SciDAC 2009 IOP Publishing
Journal of Physics: Conference Series **180** (2009) 012089 doi:10.1088/1742-6596/180/1/012089

Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated ViSUS-CDAT systems

Kristin Potter[1], Andrew Wilson[2], Peer-Timo Bremer[3], Dean Williams[3], Charles Doutriaux[3], Valerio Pascucci[1], and Chris Johhson[1]

[1]Scientific Computing and Imaging Institue, University of Utah
[2]Sandia National Laboratories
[3]Lawrence Livermore National Laboratory

E-mail: `kpotter@cs.utah.edu`, `atwilso@sandia.gov`, `{bremer5,williams13,doutriaux1}@llnl.gov`, `{pascucci,crj}@sci.utah.edu`

Abstract. Climate scientists and meteorologists are working towards a better understanding of atmospheric conditions and global climate change. To explore the relationships present in numerical predictions of the atmosphere, *ensemble datasets* are produced that combine time- and spatially-varying simulations generated using multiple numeric models, sampled input conditions, and perturbed parameters. These data sets mitigate as well as describe the uncertainty present in the data by providing insight into the effects of parameter perturbation, sensitivity to initial conditions, and inconsistencies in model outcomes. As such, massive amounts of data are produced, creating challenges both in data analysis and in visualization. This work presents an approach to understanding ensembles by using a collection of statistical descriptors to summarize the data, and displaying these descriptors using variety of visualization techniques which are familiar to domain experts. The resulting techniques are integrated into the ViSUS/Climate Data and Analysis Tools (CDAT) system designed to provide a directly accessible, complex visualization framework to atmospheric researchers.

1. Introduction

In order to better understand the long-term trends in climate change or the reliability of short-term weather predictions, it is essential to characterize the uncertainty of the data underlying scientific conclusions [9]. To this end, researchers are generating *ensemble data sets* which combine multiple runs of spatio-temporal simulations, using perturbed parameters and variations on initial conditions [6]. The resulting data sets provide a collection of estimations for each simulation variable, which in turn provides an understanding of the possible outcomes and the corresponding uncertainty. The main challenge to this work is addressing the increasing need to visualize uncertainty information [7, 12] in a manner that leads to improved understanding of the data.

The driving applications of this work reside in the fields of climate modeling and short term weather forecasting. Both areas combine processes underlying atmospheric behavior such as the global carbon cycle, atmospheric chemistry, vegetation, and ocean dynamics. The main

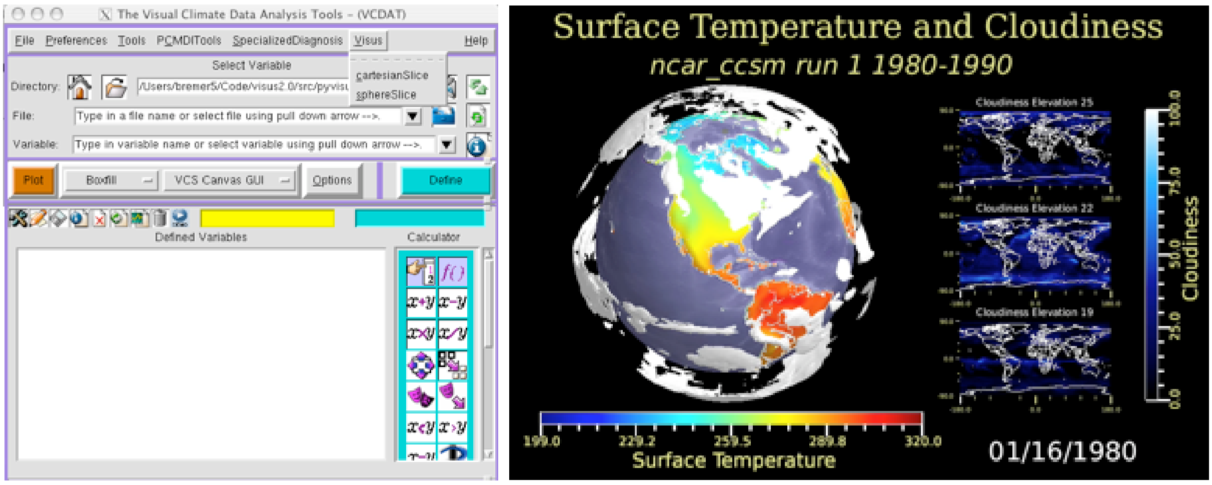

Figure 1. Screen shot of ViSUS 2.0 executed within CDAT. ViSUS/CDAT provides the user with a complex interface, multiple views, and a variety of visualization techniques for data exploration and analysis. (Left) The user interface of the VCDAT system. (Right) The data is visualized using 3D iso-contouring, slicing, and multiple 2D plots.

