Visual Analytics of Large-Scale Climate Model Data

Pak Chung Wong[1,*], Han-Wei Shen[2], Ruby Leung[1], Samson Hagos[1], Teng-Yok Lee[2], Xin Tong[2], Kewei Lu[2]

[1]Pacific Northwest National Laboratory [2]The Ohio State University

ABSTRACT

This application paper presents a visual analytics tool designed to explore large-scale scientific data modeled after a natural climate phenomenon. The data are modeled on a high-performance computer and exported to a personal computer for interactive visualization. The system is co-designed by visual analytics researchers and domain scientists after a year of rapid prototyping and evaluation of multiple information and scientific visualization techniques using a model dataset that includes both scalar fields and flow fields. Five information-visualization and one scientific-visualization techniques are included in the visual analytics system to balance analytical effectiveness and computation time for large-scale interactive exploration. The paper discusses the system design, explains the design rationale, and shares computation performance and results of different visualization techniques. The primary contribution of this application paper is to show that we can interactively and effectively visualize a large amount of scientific model data on a modest desktop computer. The computation performance results of the individual visualization techniques and the overall system also provide benchmark references for other large-scale visualization development efforts.

Keywords: Visual analytics application, large data analytics and visualization, climate analytics, scientific modeling.

1 INTRODUCTION

This application paper presents a large-scale visual analytics tool co-developed by visual analytics researchers and climate domain scientists to analyze large-scale scientific model data. Among the design and analytical challenges of the development are the size of the underlying data and the associated impacts to large-scale climate model analytics.

The design of our tool uses a multifaceted visualization approach that integrates multiple visualization techniques through interactive brushing and linking. Because our coordinated-visualization design uses multiple techniques to show different aspects of the dataset, the number of display values stored in different data structures in the memory for optimal visualization can be multiple times larger than the number of actual data values. For example, when we analyze hundreds of millions of floating point numbers in our discussion, over two billion numeric values are navigated and displayed on a dual-panel screen during interactive exploration.

The large number of data values creates an obvious list of hurdles from display-pixel limitation to delayed system-response time. As we describe later, some of our visualization solutions that address the pixel limitation problems would actually create new computation requirements that further challenge the interactive time requirement.

The problem of climate modeling itself is indeed a much larger data problem than what we describe in the paper. Our work is a part of the U.S. Department of Energy (DOE) Exabyte Data Visualization and Management program [3] that promotes scientific discovery through analytical computation. We generate our high-resolution climate model data on a Leadership Computing Facility computer at the National Energy Research Scientific Computing Center (NERSC) [9]. Interactively visualizing a dataset on a supercomputer with hundreds of thousands of cores is theoretically possible but practically not feasible for the domain scientists. Visualizing a scaled-down version of an extreme-scale dataset on a desktop computer is one of the few viable options today. We further elaborate on this limitation and some of the other extreme-scale data challenges in [17].

The visualization techniques presented in the paper have been selected by the domain scientists from a series of both scientific and information visualization technique candidates through a year-long rapid prototyping effort. In the end, we include 3D scatterplot visualization, parallel coordinates, matrix of pairwise scatterplots, multidimensional scaling (MDS) scatterplot, a global visualization with particle tracer visualization, as well as additional profile visualizations to study physics and math of the model in the tool. Notably, information visualization techniques excluded are force-direct graph layout, heatmap, and many other pixel-filling-based visualization techniques. The selected techniques, while popular among information visualization researchers, have rarely been found in prevailing large-scale climate analytics tools such as those discussed in Section 2.1.

The primary contribution of this application paper is to show that we can interactively and productively visualize such a large amount of scientific model data using a set of prevailing and intuitive information visualization techniques on a modest desktop computer. It is uncommon to find a similar combination of visualization techniques to climate model data of this scale in the climate analytics community. Other contributions of the paper include the computation and analytical performance benchmark results that can serve as reference guides for other researchers when visualizing large-scale data on a desktop computer.

2 RELATED WORK

We highlight selected work found in large-scale climate, geo, and earth visualization areas.

2.1 Climate Modeling Visualization Tools

Visualization has played a critical role in climate change model studies for many years. There are a number of prevailing visualization tools in the public domain developed by various research groups; these tools are widely accepted by the climate modeling community.

Ncview [14] is a visualization tool customized for quick and simple netCDF data visualization. The tool supports mainly 2D techniques and line plots. The tool is used by many to visualize modest-sized climate model datasets. Ncview is developed at and supported by Scripps Institution of Oceanography.

Another popular climate visualization tool is GrADS [6]. Although it has limited capability in both computation and visualization, the lightweight system is popular for modest-sized dataset visualization. GrADS is developed at and supported by the Center for Ocean-Land-Atmosphere Studies (COLA).

VAPOR [15] is a newer and more powerful climate visualization tool that supports large-scale scientific visualization techniques for

* pak.wong@pnnl.gov

IEEE Symposium on Large Data Analysis and Visualization 2014
October 9–10, Paris, France
978-1-4799-5215-1/14/$31.00 ©2014 IEEE

the astro-geo-science computation fluid dynamic (CFD) community. Both VAPOR and a specialized language for climate visualization known as NCL [12] are developed at and supported by National Center for Atmospheric Research (NCAR). The NCL website [12] has a list of additional visualization tools available for climate visualization.

Ferret [5] is another prevailing visualization tool that supports interactive explorations of climate model datasets with three to four data parameters at a time. Among the unique features of the tool is a Mathematica-like mathematical expression language for calculation and analysis. Ferret is developed at and supported by the National Oceanic and Atmospheric Administration (NOAA)'s Pacific Marine Environmental Lab.

2.2 Geo- and Earth-Science Visualization Research

Individual visualization techniques described in this paper have also been used to study other geo- and earth-science problems. While not described as analytical tools like those discussed in Section 2.1, they represent cutting-edge innovations in applying visualization to geo- and earth-science related problems.

