Time-Varying Multivariate Visualization for Understanding Terrestrial Biogeochemistry

Robert Sisneros[1], Markus Glatter [1], Brandon Langley[1], Jian Huang[1], Forrest Hoffman[2], and David J. Erickson III[2]

[1] University of Tennessee, Knoxville, TN, [2] Oak Ridge National Laboratory, Oak Ridge, TN

Email: {sisneros, glatter, langley, huangj}@cs.utk.edu, {hoffmanfm, ericksondj}@ornl.gov

Abstract. Petascale computing has brought forth a transformational way of science. To the global effort on studying climate change, this shift has enabled not only tools more functional and more powerful than before, but also a scientific exploration more comprehensive than before. In this work, we report our efforts to employ recent ultrascale visualization technologies (SciDAC Ultravis) to study model comparison in terrestrial biogeochemistry datasets produced by computation (SciDAC C-LAMP). While many of the current efforts are specific to climate modeling research, our method of location-specific summarizing visualization of extreme and normal relative distribution patterns is generally applicable to other fields of computational sciences.

1. Introduction

Our driving application is to visualize SciDAC [1] caliber datasets from global coupled climate-carbon cycle model simulations produced by different models. We draw our research motivation from the current thrust to generate accurate simulations of the global carbon cycle that model the interactions and feedbacks between the terrestrial biosphere and the climate system.

The Carbon-Land Model Intercomparison Project (C-LAMP), www.climatemodeling.org/c-lamp [2] was initiated to allow the international scientific community to thoroughly test and compare such terrestrial biogeochemistry models through a set of carefully crafted experiments. Well-defined metrics have been established for comparison of model results against best-available observational datasets, and models are graded on their scientific performance with respect to these metrics. Visualization tools and diagnostics are particularly helpful in uncovering model differences and discovering ways for improving individual models.

The visualization aspects of this task are very demanding for several reasons. Firstly, there are a large number of variables involved in each simulation. Exacerbated by the need to study multiple runs in a cohesive manner, the combinatorial space that needs to be explored is overwhelming, even just the task of studying two variables from two simulation runs. For instance, scientists already have some empiric understanding of how net ecosystem exchange relates to net primary productivity. Does the relationship exist as expected in a peta-scale simulation, including across different time spans of different runs? This model of investigative study and the need of high interactive rates currently present a challenge for large data visualization.

2. C-LAMP

The purpose of this model-measurement inter comparison is to allow the international scientific community to evaluate the performance of biogeochemical models normally coupled to general circulation models (GCMs) [2]. Terrestrial models are scored based on their performance as compared to best-available site, field, and satellite observations through a rigorous set of metrics. To this end, we encourage you to provide feedback on the experimental protocol, the metrics used to evaluate model performance, and the observational datasets available for use in the inter comparison. The C-LAMP project bridges the climate modeling community and the measurement community. This role of C-LAMP enables significant potential of model improvement and more comprehensive measurement campaigns.

The C-LAMP project conducts two types of experiments. In the first type of experiments, biogeochemical land surface models are forced with an improved NCEP/NCAR reanalysis climate dataset. In these offline runs, the objective is to examine the ability of the models to reproduce surface carbon and energy fluxes at multiple sites, and to examine the influence of climate variability, prescribed atmospheric CO2 concentrations, and land cover change on terrestrial carbon fluxes during the 20th century, and specifically during the period for which the reanalysis data are available (1948-2004).

In the second type of experiments, an active atmosphere model is used to couple energy flows between the atmosphere and the terrestrial biosphere. However, atmospheric CO2 follows a prescribed trajectory for both steady state and transient components of the experiment. The prescribed CO2 is radioactively active and sea surface temperatures (SSTs) and ocean carbon fluxes are prescribed. The objective of these simulations is to examine the effect of a coupled biosphere-atmosphere on carbon fluxes and climate during the 20th century.

3. Parallel Query-Driven Visualization

Decadal to century time scale climate simulations typically output a large number of two- and three-dimensional variables at regular intervals, usually monthly. While identifying features is difficult, it is actually intuitive and practical to select only a subset of the data from within that high-dimensional variable space, to obtain a qualitative understanding of the overall results. This approach can be easily implemented if the resulting dataset can be stored entirely in-core. However, to handle larger datasets, a more sophisticated solution is necessary.

Our solution [3] involves designing specialized scalable visualization data servers with large-scale parallelism. Our index system indexes general data items, including vertices, voxels, or particles, and is independent of grid type. The core data structure for indexing is an optimized M-ary search tree. The tree structure only amounts to ~1% the size of the whole dataset. The dataset can be stored externally on hard drives in a compressed manner. Only parts of the data that are used by the scientist are decompressed (and cached). The compression rates vary from dataset to dataset. For some typical datasets, we obtained a 20x compression rates, while we could obtain as low as a 4x compression on highly turbulent or noisy datasets. Using these rates, we conservatively estimate that a mid-sized cluster could already support parallel visualization of a dynamically queried dataset of 1 TeraByte (TB) size.

The M-ary tree uses a large branching factor, and serves the role of metadata to guide a search process. The branching factor is one of the primary differences between this search data structure and previous data structures, such as interval tree, k-d tree, quadtree and octree. Due to the large branching factor, M, the M-ary search tree requires little storage space. The data is not stored in the tree, but in a linear list sorted by a key function. The leaf nodes of the tree store only pointers to the respective data records in a sorted linear list. We have discovered that conventional methods to access records by traversing the tree are too expensive, both in terms of caching performance and the large number of addressing operations. We use a novel method to accelerate range searches in an M-ary tree to address this, optimized specifically for multivariate datasets [3].

Data items are partitioned into groups by round-robin assignment according to high-dimensional space filling curve [8] order in attribute space. We use this type of data partition to distribute data amongst all visualization data servers equally. Thereby, we are able to achieve a nearly optimal load-balance for almost all kinds of queries. The M-ary search tree is then used to manage the data on each server. Our approach is relatively easy to deploy on networked commodity computers, whether clustered or not. The necessary number of parallel data servers depends on the size of the dataset.

4. Concurrent Visualization of Multiple Patterns

It is common for scientific visualization production tools to provide side-by-side images showing various results of significance. This is particularly true when handling time-varying datasets with a large number of variables. However, it is often desirable to have a general scientific visualization method to summarize the data into the fewest possible images. As shown in [5] this is a hard problem. It is often impossible without reducing the data of each variable. One way to do so is by feature detection and extraction. Previous authors [6-7] have presented a number of successful example methods based on feature extraction. A common limitation of these methods, however, is the requirement to have features accurately defined before a visualization can be created. This is particularly hard when the purpose is to provide initial concepts of unknown features involving unprecedented numbers of concurrent variables.

For comparative visualization of unknown model differences in C-LAMP, we use a location-specific summarizing method to visualize extreme and normal relative patterns among multiple concurrent attributes [4].

With each attribute representing a different physical variable, a different model or simulation run, we provide a novel comparative perspective of the data. Our motivation is to guide a user's attention to a much-reduced subset of a large and otherwise incomprehensible multivariate dataset. In order to do such guidance without requiring too much user input, we base our methods on relative distribution patterns of multiple variables. In the overall problem space formed by all concurrent attributes, we term each subset of interest in our framework an attribute-specific subspace, because each subspace can eventually be represented by one tag that is in direct correspondence with an attribute.

The input to our method [4] is a general dataset that has, at each location, an associated attribute set. Each attribute must have a value at each location (point). We normalize all attributes to a canonical range. Also, the user supplies attribute target values that allow us to determine if the value at any location is "good", i.e. close to its target value. Common examples are: highest possible temperature around the globe, heaviest seasonal rainfall around the globe or longest period of drought around the globe. At each location, classification is based on which attribute at this location is the closest to its target value. Each location is classified and assigned to at most one attribute-specific subspace. When all values on a location are sufficiently far away from the target values, this location is thresholded and assigned to no subspace. If it is a dataset with 20 concurrent variables, after classification we always have 20 attribute-specific subspaces. To render, each subspace gets a different color. In renderings, thresholded locations appear as empty spaces.

Let us illustrate how we intend the technique to be used in the following. A typical time-varying dataset contains a set of variables associated with each point. We can simply treat the temporal values of a single variable associated with each spatial location as different attributes. Please note in our method, extreme vs. normal are equivalent because it's just a matter of picking an extreme value or an average value as the target. In this case, we show, in a location specific manner, for this one variable the timesteps in which interesting features arise. This could be used to discern which timesteps would be most valuable in that variable's evolution over time.

5. Results and Discussion

To application scientists, parallel query-driven visualization serves as a data reduction, which we have accelerated to interactive rates [3]. The concurrent visualization method based on attribute subspaces is a subsequent step to summarize multiple patterns according to their relative strengths so that overall trends can be visually compared, analyzed and comprehended [4].

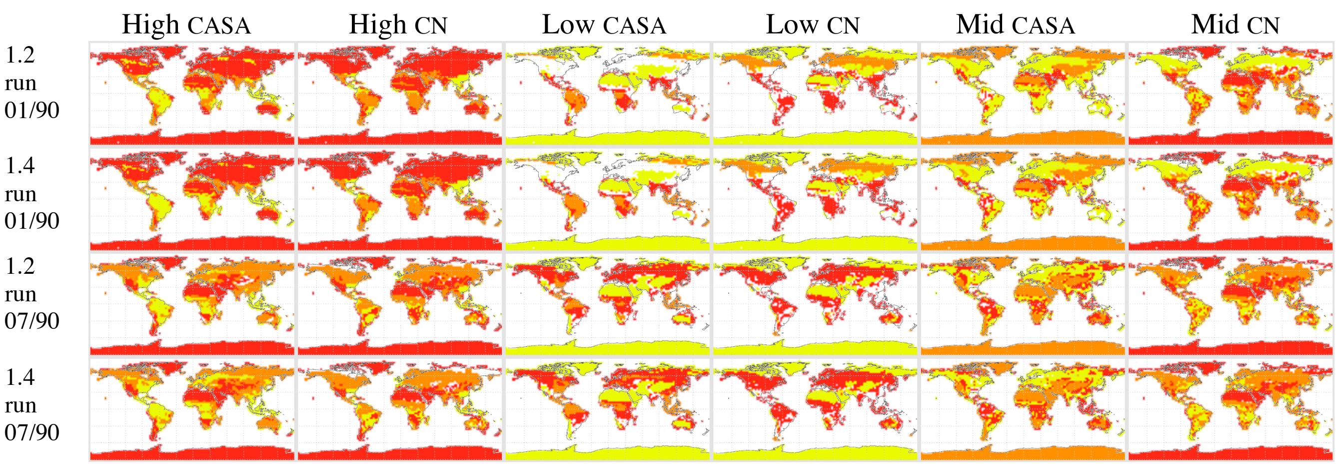

Figure 1. A location-specific summarizing visualization of extreme and normal relative distribution patterns among CASA and CN models in C-LAMP 1.2 and 1.4 runs. The variables considered are NEE (red color), NPP (orange color) and TLAI (yellow color) in January and July of 1990.

As typical with any climate modeling efforts, the same models would be run under different conditions. To demonstrate our approach and typical use cases, we show three different types of runs: (i) control run (C-LAMP 1.2), (ii) varying climate transient run (C-LAMP 1.3) and (iii) varying climate, CO_2 and N deposition transient run (C-LAMP 1.3). Under each type of run, there are two models from NCAR Community Climate System Model Version 3 (CCSM3), specifically CCSM3-CASA and CCSM3-CN. Hence we have in total 6 different simulation scenarios to visualize. The time span considered is the decade from 1990 till 1999. In this section, we show major patterns of three variables: net ecosystem exchange of carbon (NEE), net primary production (NPP) and total leaf area index (TLAI). While we find that, in reference to the general climate trends of these three variables, the differences among models are visually discernible in several cases, within both models the results from 1.2 and 1.3 runs seem visually indistinguishable. For reason of space, we do not show results from the 1.3 run.

When viewing more than one variable in one setting, differences in models become apparent. In Figure 1, "high," "low" and "mid" mean that the target values for all three variables are chosen to be global maximum, global minimum and mid-way between global extremes, respectively. In terms of relative strength (or relative variation patterns) among NEE, NPP and TLAI, CASA and CN largely agree with high and low distributions in both January and July in 1990 (peak summer and peak winter months). The mid range distributions do reveal differences in high latitude areas, large deserts and mountainous (high elevation) regions.

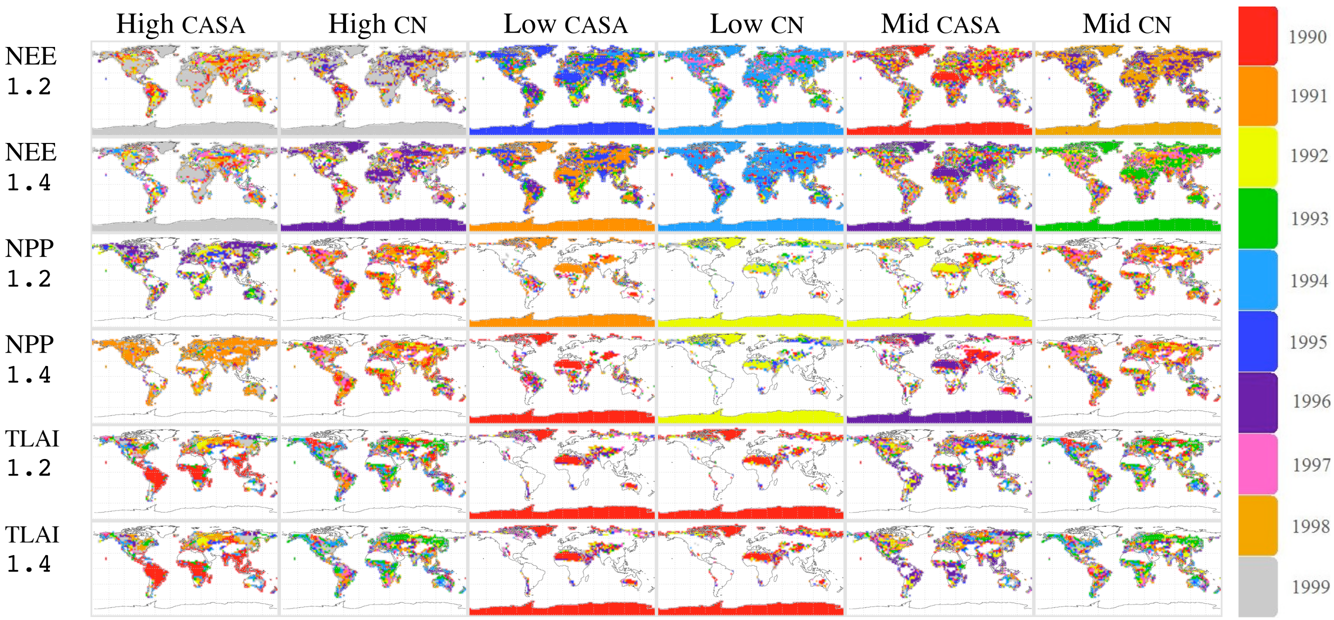

Figure 2. Year long average computed over 1990-1999. For each of NEE, NPP and TLAI and in 1.2 and 1.4 runs, which year is the closest to the decadal maximum (High), mid-point of the possible range (Mid) and the decadal minimum value (Low)?

In Figure 2 with TLAI, 1.2 and 1.4 are visually similar. The heavier coupling in 1.4 does not seem to change relative strength in temporal distribution of yearly averages in the decade. Both CASA and CN seem to agree that 1990 has the lowest TLAI in the decade and in the geographic regions of barren lands (Arctic, Antarctic, Sahara and Tibet). As to relatively high TLAI in the decade, CASA still favors 1990 (closer to decadal-global extremes) in South America and sub-Sahara Africa. Also the earlier three years have high TLAI in all other areas of the globe. This pattern of earlier years in 90's showing high relative TLAI is not in CN runs. With both NEE and NPP, CASA and CN largely agree, but are, however, off by a year or two (both for high and lows) in some cases. For the NPP modeling results (both High-CASA and Low-CASA), the Amazon mostly shows the first half of the 90's to have lower values. In CN, the Amazon shows high NPP more uniformly distributed throughout the decade. No year during the 90s showed "abnormally" low NPP in CN models.

As another example, we calculate Pearson's correlation coefficient between pairs of variables for each simulation scenario over 1990-1999 (Figure 3). By viewing all models concurrently, some model trends are clearly visible. Since 1.2 and 1.3 are so similar, only 1.2 is represented in this image, i.e. no visible green or yellow. C-LAMP 1.4 is represented as blue and light blue (CASA and CN), so blue in an image represents the more coupled the run. NEE-NPP shows stronger positive correlation in more coupled runs in Eurasia, whereas NEE-TLAI shows stronger positive correlation more in less coupled runs in the same area. With respect to correlation we would expect this viewing to appear completely random, but the presence of contiguous areas of identical color highlight geographical regions of interest with a model.

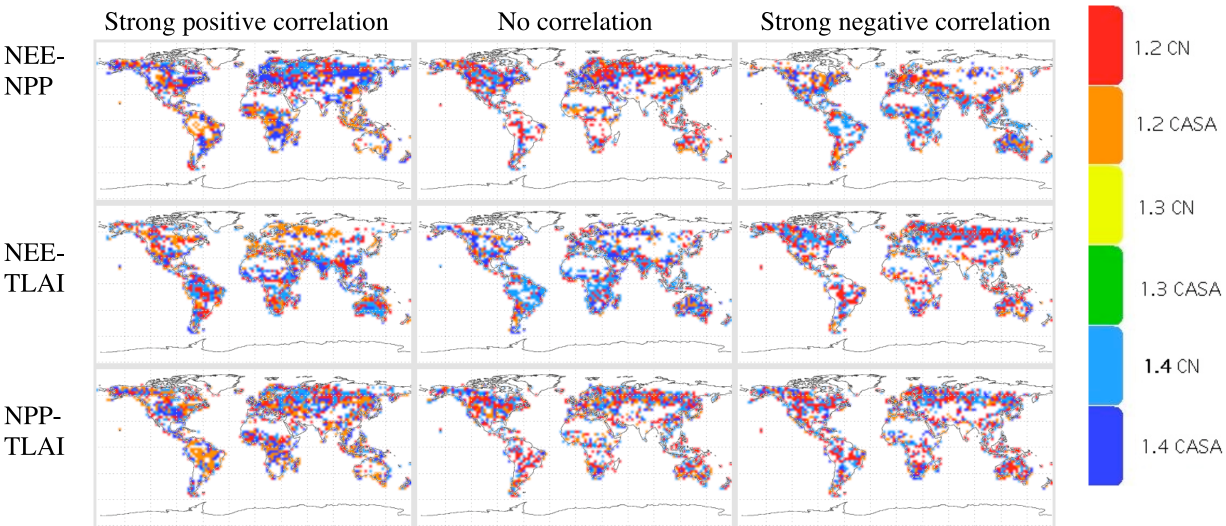

Figure 3. Model differences in bi-variate Pearson's correlation coefficient, computed over 1990-1999.

In the future, we plan to engage in even closer collaboration with C-LAMP researchers because few of the model differences we have identified so far have clear explanations from the modeling community. We also plan more research for evaluating model differences in the context of model biases from real observational data.

Acknowledgement

The primary source of funding for this work is provided through Institute of Ultra-Scale Visualization (www.ultravis.org) under the auspices of the SciDAC program within the U.S. Department of Energy (DOE). This work is also supported in part by NSF grant CNS-0437508 and a DOE Early Career PI grant awarded to Jian Huang (No. DE-FG02-04ER25610). The Carbon-Land Model Intercomparison Project (C-LAMP) are partially sponsored by the Climate Change Research Division (CCRD) of the Office of Biological and Environmental Research (OBER) and the Computational Science Research and Partnerships (SciDAC) Division of the Office of Advanced Scientific Computing Research (OASCR) within DOE Office of Science. This research used resources of the National Center for Computational Science (NCCS) at Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle, LLC, for DOE under Contract No. DE-AC05-00OR22725.

References

[1] "Scientific Discovery through Advanced Scientific Computing," Office of Science, U. S. Department of Energy, Washington, D.C. March 24 2000.

[2] Hoffman, F., C. Covey, I. Fung, J. Randerson, P. Thornton, Y-H Lee, N. Rosenbloom, R. Stockli, S. Running, D. E. Bernholdt, and D. Williams. "Results from the Carbon-Land Model Intercomparison Project (C-LAMP) and availability of the data on the Earth System Grid (ESG)". SciDAC 2007, June 2007, Boston. *Journal of Physics: Conference Series*, Vol. 78, page 012026. Institute of Physics, 2007.

[3] Glatter, M., C. Mollenhour, J. Huang and J. Gao, "Scalable Data Servers for Large Multivariate Volume Visualization", IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No. 5, pp. 1291-1299, 2006.

[4] Sisneros, R., C. R. Johnson and J. Huang, "Concurrent Viewing of Multiple Attribute-Specific Subspaces", Proc. of EuroVis'08 (Eurographics/IEEE VGTC Symposium on Visualization), Eindhoven, NL, May 2008.

[5] Taylor, R., "Visualizing multiple fields on the same surface," IEEE Computer Graphics and Applications, 22(3), pp. 6–10, 2002.

[6] Weiler, M., Botchen R. P., Stegmaier S., Ertl T., Huang J., Jang Y., Ebert D. S., Gaither K. P., "Hardware-assisted feature analysis and visualization of procedurally encoded multifield volumetric data", IEEE Computer Graphics and Applications 25, 5 (2005), 72–81.

[7] Walter ,J. D., Healey C. G., "Attribute preserving dataset simplification", Proc. of *IEEE Visualization 2001* (Oct. 2001), pp. 113–120.

[8] Pascucci, V. and R. Frank, "Global static indexing for real-time exploration of very large regular grids", In *ACM/IEEE Conference on Supercomputing (SC'01)*, pages 2–2, 2001.

[9] Hoffman, F., William, W. Hargrove, D. J. Erickson and R. J. Oglesby, "Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models." