$:

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

where n is number of non-base runs, and $d_i(\mathbf{x})$ represents *behavior* of i-th ensemble member. For the driven application, our definition to *behavior* $d_i(\mathbf{x})$ is 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:

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

where $\hat{C}(\mathbf{x})$ and $C_i(\mathbf{x})$ are values of the base run and the i-th run, respectively. Thus, each spatiotemporal location has a high-dimension 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 have similar behaviors. In our driven application, this means that ozone over two places is influenced by a similar combination of emission sources. If the **v** is similar over a continuous region, we call it a *behavior pattern*.

4.2 Spatial Domain Partitioning for Ensemble Data

In order to give description to how behavior patterns distributes in ensemble domain, we do classification for all behavior vectors. *k*-means clustering is one of the most commonly used methods for vector quantization. It assigns high-dimensional data points to several groups. The clustering result keeps points in the same groups close to each other, while points in different groups distinct from each other. The *k*-means clustering algorithm starts from *k* randomly selecting seed points as centers of the groups, and assigns all data points to the nearest groups. It then finds new centers for each group, and executes another iteration until the assignment becomes stable. We execute *k*-means clustering for behavior pattern vectors, and use this result to label the location, so that locations with the same labels have close behavior patterns. Such classification step is repeated for each single time steps.

Two essential factors influence *k*-means clustering results: *k* value selection, and initial seed selection [23]. Too large *k* value will cause over partitioning, while too small *k* value will blur classification outcome. In our implementation, we set a initial *k* and let user change this value to launch new preprocessing pass for refinement. Since our clustering algorithm does not have a ground truth for validation, partitioning method is acceptable if the result shows close member values over same sub-regions. In our future work, we plan to introduce domain specific algorithms to let the partitioning method more correlated to application requirements. As for initial seeds, random selection causes unsatisfying result: too close seeds make clusters overlap. In order to keep initial seeds far away from each other, our solution refers labeling result from last time step as initial input for the current *k*-means procedure. This makes clusters more distinct, meanwhile improves the consistency of results for the neighboring time steps.

We then group locations with same clustering labels to detect continuous regions with similar behavior vectors. Therefore we obtain partitioned result for ensemble domain. For each partition, we use the centroid of the behavior vectors in this region as representative behavior for this region.

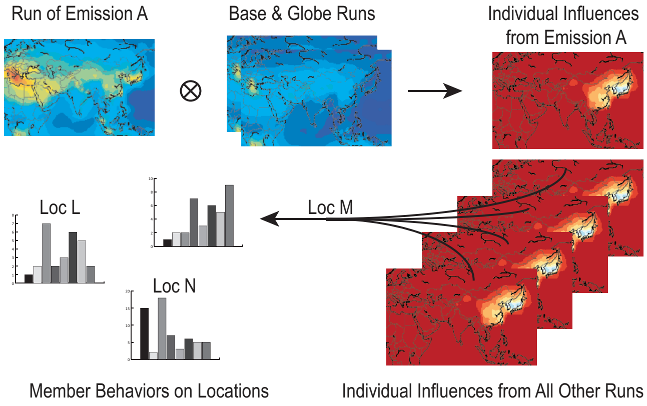

Run of Emission A
Base & Globe Runs
Individual Influences from Emission A
Loc L
Loc M
Loc N
Member Behaviors on Locations
Individual Influences from All Other Runs

Figure 5: Partitioning ensemble domain according to the behavior vectors on each location. Firstly it calculates the on each location using the difference between the base run and the perturbation runs, and quantifies the behavior pattern of each location using a vector, then clusters for all behavior vectors, and uses the result for spatial domain partition.

4.3 Partitioned Region Connection

Regions with similar patterns are summarized for each frame by classification and region grouping. Those patterns may change their location and values over time. In order to track those changes, we correspond regions over time by feature tracking methods. The algorithms of feature tracking have been well studied over decades. One effective way for tracking related features is to correspond neighboring features according to their spatial intersection size [30]. In our approach, connections between two regions are established as long as they have a certain amount of spatial overlap. We detect all connections from the beginning time to the end, and then obtain a graph structure. Regions are regarded as nodes, links between nodes represent their connectivity. The result is an acyclic directed graph, direction means time increasing. Following along nodes in a path means tracking the ensemble behavior of sub-region over time.

We run these procedures by server side. A control panel interface provides parameter setting access for user at front-end. Adjustable parameters include: *k* value for *k*-means clustering, the region size threshold R_r and the overlap size threshold R_o. Regions with rather small sizes less than R_r are skipped so as to avoid uninteresting noisy regions. We set initial R_r as S/k, S is the size of spatial domain. Users can tune it down to tell server side to include more smaller regions. The overlap size threshold R_o is for region connectivity, and is initialized as $min(R_a, R_b)/2$. Lower down this value will get more connectivity between neighboring partitions. The server process keeps running in the background, updates the new data structure and replies requests from the frond-end interfaces, which we discuss in the next section.

5 Visualization and Interaction Design

The interface of our work has three main components: temporal view, spatial view, and comparison view. The temporal view provides a visual summarization of all regions with similar behaviors over time of ensemble members. The spatial view shows the domain paritioning results. The comparison view contains detail visualization for highlighted sub-regions. The design of these components are introduced in the following subsections.

5.1 Temporal View

In temporal view, a graph-based interface (*EnsembleGraph*) provides overview of behavior patterns of ensemble members. In the *EnsembleGraph*, the vertices indicate the behavior patterns in each time frame, and the edges are the temporal connectivity of difference behavior patterns. In order to show information of counterpart behavior patterns, we use special graph layout and color appearances for this graph interface.

The layout of the graph follows several principles. First, the nodes are aligned with their time of occurrence, thus the layout is in a "streaming" style. Second, edge crossings are reduced through computation as much as possible, in order to enhance readability. Third, the paths should be as straight as possible, if they have no branches. In our implementation, we use *dot* algorithm to achieve above goals. As the input graph structure for dot interface, nodes occurring at the same time are set with same rank values, so that they are horizontally aligned. Edges on non-branch paths are marked by higher weights. The non-branch paths are straight. We also wrap these non-branch paths, in order to emphasize linked sub-regions, for easier tracking. Sizes of the nodes are proportional to the counterpart region sizes. Node colors are assigned according to their behavior patterns.

By clicking on nodes, users can highlight counterpart sub-regions. Dashboard on the right side shows all highlighted nodes in star glyph, depicting behavior of each ensemble member over this sub-region. In this application case we use seven axes to represent influences of emission from corresponding number of sources.

Each highlighted node is related to a polygon. The position on each axis represents how much influence introduced by the source emission. In this case, user can easily compare influence differences among the highlighted nodes.

The temporal view not only displays behavior patterns and their connections, but also can serve as an entrance for exploring ensemble of sub-regions of interests. Once being highlighted, the corresponding sub-regions will also displayed in other views, showing their location, origin ozone concentration and detail comparison.

5.2 Spatial View

Spatial view (Figure 6) provides visualization for original values and the domain partitioning results. We provide data mode, partition mode, and selection mode in this view. Data mode shows actual ozone concentration for all ensemble members. Users choose ensemble members, and explores their origin values (in blue-yellow-red color map), or influence values (in blue-white-red color map, representing difference between each individual run with the base run). Users drag the slider bar to change current time step in display, and effectivey obtain animation of maps change over time. Partition mode shows sub-regions with colored segments. Selection mode enables editing and submitting regions of interests. Users create sub-region by double-clicking a region on map, and then submit to the server side. The targeted sub-region will also be broadcasted to graph views and comparison views, updating the current exploration status.

In a typical analysis process, scientists select rectangular regions by hands. Our automatically partitioned sub-regions provide region selections not limited to rectangle shapes. Such partitioning method is more flexible and informative from the view of behavior. Besides, it is more efficient than manually choosing destination regions if not knowing where to start observing.

5.3 Comparison View

In our visualization, we design a view for detail values of ensemble members over highlighted sub-regions (showing in the upper left in Figure 6). It is desirable to provide capability of comparison of both inside one region, and among different regions. This comparison view is designed in list style, with each list item containing an element for one highlighted sub-region. User can extract detailed information from each item, and make conclusion through exploration and comparison.

Figure 6 shows the interface. Thumbnails on the left side indicate corresponding sub-regions. The same color scheme is used as that in the temporal view and spatial view. The behavior patterns from seven regions is are encoded by a small horizontal bar chart encoded and shown on the right side of the view. Those bare charts in the middle show original data values or aggregated values in this highlighted sub-region. Following the basic analysis process of domain scientists, Our visualization system provides three viewing modes in this part: Natural-anthropogenic mode compares natural influence and anthropogenic influence; Domestic-foreign mode compares emission influences from one area with ones from other areas; 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 plot the anthropogenic influence (defined by difference of the base run and the globe run) as the foreground. This lets users gain knowledge about overall ozone concentration and the influence fraction by human activities. User could also switch the foreground chart to globe run, to focus his/her interests over the natural influence. We apply a white-to-blue color scale for globe run values to indicate ozone concentration without any anthropogenic influences. The domestic-foreign mode visualize the domain-foreign ratio of emission influences, which is calcualted by dividing domain emission influences (differences between base run and individula runs) by foerign emission influences (differences between individual runs and globe runs). We use red-to-blue color map for this ratio as foreground, and white-to-gray color map for total anthropofenic influences in the background. The individual influence mode visualizes temporal distribution of influence from all source regions. In the pixel based table style visual representation, each row is one chart for corresponding source region, while each column represents corresponding time step. Color of red encoding positive values and blue encoding negative values.

Our system is implemented in C++ with OpenGL and Qt libraries. We separately run graph interface remotely on a work station, and run server side on a cluster machine which is handling data storage with more powerful computation capabilities.

6 Results

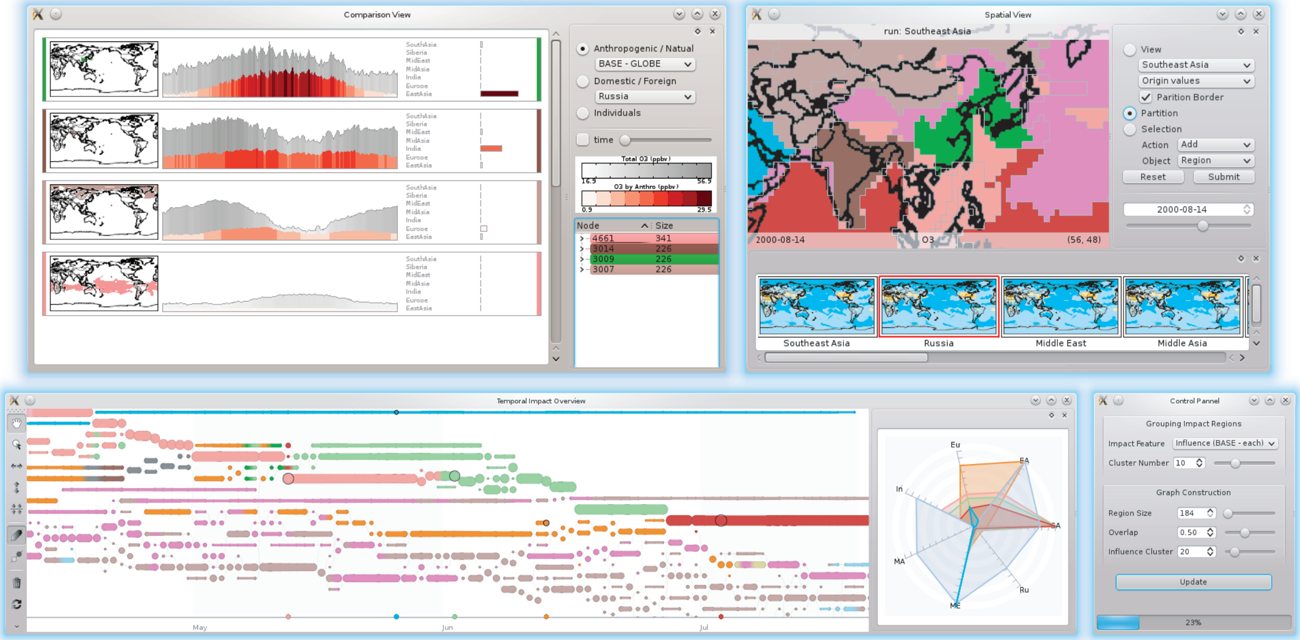
This section demonstrates two case studies with *EnsembleGraph*, identifying emission influences of tropospheric ozone over China in Case I, and investigating spatial patterns in Southern Hemisphere in Case II. Both case studies use daily output from nine simulation runs from MOZART-4 model, in year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

6.1 Case I: Observing and Comparing Relative Regional Emission Influences of Tropospheric Ozone over China

Scientists would like to analyze how ozone over China is influenced by anthropogenic emissions from the Eurasian Continent. Specifically, they intend to find out which places are more influenced by domestic emissions, and which places are influenced by foreign regions, and how influence changed over time.

Instead of manually probing a lat-lon box as target range for analysis, our tool enables user to directly observe how this area is partitioned into continuous regions with similar member behaviors. With *EnsembleGraph*, scientists explore the spatial view with panning and zoom into the area around China and review the partitioning results. The partition map (in Figure 1 and Figure 6) shows that, area around China is divided into three parts according to member behaviors. The first one is the East China (the first row in Figure 1), which also includes neighboring regions such as Japan and Korean Peninsula. The second one is the Southwest China (the second row in Figure 1), being connected to India. The last one is the Northwest China (the third row in Figure 1), connecting to Middle Asia and Russia. Such partitioning result agrees with the geographical terrain of China: Southwest China is Qinghai-Tibet Plateau, and Northwest China is separated from Eastern China by mountains. The automatic results are similar to the manually chosen areas in previous research work [14], in which the researchers chose Xinjiang Province and Qinghai-Tibet Plateau as two lat-lon boxes (40°N-45°N, 84°E-90°E over Xinjiang and 29°N-34°N, 86°E-92°E over Tibet). *EnsembleGraph* successfully helps scientists to choose destination area for analyze, based on automatic partitioning method.

We choose one node for each region, and observe relative emission influences in comparison view. From the results (Figure 1) we find that in Northwest China, most anthropogenic influences appear in spring and summer, but in Southwest China the anthropogenic influences almost lasts for the whole year. Meanwhile in Eastern China, influences happen mostly in summer and almost disappear in winter. Switching to individual influence mode (Figure 1) allows comparison for temporal distribution of each regional emission. In Northwest China (second row), Europe is the dominant emission source region in spring and summer, followed by East Asia. Emission influences over Southwest China (the third row) appear totally differently: It is largely affected by India throughout the whole year, and occasionally affected by Middle East during Spring, while the



● Anthropogenic / Natual
BASE - GLOBE
○ Domestic / Foreign
Russia
○ Individuals
□ time
Total O3 (ppbv) 16.9 – 56.9
O3 by Anthro (ppbv) 0.9 – 29.5

Node	Size
4661	341
3014	226
3009	226
3007	226

○ View
Southeast Asia
Origin values
☑ Paritions Border
● Partition
○ Selection
Action: Add
Object: Region
Reset
Submit
2000-08-14

Grouping Impact Regions
Impact Feature: Influence (BASE - each)
Cluster Number: 10

Graph Construction
Region Size: 184
Overlap: 0.50
Influence Cluster: 20
Update
23%

Figure 6: The visual analysis framework includes interface: Bottom left: Temporal view. Bottom right: Control panel for parameter settings. Upper right: Spatial View. Upper left: Comparison view. In this case, we focus on ozone over China and neighboring regions.

