Computers & Geosciences 61 (2013) 71–82

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Big data visual analytics for exploratory earth system simulation analysis

Chad A. Steed [a,*], Daniel M. Ricciuto [a], Galen Shipman [a], Brian Smith [a], Peter E. Thornton [a], Dali Wang [a], Xiaoying Shi [a], Dean N. Williams [b]

[a] Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[b] Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

ARTICLE INFO

Article history:
Received 4 March 2013
Received in revised form
14 June 2013
Accepted 31 July 2013
Available online 14 August 2013

Keywords:
Visualization
Parallel coordinates
Climate
Sensitivity analysis
Data intensive computing
Data mining
Statistical visualization
Multivariate
Big data

ABSTRACT

Rapid increases in high performance computing are feeding the development of larger and more complex data sets in climate research, which sets the stage for so-called "big data" analysis challenges. However, conventional climate analysis techniques are inadequate in dealing with the complexities of today's data. In this paper, we describe and demonstrate a visual analytics system, called the Exploratory Data analysis ENvironment (EDEN), with specific application to the analysis of complex earth system simulation data sets. EDEN represents the type of interactive visual analysis tools that are necessary to transform data into insight, thereby improving critical comprehension of earth system processes. In addition to providing an overview of EDEN, we describe real-world studies using both point ensembles and global Community Land Model Version 4 (CLM4) simulations.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Environmental variability and change stimulates our fervency for understanding past climate patterns and forecasting the future. Improved comprehension of the earth system process through simulated data analysis will facilitate well-informed decisions for critical climate challenges at local and global scales. Due to unprecedented technological increases in high performance computing (Gent et al., 2011; Lawrence et al., 2011; Overpeck et al., 2011; U.S. Department of Energy, 2012), simulations are evolving toward higher numerical fidelity and complexity. However, techniques to efficiently analyze the data, particularly interactive visual techniques, have not kept pace with the growth. Consequently, climate scientists grapple with so-called "big data" challenges related to the discovery of significant spatiotemporal associations among interrelated variables. The scientist has an understanding of expected relationships, based on intuition and experience, but serendipitous discoveries are nearly impossible with conventional climate analysis tools.

Climate scientists typically rely on basic, static plots (e.g., trend plots, histograms) that require the use of multiple views since the techniques are limited to at most three variables; but using multiple, non-coordinated views is not an ideal approach due to the limited human memory for information that can be gained from one glance to the next (Rensink, 2002). In addition, statistical analysis methods are typically not integrated with these plots, which further inhibits knowledge discovery. Although many new multivariate, visual analysis techniques have been introduced in recent years, few of these approaches have been brought to bear in climate science. The approaches that do target climate are usually not adopted into practice because of issues related to non-intuitive interfaces and/or a failure to respond to the scientists' needs. Consequently, there is a growing gap between viable visualization techniques and real-world climate analysis. To bridge this gap, experts from both areas must work closely together to create practical systems for today's most pressing problems.

In response to said challenges, we formed a team of researchers with expertise in climate modeling, visualization, and high performance computing across multiple research institutions under

* Corresponding author. Tel.: +1 865 574 7168.
E-mail addresses: csteed@acm.org (C.A. Steed), ricciutodm@ornl.gov (D.M. Ricciuto), gshipman@ornl.gov (G. Shipman), smithbe@ornl.gov (B. Smith), thorntonpe@ornl.gov (P.E. Thornton), wangd@ornl.gov (D. Wang), shix@ornl.gov (X. Shi), williams13@llnl.gov (D.N. Williams).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.07.025

the Climate Science for a Sustainable Energy Future (CSSEF)[1] project to improve the visual analysis of Community Land Model version 4 (CLM4) (Lawrence et al., 2011) simulation data. Our new system, called the Exploratory Data analysis ENvironment (EDEN),[2] is freely available and facilitates interactive knowledge discovery and hypothesis generation for more productive exploratory analysis of climate simulation data. As shown in Fig. 1, EDEN harnesses the high bandwidth human visual channel with interactive parallel coordinates and other coordinated views that guide the scientist to significant associations in the data. EDEN fulfills the requirement for an information visualization centric capability within the context of a broader suite of scientific visualization and analysis tools called the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT).[3] Funded by the Department of Energy (DOE) Office of Science, UV-CDAT provides a number of "big data" analysis tools for climate data such as volume visualizations and other 3-dimensional views.

Through several practical evaluations of EDEN in real-world climate studies, we corroborate the notion that an interactive visual analytics framework yields a more efficient process for climate analysis as compared to conventional tools. Furthermore, our research addresses an important point brought out in the NIH/NSF Visualization Challenges Report (Johnson et al., 2006) which encourages visualization researchers to "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs" – a challenge that is echoed in the more recent strategic vision for DOE Climate and Environmental Sciences Division (CESD) (U.S. Department of Energy, 2012). The tool in the current work is EDEN, the techniques are interactive information visualizations and statistical analytics, and the real-world need is the understanding of earth system simulations and climate change.

2. Related work

In the literature, we find several efforts to improve visual climate data analysis although we note that it is rare to find such systems in practice. For example, Potter et al. (2009) introduced the Ensemble-Vis framework for generating maps, trend charts, and visualizations of climate ensemble data sets. The effectiveness of Ensemble-Vis hinges upon coordinated multiple views (CMV) – a popular approach that has been shown to foster more creative and efficient analysis (Roberts, 2004). With EDEN, CMV is also a key catalyst in the interaction model. However, Ensemble-Vis is apparently devoid of multivariate visualization techniques. Perhaps the most similar approach to EDEN is the visual multivariate data exploration system described by Kehrer et al. (2008). Like EDEN, this system is designed to assist the climate scientist with hypothesis generation for simulation and observational data sets using CMV. The system focuses on brushing extensions that facilitate knowledge discovery using data aggregation and degree of interest functions with promising results. In a follow-on to this work, Ladstädter et al. (2010) add a variant of the parallel coordinates visualization technique to the system, but it is not the focus of the system. In Sips et al. (2012), a matrix visualization technique that supports visual pattern detection in multi-scale, environmental time series data is described. The focus is on a unique visualization technique, called Pinus, with case studies related to the analysis of ocean modeling data sets. Although Pinus does not offer a multivariate visualization technique like parallel coordinates, it accommodates multi-scale analysis via a novel graphical representation. EDEN differs from the above-mentioned systems in its focus on full spectrum analysis – from high level overviews to intermediate views to detailed parallel coordinates plots. EDEN is highly interactive and although aggregation and statistical summaries are provided, access to the individual data elements remain accessible on-demand. Furthermore, the focus of the detailed views is a highly interactive and unique parallel coordinates implementation that is powerful, yet practical for use in climate hypothesis formulation. EDEN provides an alternative visual query interface to the data and is intended to work in conjunction with, rather than to replace, the standard tools that are deeply engrained in the climate scientists toolbox, such as IDL and MatLab. Designed in close collaboration with climate experts, EDEN's intuitive interface has facilitated its early adoption by scientists in ongoing climate studies, thereby overcoming a reluctance to employ unfamiliar techniques that are often difficult to grasp and subject to significant trust issues.

In practice, climate researchers commonly rely on non-interactive, static graphics using decades old techniques (e.g. histograms, trend line charts, and scatter plots); and it is questionable whether these techniques can cope with the complexity of today's "big data" challenges. One approach often used in general multivariate analysis is the scatterplot matrix (SPLOM), which represents multiple adjacent scatterplots for all the variable comparisons in a single display with a matrix configuration (Wong and Bergeron, 1997); but the SPLOM requires a large amount of screen space and forming multivariate associations is still challenging. Wilkinson et al. (2006) used statistical measures for organizing both the SPLOM and parallel coordinates plots to guide the viewer through an exploratory analysis of high-dimensional data sets. Although the organization methods improve the analysis, the previously mentioned perceptual issues with SPLOMs remain. Another alternative is to use layered plots, which condense the information into a single display; but there are significant issues due to layer occlusion and interference as demonstrated by Healey et al. (2004).