distinctions between the two applications stem from the spatial and temporal domains [5]. Climate models encompass the entire globe and are run over hundreds of years. The main goal is to understand global phenomena such as the impact of human activity on land and ocean temperatures or trends in natural disasters [1]. In contrast, short-range ensemble forecasting (SREF) data [2] focuses on North America, and predicts atmospheric variables such as temperature and humidity out to 87 hours (roughly three days). This data is used to predict near future weather patterns and is run on a daily basis, updating the initial conditions with observed data. While this work focuses on the challenges specific to the analysis and visualization of atmospheric data [11], the approaches presented here can be generalized to the many ensemble data sets becoming prevalent today.

ViSUS/CDAT, shown in Figure 1, is the framework used to develop the techniques for visualizing ensemble data and its corresponding uncertainty. This tool set is developed as a collaboration with the Earth System Grid Center for Enabling Technologies (ESG-CET), A Scalable and Extensible Earth System Model for Climate Change Science, and the climate modeling community in general. CDAT, or the Climate Data Analysis Tools, are a set of utilities specifically designed for the needs of climate researchers, providing advanced data analysis combined with the ability of reading specific data formats, supplying geospatial information and supporting a high level user interface. ViSUS is an integrated 3D visualization package, providing sophisticated visualization techniques seamlessly to the climate scientists. The incorporation of ViSUS with CDAT provides (among other things) a flexible system for the visualization of large ensemble datasets and is founded on the tight integration of 3D visualization tools, statistical data analysis techniques and compelling metaphors that facilitate the user intuition about the uncertainty of large, multidimensional, multivariate, time-varying datasets.

2. Ensembles and Uncertainty

An *ensemble data set* can be defined as a collection of multiple time-varying data sets (called *ensemble members*) that are generated by computational simulations of one or more state variables across space. The variation among the ensemble members arises from the use of different input conditions, simulation models, and simulation parameters. The generated data

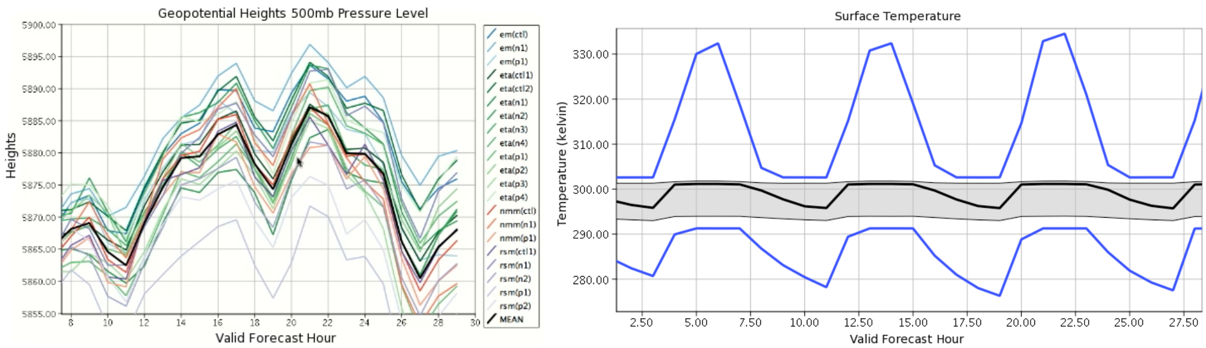

Figure 2. (Left) An example of the complexity of an ensemble data set. Here, the geopotential heights of the 500mb pressure level are shown at a single weather station across all valid forecast hours. Each model/perturbation group is shown using a different color, and the mean is shown in black. (Right) A quartile chart showing the range of surface temperatures across the valid forecast hours. The median of the data is shown in black, the inner quartile range shaded in grey, and the minimum and maximum ensemble members are in blue.

sets are multidimensional, since the spatial domain of the simulations can be 2 or 3D (or higher), multivariate, due to the possibly hundreds of state variables, and multivalued by collecting several values for each variable in space.

Ensemble data sets are chiefly useful as a tool to quantify and mitigate uncertainty and error in simulation results. Uncertainty can be described as the accuracy or confidence associated with the data. Errors can arise in the simulation through faulty estimations or measurements of the initial conditions, finite resolution and precision of the numerical model, sensitivity to input parameters, and from the nature of a numerical simulation as an approximate model of an incompletely understood real-world phenomenon. Ensembles alleviate the uncertainty present in a single, deterministic data set by providing a collection of outcomes, from which a best guess can be dervied with some level of confidence.

2.1. Challenges for Analysis and Visualization

The complex nature of the ensemble data creates challenges in data display due to the fact that it is multidimensional [3], multivariate [4, 10, 14], and multivalued [8]. For example, each of the four daily runs of the SREF ensemble contains 21 members comprising four models and eleven sets of input conditions. Each member contains 624 state variables at each of 24,000 grid points and includes 30 time steps. A single day's output contains 84 members, each of which is a complex data set that poses visualization challenges in its own right. When information from all members is displayed together, as in the plume chart in Figure 2, the result is visual chaos that conveys only a general notion of the behavior of the predicted variable. These challenges are exacerbated in more complex data sets such as the climate simulations that incorporate 24 different models instead of four.

3. Visualization of Ensembles

The most straightforward approach to reducing the complexity of ensemble data is to summarize through mean and standard deviation. Mean provides a general overview of the outcome of the ensemble by averaging across the members. Standard deviation reflects the variation of the ensemble and is representative of how well the mean characterizes the data. The combination of these two measures provides a strong indication of the behavior of the ensemble, and while they may not capture nuances of the underlying distribution, they are often appropriate since many observed quantities and natural phenomena are well modeled by a normal distribution [13].

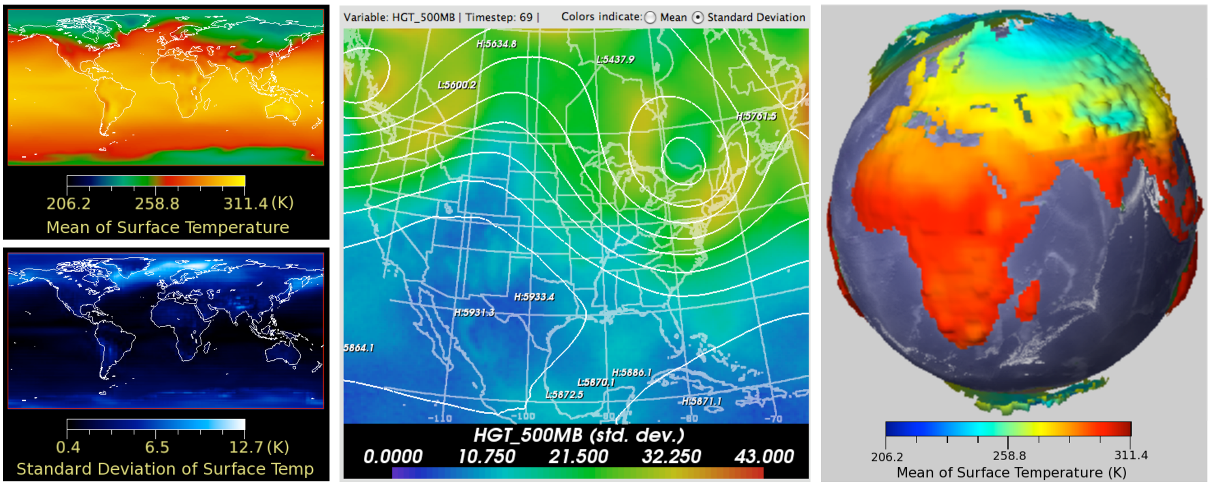

Figure 3. Visualizations of mean and standard deviation. (Left) Mean and standard deviation are visualized independently using color maps. (Center) Mean is presented through a color map, and standard deviation is shown as an overlaid contour. (Right) Standard deviation is mapped to a height field and mean is color mapped.

The ViSUS/CDAT framework provides a variety of techniques for visualizing mean and standard deviation, some of which are shown in Figure 3. Color maps, contours, and height fields are used independently or in combination to indicate the ensemble estimate as well as the confidence of that estimation. These visualizations provide a global understanding of the ensemble behavior, and identify regions where the results of the ensemble runs are questionable. In addition, ViSUS annotates the data with domain specific information, such as geographical and political maps, and latitude and longitude lines.

While mean and standard deviation are excellent tools for summarization, these measures discard large amounts of information available from ensembles data sets. Model biases, extrema, clusters, ranges, and other characteristics are hidden through this type of abstraction and, depending on the motivations of the driving application, may be of great importance. For this reason we provide additional visualization techniques to allow the user to further explore the ensemble.

Two dimensional charts present the data by removing the spatial component, either through summarization across space, or the selection of specific locations. This is helpful when specific locations in the data, or trends through time are of interest. The plume chart, Figure 2, left, shows the visualization of all ensemble members across time. While this plot reduces the overall data, it is still too visually cluttered to assist in data analysis beyond giving a notion of the general outcome. However, using a plot of this type to explore a specific model, or collection of model perturbations can help the user determine the sensitivity of a model to changes in parameters or if a model commonly predicts low or high values. The quartile chart, shown in Figure 2, right, simplifies the plume chart by expressing only the quartiles of the data. This type of chart conveys the range of the data and indicates where the central 50% of the data occurs. The median, or central value of the data is shown as a thick black line. This value is distinctive from the mean in that it is not affected by outlying data values and can indicate the most centrally positioned member.

Each of the visualization techniques presented here reduce the complexity of the data in some dimension. Mean/standard deviation displays show a single instance of time, and the charts compact the spatial domain. Depending on the needs of the researcher, the questions being asked of the data, and the scientific goals behind the ensemble, one paradigm may be more appropriate than another. The benefit of working within the VCAT framework is the ability to use more than one visualization method, thus exploiting the advantages of each. Within this system, the user can select a variable, choose the visualization technique, and interactively explore the data. Such a system provides a flexible environment for the investigation and visual analysis of ensemble data.

4. Conclusion

The rapid increase in computational capacity over the past decade has rendered ensemble data sets a viable tool for mitigating uncertainty. We believe that our work constitutes early progress toward the many new challenges posed by these large, complex and rich data sets. Future work will include the visualization of higher dimensional probability distribution functions, the use of non-normal distributions and higher order descriptive statistics, and the integration of statistical measures into more sophisticated visualization techniques.

5. Acknowledgements

This work was funded in part by the DOE SciDAC Visualization and Analytics Center for Enabling Technologies (www.vacet.org) and the NIH NCRR Center for Integrative Biomedical Computing (www.sci.utah.edu/cibc), NIH NCRR Grant No. 5P41RR012553-02.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] Climate of the 20th century experiment (20c3m). https://esg.llnl.gov:8443/index.jsp.
[2] Short-range ensemble forecasting. http://wwwt.emc.ncep.noaa.gov/mmb/SREF/SREF.html.
[3] A. Buja, D. Cook, and D. F. Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5(1):78–99, March 1996.
[4] R. Bürger and H. Hauser. Visualization of multi-variate scientific data. In *Eurographics 2007 STAR*, pages 117–134, 2007.
[5] G. Compo, J. Whitaker, and P. Sardeshmukh. Bridging the gap between climate and weather. http://www.scidacreview.org/0801/html/climate.html, 2008.
[6] T. Gneiting and A. Raftery. Atmospheric science: Weather forecasting with ensemble methods. *Science*, 310:248–249, October 2005.
[7] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE CG & A*, 23(5):6–10, 2003.
[8] A. Love, A. Pang, and D. Kao. Visualizing spatial multivalue data. *IEEE CG & A*, 25(3):69–79, May 2005.
[9] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, July 2005.
[10] T. Mihalisin, J. Timlin, and J. Schwegler. Visualization and analysis of multi-variate data: a technique for all fields. In *IEEE Vis '91*, pages 171–178, 1991.
[11] T. Nocke, M. Fleshig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *IEEE 2007 Water Simulation Conference*, pages 703–711, 2007.
[12] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, Nov 1997.
[13] J. Sivillo, J. Ahlquist, and Z. Toth. An ensemble forecasting primer. *Weather Forecasting*, 12:809–817, 1997.
[14] Z. Xie, S. Huang, M. Ward, and E. Rundensteiner. Exploratory visualization of multivariate data with variable quality. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 183–190, 2006.