Andrienko and Andrienko [1] use Sammon Projection, which is a non-linear multi-dimensional scaling technique, to cluster geographical objects. Dransch et al. [4] present a visual analytics design study to analyze geo-science and geo-information models. Kehrer et al. [8] discuss the heterogeneous scientific data challenges in interactive visual analytics. Blaas et al. [2] use the parallel-coordinates technique to visualize hurricane models. Muigg et al. [11] also use the parallel-coordinates technique to visualize fluid dynamic data, which are commonly found in earth-science models such as ground water or atmospheric simulations. All these apply information visualization techniques to analyze scientific data.

3 Modeling, Analytics, and Computation

This section provides information on the scientific data model, analytical problem, and hardware platform involved in the large-scale visual analytics study.

3.1 Data Modeling and Pre-Processing

The climate model data used in this study are time-varying 3D curvilinear model data containing both scalar fields (such as temperature and moisture) and flow fields (such as wind) parameters. Generating the model data took about 250,000 CPU hours on a Leadership Computing Facility computer at NERSC [9]. For analysis on a desktop computer, the data were scaled down using multiple temporal and spatial compression schemes on the same computer before they were exported to the desktop computer. Only thirteen parameters—time, elevation, longitude, latitude, velocities (in upward, downward, north, south, east, and west directions), ice, vapor, and cloud moisture—are used in this visual analytics study.

Structure-wise, the climate model data contain 120 time steps of 3D volume (64×32×50 = 102,400) data for each of the 13 data parameters (i.e., 102,400×120×13 = ~160 million). Because our design uses a coordinated visualization approach with multiple visualization techniques that include 1) 3D volume display (1 time), 2) parallel coordinates (1 time), 3) pairwise scatterplot matrix with 5 parameters (15 times), and 4) a multidimensional scaling (MDS)-based scatterplot (1 time), our tool can interactively display and manipulate over 160 million×18 = ~2.9 billion numeric values stored in *different memory structures* during the exploration.

3.2 Analytical Problem

Although our system is designed for general climate model analytics, the paper uses a climate phenomenon known as the Madden-Julian oscillation (MJO) [10] for demonstration purposes. In regional climate studies, MJO is the primary mode of large-scale intraseasonal variability in the tropics. Our regional climate model [7] realistically simulates the two MJO episodes observed in Asia during the winter of 2007/08 for 120 days. Analysis of these two cases shows that instabilities and damping associated with variations in diabatic heating and energy transport work in concert to provide the MJO with its observed characteristics.

From a visual analytics point of view, MJO involves a sequence of very slowly time-evolving (time varying) climate features, which include certain patterns of convection, ice, cloud, and vapor moisture (weighted semantic combination), and move across an open geographic space (geospatial) around mid-atmospheric levels (spatial). We show later that no one individual visualization technique can fully address multiple analytical aspects of the underlying problem.

3.3 Desktop Computer Hardware

For the visual analytics study, we use a Dell Precision T7500 Quad-core Xeon workstation with 48 GB of memory running Windows 7. With the exception of data I/O, our C++/OpenGL/Qt based program is widely parallelized among the eight Xeon cores. We use a dual 30-in desktop panel display with 2×2560×1600 pixels in our study.

4 Visual Analytics of Large-Scale Climate Model

This section describes the design and development of our large-scale climate model visual analytics system.

4.1 Design Criteria, Requirements, and Rationale

We present a number of design criteria and requirements followed by a brief discussion of our solutions. Supporting *interactive response time* and *maximizing analytical capability* are at the top of our development priorities. Because the system is co-designed by both visual analytics researchers and domain scientists, balancing the time and analytics demands on a desktop computer often becomes a challenge itself.

Multifaceted Analytics: We mentioned earlier in Section 3.2 that the temporal, spatial, geo-spatial, and semantics facets of the analysis require us to use multiple techniques to address different analytical aspects. The multifaceted requirements demand multiple visualization techniques that coordinately support aggregations and reasoning of all different analytical perspectives—time, weighted semantic combination, geospatial, and spatial—simultaneously.

Interactive Brushing: Climate modeling scientists often study complicated physics and their compound effects. Being able to visually aggregate potentially "every" individual data entity is highly desirable and sometime critical for data exploration. Data brushing, by its interactive nature, mostly satisfies the aggregation demand. In our design, the scientists can brush any visualization, including data shown in 3D volumes, using mouse selections.

Information Visualization: Many scientific visualization techniques do not support interactive data brushing directly. The above visual brushing requirement leads us to extensive use of information visualization techniques, which often show discrete data values explicitly for effective data brushing and aggregation. However, we have found no simple alternatives that visualize flow field data without using scientific visualization techniques such as a 3D flow field particle tracer [13].

3D Time Series Visualization: MJO is a slowly moving localized feature that evolves over an extended period of time. After experimenting with different visualization options, we found that displaying a series of 3D volumes, which show the aggregated entities spatially in each volume and temporally in different volumes, is the most effective analytics technique in our

application. We have tried animation in our study but the attempt failed to capture the attention of the domain scientists. It was cumbersome for the scientists to interact with an animation and conduct multifaceted exploration all at the same time.

Semantic Brushing: Visual brushing can only brush data that are arranged contiguously in certain orders on a display. One option to overcome this limitation is to semantically change the arrangement of the data visualization layout. In our system, we represent the multivariate data records as "weighted" data vectors and then use MDS to generate a scatterplot that clusters similar data vectors together for weighted semantic combination brushing. This is a powerful but also computationally expensive visualization technique.

Binning and Aggregation: The number of display values in our study (~2.9 billion) is far more than the number of available display pixels (8 to 15 million). We use dynamic binning extensively to collapse neighboring values and subsequently speed up the drawing and visualization times. However, brushing binned data also causes additional computation requirements. Because we bin individual data parameters separately (for optimal parallelization), we have to put extra computational effort in determining if a record is selected.

Interactive Response Time: We first identify a set of working visualization techniques for effective climate analytics and then determine the implementation of the parallel algorithms that maximize the multithreading computation capability of the Quad Zeon processors. The input data size plays a critical role in determining the interactive response time requirement. We conduct a series of performance studies (as described in Section 6) to determine the optimal or preferred data size that our visual analytics tool can perform interactively. The result becomes a benchmark data size for our high-performance computing (HPC) model program to decide the amount of data exported to the desktop computer.