At the heart of EDEN is a highly interactive variant of parallel coordinates – a popular multivariate visualization technique that is well-suited to the analysis of large multivariate data sets. The parallel coordinates technique was initially popularized by Inselberg (1985) as an approach for representing hyper-dimensional geometries, and later demonstrated in multivariate analysis by Wegman (1990). In general, the technique yields a compact 2-dimensional representation of even large multidimensional data sets by representing the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ by points on N parallel axes which are joined with a polyline (Inselberg, 2009). In theory, the number of attributes that can be represented in parallel coordinates is only limited by the horizontal resolution of the display device (in Fig. 2 we have a parallel coordinates display that accommodates the simultaneous display of 88 variable axes). But in a practical sense, the axes that are immediately adjacent to one another yield the most obvious information about relationships between attributes. In order to analyze attributes that are separated by one or more axes, interactions and graphical indicators are required. Several innovative extensions that seek to improve interaction and cognition with parallel coordinates have been described in the visualization research literature. For example, Hauser et al. (2002) described a histogram display, dynamic axis re-ordering, axis inversion, and details-on-demand capabilities for parallel coordinates. In addition, Siirtola (2000) presented a rich set of dynamic interaction techniques. The literature covering parallel coordinates is vast and covers multiple domains as recently surveyed by Heinrich and Weiskopf (2013).

EDEN augments the classical parallel coordinates axis by providing cues that guide and refine the analyst's exploration of the information space. This approach is akin to the concept of the scented widget described by Willett et al. (2007). Scented widgets are graphical user interface components that are augmented with

[1] CSSEF website: http://climate.llnl.gov/cssef/
[2] EDEN website: http://cda.ornl.gov/projects/eden/
[3] UV-CDAT website: http://uv-cdat.llnl.gov

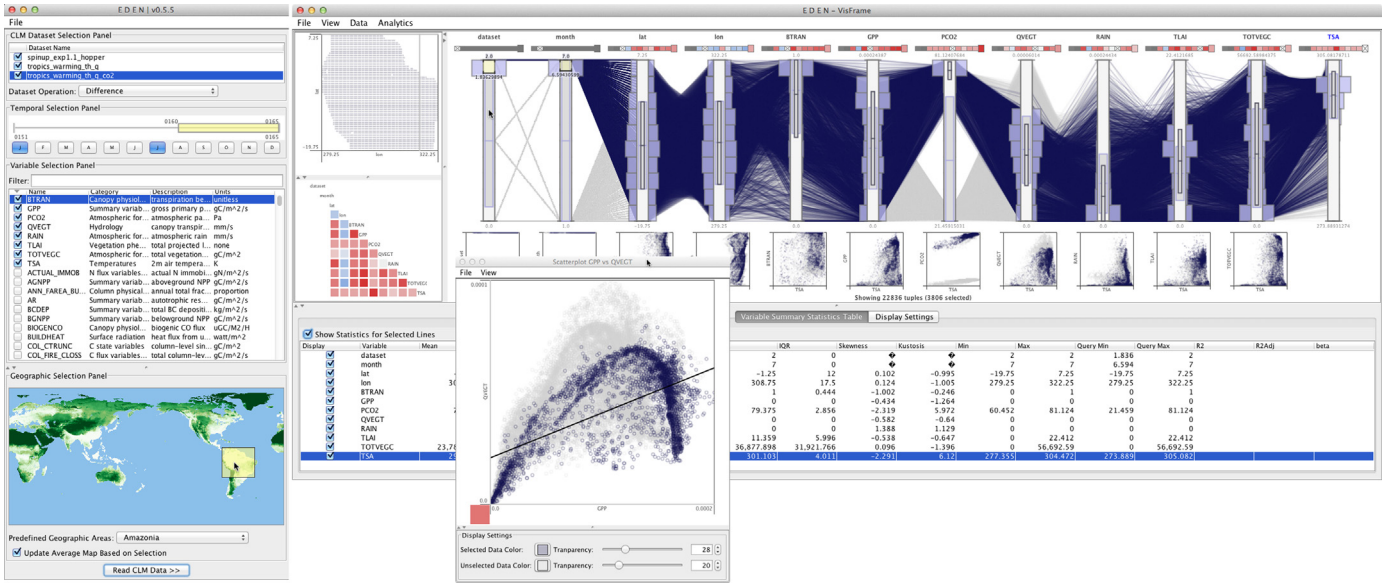

Fig. 1. This figure provides an overview of EDEN during analysis of a global CLM4 data set. The CLM4 filter panel (left) facilitates interactive queries into large CLM4 data sets. The VisFrame (right) offers a highly interactive, visual interface to explore multivariate relationships via linked parallel coordinates, scatterplots, correlation matrix, and geographic scatterplot visualizations. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Fig. 2. An early version of EDEN is used to visually analyze a 1000 simulation CLM4 point ensemble data set with 81 parameters and 7 output variables on ORNL's EVEREST power wall facility which offers 11,520 × 3072 (35 million) pixels.

an embedded visualization to enable efficient navigation in the information space of the data items. The concept arises from the information foraging theory described by Pirolli and Card (1999) which relates human information gathering to the food foraging activities of animals. In this model, the concept of information scent is identified as the "user perception of the value, cost, or access path of information sources obtained by proximal cues" (Pirolli and Card, 1999). The scented axis widgets are also assisted by automated data mining processes that reduce knowledge discovery timelines. In Seo and Shneiderman (2005), a framework is used to explore and comprehend multidimensional data using a powerful rank-by-feature system that guides the user and supports confirmation of discoveries. Piringer et al. (2008) expanded this rank-by-feature approach with a specific focus on comparing subsets in high-dimensional data sets. EDEN is designed to support a similar rank-by-feature framework with subset selection capabilities, correlation mining, and interactive visual analysis.

The parallel coordinate plot is ideal for visual analysis of climate model data because it accommodates the simultaneous display of a large number of variables in a 2-dimensional representation. In EDEN, the parallel coordinates plot is extended with a number of capabilities that facilitate exploratory data analysis and guide the scientist to the most significant relationships in the data. In the following sections, these features are summarized to provide context for the following case studies, but the reader is encouraged to explore our prior publications for more detailed explanations of our multivariate analysis techniques (Steed et al., 2009a, 2009b, 2012). In the current work, these techniques are expanded to address large scale data analysis on a variety of platforms with new evaluations that reveal significant findings.

3. Community land model (CLM) data

Although EDEN is designed for analyzing any multivariate data set, in the current work we focus our attention on CLM4 data sets. CLM4 is the land component of the Community Climate System Model version 4 (CCSM4) (Gent et al., 2011). We have analyzed both $\frac{1}{2}$ degree, global simulations and single location ensemble data. Our global CLM4 simulations contain 360 output variables most of which are 2-dimensional, with some being 3-dimensional. Simulations consist of monthly output files that are typically about 415 megabytes each. For a 100 year simulation, we produce 1200 files totaling about 500 gigabytes. The scientists usually produce multiple simulations, with one control run and several instrumented runs with parameter variations designed to support intercomparisons. Assuming a single control and two additional instrumented simulations, the amount of data to be processed triples and the intercomparison combinations for variables, spatial regions, and temporal ranges grow rapidly, exceeding the capacity of existing tools.

The scientists also produce simulation ensembles using models like CLM4 for sensitivity analysis and uncertainty quantification. Such analysis may produce thousands of different simulations (see Fig. 2). Due to computational costs of running the simulations, these ensemble runs are usually restricted to a single location (or a modest selection of locations) over some time range instead of global results. However, as computing capabilities continue to increase, global ensemble analysis will become more common.

Although the point ensembles are often much smaller than the global simulations, the number of intercomparisons over time,

space, and variable combinations highlights a critical limitation in traditional climate analysis tools. That is, file size is not the only consideration for “big data” analysis. Data complexities such as the number of different intercomparisons in the simulation ensembles, all of which should be considered, can quickly render traditional tools inadequate, even with modest file sizes.

4. Challenges in exploratory earth system simulation analysis

Climate scientists use simulations to explore climate change impacts, attribute these impacts to specific factors, and identify potential abrupt systems changes. Advances in high performance computing continue to feed the development of new high- and variable-resolution simulations. Consequently, the size and complexity of the resulting data sets are exploding (Overpeck et al., 2011). During analysis, these data sets are inevitably reduced, based on the scientists' intuition and experience, and analyzed in isolation. The reductions and isolated investigations hinder holistic intercomparisons with the full data set, but these steps are necessary due to the limitations of conventional climate study tools when confronted with “big data” challenges. Nevertheless, this reduction is troubling because detail is inevitably lost, and the portions of the data that are filtered away may hold unexpected and profound insights about earth system processes. Consequently, we have an unfortunate situation in which climate scientists are forced to reduce the data to fit inadequate tools.

Scientists are understandably cautious about adding to a growing backlog of simulation data. And as they diligently work to analyze today's data sets, which are typically at the gigabyte to terabyte scale, it is clear that when data sizes reach petabytes and beyond, the struggle to effectively analyze and understand the data will be greatly exasperated. To answer pressing scientific questions, we must turn this data overload into opportunity by creating new approaches that effectively blend automated analytics with interactive visualizations in a visual analytics framework. The visual analytics process differs from ordinary visualization in the active role of the computer and the human in guiding the scientist and steering analytical models, respectively (Thomas and Cook, 2005). The active involvement of a human in the analysis task makes visual analytics a supervised or semi-supervised process involving real-time interactions between the computer and the scientist. Furthermore, dynamic summarization and interactive visual queries connect the scientist to the data behind the visualization.

In order to engage the human and machine in data intensive climate analysis, it is imperative that we harness the power of high-performance computing platforms (e.g., ORNL's Titan) and parallelism to efficiently compute statistical summaries and execute data mining algorithms during the analysis session in a manner that encourages human participation. These characteristics paint a

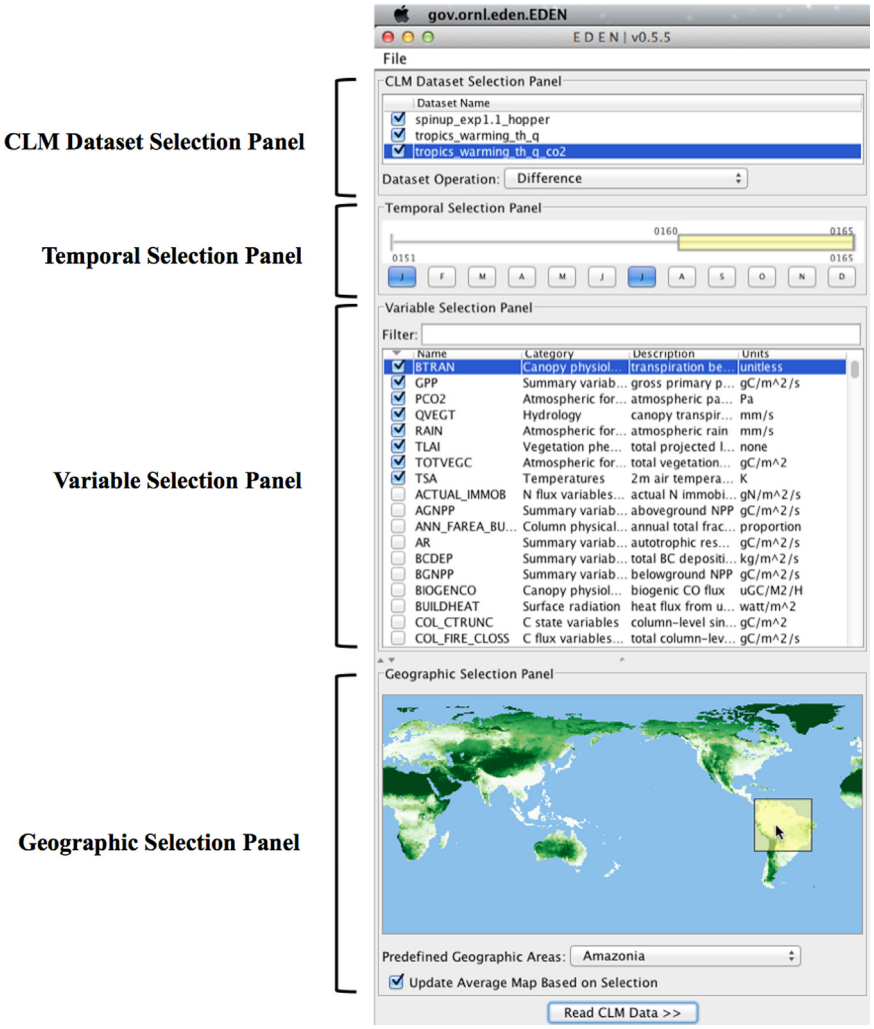

Fig. 3. EDEN provides a filter panel specifically designed for intelligent drill-down to detailed investigations with CLM4 data sets. The filter panel provides access to high-level data set operations, temporal filtering, variable selection, and geographical filtering.

compelling picture of the methodology necessary to enable knowledge discovery for "big data" analysis in earth system simulations, and EDEN is a promising realization of this vision.

5. ParCAT: parallel climate analysis tools

As models and simulations increase in numerical fidelity, parallel computing tools are becoming a necessity for interactive analysis. Originally developed to support EDEN, ParCAT is an independent suite of MPI-based routines that efficiently use parallel computing techniques to facilitate interactive exploratory data analysis at scale (Smith et al., 2013). ParCAT can be used on leadership class supercomputers, such as ORNL's Titan, or local workstations. The toolkit provides the ability to compute spatio-temporal means and variances, differences between simulation data sets, and frequency distributions of the data sets.

ParCAT is designed to execute the "heavy lifting" that is required for large, multidimensional data sets. The toolkit does not focus on performing the final visualization or presentation of results. Instead, it helps reduces large data sets to smaller, more manageable statistical summaries. If ParCAT is installed, EDEN will use it to calculate summary statistics and high-level comparative analysis to augment the graphical displays. If ParCAT is not available, EDEN will default to use a suite of multi-threaded Java routines to offer similar parallelization, although ParCAT yields greater efficiency.

6. EDEN: an exploratory visual analysis framework

EDEN is a robust visual analytics framework that fosters interactive visual queries. EDEN features a multi-faceted filter panel, as well as a highly interactive visual data analysis canvas that integrates parallel coordinates with coordinated, multiple views of the data in the form of interactive scatterplots, a correlation matrix, and a geographic scatterplot. In this section, we will describe several key capabilities in EDEN.

6.1. Visual filtering supported by information scent

As shown in Fig. 3, EDEN's global CLM4 filter panel provides user interface components for visually forming multi-faceted selections using information scent (Willett et al., 2007) to effectively guide the scientist to the most promising relationships. Each panel facilitates interactive queries in terms of the simulation, temporal ranges, variables, and geographic region of interest.

Once loaded, the data sets are listed in the data set selection panel and can be processed in various ways (e.g., differencing, averaging) using the operations combo box. These calculations are performed over the full temporal and geospatial ranges of the data set(s) using ParCAT, if available.

The temporal filter panel facilitates the selection of a span of years and months of interest for the selected years. The user selects the years of interest by dragging a range box in the year timeline. Specific months of interest are selected by using the month toggle buttons below the year timeline. For instance, in Fig. 3 the temporal selection encompasses 5 years of the simulation and the months January and July.

The variable selection panel lists the variable names and associated metadata. With 360 variables in global CLM4 simulations, selecting the variables of interest can be a laborious task from a user interface perspective. To alleviate this challenge, metadata are shown in separate, sortable columns revealing the units, description, and the variable category. In addition, a variable

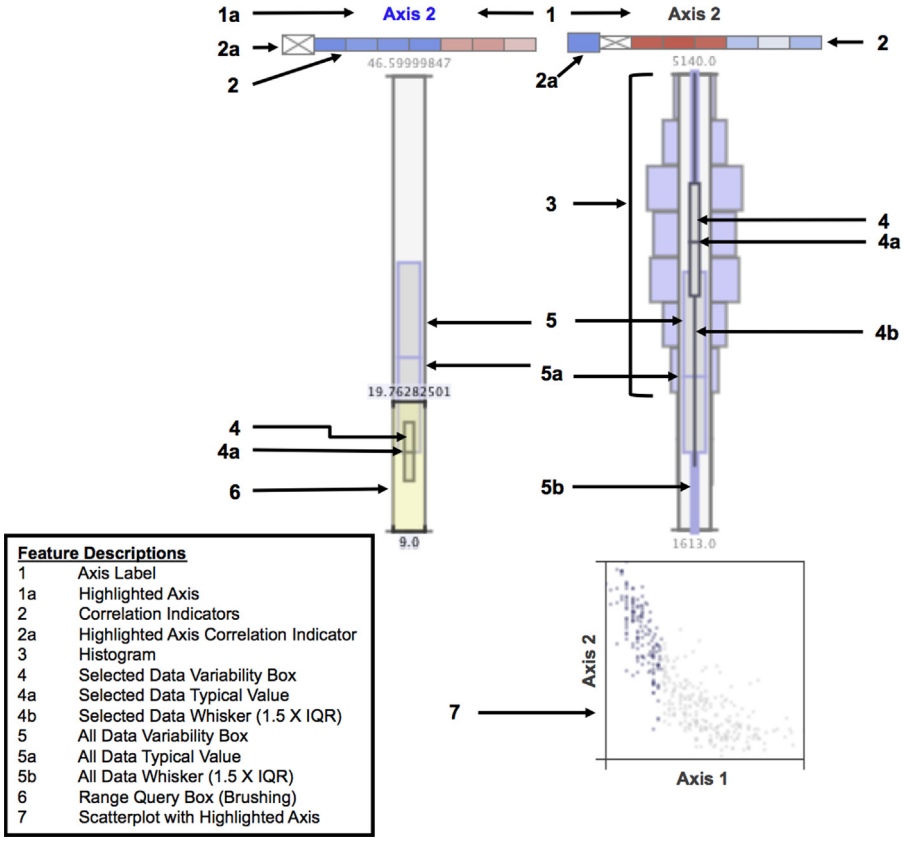

Fig. 4. The parallel coordinate axes are augmented with key descriptive statistics for dynamic summaries, correlation information, and scatterplot displays. Here the parallel coordinate polylines are not shown to emphasize these graphical cues.

filter text field is available to interactively search for keywords from the variable fields. The scientist selects the left-hand check box for the variable(s) of interest for subsequent detailed analysis.

In the geographic panel, the scientist defines an area of interest by dragging a box in the map view or selecting one of the predefined regions. The map image shows the average values for each grid cell of the currently highlighted variable as a color map. In Fig. 3, the average map is shown for the BTRAN (transpiration beta factor) variable. If the "Update Average Map Based on Selection" option is checked, the average map will be generated on-the-fly based on the currently selected time range, data set, and variable of interest. If this option is not enabled, the map will represent the average for all years in the data set for the selected variable and data set.

When the filter criteria have been selected, we click the "Read CLM Data" button to read the data from the simulation(s). When these data are read into the system, a new visual analysis canvas, called the VisFrame (see Fig. 1), is displayed for subsequent detailed investigation.

6.2. Visual exploratory data analysis

The VisFrame is an interactive multivariate visual analysis canvas that is composed of a number of inferential information visualization techniques that are connected together in a coordinated model. The VisFrame is built around a highly interactive variant of the parallel coordinates visualization technique (Inselberg, 1985). A common set of parallel coordinates features are available in EDEN, such as

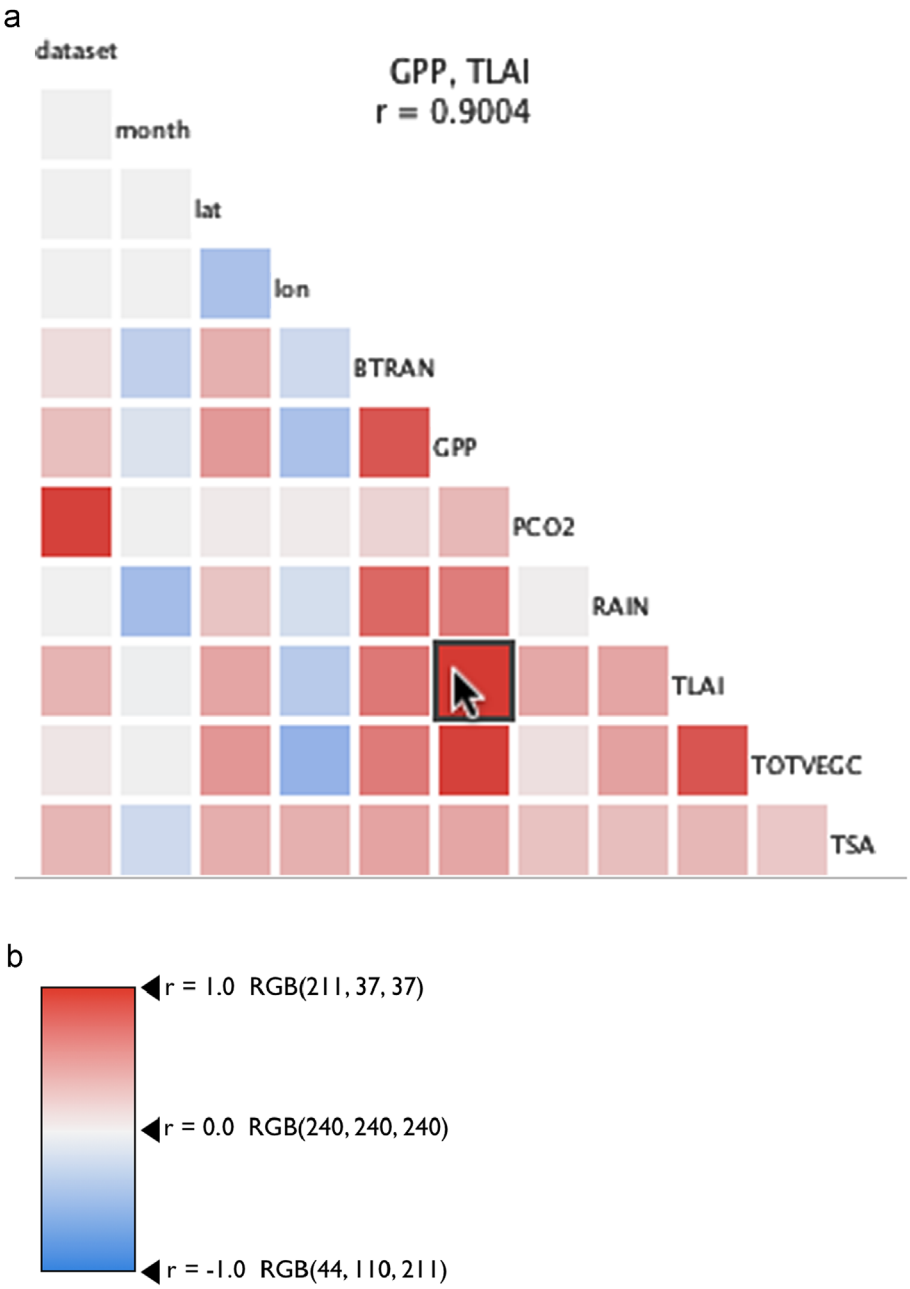

Fig. 5. As the user forms visual queries in the VisFrame, a correlation matrix (a) is dynamically calculated and displayed graphically using red and blue color-filled boxes for negative and positive correlations, respectively. We use a saturation color scale (b) that maps stronger correlations to more saturated colors making these associations more visually salient. (a) Correlation matrix and (b) color scale. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

movable axes, details-on-demand, polyline brushings, and automated axis arrangement. We have extended the parallel coordinate plot with a number of techniques that use statistical analytics to augment the display for guided analysis.

