Web-based Visual Analytics for Extreme Scale Climate Science

Chad A. Steed*, Katherine J. Evans* John F. Harney*, Brian C. Jewell*,
Galen Shipman*, Brian E. Smith*, Peter E. Thornton*, and Dean N. Williams†

*Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831
Email: csteed@acm.org, {evanskj, harneyjf, jewellbc, gshipman, smithbe, thorntonpe}@ornl.gov
† Lawrence Livermore National Laboratory, Livermore, California 94550
Email: williams13@llnl.gov

***Abstract*—In this paper, we introduce a Web-based visual analytics framework for democratizing advanced visualization and analysis capabilities pertinent to large-scale earth system simulations. We address significant limitations of present climate data analysis tools such as tightly coupled dependencies, inefficient data movements, complex user interfaces, and static visualizations. Our Web-based visual analytics framework removes critical barriers to the widespread accessibility and adoption of advanced scientific techniques. Using distributed connections to back-end diagnostics, we minimize data movements and leverage HPC platforms. We also mitigate system dependency issues by employing a RESTful interface. Our framework embraces the visual analytics paradigm via new visual navigation techniques for hierarchical parameter spaces, multi-scale representations, and interactive spatio-temporal data mining methods that retain details. Although generalizable to other science domains, the current work focuses on improving exploratory analysis of large-scale Community Land Model (CLM) and Community Atmosphere Model (CAM) simulations.**

I. Introduction

Rapid advances in extreme scale computing feed the development of increasingly complex and higher fidelity climate simulations. These simulations promise a more comprehensive and potentially revolutionary understanding of complex climate processes. However, the full potential of such advances are delayed since exploratory analysis tools disproportionately lag behind the volume and complexity of the data. Consequently, the data are drastically reduced to high-level statistical summaries or subsets, which precludes new hypothesis generation. Scientists need a new class of scalable visual analytics tools that harness emerging high performance computing (HPC) architectures. These tools must deliver new interaction schemes and information visualization techniques that address both multi-scale and hyper-variate analysis requirements. Furthermore, barriers to the accessibility of advanced visual analysis capabilities must be torn down using loosely coupled, Web-enabled frameworks that streamline deployment and increase adoption rates among domain experts.

In response to these challenges, we introduce a new Web-based visual analytics framework (see Figure 1) that allows scientists to interactively explore large-scale climate simulation data using interactive information visualization techniques. Although HPC architectures and algorithms are a critical component of our overall system, the current work emphasizes our efforts to harness dynamic human interaction and democratize advanced visual analysis for climate science. Leveraging statistical analytics that execute on a high performance Web server, the framework uses a thin-client approach exposing both diagnostic computation results and raw data access via a flexible interface. On the front-end, the scientists use an intuitive Web-based interface to interactively explore simulation data via dynamic visual queries and representations. Furthermore, we designed the framework in close collaboration with climate scientists (who are also co-authors on this paper) to ensure an efficacious response to both their present and future data analysis needs.

The framework addresses two pressing Big Data challenges recently highlighted by the White House[1]: (1) scalable algorithms for working with imperfect data and (2) effective human-computer interaction tools for facilitating visual reasoning. Specifically, the current work offers four main contributions to the scientific computing domain:

1) New Web-based interactive information visualization techniques for large-scale simulation data analysis,
2) A tree-based navigation method for visually exploring hierarchical diagnostic parameter spaces,
3) A level-of-detail (LOD) visual analysis system that allows interactive drill-down, and
4) A loosely-coupled visual analytics architecture that connects Web-based visualizations to fast diagnostic algorithms executing on HPC architectures.

The remainder of this paper is organized as follows: Following an overview of related work in Section II, Section III introduces the climate simulation data and related challenges. In Section IV, we describe the scientific workflow and diagnostics stack utilized by the visual analytics system. In Section V, we introduce the Web-based visual analytics techniques. Then, in Section VI, we discuss advantages and challenges for thin- and thick-client solutions with respect to visual scientific analysis, followed by the concluding Section VII.

[1]White House "Big Data Initiative" press release: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

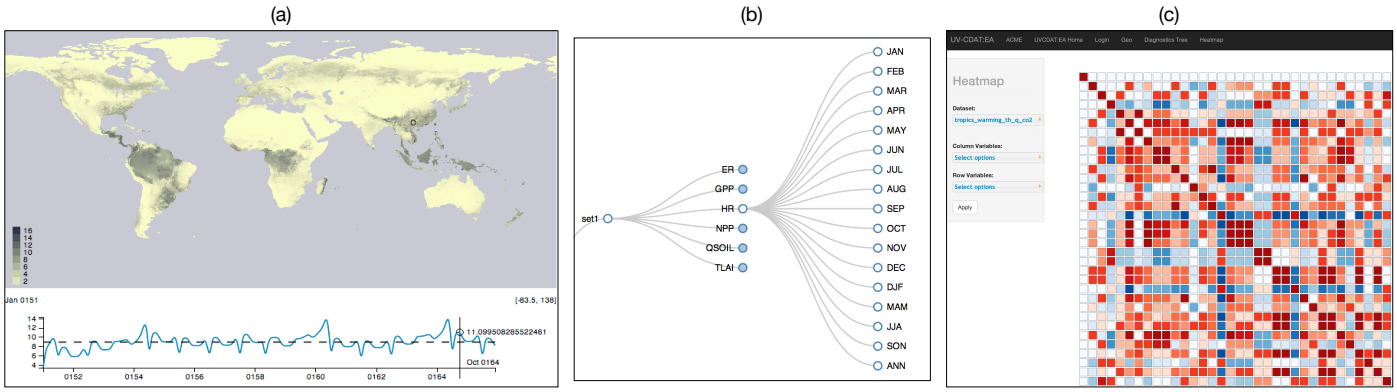

Fig. 1. The Web-based visual analytics framework introduces several key contributions to the scientific computing domain: (a) Web-based visualization techniques for large-scale data analysis, (b) a unique tree-based visual analysis method for climate diagnostics, and (c) a multi-scale correlation mining technique.

II. Related Work

A. Climate Visualization and Analysis

In the literature, the majority of efforts that seek to improve climate data analysis can be subsumed under the standalone, thick-client philosophy. For example, Potter et al. [16] introduced the Ensemble-Vis framework making use of coordinated multiple views (CMV)—a popular methodology involving linked brushings spread across multiple visualizations, which has demonstrated more creative and efficient analysis [17]. Using this CMV interface, Ensemble-Vis is capable of producing geographic plots, trend charts, and climate ensemble visualizations. In other similar work, the focus is on addressing the challenges of multivariate visualization and analysis of climate data [10], [12], [18].

UV-CDAT is a modern system that integrates several existing scientific data visualization tools into a thick-client framework [29]. However, UV-CDAT focuses on traditional scientific visualization techniques such as 3-dimensional volumes and geographic plots. The Exploratory Data analysis ENvironment (EDEN) is a related thick-client system, which uses a parallel coordinates based canvas with coordinated statistical visualizations for general purpose exploratory analysis [20] and more specific climate simulation studies [21], [22].

