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.



## 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