6.2.1. Dynamic dimensional summarization via embedded visualizations

In the parallel coordinates visualization, each vertical axis represents one of the variables selected in the CLM filter panel. Five additional axes are added to the selection: data set id, year, month, latitude, and longitude. The axes are augmented with embedded visual cues that guide the scientists' exploration of the information space (Willett et al., 2007). Scientists can rapidly build visual queries by brushing regions of the axes to select lines of interest (see 6 in Fig. 4) which represent multivariate tuples of the selected data. In this way, multiple queries on separate axes are used to construct conjunctive selections, as shown in Fig. 1, where polylines from data set 2 ('tropics_warming_th_q_co2') and month 7 (July) are selected. The selected polylines are shaded dark gray while the non-selected lines are shaded light gray.

Certain key descriptive statistics are graphically represented in the interior boxes for each axis. The wide boxes (see 5 in Fig. 4) represent the statistics for all axis samples, while the more narrow boxes (see 4 in Fig. 4) capture the statistics for the samples that are currently selected. The statistical displays can be configured to show the mean-centered standard deviations (see left axis in Fig. 4) or a box plot with whiskers (see right axis in Fig. 4). In the standard deviation mode, the height of the box is equal to two standard deviations centered about the mean value which is represented by the thick horizontal line dividing the box (see 4a, 5a in Fig. 4). In the box plot mode, the box height represents the interquartile range (IQR) and the thick horizontal line is the median value. Additionally, the whisker lines (see 4b, 5b in Fig. 4) are shown in the box plot mode. The frequency information can also be displayed on each axis as shaded histogram bins (see 3 in Fig. 4) with widths that are indicative of the number of polylines that pass through the bins sector on the axis.

6.2.2. Detailed exploratory analysis with coordinated scatterplots

As the scientist forms visual queries in parallel coordinates, the interactions are propagated to the other views of the data. One of these views includes a panel of scatterplots shown below the axes. In Fig. 1, the TSA (2-meter air temperature) axis is highlighted resulting in scatterplots that map TSA to the scatterplot's x-axis and the variable representing the axis above the scatter plot to the scatterplot's y-axis (see 7 in Fig. 4). These scatterplots complement the parallel coordinates visualization by providing additional detail such as nonlinear trends, thresholds, and clusters. The scatterplots are linked to the other visualizations so that the shading configuration of the points reflects the current multivariate query in the parallel coordinates display and vice versa. Double clicking one of these scatterplots will display a separate scatterplot window with more detail as shown in Fig. 7. Furthermore, the user can select data points and these selections are propagated to the other views.

6.2.3. Guided visual analysis with graphical correlation indicators

EDEN facilitates visual correlation mining to judge the strength of interrelated variables and visually highlight significant associations. The correlation statistics are updated, based on user selections, and used to augment the displays. For each possible pairing of axes, the system automatically calculates the Pearson product-moment correlation coefficient, r. Given a series of n measurements of the variables X and Y written as x_i and y_i where $i=1,2,\ldots,n$, r is given by evaluating the following equation:

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

This computation yields a correlation matrix where each i, j element is equal to the r value between the i and j variables. This matrix is displayed graphically as shown in Fig. 5. The blocks are encoded with color to indicate the type (blue for negative and red for positive) and strength (stronger correlations receive more saturated colors) of the correlation. Double-clicking a correlation matrix block will cause a separate scatterplot window to display with the variables of interest.

In the parallel coordinates plot, the display can be configured to show rows (which correspond to a variable axis) from the correlation matrix beneath the corresponding axis label (see 2 in Fig. 4). The correlation indicator for the currently highlighted axis is enlarged for each axis' correlation indicator row (see 2a in Fig. 4) to make the relationship more visually salient. For example, in Fig. 1 the TSA axis is highlighted which enlarges the last correlation indicator block for each axis. The strongest correlation with TSA can be determined by seeing the highly saturated red correlation indicator on the PCO2 (atmospheric partial pressure of CO_2) axis.

7. Practical evaluations of EDEN for climate analysis

EDEN is currently used by climate researchers, some of which are co-authors on this paper, to analyze real world CLM4 data sets. Depending on the simulation time range and fidelity, these data sets are typically in the range of terabytes in current experiments with 360 variables. Despite the challenges, EDEN performs well, offering interactive frame rates and efficient filtering and interactive summarization. EDEN accommodates larger scale data sets on high-performance platforms, such as ORNL's Titan supercomputer (the fastest supercomputer in the world at the time of this writing[4]). Climate researchers use EDEN on supercomputers, desktop workstations, and even laptops using a variety of operating systems.

Before EDEN, climate researchers executed the CLM diagnostics package, which consists of scripts that generate several hundred static plots from the data set. The scientists then manually look at the plots to glean interesting associations. EDEN improves on this process by not only providing new multivariate views of the data, but also by providing interactive exploration of the parameter space with dynamic visual queries. In this section, we provide two illustrative case studies demonstrating the power of EDEN in a global case study and a smaller point-based ensemble analysis. These case studies involve real-world simulations and the analysis is driven by the climate scientists who co-authored this paper.

7.1. CLM4 global case study

Using the Global CLM Filter Panel, we analyze global data from an actual experiment to study the sensitivity of tropical carbon fluxes to potential climate change. In this study, the model configurations followed Mao et al. (2013); Shi et al. (2013) and the following analysis involves three CLM4 simulations – one control and two with prescribed increases in temperature and CO_2 input variables. As shown in Fig. 1, we select all three simulations, January and July for 5 years, the Amazonia region, and variables BTRAN, GPP, PCO2, QVEGT, RAIN, TLAI, TOTVEGC, and TSA. Fig. 6 (a) shows the VisFrame after reading the data and averaging the months over the selected years and with the TSA axis selected. The 'tropics_warming_th_q_co2' simulation (data set identifier=2) is

[4] November 2012 Top 500 List: http://www.top500.org/lists/2012/11/

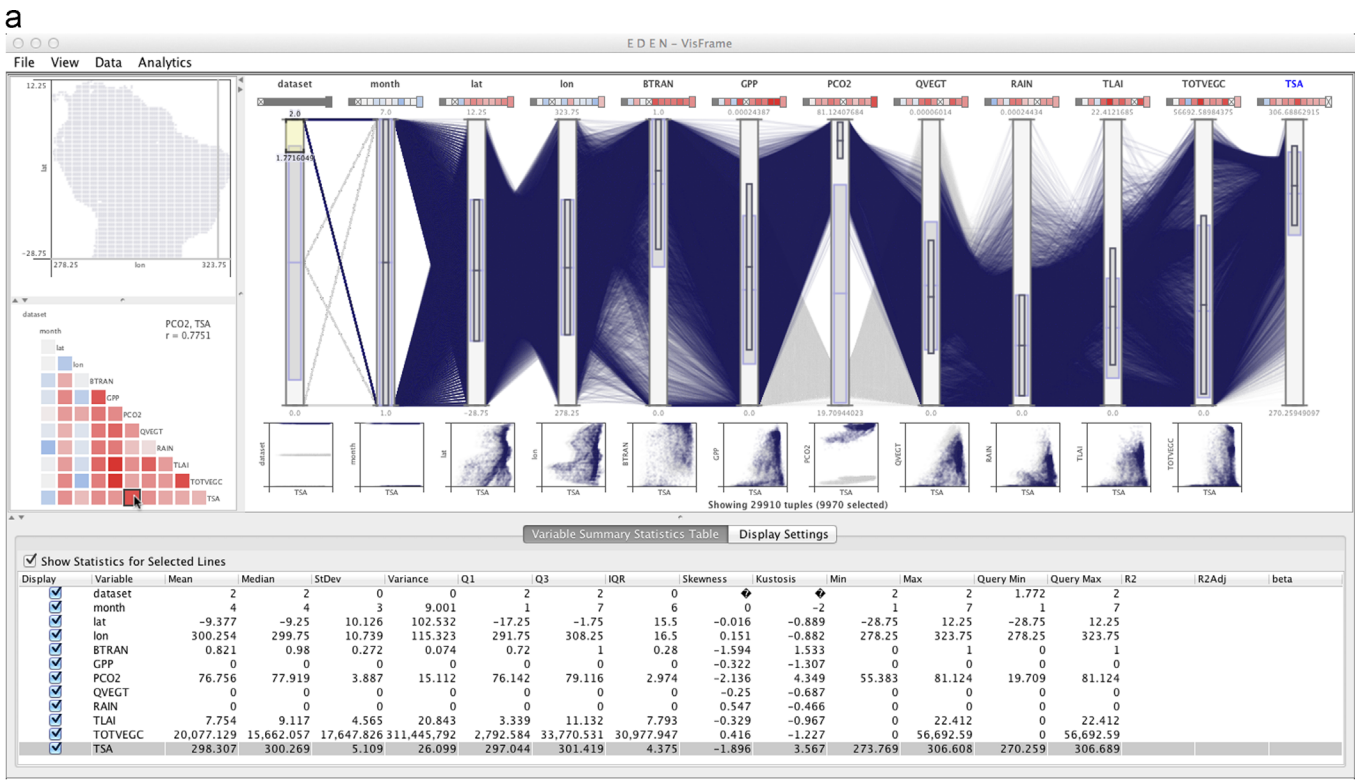

Fig. 6. Data from the Amazonia region is investigated for three CLM4 simulations. We have selected the polylines for a simulation in which CO_2 parameters have been increased. In (a), the TSA axis is highlighted. We observe the PCO2 vs. TSA scatterplot and the parallel coordinate plot to verify that CO_2 has been increased in the selected simulation. In (b), the RAIN axis is highlighted. Also, we have selected polylines for the month of July and the more northern latitudes. The scatterplots for the selection reveal sharp increases that reach a threshold and then either drop or level off.

selected (see the first axis brushing) which shows that the 9970 selected polylines cluster on the upper range of the PCO2 axis (see the PCO2 axis and the scatterplot below it). This observation validates the instrumentation of the selected simulation with increased CO_2 levels. Also, we see that BTRAN, GPP, TLAI, TOTVEGC, and TSA values are increased to a lesser degree in this simulation by observing differences between the axes variability boxes for the selected (narrow boxes) and all (wider boxes)

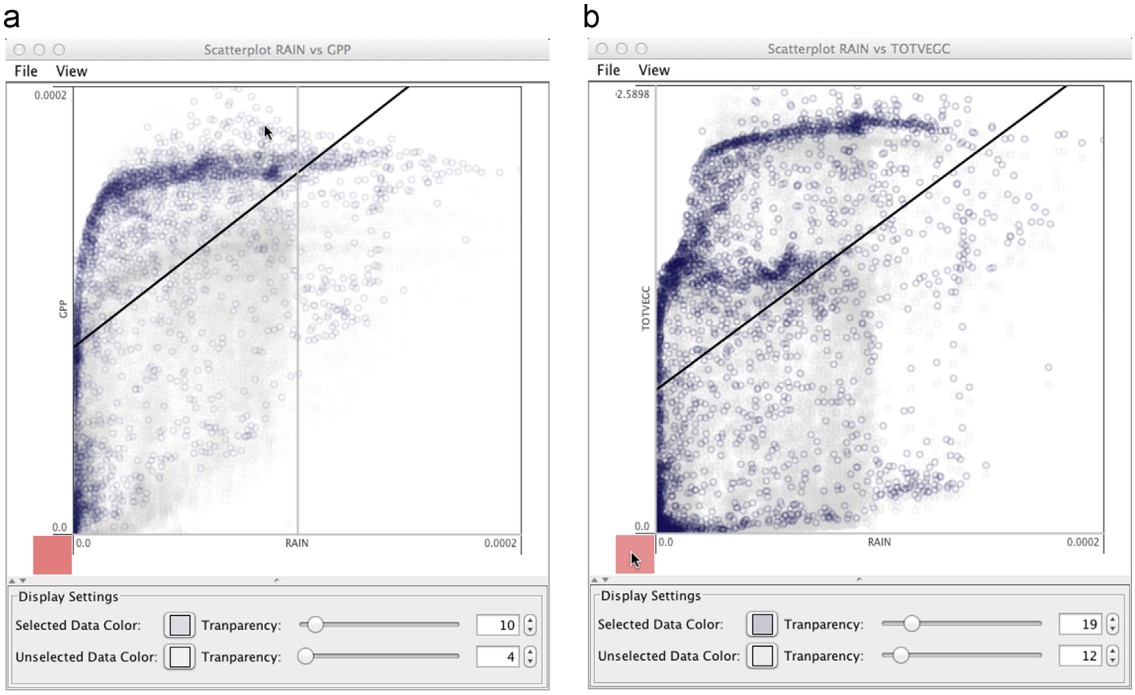

Fig. 7. In these figures, we examine the GPP vs. RAIN (a) and TOTVEGC vs. RAIN (b) relationships from Fig. 6(b) in more detailed scatterplots. In (a), we find that GPP increases very rapidly as RAIN increases up to a certain threshold (RAIN=0.000024 mm/s) above which the slope levels off and GPP increases at a much slower rate. In (b), we see the relationship between TOTVEGC and RAIN and for the CO_2 simulation for the month of January. The plot reveals a strong positively correlated relationship with similar trends as in the GPP vs. RAIN plot, but here we see three clusters of points. The clusters follow the same trends but with different threshold values for RAIN. (a) GPP vs. RAIN and (b) TOTVEGC vs. RAIN.

polylines. The variability boxes also show that transpiration (QVEGT) values are typically lower and RAIN values are nearly equal for the selected simulation. Precise numerical listings of the summary statistics are shown in the table beneath the parallel coordinates plot. For example, the TSA average is about 298.3 K for the 'tropics_warming_th_q_co2' simulation (see Fig. 6(a)) which is about 3 K warmer than the control simulation (not shown). These observations reveal one of the major findings from the experiment: CO_2 and temperature increases cause an increase in gross primary production (GPP). The ability to derive this information shows how the EDEN framework facilitates not only intuitive exploratory data analysis, but also validation and verification of the parameter instrumentation.

In Fig. 6(b), we have the same data shown in Fig. 6(a), except now the RAIN axis is highlighted and the mid to upper latitude range is selected (see the geographic plot and 'lat' axis). We can identify interesting features in the parallel coordinate plot, such as the clustering of three groups of data on the TOTVEGC axis. But we also glean insight from the shapes in the scatterplots below the axes. In nearly all the selected variable axes, we see that as RAIN increases it produces sharp increases in the *y*-axis variables up to a certain threshold, above which the *y*-axis variables either level off or decrease slightly.

In Fig. 7, we have double-clicked on the GPP vs. RAIN and the TOTVEGC vs. RAIN scatterplots to show a detailed view. In Fig. 7(a) we see a sharp increase in GPP until RAIN reaches a value of about 0.000024 mm/s, above which GPP seems to level off significantly. In Fig. 7(b) we see similar trends in the shapes of the point profiles, but here we see three clusters of points, as noted above, which correspond to three different thresholds but with similar overall trends. It is also significant to note that the same threshold for GPP vs. RAIN noted above appears to be a threshold in the TOTVEGC vs. RAIN relationship.

In Fig. 1, the scatterplot of GPP vs. QVEGT reveals an interesting nonlinear relationship for 'topics_warming_th_q_co2' data set and the month of July. As GPP increases, QVEGT increases sharply and then begins to level off until a threshold is reached, above which QVEGT begins to decrease sharply. These visually detectable thresholds and profiles are examples of unexpected discoveries that the climate scientists were not aware of until exploring the data in EDEN. The ability to interactively explore relationships with visual guidance fed by statistical analytics is a profound concept that paves the way for more extensive scientific inquiries.

7.2. *CLM4 point ensemble case study*

EDEN is also used for detailed analysis of CLM4 point ensembles. The data set is for the location of the Harvard Forest eddy covariance flux tower (42.5378N, 721715W) and includes 11 variables (3 parameters and 8 output variables) averaged for the month of May over 10 years (1995–2004). This data set contains 134 model simulations in which the 3 parameters were varied to examine parameter sensitivities in the 8 output variables (Hou et al., 2012).

In Fig. 8(a), the initial parallel coordinates plot is shown after loading the data set. From this plot, several interesting features in the data are revealed. First of all, the variability boxes on the first 3 axes (these are the parameter axes) reveal uniform distributions, reflecting the sampling strategy used to generate the model parameters. The statistical variation boxes on the other eight axes suggest skewed distributions toward either the minimum (e.g., runoff and SH) or maximum values. In Fig. 8(a), the SH (sensible heat flux) axis is highlighted, revealing strong correlations (notice the highly saturated correlation indicator blocks that are enlarged below each axis label) with the other 7 output variables (runoff, LH, SH, GPP, TLAI, TOTVEGC, and TOTSOMC). These strong correlations reflect the interrelated nature of the model outputs. For example, low values of total runoff (runoff) and SH as described above occur in simulations in which there are low amounts of foliage (low TLAI and TOTVEGC). Low foliage means low trans-

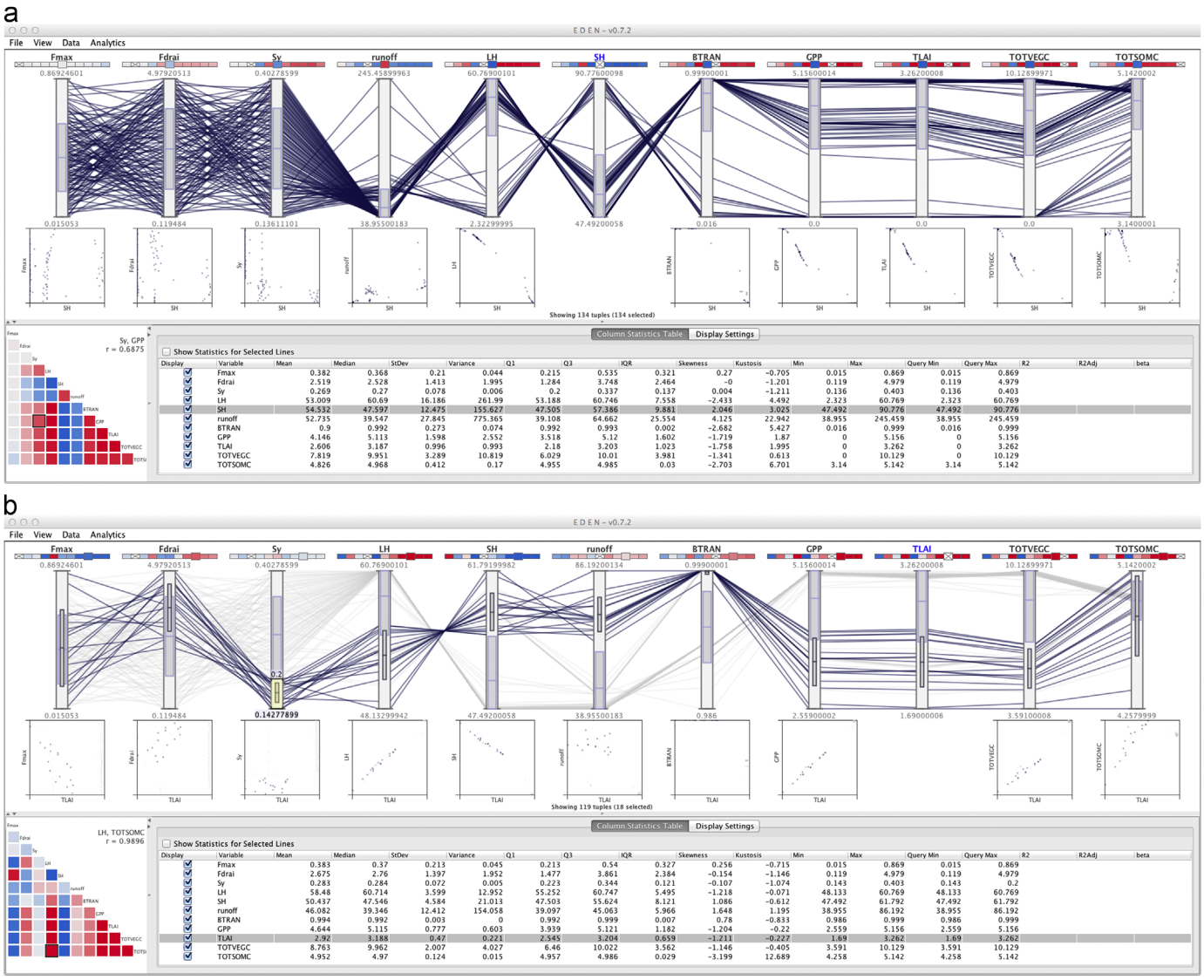

Fig. 8. In (a), the initial view of the Harvard point ensemble data set is shown after loading into EDEN. With the SH axis highlighted, this view reveals strong correlations with the other 6 output variables and a moderately strong correlation with the runoff variable. In (b), we have selected polylines below the Sy parameter threshold of 0.2 (see brushing on Sy axis). These tuples all fall into one cluster of polylines, which indicates a particular sensitivity for this parameter. This is the central promise of such exploratory tools as EDEN and the exact kind of insight scientists seek to glean from these data sets.

piration (a component of LH), leaving more soil water available for runoff and more net energy available for sensible heat. These strong feedbacks between vegetation and soil hydrology are one example of many where potentially unexpected relationships exist between indirectly connected CLM4 model parameters and outputs.

Although Fig. 8(a) is useful for representing broad relationships among the variables, outliers exercise too much influence on the display making it difficult to form more detailed judgments on the correlation and distribution patterns. We select the 15 outlier polylines and remove them from the display (see Fig. 8(b)). With the outliers removed, we see more structure in both the parallel coordinates and the scatterplots. Next, we highlight the polylines below the Sy (average specific yield) parameter value of 0.2. We find that simulations with high SH correspond to low values for LH and high values for BTRAN. Conversely, the lines crossing the upper range of SH have a negative correlation pattern with LH (notice the 'X' shaped crossing for the selected lines between the SH and LH axes in Fig. 8(b)). The bimodality appears to be driven by the threshold value of Sy near 0.2, above which runoff is near the minimum value and below which there is increased runoff and also increased variability of this and other output variables. In this study, EDEN has helped to identify a particular parameter sensitivity with the Sy parameter, thus, corroborating the notion that EDEN improves this type of exploratory analysis. This type of study is vital to climate researchers in their identification of thresholds and tipping points, uncertainty quantification, and other similar tasks that will benefit from interactive exploration.

8. Discussion

The general success of EDEN and its broader adoption in the climate community can be largely attributed to the fact that the domain experts were tightly integrated in the development process and subsequent design iterations. This was a intentional strategy from the start, and we believe that it mutually benefited all parties involved. Initially, we focused our attention on the general techniques and overall visual design, but our effort reached a critical mass, as it were, when we identified a clear case

study to focus our efforts. As we witness the evolution of EDEN, it is clear that our frequent interactions guarantee that EDEN responds to the actual needs of the end-users, thus enhancing adoption in the larger community. It is also important to note that although the climate scientists were not familiar with parallel coordinates or scented widgets, they quickly learned the concepts and foresaw the benefits more rapidly as members of the development team. Furthermore, the flexibility of the interface to accommodate dynamic visual queries is a major improvement over the laborious process of sifting through hundreds of plots from pre-determined queries executed in external scripts. With interactive feedback, the scientist is effectively connected to the data behind the visualization, thereby enhancing the process of formulating and confirming hypotheses.

Because EDEN is an ongoing development, we have identified some areas to further improve its capacity to surmount challenges of climate analysis. The inability to support direct comparisons between different queries, in time, space, and between different simulations, is a key feature that is not currently supported in EDEN. However, some of the basic building blocks are available in EDEN, and we are currently formulating new methods to facilitate these comparisons as well as support for observational data. It has been noted that the VisFrame map, a geographic scatterplot, is limited in its functionality. We are currently expanding the map to facilitate navigation and new multivariate visualizations. Because EDEN's visual query interface is very flexible, it can be difficult to resume analysis after closing and restarting the application. Thus, we are investigating mechanisms to save configurations and track workflow provenance. One option for doing this would be to take advantage of the VisTrails mechanism which is embedded in UV-CDAT (Bavoil et al., 2005). We are also interested in exploring *in situ* analysis of the climate simulation data during the actual model execution to streamline and reduce knowledge discovery timeframes. Finally, we are adding new data mining and machine learning algorithms that can automatically suggest associations and intuitively encourage human participation in analytical models.

9. Conclusion

EDEN is a unique multivariate analysis tool that has been designed in close collaboration with climate scientists to address the exploratory analysis and “big data” challenges inherent to today's climate science. EDEN is an ongoing project that is in use today for several real-world climate studies and it is freely available for download. In our case studies and through EDEN's practical use, we demonstrate the promise of an interactive visual analytics approach for more productive climate analysis. EDEN and other tools developed in this same spirit deliver on the central promise of visual analytics. Such techniques will be key enablers for fielding test-bed environments that climate scientists desperately need.

Acknowledgments

We wish to express our gratitude to the reviewers and editorial staff of Computers & Geosciences for their valuable feedback. We would also like to thank Jiafu Mao (ORNL) and Zhangshaun Hou (PNNL) for generating the global simulation data sets and parameter samples for the point simulations, respectively. This research is sponsored by the Office of Biological and Environmental Research; U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the Department of Energy, under Contract no. DE-AC05-00OR22725. This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract no. DE-AC0500OR22725.

References

Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T., Vo, H.T., 2005. Vistrails: enabling interactive multiple-view visualizations. In: IEEE Visualization 2005, pp. 135–142.

Gent, P.R., Danabasoglu, G., Donner, L.J., Holland, M.M., Hunke, E.C., Jayne, S.R., Lawrence, D.M., Neale, R.B., Rasch, P.J., Vertenstein, M., Worley, P.H., Zong-Liang, Yang, Zhang, M., 2011. The community climate system model version 4. Journal of Climate 24 (19), 4973–4991.

Hauser, H., Ledermann, F., Doleisch, H., October 2002. Angular brushing of extended parallel coordinates. In: Proceedings of IEEE Symposium on Information Visualization, pp. 127–130.

Healey, C.G., Tateosian, L., Enns, J.T., Remple, M., 2004. Perceptually-based brush strokes for nonphotorealistic visualization. ACM Transactions on Graphics 23 (1), 64–96.

Heinrich, J., Weiskopf, D., 2013. State of the art of parallel coordinates. In: Proceedings of the European Association for Computer Graphics, Eurographics, Girona, Spain, pp. 95–116.

Hou, Z., Huang, M., Leung, L.R., Lin, G., Ricciuto, D.M., 2012. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the community land model. Journal of Geophysical Research: Atmospheres 117 (D15), 1–18.

Inselberg, A., 1985. The plane with parallel coordinates. The Visual Computer 1 (4), 69–91.

Inselberg, A., 2009. Parallel coordinates: interactive visualization for high dimensions. In: Zudilova-Seinstra, E., Adriaansen, T., Liere, R. (Eds.), Trends in Interactive Visualization. Springer-Verlag, London, UK, pp. 49–78.

Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P., Yoo, T.S. (Eds.), 2006. NIH/NSF Visualization Research Challenges. IEEE Press.

Kehrer, J., Ladstädter, F., Muigg, P., Doleisch, H., Steiner, A., Hauser, H., 2008. Hypothesis generation in climate research with interactive visual data exploration. IEEE Transactions on Visualization and Computer Graphics 14 (6), 1579–1586.

Ladstädter, F., Unger, A., Lackner, B.C., Pirscher, B., Kirchengast, G., Kehrer, J., Hauser, H., Muigg, P., Doleisch, H., 2010. Exploration of climate data using interactive visualization. Journal of Atmospheric and Oceanic Technology 27 (4), 667–679.

Lawrence, D.M., Oleson, K.W., Flanner, M.G., Thornton, P.E., Swenson, S.C., Lawrence, P.J., Zong-Liang, Yang, Levis, S., Sakaguchi, K., Bonan, G.B., Slater, A.G., 2011. Parameterization improvements and functional and structural advances in version 4 of the community land model. Journal of Advances in Modeling Earth Systems 3 (M03001), 27.

Mao, J., Shi, X., Thornton, P.E., Hoffman, F.M., Zhu, Z., Myneni, R.B., 2013. Global latitudinal-asymmetric vegetation growth trends and their driving mechanisms: 1982–2009. Remote Sensing 5 (3), 1484–1497.

Overpeck, J.T., Meehl, G.A., Bony, S., Easterling, D.R., 2011. Climate data challenges in the 21st century. Science 331 (6018), 700–702.

Piringer, H., Berger, W., Hauser, H., July 2008. Quantifying and comparing features in high-dimensional datasets. In: International Conference on Information Visualization. IEEE Computer Society, London, UK, pp. 240–245.

Pirolli, P., Card, S.K., 1999. Information foraging. Psychological Review 106 (4), 643–675.

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

Rensink, R.A., 2002. Change detection. Annual Review of Psychology 53, 245–577.

Roberts, J.C., 2004. Exploratory visualization with multiple linked views. In: Exploring Geovisualization. Elseviers, pp. 159–180.

Seo, J., Shneiderman, B., 2005. A rank-by-feature framework for interactive exploration of multidimensional data. Information Visualization 4 (2), 96–113.

Shi, X., Mao, J., Thornton, P.E., Huang, M., 2013. Spatiotemporal patterns of evapotranspiration in response to multiple environmental factors simulated by the community land model. Environmental Research Letters 8 (2), 1–12.

Siirtola, H., 2000. Direct manipulation of parallel coordinates. In: Proceedings of the International Conference on Information Visualisation, IEEE Computer Society, London, England, pp. 373–378.

Sips, M., Köthur, P., Unger, A., Hege, H.-C., Dransch, D., 2012. A visual analytics approach to multiscale exploration of environmental time series. IEEE Transactions on Visualization and Computer Graphics 18 (12), 2899–2907.

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

Steed, C.A., Fitzpatrick, P.J., Jankun-Kelly, T.J., Yancey, A.N., Edward Swan II, J., 2009a. An interactive parallel coordinates technique applied to a tropical cyclone climate analysis. Computers & Geosciences 35 (July (7)), 1529–1539.

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

Steed, C.A., Jankun-Kelly, T.J., Fitzpatrick, P.J., October 2009b. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In: IEEE Symposium on Visual Analytics Science and Technology. IEEE Computer Society, Atlantic City, NJ, pp. 19–26.

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

U.S. Department of Energy, 2012. Biological and Environmental Research, Climate and Environmental Sciences Division: Strategic plan. Technical Report, Office of Science.

Wegman, E.J., 1990. Hyperdimensional data analysis using parallel coordinates. Journal of the American Statistical Association 85 (411), 664–675.

Wilkinson, L., Anand, A., Grossman, R., 2006. High-dimensional visual analytics: interactive exploration guided by pairwise views of point distributions. IEEE Transactions on Visualization and Computer Graphics 12 (November (6)), 1366–1372.

Willett, W., Heer, J., Agrawala, M., 2007. Scented widgets: improving navigation cues with embedded visualizations. IEEE Transactions on Visualization and Computer Graphics 13 (November–December (6)), 1129–1136.

Wong, P.C., Bergeron, R.D., 1997. 30 years of multidimensional multivariate visualization. In: Scientific Visualization—Overviews, Methodologies, and Techniques. IEEE Computer Society Press, pp. 3–33.