The above mentioned works belong to a small class of interactive visualization systems that deliver analysis capabilities via a thick-client, standalone application approach. Such techniques are valuable for expanding our exploratory data analysis capabilities, but they also tend to suffer, at varying degrees, from usability, deployment, and adoption issues. Furthermore, these techniques are often difficult to configure and use on the scientists' workstations—a significant barrier to widespread adoption at the point where such capabilities are most needed. Consequently, these solutions never fully replace the familiar tools of the scientists that were originally developed to deal with data scales from past decades. In light of these issues, the objective of the current work is to bring the advanced visual analytics tools to the more accessible Web environment.

B. Web-based Climate Analysis

Just as hardware advances drive extreme scale computing to exascale, the recent surge in Web-based application technologies feeds a growing trend toward Internet application deployment, which may eventually phase out standalone, desktop applications. Using loosely-coupled and flexible components, such as d3.js [4], jQuery, Bootstrap, and Django, end-user deployment and accessibility issues are minimized since applications are directly executed in a Web browser via a URL. Furthermore, intense processing and large-scale data storage can be hosted on high performance back-end servers and remotely accessed by the clients interfaces.

Web-based applications permeate many domains, including climate data analysis. A literature review of Web-based climate analysis systems reveals several geoprocessing frameworks [31], [32], but the typical focus is on the Application Programming Interfaces (APIs) and standards for connecting back-end processing modules. Typically, they do not address the development of interactive visualization and analysis techniques for large scientific data sets, particularly those used in climate studies. Of the works that do consider Web-based visualizations, the majority of works primarily rely on the Google Earth application or similar geographic visualization systems [2], [3], [9], [23], [27].

The distinguishing feature of our framework is the emphasis on providing new information visualization techniques, including novel interaction techniques, that permit dynamic visual queries of large-scale climate simulation data sets via distributed connections to a high performance diagnostics server. We focus on harnessing distributed diagnostic algorithms for extreme scale science using interactive information visualization techniques to achieve greater efficacy in scientific knowledge discovery. We address the need for multi-scale and hyper-variate exploratory data analysis by using unique visualization techniques and interfaces to efficiently navigate the information space. Our work coalesces these concepts into a loosely-coupled framework that is deployed via the Internet, thereby increasing the accessibility of advanced visualization tools and, ultimately, equipping climate scientists with more advanced visualization and analysis capabilities.

C. Exploratory Data Analysis

Visual analytics is a modern take on the concept of exploratory data analysis (EDA), which was introduced by Tukey [25]. EDA is a philosophy for data analysis that

emphasizes the involvement of both visual and statistical understanding in the analysis process. Tukey likened EDA to detective work—a creative process for finding and revealing clues. Since the introduction of EDA, subsequent advances in interactive computer graphics spawned the formation of the visual analytics field, which is generally defined as the "science of analytical reasoning facilitated by interactive visual interfaces" [24]. Visual analytics is a multi-disciplinary approach for designing visual analysis techniques that efficiently combine the strengths of machines with those of humans [11]. In the current work, we embrace the visual analytics paradigm, by integrating both human and high performance computation through intuitive interaction schemes and human-centered interfaces. Although human-centered computing is not typically the focus of high performance computing, we believe the intelligent combination of these two components, each having unique and powerful features, is the key for realizing the full potential of extreme scale science.

Visual analytics and EDA techniques are the key ingredients for dealing with extreme scale data because they foster serendipitous discoveries and answer questions unasked [6]. Scientific progress heavily depends on formulating and investigating new hypotheses. As Cleveland states, the analysis, however, "should not narrowly focus on just those hypotheses that led to collection" which will "inhibit finding surprises in the data" [6]. This philosophy is important for climate simulation analysis and motivates our quest to create effective systems that help scientists generate and test new hypotheses.

Presently, scientific analysis tools cannot deal with the increasing scale and complexity of extreme scale data without drastic data reductions that precede the analysis process. By reducing these rich data sets to a few statistics such as means, standard deviations, variance components, and correlation coefficients, inferences that follow are based on a very limited collection of values, which is too restricting since information in the data can be lost. As Cleveland states: "We cannot expect a small number of numerical values to consistently convey the wealth of information that exists in the data" [6].

In the current work, our objective is to create visual analytics tools that blend statistical summaries with detailed views via multi-scale, interactive visualizations. Our approach respects the benefits of summary statistics, especially for extreme scale data exploration, while also integrating drill-down capabilities to access the raw data behind statistical summaries.

III. Earth System Simulation Data

Although generalizable to any large-scale scientific data set, the current work focuses on improving the analysis of data sets produced by the Community Land Model version 4 (CLM4) [13] and the Community Atmosphere Model version 4 (CAM4) [14]. CLM4 and CAM4 are the land and atmosphere components of the Community Climate System Model version 4 (CCSM4) [8]. In particular, we focus on $1/2$ degree global CLM4 data sets and $1/4$ degree global CAM4 data sets to address our active climate studies. These data sets contain over 440 output variables (CLM4 plus CAM4) which can be either 2-dimensional or 3-dimensional (atmospheric variables are commonly 3-dimensional). Simulation data sets consist of monthly output files, which are about 415 megabytes for CLM4 ($1/2$ degree) and about 4.3 gigabytes for CAM4 ($1/4$ degree). For a 100-year simulation, 1,200 files are produced from each component, totaling about 5.6 terabytes. Typically, scientists produce multiple simulations, including a control run and several instrumented runs with parameter variations designed to support inter-comparisons of the simulation results. Assuming a modest study comprised of a single control and two additional instrumented simulations, the amount of data to be processed triples and the inter-comparison combinations for variables, spatial regions, and temporal ranges grow exponentially, far exceeding the capacity of existing tools.

Climate scientists also generate ensembles of simulations for conducting sensitivity analysis and uncertainty quantification. Such analysis may produce thousands of different simulations. Due to computational costs of running the simulations, these ensembles are usually restricted to single geographic locations (or some modest selection of locations) over some time range. However, with exascale computing architectures on the horizon, global ensemble analysis featuring hundreds or thousands of simulations will soon be possible, producing an overwhelming increase in the volume of data to be processed. Therefore, now is the time to develop new systems that minimize data movements via distributed frameworks and seamlessly scale to the volume and complexities of the data as well as the perceptual capacity of humans.

IV. Technical Approach

Our Web-based visual analytics infrastructure functions as an embedded, post-processing component in larger architectures that encompass all aspects of an HPC environment. Figure 2 depicts a notional view of our end-to-end framework featuring the Community Earth System Model (CESM) workflow within the Oak Ridge Leadership Computing Facility (OLCF) and facilities such as the ORNL Compute and Data Environment for Science (CADES)[2]. A pre-determined workflow is considered (labeled as A in Figure 2) where a climate model is configured, verified, and executed using standard procedures on a supercomputer (e.g., OLCF's Titan). The raw output of the model may then be migrated to a high performance cluster for preliminary large-scale scientific discovery via data analysis and visualization of the output. The model and analysis results are subsequently archived into either tertiary storage devices such as the High Performance Storage System (HPSS), or large data management enterprise systems such as the Earth System Grid Federation (ESGF). These data objects are combined with external sources such as observational data to provide the underlying data of our analysis framework. The data are queried and manipulated using our high performance climate diagnostics server (B in Figure 2), which includes analysis toolkits and an externally-facing Web service API (C in Figure 2). The front-end clients (D in Figure 2), which may entail desktop or mobile Web browsers, enterprise clients, or simple command line utilities, send requests to the Web application using these APIs. These requests trigger interactions with a series of scripts that facilitate retrieval of raw data, generate diagnostics plots, extract time series, and calculate various statistical summaries. These summaries include measures like variances and climatologies

[2]http://computing.ornl.gov/cades

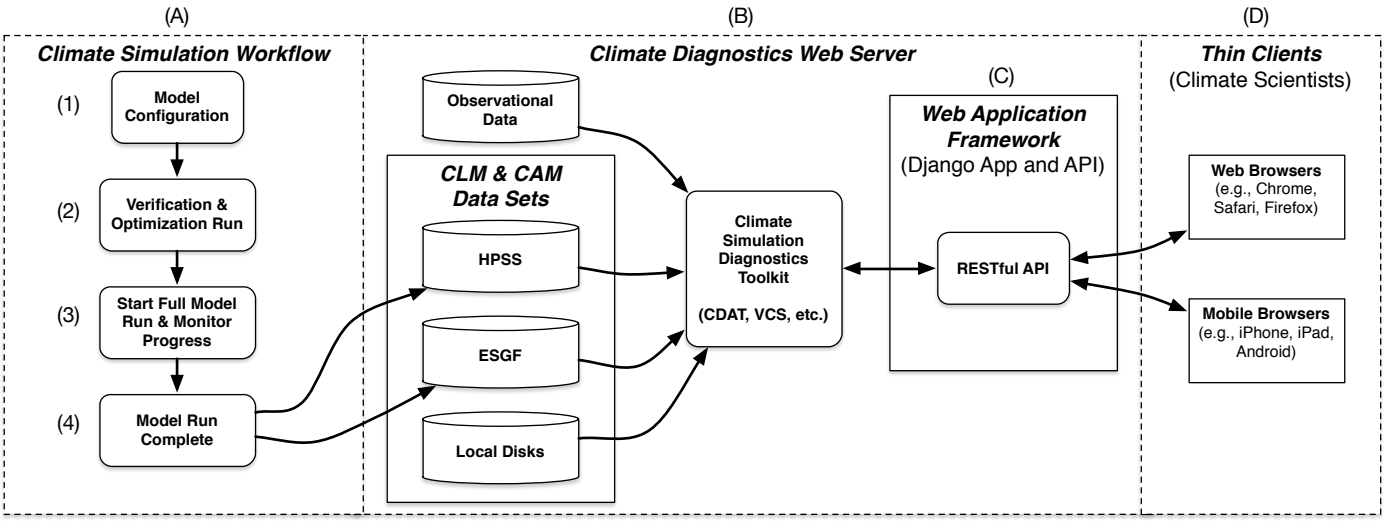

Fig. 2. This figure provides a notional view of the Web-based visual analytics framework in the context of the larger processes and architectures encompassing the full spectrum Community Earth System Modeling (CESM) within a HPC environment. Loose connections between the interactive visualizations and the back-end diagnostics toolkit ease deployment issues and can increase adoption among scientists for post-processing analysis. Thin clients are flexible and lightweight running on both desktop and mobile Web browsers.

(averages). The responses are transmitted using a lightweight interchange format and consumed by the clients to produce the visualizations detailed in Section V.

A. Scientific Workflow

The process described above is driven by the requirements of climate scientists who utilize HPC facilities to perform computationally-intensive experiments and simulations. Large-scale climate modelers, like many other computational scientists, use a diversity of workflow processes depending on the sub-area of research, stage of development, and diversity of collaborators. With this diversity comes a range of expectations and processes, but we focus on three levels of workflow that enable climate model development and analysis.

The first stage is a developmental workflow, which spans the process altering code and features that affect the solution and performance of the model. In this stage, lightweight workflow tools are employed, which enable fast but less comprehensive analyses for verification purposes. The second stage of workflow is exploratory, and involves the most diverse set of tasks and processes because it covers the extension of the model development results to more rigorous and comprehensive testing. The focus at this stage of research is on the flexibility to explore results across one or multiple components by looking at a standard set of broad, simple measures coupled to the ability to 'dive' into more details where interesting results are discovered. Usually a shorter run is completed, bursts of analysis are performed, and, if issues arise, a return to the developmental stage occurs.

The third and most mature workflow occurs when model development is complete and a climate modeler has designed an experiment using a released climate model. The workflow steps are codified and involve: (1) creation of the desired model configuration, (2) a short run and evaluation to verify the set up and check for optimized performance, (3) run execution with some informal checking of the output, and (4) substantial post-processing that varies based on the experiment. Recent efforts by climate and computational scientists are focused on addressing these key steps. The first three tasks involve intricate system integration schemes and the introduction of configuration and model scripting engines and are beyond the scope of this particular work. Instead, we focus on the fourth step of post-processing—a key component in verification, analysis, and dissemination of model output.

B. Climate Diagnostics Server

Once the model output is archived (e.g., ESGF, HPSS, shared file system, or local disk), it can be accessed by our climate simulation diagnostics toolkit (see Figure 2). This toolkit consists of scripts and library routines that create diagnostic output plots, extract time series information, and perform statistical calculations. One major advantage of having the scripts on the back-end is that the data will typically reside on the same network or machine as the processing scripts. Thus, only the relevant data is transmitted to the client.

In the current implementation, diagnostic output plots are based on existing land and atmospheric routines, with extensibility to other diagnostic packages (see section V-A). Typical diagnostic operations involve climatology calculations and producing horizontal contour plots for seasonal means, line plots showing the global average of a variable over a time range of the data set, vertical contour plots for zonal means, and horizontal vector plots for seasonal means. Pre-computed plots for the most commonly used variables and seasons are generated by a batch process as part of the model workflow. The framework also allows ad hoc plots to be generated by the user with the diagnostics tree viewer (see section V-A) for arbitrary variables and time ranges—a feature not currently available in existing static analysis tools. Once a plot is generated, it is cached for fast future retrievals.

The back-end scripts also enable easy access to the statistical summary capabilities of external resources outside of

the diagnostics framework. These capabilities are used for calculating average maps over spatio-temporal slices of the data for the geospatial view (see Section V-B) and calculating statistics for the heatmaps (see Section V-C). The calculations are fast (even in serial) taking approximately 15 seconds to calculate one climatology for 15 years of CLM data (about 100 GB) on a laptop. To address larger scales, we are increasing the parallel capabilities of the back-end processes. Many of these operations are trivially parallelizable but require careful implementations to avoid significant I/O contention. Techniques such as those utilized in ParCAT [19] will be integrated into CDAT to increase performance. Additionally, the workflow process will cache climatologies for the most common operations. Thus, the system can retrieve data very quickly since it resides in the same space as the server with little overhead beyond the Web query and transmission of results to the client.