4.2 Scalable Visualization & Flow Field Discretization

We mentioned in Section 4.1 that we use binning to reduce the data size for visual analytics. But conventional binning approaches are normally applied to *discrete* scalars such as time or sensor readings. One option to analyze *continuous* flow field data using traditional information visualization techniques is to discretize the flow data and project it into a scalar data structure.

For example, Figure 1 shows three stages of progressive abstractions that transform a 3D flow field dataset (in Figure 1a) to a flow graph structure (in Figure 1b), then a set of numeric graph signature vectors [16] that characterize the connectivity of the flow graph in Figure 1c, and finally an MDS-based scatterplot that depicts the similarities of the flow graph signatures in Figure 1d.

The flow field discretization process shown in Figure 1 has multiple advantages for big data visual analytics:

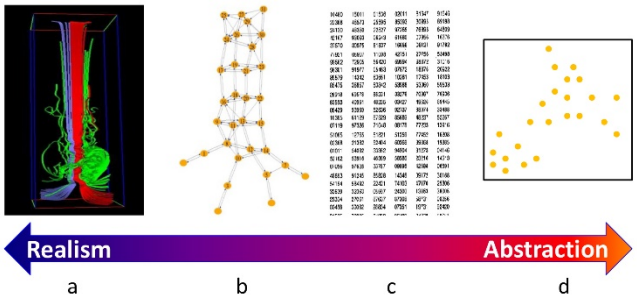


Figure 1: Transformation of a 3D flow field in a) to a 3D flow graph in b), to a set of numeric flow graph vectors in c), and finally to a scaling-based visualization in d).

- **Memory Footprint**: The memory footprints of the underlying structures consistently reduce from Figures 1a to 1d—good news for big data visual analytics in general.
- **Newly Gained Semantics**: Figure 1b provides additional graph topology information such as the longest flow line of the 3D flow field. Figure 1c allows users to compare the similarities among the graph signature vectors based on their numerical values and user preferences. Figure 1d supports, for example, clustering analysis of the signature vectors.
- **Interactivity**: In Figure 1a, we can only pan and zoom the graph structure. In Figure 1b, we can also highlight a particular node or set of nodes in the 3D space. In Figure 1c, we can now sort or classify different types of local graph structures (or flow features) and highlight their similarities or dissimilarities. Perhaps the most important gain of Figure 1d is that we can now precisely brush a set of similar structures and study their common characteristics.

We use a simplified 2D example in Figure 2 to illustrate the concept of transforming a 2D flow field to the 2D flow graph. A flow area (or volume in 3D) is divided into neighboring blocks. One of these blocks is shown in red in Figure 2. A weighted edge is created between every pair of adjacent nodes (or blocks) in case a rendering seed moves from one block to the adjacent one. In the example highlighted in green in Figure 2, eight seeds are in the source block (along the block edges) and six of the eight seeds move to the destination block on the right (and thus a weight of 6/8). Additional weight definitions can further be included for different applications. By using different block sizes, we create a coarse-to-fine data hierarchy that supports multiscale visual analytics.

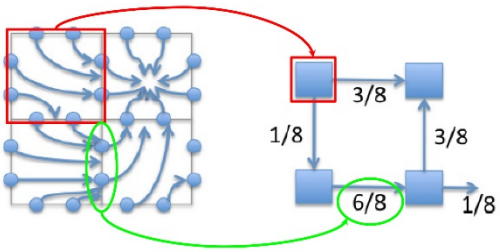


Figure 2: Transforming a flow field (left) to a flow graph (right).

4.3 Multifaceted Coordinated Visualization

Six different visualization techniques are included in the system. This section gives a "very" brief overview of the individual visualization tools. Except for the 3D flow field visualization, all of the tools support interactive brushing through either mouse clicking (for individual values) or freehand lasso (for a set of values).

3D Volume Time Series: This visualization contains a series of 3D volume cubes that shows the brushed data in 3D spaces arranged along the timeline. Users can rotate, zoom, and pan the flow graphs within the cubes using a mouse. Figure 3A shows a snapshot of 120 3D volume cubes that represent 120 model time steps. Each 3D volume cube shows the corresponding flow graphs (with the graph links turned off in this example). The design supports both spatial and temporal explorations in one visualization.

Parallel Coordinates: The parallel-coordinates visualization (Figure 3F) supports multiple brushes of an individual parameter or aggregate query of multiple parameters. Color blending is used to show the overlaps between different brushed values. For reference, a color-blending chart is shown in Figure 3D. The parallel-coordinates visualization supports both data brushing (query) and linking (results).

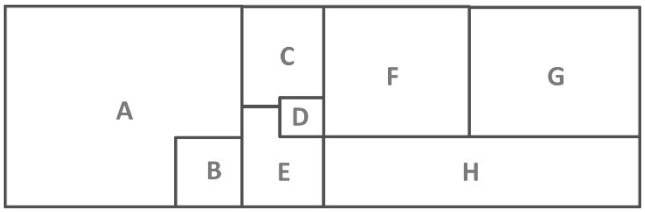


Figure 3: (Right) A landscape view of the system screenshot. (Above) An annotated sketch that shows the identities of the visualization windows in the screenshot on the left. A: 3D volume time series. B: System control panel. C: Multidimensional scaling scatterplot. D: Color blending chart. E: Matrix of scatterplots. F: Parallel coordinates. G: 3D Flow field and globe visualization. H: Profile plot.

Matrix of Scatterplots (MOS): The MOS visualization displays pairwise 2D plots of the multivariate data. Figure 3E shows 5 parameters and 21 scatterplots. The diagonal scatterplots are histograms that show the distribution of the individual parameters. We will revisit the histogram visualization in Section 4.4. The MOS visualization supports pairwise comparisons of all parameters.

Multidimensional Scaling Scatterplot: The MDS scatterplot visualization (Figure 3C) projects a high-dimensional dataset into a 2D scatterplot, which approximates the similarities among the data records using the pairwise distances among the corresponding dots in the scatterplot. Our tool allows users to control the weight of individual parameters (highlighted in green in the lower left of Figure 4), contributing to the computing of the numerical vectors for scaling calculation.

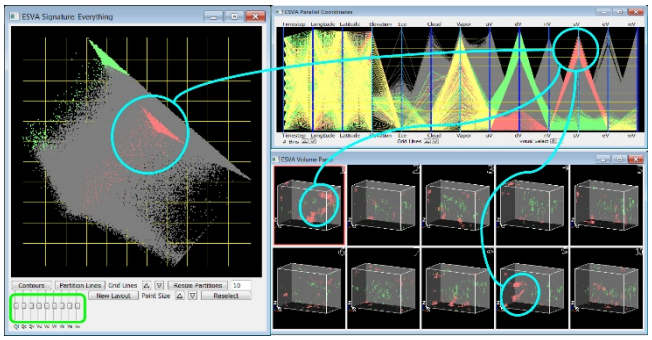
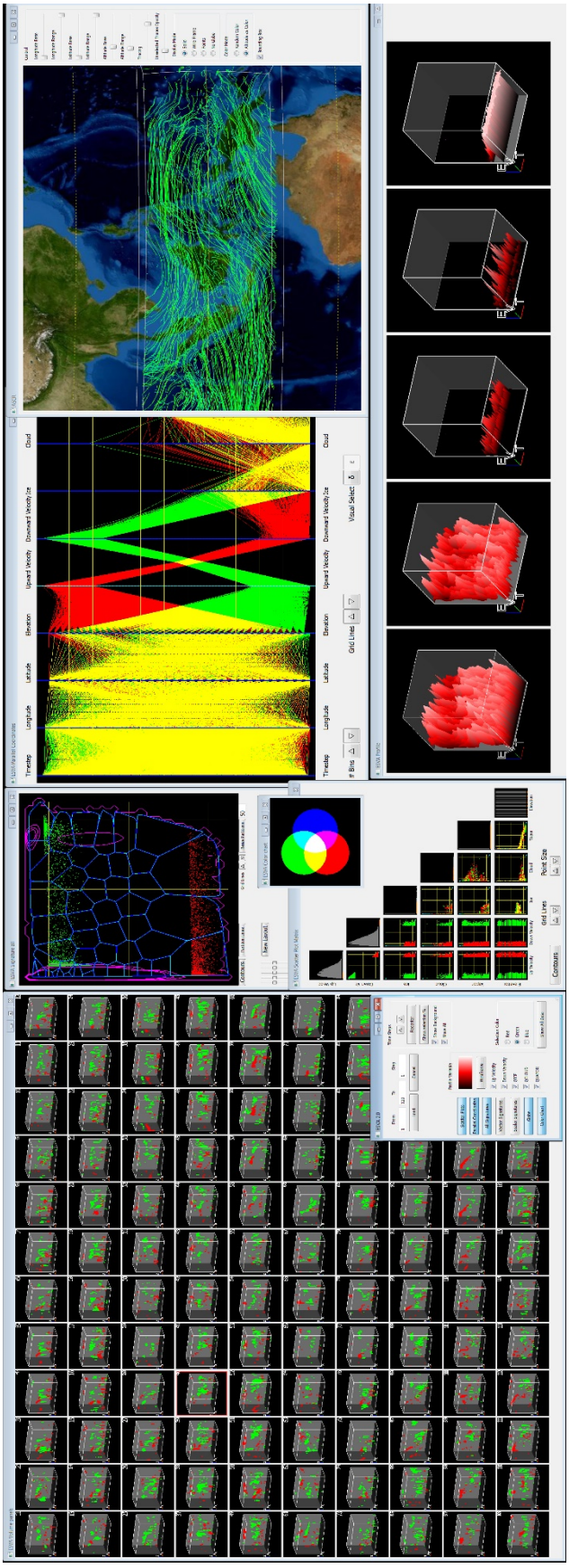


Figure 4: A snap-together example that demonstrates the interactions among the MDS plot and the two information visualization tools.

Figure 4 depicts a simplified example that snaps together three visualization tools. The MDS plot on the left shows the relative locations and density of the high southward velocity (red) values as compared to the high downward velocity (green) values.

Profile Plot: The profile plot visualization (Figure 3H) is a 3D line or shading plot that depicts the changes of different scalar values of a brushed geospatial location (longitude and latitude coordinates) versus the elevation (atmospheric levels) versus the model time steps. We have multiple examples of using this visualization in Section 5.1.

Flow Field and Globe: This visualization contains both a flow field particle visualization in 3D curvilinear format and a globe with map information and other geographic references (see Figure 3G). Users can zoom, pan, and rotate the globe. They can also control the elevation, longitude, and latitude coordinates of the flow visualization area. We develop a customized flow field visualization library [13] for fast rendering of very large 3D flow data. Figure 5 depicts the locations of two circular flows (in yellow) as shown in the globe visualization and the 3D time-volume



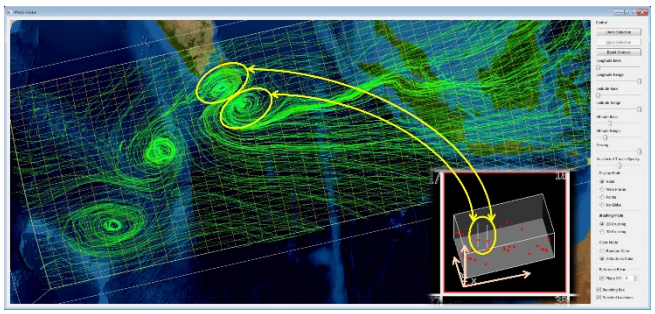


Figure 5: The two yellow-highlighted circular flows suggest that there may be a vertical convection at different atmospheric levels of the location, which can be a clue for identifying a MJO.

visualization. We can interactively tie the two features together because of the flow field to flow graph transformation step (Section 4.2) that provides direct linkage between the two features.

4.4 Drawbacks of Brushing and Linking

When we have more data items than display pixels, overlapping in visualizations (especially ones employing information visualization techniques) is inevitable. When we create a visualization based on data value mapping, an unevenly distributed data layout is also unavoidable. Putting these two visualization realities together often creates unintended problems for interactive visualization of large-scale data.

Because some pixels represent more data items than the others, the unevenly distributed visualization pattern can potentially cause unpredictable response time patterns in both brushing (asking questions) and linking (showing answers). This is particularly true in scientific data with a highly skewed data distribution pattern.

One remedy to alleviate the problem is to provide visual cues that show the data distribution density in the visualization. In our system, we include 1) a histogram to complement a single variate visualization (see diagonal plot in Figure 3E) and 2) a contour map for a bivariate visualization, such as a 2D scatterplot. Figure 3C shows a contour map (in magenta) on top of an MDS scatterplot. The visual cue feature is a welcome addition to the climate scientists because there is often a high percentage of uninteresting data with very similar values within a large-scale model dataset.

5 Climate Analytics Exploration

We show how domain scientists apply our visual analytics tool to explore a large-scale climate model data and identify different features associated with MJO. The goal is to demonstrate the practical applicability of the tool in large-scale data analytics and evaluate the user experience for future improvement.

5.1 A Large-Scale Climate Model Exploration Session

The exploration session involved two climate scientists and two visual analytics researchers using two 30" desktop display panels. The role of the visual analytics researchers in the session was to observe, record, and occasionally answer operation-related questions. The two domain scientists designed the underlying climate model and have a theoretical understanding of the science behind the underlying data. However, they had never seen a complete visualization of their data, especially when using a multifaceted interactive exploration tool.

The session lasted for about two hours. The scientists spent the first few minutes exploring different visualization techniques and interaction options. During the visualization exploration, the domain scientists also used a whiteboard to plan their exploration strategies. We highlight three particular instances that show how the domain scientists used the tool to explore their model data.

Support or Refute Arguments: Among the many characteristics of an MJO is the appearance of slowly evolving features with high convection flow fields and high moisture content. The domain scientists first used the parallel-coordinates visualization to brush the high upward velocity and high moisture variates and then searched for consistent spatial and temporal patterns in the 3D volume time series visualization. An interesting pattern instantly appeared between time steps 103 and 111 as shown in Figure 6a. As we can see, the trail starts in the second panel, moves slowly towards the center of the 3D volume, and finally dissolves in the last panel of Figure 6a.

However, the visual evidence was soon refuted by the scientists. By interactively manipulating the volume visualization, they realized that the location of the trail pattern was too close to the boundary in the 3D volume, as shown in Figure 6b. An MJO feature should not be found in those geographic locations.

Multifaceted Exploration: Another exploration strategy of the MJO problem is to examine the high moisture content in different atmospheric levels. In Figure 7a, mid-range elevation and high ice content were selected. This time, instead of comparing the temporal and spatial patterns from the 3D volume time series visualization, the scientists inspected the individual brushed data point by visualizing the 3D volume from different perspectives and confirmed that the location of the brushed data point was far away from the boundaries. The scientists also locked in longitude and latitude coordinates for further multifaceted exploration using additional visualization techniques.

Rather than detecting temporal features directly from the 3D volume time series as in the previous example, the scientists used

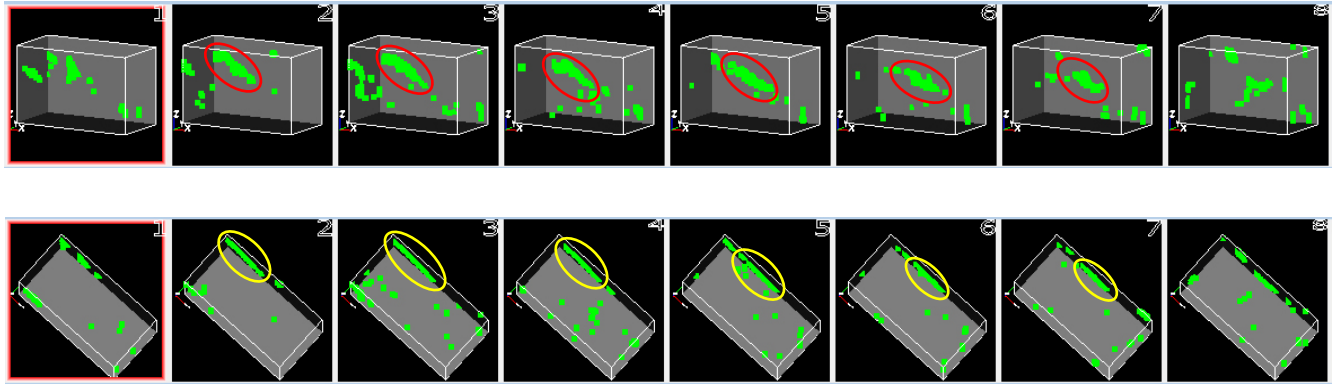


Figure 6: a) An evolving pattern that starts in the second panel, moves slowly from northwest to central in the next six panels, and finally dissolves in the last panel. The patterns are highlighted in red. b) After rotating the 3D rectangular volume, the scientists realize the pattern was detected along volume boundaries. An MJO feature should not be found in those areas, which are highlighted in yellow.

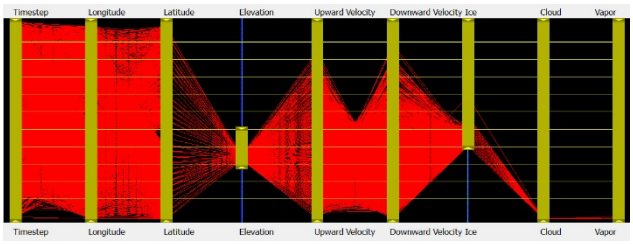
the profile visualization to examine the ice content of the same longitude and latitude coordinates of the model throughout the entire 120 time steps. Their goal was to evaluate the trend of changes during the entire model period. Figures 7b and 7c show two different views of the corresponding profile visualization. In both Figures 7b and 7c, the scientists noticed two peaks separated by a valley in the profile visualization, which suggested two potential episodes of MJO during the model period.