East China is almost only influenced by East Asia. The findings above agree with the ones in the previous studies [14].

6.2 Case II: Observing Spatial Patterns in Southern Hemisphere Based on Different Regional Emission Patterns

This case shows how scientists use *EnsembleGraph* to study spatial patterns over the Southern Hemisphere by different regional emission patterns. Firstly, by exploring the temporal view using panning and zooming, scientists could discover an obvious node chain that stays throughout the whole year. According to the size, it seems related to a very large region. Partition map in the spatial view shows that it exactly belongs to the largest segment covering the whole Southern Hemisphere. It explains that majority parts of southern hemisphere have similar ensemble behaviors. By double clicking on the Southern Hemisphere partition, and submiting this selected to server side, they can filter out all other unrelated regions, and only left a trunk with several branches in the graph, which indicates that the Southern Hemisphere could be separated into a whole regions or separate into two or three regions according to the different member behaviors. The counterpart subregions are shown in the partitioning map in Figure 7 (thumbnails in the left side).

If we watch them closely in the comparison view, we can find their temporal member behavior differences: Most southern and most northern areas of these three subregions have higher ozone concentration in January and December. However, through the natural-anthropogenic comparison mode (Figure 1) we found that, ozone over the most southern subregion is caused by natural emissions, while ozone over the latter subregion is affected by anthropogenic emissions (see Figure **??**. The most northern strip, although appears to have lower total ozone concentration, keeps suffering from a relative higher anthropogenic emission influences during the whole year. The individual influence comparison mode shows relative importance: for the most northern sub-region, South Asia emission has been the dominant source throughout the year, followed by East Asia and Middle East during Summer in Northern Hemisphere. The other two sub-regions in the south are similarly more affected by South Asia, East Asia and Siberia during the Summer time.

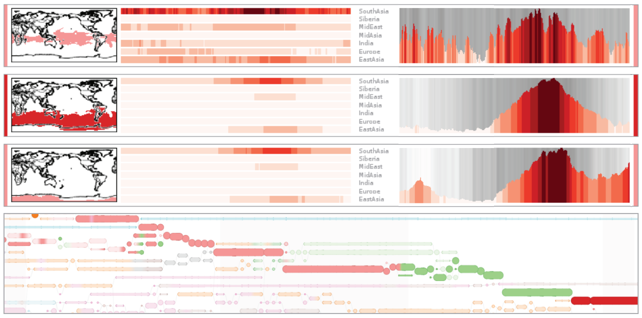


Figure 7: User track regions with similar ensemble behaviors in Southern Hemisphere. In the last row, unrelated nodes fade out, user highlight region by clicking on nodes, and compare them in detail comparison view. Here shows when Southern Hemisphere is divided into three regions according to ensemble behaviors (thumbnails on the left side). User observe differences using individual comparison mode and domestic-foreign comparison mode in interface (third row).

6.3 Domain Scientists Review

We discussed with our domain scientists and data provider on our results and got feedbacks from them . The scientists showed highly interests in the partitioning results, and explained the discovered member behaviors according to their knowledge. For the first case, they confirmed our findings, Eastern China is significantly influenced by its domain emission, while Northwestern and Southwestern China are isolated from those emissions due to mountains and plateaus, thus is more affected by the upwind areas. For the second case, the partitioning result is explained as follows: The middle part is located in the westerly of Southern Hemisphere, which almost has no land as obstacles, thus exists strong wind convection. This leads similar air around this latitude, and also isolates the air above the Antarctic Continent, which is exactly 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 into our framework. In general, the scientists stated that the partitioning method gives reasonable results, and it is effective to analyze by automatically dividing regions. At last, the scientists suggested us to provide more specific operators and flexible manipulators for controlling measurement during the partitioning procedure. We plan to include this in our future work.

7 Conclusions and Future Work

EnsembleGraph presented in this paper provides a visual analysis framework for ensemble data. The goal of this work is to provide interactive exploration for behavior patterns in spatiotemporal ensemble domain. Particularly, with emission simulation data, EnsembleGraph supports scientists to evaluate and compare regional anthropogenic emission impacts on global tropospheric ozone. The new approach develops the impact-based method, which automatically does data partition to whole domain, so that the impact patterns are concluded. A seamless design of graph based interface with linked interaction enables spatio-temporal explration efficiently for all impact patterns, and further allows users to compare them. Case studies shows *EnsembleGraph* is capable of facilitating scientists to understand the spatio-temporal distribution, as well as compare emission impacts efficiently. In the future, we plan to develop more domain specific algorithms and customized interface for ensemble domain partitioning, in order to support more application requirements to understand ensemble data. With futher improvements, this new visual analysis framework is envisioned to be adopted into a wider range of spatialtemporal data in different domains.

Acknowledgements

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

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*, page 8294, 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. 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.

[6] 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.

[7] 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.

[8] 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. Pacific Visualization Symposium*, pages 185–192, 2013.

[9] 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.

[10] 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.

[11] H. Jänicke and G. Scheuermann. Visual analysis of flow features using information theory. *IEEE Comput. Graph. Appl.*, 30(1):40–49, 2010.

[12] 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.

[13] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.*, 19(3):495–513, 2013.

[14] 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.

[15] 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.

[16] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *VisSym'03: Porc. Symp. Data Visualization*, pages 29–38, 2003.

[17] A. Luo, A. Pang, and D. Kao. Visualizing spatial multivalue data. *IEEE Comput. Graph. Appl.*, 25(3):69–79, 2005.

[18] 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.

[19] J. Ma, C. Wang, and C.-K. Shene. FlowGraph: A compound hierarchical graph for flow field exploration. *Proc. Pacific Visualization Symposium*, 8174:233–240, 2013.

[20] 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.

[21] T. Nocke, M. Flechsig, and U. Bohm. Visual exploration and evaluation of climate-related simulation data. In *Proc. Simulation Conference 2007*, pages 703–711, 2007.

[22] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *IEEE Comput. Graph. Appl.*, 34(3):8–11, 2014.

[23] D. T. Pham, S. S. Dimov, and C. Nguyen. Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1):103–119, 2005.

[24] H. Piringer, S. Pajer, W. Berger, and H. Teichmann. Comparative visual analysis of 2D function ensembles. *Comput. Graph. Forum*, 31(3.3):1195–1204, 2012.

[25] 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.

[26] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. SimilarityExplorer: A visual intercomparison tool for multifaceted climate data. *Comput. Graph. Forum*, 33(3):341–350, 2014.

[27] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. *Comput. Graph. Forum*, 29(3):823–832, 2010.

[28] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johhson. 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(012089):1–5, 2009.

[29] 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.

[30] R. Samtaney, D. Silver, N. Zabusky, and J. Cao. Visualizing features

and tracking their evolution. *Computer*, 27(7):20–27, 1994.

[31] 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.

[32] 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.

[33] D. Thompson, J. A. Levine, J. C. Bennett, P.-T. Bremer, A. Gyulassy, V. Pascucci, and P. P. Pébay. Analysis of large-scale scalar data using hixels. In *LDAV'11: Proc. IEEE Symposium on Large Data Analysis and Visualization*, pages 23–30, 2011.

[34] C. Wang. A survey of graph-based representations and techniques for scientific visualization. *Comput. Graph. Forum*, 2015.

[35] 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.

[36] 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.

[37] 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 2010*, page 75300, 2010.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under contract number DE-AC02-06CH11357.