C. Web-Based Application Framework

The back-end climate diagnostic tools form the fundamental computational and data manipulation components utilized by our infrastructure middleware. The middleware is a critical component for feeding data to our extreme scale visual analytics clients. As such, careful design is required to enable an efficient and scalable experience for scientists. To this end, we use Django, a Python-based scalable Web application framework that provides a rich suite of capabilities and fosters rapid-prototyping through its conformity to the model-view-controller (MVC) software design pattern. Django enjoys a vibrant community, widespread adoption in both industry and academia, and it can easily be deployed in most environments.

The Web service APIs in our framework are built in the style of REST (Representational State Transfer) [7]. REST is a high-level architectural term describing an approach for providing guidance about making the best use of the Web's existing technologies, rather than reinventing new strategies of Web-based communication. The exchange format is the well structured and compact JavaScript Object Notation (JSON) format, which may be quickly written and read into the payloads on both the client and the server side. The RESTful methodology has been proven to be a successful foundation in HPC environments, as demonstrated in the NEWT framework [5]. While a discussion of the benefits of REST (e.g., the promotion of scalability through Web caches) is beyond the scope of this paper [15], it is important to note that it works naturally with our visualizations (see Section V) and it seamlessly integrates with the climate diagnostics back-end layer.

V. Web-based Visual Analytics

Interactive visualization is the indispensable interface between the human visual system and computational resources acting on the data [28]. Therefore, our objective is to improve this interface, using representation and interaction, to enable rapid knowledge discovery in the overall human cognitive system. In the remainder of this section, we will detail three information visualization techniques utilizing our framework: a tree-based visualization of the diagnostic parameters space, an interactive spatio-temporal visualization, and a multi-scale correlation heatmap. Although we limit the discussion to three visualizations, the framework supports and enables other views, which are currently under development.

A. Tree-based Navigation of Diagnostic Parameter Space

Through experience, climate scientists have identified several diagnostic techniques for evaluating key patterns in simulation data sets using statistical calculations and static plots. For CLM4 and CAM4, these diagnostic techniques are captured in the National Center for Atmospheric Research (NCAR) Land Model Diagnostics Package (*lnd_diag*[3]) and Atmosphere Model Diagnostics Package (*atm_diag*[4]). Like many traditional climate analysis packages, the diagnostics packages are executed via non-interactive, command line scripts. The scripts accept configuration parameters, process the data sets, and generate a hierarchy of hundreds of static plot images. The hierarchy is defined by organizing the plot parameters according to geographic regions, temporal ranges, and plot types for each variable of interest. To facilitate viewing the images file, a set of HTML files (see Figure 3c) are generated with hyperlinks for navigating the hierarchical parameter space. Although the hyperlinks offer a universal mechanism to explore the images in a Web browser, it is a laborious and time-consuming tree traversal exercise. Given the sheer number of resulting plots, it is nearly impossible to consider all the information. With simulations rapidly increasing in volume and complexity, this approach will soon fail altogether. Furthermore, the current approach supports neither comparative analysis of multiple plots nor context preservation of the plot parameter hierarchy.

To overcome these issues, we used the d3.js [4] tree layout API to develop a tree-based visualization (see Figure 3a). The scientist can interactively explore the complex parameter space of the diagnostic plots via a single page Web-based interface. The visualization allows the scientist to consider the entire hierarchy while maintaining a visual context of the location in the overall parameter space. By clicking the tree nodes, the scientist can expand sub-trees and drill-down to the leaf nodes, which represent the diagnostic plots. When a node is clicked, the application sends a request to the Web server. The server processes the request by either dynamically generating the plot using the back-end diagnostics engine (section IV-B) or retrieving pre-generated images from the server cache.

As shown in Figure 3a, the plot image is transmitted to the visualization application for display in the right-hand plot panel. This plot panel is scrollable showing a descending history of plots viewed from top to bottom. Each plot thumbnail includes two buttons that enable removal from the panel and full plot viewing. In the full plot view (see Figure 3b), the image can be annotated with comments, which are persisted on the server to allow interactive note taking and sharing. Directly above the tree visualization, key interface components allow the scientist to save the current tree navigation state to support reproducibility. The interface components also enable accessing both the saved tree states and plot images. The ability to save and retrieve tree and plot configurations form the basis for future expansion of the framework to support social

[3]NCAR Land Diagnostics website: http://www.cgd.ucar.edu/tss/clm/diagnostics/webDir/lnd_diag4.1.htm

[4]NCAR Atmosphere Diagnostics website: http://http://www.cgd.ucar.edu/amp/amwg/diagnostics/