Validation and Interpretation: To validate the finding of two potential episodes of MJO, the domain scientists needed to consider additional scalar parameters in the model. The climate physics of the MJO suggests that a consistent change of ice content within an area would likely be followed by a similar change of vapor in the area.

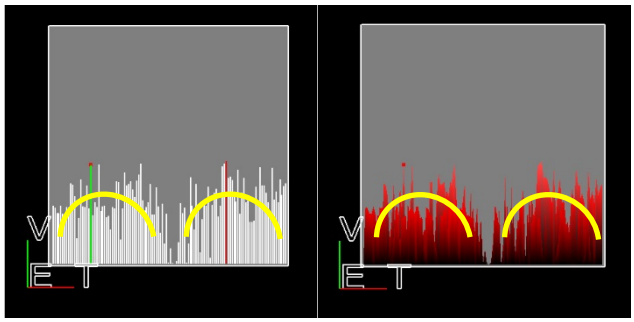
Figures 8a and 8b show the profile visualizations of vapor (vertical V-axis) versus time (horizontal T-axis) of the same brushed location using line and shading plots. Unfortunately, both figures show consistent vapor contents (i.e., no major peaks or valleys) throughout the entire model period.

By further manipulating the 3D profile plots, the domain scientists discover that heavy vapor is found at the lower atmospheric levels. As shown in Figure 8c, the low-level vapor (highlighted by a yellow arrow) blocks the more interesting vapor pattern (highlighted by a cyan arrow) found in higher atmospheric levels (vertical E-axis). That is the reason for the flat, uninteresting patterns found in Figures 8a and 8b.

Our visual analytics system has a customized feature that removes the low-level climatology of the data. This is done by subtracting the time-varying vapor values with the average vapor values at the same atmospheric level, i.e., $Vapor\ (t) - \overline{Vapor}$. The results are plotted in Figures 8d and 8e. As shown in the figures, the two potential MJO features (annotated by the yellow arcs) instantly appear in both figures. Because both ice (Figure 7) and vapor (Figure 8) exhibit consistent data-varying patterns, together they provide keen evidence that there are indeed two episodes of MJOs in the time-varying climate model dataset.

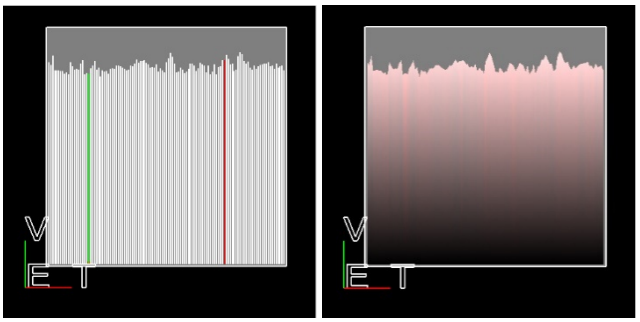


a

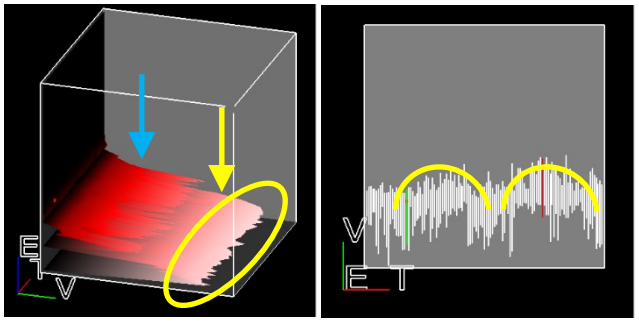


b c

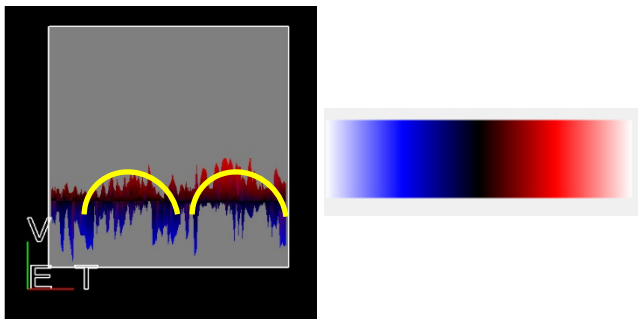
Figure 7: a) Mid-range elevation and high ice content were selected as shown by the yellow bars. Two profile plots show the ice values (vertical V-axis) versus time (horizontal T-axis) of a brushed area in both line plot (b) and shading plot (c). The yellow arcs highlight the peaks and valleys of the two potential MJO episodes.



a b



c d



e f

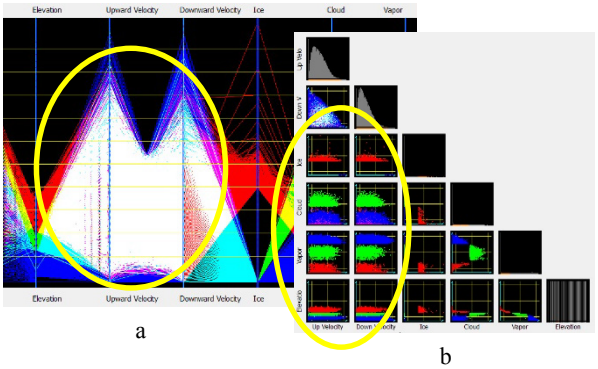
Figure 8: a) A profile plot that shows the vapor values (vertical V-axis) versus time (horizontal T-axis) of the same brushed geographic location as Figure 4 using line plot. b) The same profile plot using shading color. c) High vapor values (V axis) at low atmospheric level (E-axis) highlighted in yellow hide the more interesting patterns at higher atmospheric level, as highlighted by the cyan arrow. d) Profile plot after the low-level climatology is removed. The two peaks appear again, as highlighted by the two yellow arcs. e) The same profile plot in shading color. f) The color map of visualization. Blue is negative and red is positive.

5.2 Applicability Evaluation by Domain Scientists

Section 5.1 summarized an exploration session that involved only a subset of visualization techniques supported by our system. The domain scientists did in fact sample all the other visualization techniques during the session. After the exploration session, both visual analytics researchers and domain scientists evaluated the applicability of the tool in large-scale information visualization. The comments, summarized below, should only be treated as a reference for a similar information visualization study.

- The domain scientists found the parallel coordinates technique "extremely" powerful and useful in both brushing (aggregating query) and linking (showing corresponding answers) multivariate data.
- Even though we didn't include the MOS technique in the discussion in Section 5.1, the domain scientists indeed used it

frequently to complement the other techniques for multifaceted exploration. For example, it is not easy to trace the red (high ice), green (high cloud), and blue (high vapor) lines in Figure 9a, especially in around the two flow velocity axes highlighted in yellow. However, the corresponding matrix of scatterplots visualization in Figure 8b clearly depicts the relationships among the two velocity parameters and the brushed red, green, and blue data for visual analysis (also highlighted in yellow).



a b

Figure 9: a) High ice, high cloud, and high vapor values are brushed in red, green, and blue respectively. It is not easy to explore the two flow velocity parameters near the yellow highlighted area. b) The same brushed data (also marked in yellow) are ready for visualization.

- As described in Section 5.1, the scientists manipulated the 3D volume extensively using the mouse to visualize the brushed data layout from different perspectives. The same 3D volume time series capability is not available in the climate visualization tools that they use daily, as shown in Section 2.1.
- We made an interesting observation that the domain scientists did not brush more than 10% of the data throughout the entire exploration session. It was later explained to us that the scientists understood the model theory (both overview and trends) well. But they don't have the same degree of understanding of local fine details.
- Both the profile plot and the globe with flow field visualization were deemed useful in the exploration session. What separates them, performance-wise, is the brushing capability in the profile plot and the lack of it in the flow field visualization. Without the brushing capability, the flow field visualization became a less appealing multifaceted information visualization for data exploration. However, we are in the process to develop a 3D virtual globe with full 3D spatial brushing capability to support climate model analytics.
- The visual analtyics researchers were surprised that the MDS-scatterplot technique was not used as often by the domain scientists as we expected in this particular case study. The scientists explained that they needed to study the behaviors and implications of "individual" parameters in the MJO exploration. The dimension reduction process of MDS hides the sensitive inter-parameter correlations and hinders the exploration goal of this particular analysis. The domain scientists were informed that there are datasets that contain tens of hundreds of parameters. Dimension reduction may be the only viable option for that kind of data. The domain scientists instantly came up with possible applications that can potentially utilize the visualization techniques in our next climate simulation. In the end, we agreed that we should keep the MDS-scatterplot technique in the system for future climate data analysis.

6 Computation Performance Studies

We showed the analytical applicability of the visual analytics tool in Section 5. Here we demonstrate the computation performance of the tool running on a commodity desktop computer. Details of the dataset and computer hardware used in this study are described in Sections 3.1 and 3.3.

We mentioned earlier that there are 120 time steps in the dataset. To generate different data sizes for the performance benchmark study, we progressively reduce the data size in the study by increasing the step sizes of the data. So when time step is equal to 1, the entire dataset is used in the study. When the time step is equal to 2, the data size is reduced by half. In other words, the larger the time step number, the smaller the size of the data.

We are particularly interested in the system response time for the interactive analytical operations, i.e., brushing and the corresponding linking processes. In this study, we investigate the *worst-case scenario* when the users bring up *all* the visualization panels and use the full-resolution dataset, as described in Section 3.1. In other words, when users brush the entire dataset, the system will filter and navigate all the scalar values. Table 2 shows the system response times of the four information visualization techniques when brushing different portion percentages of the dataset (from 100% down to 5%). Because the system response times of the globe and 3D flow field data visualization and the profile visualization take less than 0.1 seconds to complete, we do not include these time results in the table.

Table 1: System response times (in wall clock seconds) of applying parallel coordinates (PC), multidimensional scaling scatterplot (MDSS), 3D volume time series (3DTS), and matrix of scatterplots (MOS) when brushing different portions of the full-resolution data.

Brushing %	PC	MDSS	3DTS	MOS
100%	0+	0.16	5.523	0.31
75%	0+	0.15	3.635	0.31
50%	0+	0.16	3.198	0.31
25%	0+	0.15	1.731	0.32
10%	0+	0.16	1.264	0.31
5%	0+	0.16	0.608	0.31

The time differences between brushing 100% (solely for benchmarking purposes) to 5% of the data are between 1 to 5 seconds. Our domain scientists generally agree that the response time performance meets their analytical needs to interactively visualize the data. We did observe in Section 5.2 that the domain scientists do not normally brush more than 10% of the data.

7 Observations, Lessons Learned, and Challenge

A key objective of this application paper is to share the lessons learned with readers. We have already reported our applicability and computational performance experiences in Sections 5 and 6. Here we present observations and lessons learned in the development of a large-scale visual analytics tool for compatible data size using a commodity desktop computer.

Fresh Model Data Exploration Experience: Except for 3D volume visualization and the globe/flow field visualization, the information visualization techniques supported by our tool are not available in any of the prevailing tools described in Section 2.1. Naturally, the tool is a welcome addition to our domain scientists' tool chest.

Additionally, this is our domain scientists' first experience visually analyzing and exploring such a large amount of model data using high-resolution visualization graphics in interactive mode on a desktop computer.

Low System Learning Curve: The few information visualization techniques included in the system design have shown to be intuitive to learn and easy to use. The tool developers didn't

provide any documentation or help files to the domain users. And yet after only a few minutes of demonstrations, the domain scientists were able to start using the tool and visually search for features and clues in their model data.

Information Visualization: The computation and applicability study results in Sections 5 and 6 suggest that many conventional information visualization techniques, like some of those discussed in the paper, can theoretically go beyond the 3 billion number mark on a similar data exploration problem. This is based partly on the fact that our desktop computer for development and testing has only 48GB of memory installed. The same machine can install up to 192GB (i.e., 3 times more) of memory.

However, our case study does not address large-scale visual analytics challenges posed by, for example, hierarchical data, like a node-link graph.

The Top Challenges Ahead: Beyond computation and visualization, there are other challenges that lie ahead in large-scale visual analytics. During the course of the investigation, we have compiled a list of general [17] and interaction [18] challenges in extreme-scale visualization and analytics. Top future challenges that are associated with desktop applications include 1) representation of evidence, 2) uncertainty quantification, 3) data fusion, 4) data summarization, 5) human cognition, and 6) several engineering development issues. Readers are referred to the above two publications for further details.

8 Conclusion and Future Work

For the desktop-based, large-scale visual analytics research and development effort, we will continue to 1) enhance and fine-tune the exploration capabilities and 2) analyze additional climate phenomena beyond MJO. Visualization and analytical issues identified in Sections 5 through 7 will be thoroughly assessed and addressed. These challenges include supporting interactive data brushing for 3D flow-field visualization, more powerful temporal and spatial analytics techniques, and interactive visualization using a touchable power wall with 12 back projectors. Figure 11 shows an early design prototype using a smaller, non-touchable wall display with 2 front projectors.

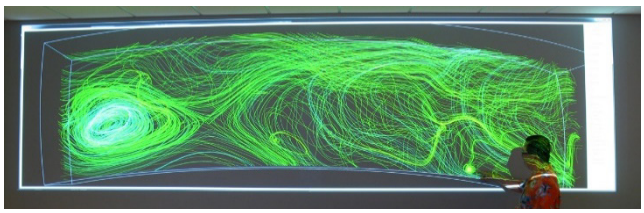


Figure 11: An early design prototype using a smaller, non-touchable wall display with two front projectors.

For the overall exabyte data visualization and analytics research effort, we are modeling a much larger and higher-resolution regional climate dataset that covers a wider area surrounding the entire globe using the same HPC described in Section 3.1. The data size and computation results reported in this paper will serve as a benchmark to determine the amount of data exported by the HPC to the desktop computer for interactive analysis.

Acknowledgments

This work was supported in parts by the U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research (award number 59172, program manager Lucy Nowell); by the National Science Foundation (NSF grants IIS-1017635 and IIS-125075); by DOE Scientific Discovery through Advanced Computing (SciDAC grant DE-FC02-06ER25779, program manager Lucy Nowell); and by DOE through the Atmospheric Systems Research (ASR) and Regional and Global Climate Modeling (RGCM) programs using resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory and the National Energy Research Scientific Computing Center (supported by Office of Science under Contract No. DE-AC05-00OR22725 and Contract No. DE-AC02-05CH11231, respectively). The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

References

[1] Gennady Andrienko and Natalia Andrienko, "Sammon's Projection for Clustering Complex Geographical Objects," *GIScience 2010,* Zurich, September 2010.

[2] Jorik Blaas, Charl P. Botha, and Frits H. Post, "Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-Timepoint Data Sets," *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1436-1443, 2008.

[3] DOE Office of Science Advanced Scientific Computing Research (ASCR) Scientific Discovery through Advanced Computing, available at http://science.energy.gov/ascr/research/scidac/, 2014.

[4] Doris Dransch, Patrick Kothur, Sven Schulte, Volker Klemann, and Henryk Dobslaw, "Assessing the Quality of Geoscientific Simulation Models with Visual Analytics Methods-A Design Study," *International Journal of Geographical Information Science*, 24(10):1459-1479, 2010.

[5] Ferret, Data Visualization and Analysis, available at http://ferret.wrc.noaa.gov/Ferret/, 2014.

[6] GrADS, Grid Analysis and Display System, available at http://www.iges.org/grads/, 2014.

[7] Samson Hagos and Ruby Leung, "Moist Thermodynamics of the Madden-Julian Oscillation in a Cloud-Resolving Simulation," *Journal of Climate*, 24(21):5571-5583, 2011.

[8] Johannes Kehrer, Philipp Muigg, Helmut Doleisch, and Helwig Hauser, "Interactive Visual Analysis of Heterogeneous Scientific Data across an Interface," *IEEE Transactions on Visualization and Computer Graphics*, 17(7):934-946, 2011.

[9] National Energy Research Scientific Computing Center (NERSC), available at http://science.energy.gov/ascr/facilities/nersc/, 2014.

[10] Madden-Julian oscillation, available at http://en.wikipedia.org/wiki/Madden-Julian_Oscillation, 2014.

[11] Philipp Muigg, Markus Hadwiger, Helmut Doleisch, and Eduard Groller, "Visual Coherence for Large-Scale Line-Plot Visualizations," *Computer Graphics Forum*, 30(3):643-652, 2011.

[12] NCL, NCAR Command Language, available at http://www.ncl.ucar.edu/overview.shtml, 2014.

[13] B. Nouanesengsy, T.-Y. Lee, and H.-W. Shen, "Load-Balanced Parallel Streamline Generation on Large Scale Vector Fields," *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1785-1794, 2011.

[14] David W. Pierce, Ncview: a netCDF visual browser, available at http://meteora.ucsd.edu/~pierce/ncview_home_page.html, 2014.

[15] VAPOR, Visualization and Analysis Platform for Ocean, Atmosphere, and Solar Researchers, available at http://www.vapor.ucar.edu/, 2014.

[16] Pak Chung Wong, Harlan Foote, George Chin, Patrick Mackey, and Ken Perrine, "Graph Signature for Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1399-1413, 2006.

[17] Pak Chung Wong, Han-Wei Shen, Christopher Johnson, Chaomei Chen, and Robert Ross, "Top Ten Challenges in Extreme-Scale Visual Analytics," *IEEE Computer Graphics and Applications*, 32(3), 2012.

[18] Pak Chung Wong, Han-Wei Shen, and Chaomei Chen, "Top Ten Interaction Challenges in Extreme-Scale Visual Analytics," *Expanding the Frontiers of Visual Analytics and Visualization*, Springer, UK, 2012.