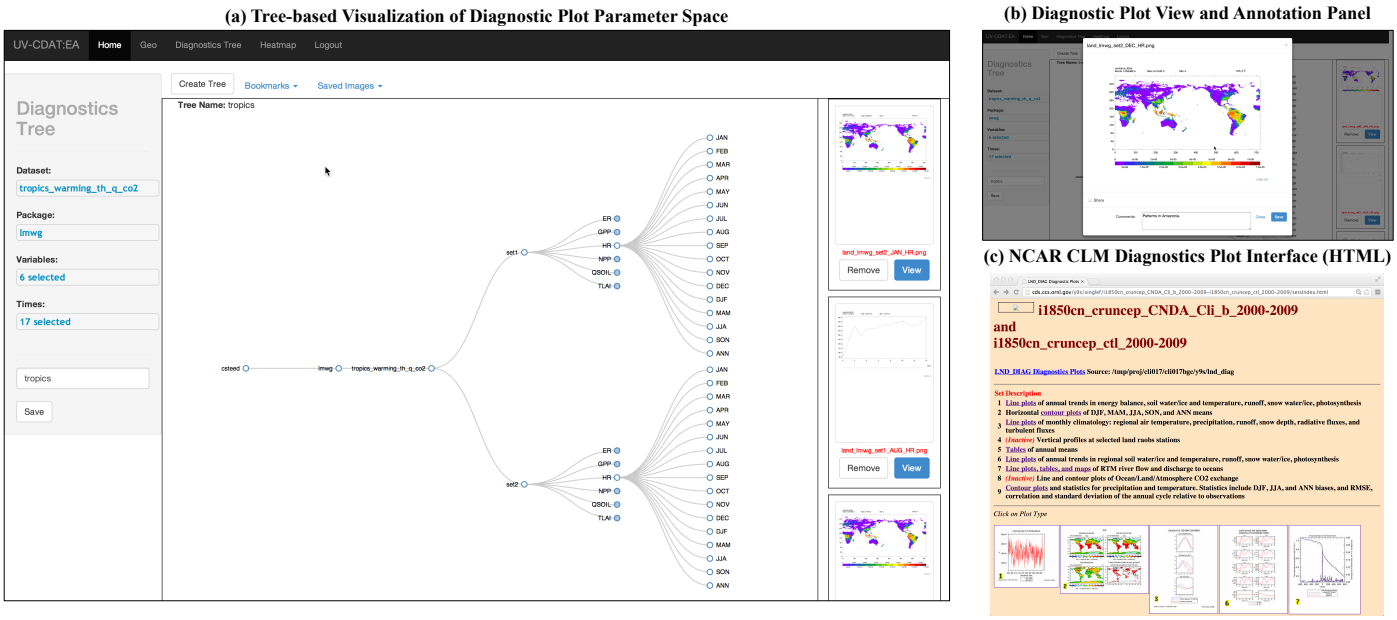

Fig. 3. We introduce a unique tree-based visualization (a) that allows interactive exploration of the complex parameter space of climate diagnostics via a Web-based interface. Scientists can expand the parameter hierarchy and click on leaf nodes to dynamically generate the appropriate views. A scrollable plot view panel keeps a history of plots (a) which can be expanded to see full views (b). Plots can be annotated for collaborative analysis and saved for future retrieval. This visualization improves the navigation experience of previous diagnostics tools (c) that rely on manual HTML-based hyperlink traversals.

computation and collaboration. Currently, more than half of the existing NCAR diagnostics have been ported to the new framework back-end, which is written in Python and utilizes the CDAT library (part of UV-CDAT [29]) for data processing.

B. Exploratory Spatio-temporal Visualizations

Spatio-temporal pattern mining is one of the most critical aspects of climate simulation analysis. As advances in high performance computing feed the creation of higher fidelity simulations, the volume and complexity of the resulting data far exceed the limits of conventional climate data analysis tools. To accommodate these deficiencies, full-scale simulation data sets are drastically reduced to statistical summaries such as means, standard deviations, and variance measures. Although these measures are informative for exploring broad trends, they are prone to mask outliers and significant associations in the raw data—the key ingredients to unexpected and potentially revolutionary scientific discoveries. To realize the full potential of extreme scale simulation data, climate scientists need visual analytics techniques that allow efficient access to both statistical summaries and raw information.

To address these challenges, we developed an interactive spatio-temporal visualization (see Figure 4a) allowing the scientist to perform dynamic visual queries at full resolution to see both summary statistical information and raw data. In the left-hand panel, the scientist selects the simulation data set and variable of interest. The Web application submits a request for the variable data to the diagnostics server. The server reads the data, calculates the statistics, and returns the results to the Web application if the data has not already been generated and cached. Then, the data are rendered in both the geographic and time series visualizations.

In the geographic map (see Figure 4a), data corresponding to the currently selected month of a particular year are rendered using a customizable color-filled level plot. As the mouse is moved in the map, the value corresponding to the geographic location under the mouse cursor is shown at the bottom right corner of the map. Furthermore, when the scientist clicks in the geographic view, the application sends a request to the diagnostics server for the time series data at the selected geographical location of the mouse via the RESTful API. The back-end server reads the data, computes summary statistics, and returns the results to the Web visualization. This computation is very quick as it is extracting a limited subset from the raw data. The data are then rendered in the time series visualization that is shown below the geographic view. Similarly, moving the mouse in the time series visualization causes the associated value to be shown with the location indicated with a vertical bar. Clicking in the time series visualization will cause the application to request the geographic data for the selected time from the diagnostics server. The server will transmit the data to the Web application and it will be rendered to the geographic view. In addition to the time series line, the mean value is also shown as a horizontal reference line.

These details-on-demand queries are supplemented with dynamic brushing capabilities in either the geographic or time series views to produce mean values for the highlighted range. For example, if the scientist drags a bounding box in the geographic view (see Figure 4b), the time series view will render the average temporal trend line for the region of interest. Likewise, if the scientist drags a box in the time series view (see Figure 4c), the geographic visualization will show the mean map for the time range of interest. Therefore, the visualization allows both summarized and detailed views for longer range trend and detailed exploration, respectively. Together, these capabilities provide an effective interface for exploring even large data sets.

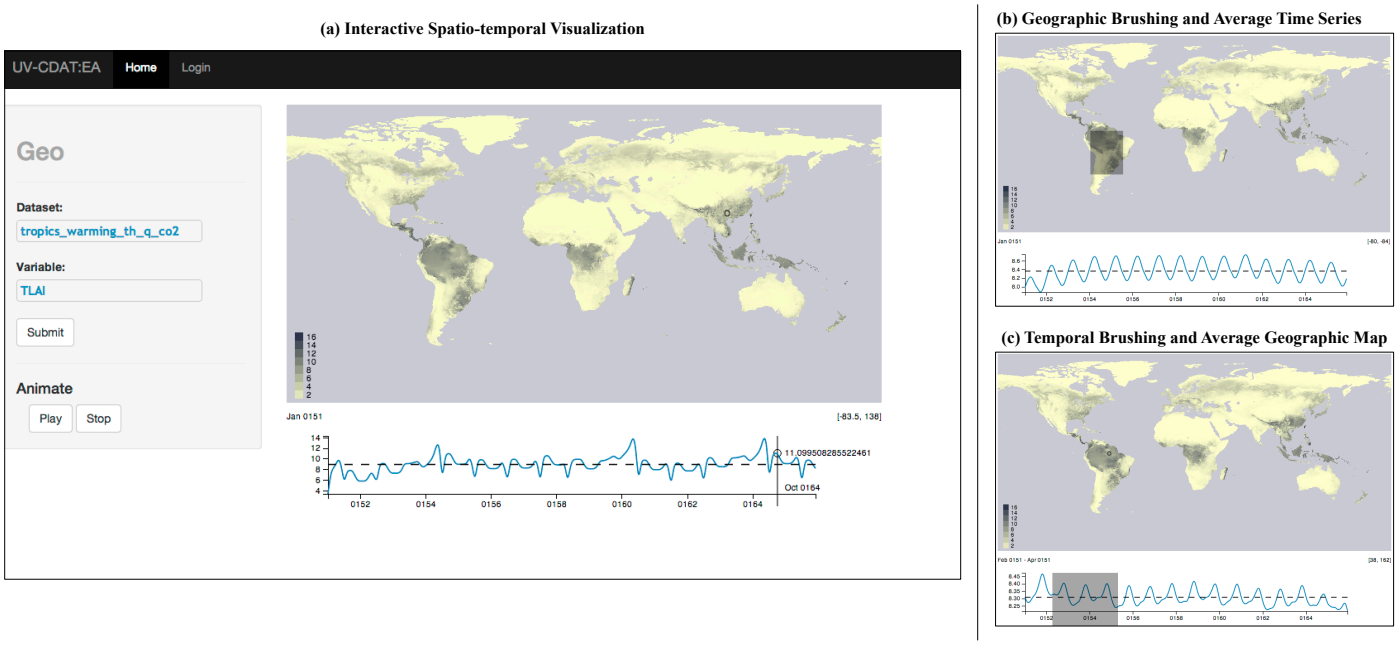

Fig. 4. The capability to interactively explore spatio-temporal patterns in climate simulation data is a fundamental requirement. Our spatio-temporal visualization allows the scientist to explore large simulations interactively via a geographic and time series view. The scientist can click locations or times of interest to query particular values (in Figure (a) a time of interest is indicated by a vertical bar). Brushings are also available in both views. In Figure (b), a region in South America (indicated by the dark shaded box) is brushed. When the brushing is performed, the data for all cells in selected region are averaged into a single time series and rendered in the temporal view below. As shown in Figure (c), a temporal range can also be brushed in the time series view to visualize the average values in the geographic view above.

To see an animation of the geographic view, the scientist can click the "Play" button in the left panel. A vertical line in the time series indicates the current time. As the line increments forward, the geographic view is updated with data for the current time step. This feature gives scientists the capability to see spatio-temporal trends and augments the interactive visual query features.

C. Multi-scale Correlation Heatmap

Another fundamental challenge of large scale scientific data analysis is enabling creative exploration of the data at multiple scales as seamlessly as possible, while maintaining contextual awareness. In particular, climate scientists need a framework that supports a high level overview, initially, and intuitive drill-down interactions to effectively zoom into increasingly detailed views of the data. This approach is analogous to tile-based map servers, such as Google Earth, which have demonstrated a superior user experience for exploring geographic information. Multi-scale interfaces, however, are not available for the statistical views that are critical to climate data analysis.

To address these challenges, we leverage a Web-based visual analytics technique that we originally developed to explore health care indicators [30]. The resulting visualization (see Figures 5a-c) provides an interactive, multi-scale technique enabling correlation mining of selected variables. The technique allows full spectrum exploration of varying scales from an initial overview of all the pair-wise correlations, to investigation of particular subsets, to individual analysis of bivariate scatterplots. Although the current implementation focuses on correlation mining, the approach also accommodates other statistical measures.

As shown in Figure 5a, the scientist begins the correlation mining task by selecting a set of variables, a time range, and a geographic area of interest. Initially, the visualization shows a correlation matrix for the selected data. The correlation matrix is a symmetric n x m matrix where each i, j element is equal to the value of the correlation coefficient, r, between the i and j variables. Specifically, the Pearson product-moment correlation coefficient is used to measure the correlation for a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \ldots, n$ [26]. The value of r is given by:

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum {x_i}^2 - (\sum x_i)^2][n \sum {y_i}^2 - (\sum y_i)^2]}} \quad (1)$$

At each intersection, r is encoded with a color-filled square, where the color is chosen from the color scale shown in Figure 5d. This color scale produces shades of blue and red for negative and positive correlations, respectively. To encode the correlation strength, a saturation scale maps the strongest correlations to more saturated and, therefore, more visually salient colors. Using this color scale, white squares indicate no correlation. The correlation of a variable with itself is always a perfect positive correlation, hence the diagonal of the matrix is represented as a series of highly saturated red squares.

From this initial correlation matrix view, the scientist can zoom into the visualization (see Figure 5b) using the mouse scroll (or similar zooming gestures) to focus on a particular set of variables. As the analyst zooms into the display, the number of visible variables decreases. When the number of variables across the row or column dimension of the visible display is reduced to 6 or fewer variables, the display transitions to

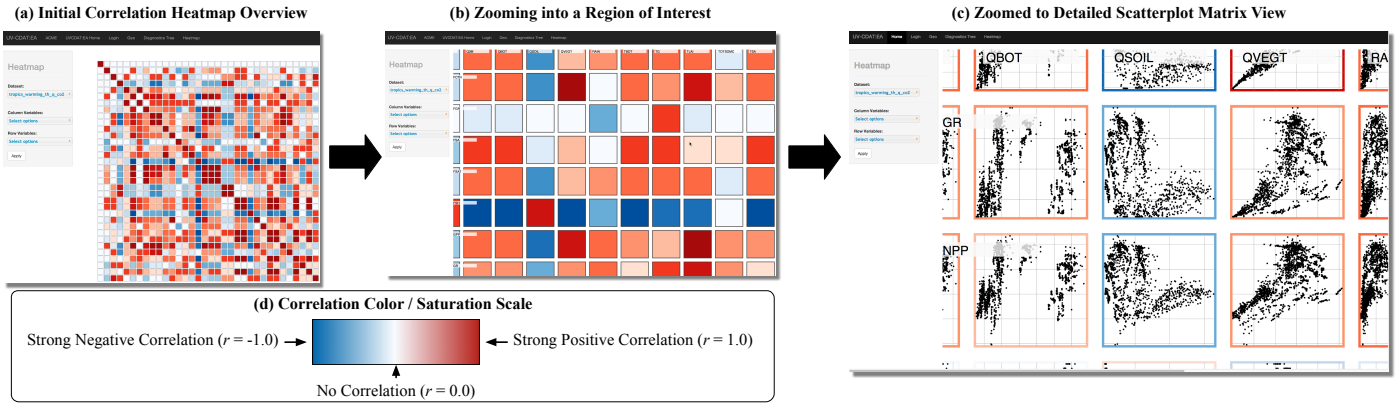

Fig. 5. The multi-scale correlation heatmap uses a unique level-of-detail algorithm to systematically stream in additional detail as the scientists zooms into the display. Initially, the visualization shows a correlation matrix for the variables of interest. The matrix is represented as a heatmap (a) where red and blue squares encode the positive and negative correlation coefficient, respectively. A saturation scale (d) is used to encode the strength of the correlations so that stronger correlations are more saturated and visually salient in the display. "Zooming" into the visualization (b) causes additional detail to display. At a defined threshold, the correlation squares transition into scatterplots (c) to reveal the most detailed information.

a more detailed scatterplot matrix as shown in Figure 5c. In this view, the frame of each scatterplot retains the original background color as a halo to maintain the context of the correlation strength, and the individual points that contribute to the correlation measure are rendered as points.

When the scientist hovers over individual points, the values of the points are shown using dynamic tooltips (see Figure 6). For each of the other scatterplots in the visible area, the points corresponding to the geographic location of the hover point are also highlighted for a coordinated multiple-view (CMV). This coordination visually links selections across views for comparative analysis. Furthermore, subtle grid lines are shown in the scatterplots to provide a reference for detecting relationships across the matrix.

The multi-scale correlation heatmap addresses the perceptual and technical scale issues associated with analyzing correlations within a large data set. The level-of-detail (LOD) algorithm only streams the raw data points when the number of scatterplots in consideration are below a certain threshold in order to manage the memory limitations of the Web browser's document object model (DOM). In addition to considering the DOM scalability issues, the LOD algorithm helps achieve perceptual scalability by avoiding an overwhelming visual representation consisting with hundreds of details views. The visualization allows the scientist to consider broad correlation patterns intuitively and, as desired, descend into detailed views seamlessly from a unified visual interface. Such a LOD approach is an exemplar model of the human-centered, multi-scale approach that is necessary to support interactive analysis of large scale scientific data in a human-centered framework.

VI. Discussion

Web-based deployment of advanced visual analytics techniques could revolutionize scientific knowledge discovery, particularly in climate sciences. However, the distributed nature of Web-based applications requires careful planning and a strategic investment into back-end development. Having developed both thick- and thin-client solutions for climate visual analytics, we have gained a valuable perspective on the challenges and advantages of Web-based applications. In the remainder of this section, we will expand upon this practical knowledge and address the importance of close collaborations with domain experts in the development cycle.

A. Challenges of Large Scale Data Analysis on the Web

Despite the promise of Web-based application development, it is difficult to develop Web-based applications, particularly systems that deal with large scale data sets. One challenge is efficiently dealing with limited memory in the DOM. From our experiments, we find that interactivity is drastically reduced when thousands of graphical items are held in the DOM. In order to ensure responsive interactions, we must restrict the amount of data sent to the browser with a LOD scheme as implemented in our correlation heatmap (see Section V-C).

Similar to geographic map systems, we allow zooming into the data, streaming increasingly detailed views as the extent of relevant data is reduced. The LOD scheme is extended beyond usage in geospatial views to permit multi-scale views for temporal, variable, and other derived dimensions of the data. A key challenge for multi-scale views is maintaining an awareness of the current position in the LOD continuum. An easily decoded graphical LOD indicator is required, which is positioned on the periphery of the display for instant reference. Furthermore, the system should preserve the context of the more complete overview of the data. That is, a focus+context display is needed to supplement the LOD indicator and reveal the gestalt [1] of the whole data set. In geographical maps systems, context is commonly shown as a world map with a box drawn around the region the scientist is currently zoomed into. In our correlation heatmap, we will show the overall heatmap, highlighting the variables that are shown in the zoomed view. Maintaining the context of the detailed view contributes to human cognition of the relationship of detailed subsets to the more complete and generalized whole.

On the back-end, parallelized diagnostics allow fast, ad hoc queries from the scientist's Web browser when dealing with extreme scale data sets. Currently, the back-end diagnostics are

primarily serial, but the operations performed are parallelizable and work is in progress to extend the algorithms. These queries must be optimized to return data to the client as fast as possible to maintain interactivity. This constraint requires a balance between parallel job start-up times and the actual amount of computation required. Improved algorithms help, but an efficient caching mechanism is vital to a good user experience. In our system, we cache query results and check the cache storage for queries before we perform expensive computations. If the query has been performed recently, we simply send the pre-cached results. We also initially populate the cache with common calculations (e.g., global average maps, temporal averages) when a new data set is loaded into the back-end system as part of the workflow process. This requires more storage on the back-end, but it pays big dividends in maintaining responsive front-end performance, especially when coupled with the level-of-detail tools presented.

B. Advantages of a Web-based Application Deployment

The decision to bring visual analytics capabilities to the Web is a direct response to the challenges of deploying advanced thick-client visualization and analysis tools for scientific applications. Despite the advanced capabilities of these systems, relatively few approaches are used in practical analysis. Some barriers to adoption of these systems include: complicated compilation and installation procedures, tightly coupled dependencies, and complex user interfaces.

A Web-based solution dramatically reduces deployment issues by running directly in the scientist's Web browser. The interface can be accessed from both desktop and mobile Web browsers yielding a high degree of accessibility for extreme scale analysis on workstations, laptops, tablets, and even smart phones. A Web-based system also enables revolutionary collaboration capabilities, which we are actively incorporating into our system. Installation procedures are reduced to having a compatible Web browser installed on the host system. Furthermore, new releases are deployed on the Web server and users have the most recent version each time they visit the site. Finally, a number of recent Web user interface frameworks, such as Bootstrap and jQueryUI, incorporate proven and efficient practices for improved accessibility and usability making a Web-based application more intuitive.

C. Importance of Multi-disciplinary Collaborations

The success of our approach can be largely attributed to the inclusion of domain experts in the design and development iterations. This intentional strategy mutually benefits all parties involved and helps ensure that we respond to the actual needs of domain experts. Our frequent interactions (weekly) guarantee the practicality of our solution and enhance broader adoption in the climate science community. Just as we learn more about the intricacies of the climate science domain, we are able to introduce new information visualization, interaction, and analytics algorithms to the scientists. As a result, the scientist have time to learn and provide feedback on new approaches by virtue of being part of the development team.

VII. Conclusion

In the current work, we introduce a new Web-based visual analytics framework that bridges the gap between advanced visualization and analysis and real-world, extreme scale climate data analysis. The framework includes human-centered information visualization techniques and novel interaction schemes in a manner that integrates the strengths of human computation with the tremendous computational power of leadership class architectures. Although it is initially difficult to develop the back-end for such distributed frameworks, the result, if designed intelligently, is revolutionary accessibility for domain scientists. Furthermore, deploying a lightweight UI built on HTML5 standards enables extreme scale analysis, powered by the world's fastest computing platforms, on a variety of platforms such as workstations, laptops, tablets, and smart phones (see Figure 6).

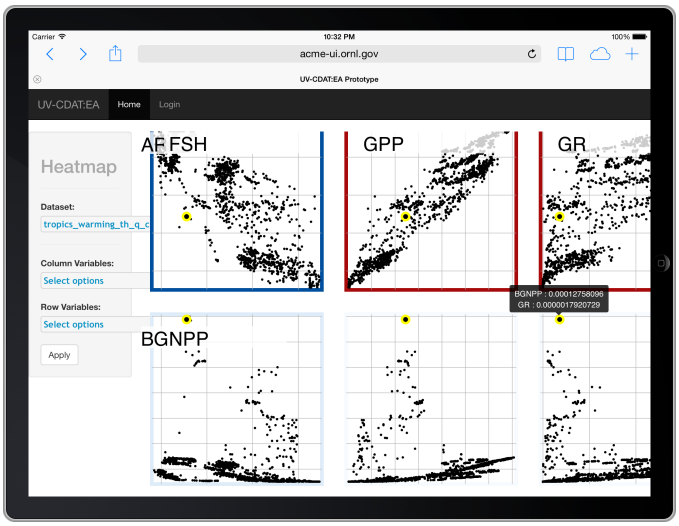

Fig. 6. Our framework UI uses HTML5 standards that make it accessible on both desktop and mobile Web browsers. In this figure, an analysis session using our correlation heatmap visualization is shown demonstrating visual analysis, backed by leadership class computing architectures, accessed via the intuitive and convenient tablet platform.

We are currently extending this framework to include additional information visualization techniques such as parallel coordinates and interactive statistical representations. Future work includes parallelizing the back-end computations and building a robust collaboration system for scientific analysis. This Web-based visual analytics framework democratizes advanced scientific visualization and analysis and it is an exemplar of the new class of techniques needed to realize the full potential of extreme scale computing.

Acknowledgments

This research is sponsored by the U.S. Department of Energy, Office of Science, Biological and Environmental Research (BER) program and performed at Oak Ridge National Laboratory (ORNL). This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] R. Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, 1974.

[2] T. G. Blenkinsop, "Visualizing structural geology: From excel to google earth," *Computers and Geosciences*, vol. 45, pp. 52–56, 2012.

[3] J. D. Blower, A. L. Gemmell, G. H. Griffiths, K. Haines, A. Santokhee, and X. Yang, "A web map service implementation for the visualization of multidimensional gridded environmental data," *Environmental Modelling and Software*, vol. 47, pp. 218–224, 2013.

[4] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[5] S. Cholia, D. Skinner, and J. Boverhof, "NEWT: A restful service for building high performance computing web applications," in *Gateway Computing Environments Workshop (GCE), 2010*, Nov 2010, pp. 1–11.

[6] W. S. Cleveland, *The Elements of Graphing Data*. Hobart Press, 1994.

[7] R. T. Fielding and R. N. Taylor, "Principled design of the modern web architecture," *ACM Trans. Internet Technol.*, vol. 2, no. 2, pp. 115–150, May 2002.

[8] P. R. Gent, G. Danabasoglu, L. J. Donner, M. M. Holland, E. C. Hunke, S. R. Jayne, D. M. Lawrence, R. B. Neale, P. J. Rasch, M. Vertenstein, P. H. Worley, Zong-Liang Yang, and M. Zhang, "The community climate system model version 4," *Journal of Climate*, vol. 24, no. 19, pp. 4973–4991, 2011.

[9] G. P. Johnson, S. A. Mock, B. M. Westing, and G. S. Johnson, "EnVision: A Web-Based Tool for Scientific Visualization," in *International Symposium on Cluster Computing and the Grid*. IEEE, 2009, pp. 603–608.

[10] J. Kehrer, F. Ladstädter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser, "Hypothesis generation in climate research with interactive visual data exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1579–1586, 2008.

[11] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual Data Mining*, S. J. Simoff, M. H. Böhelen, and A. Mazeika, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 76–90.

[12] F. Ladstädter, A. Unger, B. C. Lackner, B. Pirscher, G. Kirchengast, J. Kehrer, H. Hauser, P. Muigg, and H. Doleisch, "Exploration of climate data using intervative visualization," *Journal of Atmospheric and Oceanic Technology*, vol. 27, no. 4, pp. 667–679, 2010.

[13] D. M. Lawrence, K. W. Oleson, M. G. Flanner, P. E. Thornton, S. C. Swenson, P. J. Lawrence, Zong-Liang Yang, S. Levis, K. Sakaguchi, G. B. Bonan, and A. G. Slater, "Parameterization improvements and functional and structural advances in version 4 of the community land model," *Journal of Advances in Modeling Earth Systems*, vol. 3, no. M03001, p. 27 pp., 2011.

[14] R. Neale, J. H. Richter, A. Conley, S. Park, P. H. Lauritzen, A. Gettelman, D. L. Williamson, P. Rasch, S. J. Vavrus, M. A. Taylor, W. Collins, M. Zhang, and S.-J. Lin, "Description of the community atmosphere model (CAM 4.0)," *NCAR*, vol. TN-485+STR, 2010.

[15] C. Pautasso, O. Zimmermann, and F. Leymann, "Restful web services vs. big web services: Making the right architectural decision," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 805–814.

[16] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johhson, "Ensemble-Vis: A framework for the statistical visualization of ensemble data," in *IEEE Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes.*, 2009, pp. 233–240.

[17] J. C. Roberts, "Exploratory visualization with multiple linked views," in *Exploring Geovisualization*. Elseviers, 2004, pp. 159–180.

[18] M. Sips, P. Köthur, A. Unger, H.-C. Hege, and D. Dransch, "A visual analytics approach to multiscale exploration of environmental time series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2899–2907, 2012.

[19] B. Smith, D. M. Ricciuto, P. E. Thornton, G. Shipman, C. Steed, D. Williams, and M. Wehner, "ParCAT: Parallel climate analysis toolkit," in *Proceedings of the International Conference on Computational Science*, Barcelona, Spain, June 2013, pp. 2367–2375.

[20] C. A. Steed, T. E. Potok, L. L. Pullum, A. Ramanthan, G. Shipman, and P. E. Thornton, "Extreme scale visual analytics," in *The 4th Workshop on Petascale (Big) Data Analytics at SuperComputing 13*, Nov. 2013.

[21] C. A. Steed, D. M. Ricciuto, G. Shipman, B. Smith, P. E. Thornton, D. Wang, X. Shi, and D. N. Williams, "Big data visual analytics for exploratory earth system simulation analysis," *Computers & Geosciences*, vol. 61, no. 0, pp. 71–82, 2013.

[22] C. A. Steed, G. Shipman, P. Thornton, D. Ricciuto, D. Erickson, and M. Branstetter, "Practical application of parallel coordinates for climate model analysis," in *Proceedings of the International Conference on Computational Science*, Omaha, NE, June 2012, pp. 877–886.

[23] X. Sun, S. Shen, G. G. Leptoukh, P. Wang, L. Di, and M. Lu, "Development of a web-based visualization platform for climate research using google earth," *Computers and Geosciences*, vol. 47, pp. 160–168, 2012.

[24] J. J. Thomas and K. A. Cook, Eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos, CA: IEEE Press, 2005.

[25] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[26] R. E. Walpole and R. H. Myers, *Probability and Statistics for Engineers and Scientists*, 5th ed. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[27] Y. Wang, G. Huynh, and C. Williamson, "Integration of google maps/earth with microscale meterology models and data visualization," *Computers & Geosciences*, vol. 61, pp. 23–31, 2013.

[28] C. Ware, *Information Visualization: Perception for Design*, 2nd ed. Morgan Kaufmann, 2004.

[29] D. Williams, T. Bremer, C. Doutriaux, J. Patchett, S. Williams, G. Shipman, R. Miller, D. Pugmire, B. Smith, C. Steed, E. Bethel, H. Childs, H. Krishnan, P. Prabhat, M. Wehner, C. Silva, E. Santos, D. Koop, T. Ellqvist, J. Poco, B. Geveci, A. Chaudhary, A. Bauer, A. Pletzer, D. Kindig, G. Potter, and T. Maxwell, "Ultrascale visualization of climate data," *Computer*, vol. 46, no. 9, pp. 68–76, 2013.

[30] S. Xu, B. Jewell, C. A. Steed, and J. Schryver, "A new collaborative tool for visually understanding national health indicators," in *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, V. G. Duffy, Ed., July 2012, pp. 91–100.

[31] P. Zhao, T. Foerster, and P. Yue, "The geoprocessing web," *Computers & Geosciences*, vol. 47, pp. 3–12, 2012.

[32] K. Zhu, H. Song, and J. Gao, "Web-based Atmospheric Nucleation Data Management and Visualization," in *International Conference on Networking and Distributed Computing (ICNDC)*. IEEE, 2011, pp. 127–131.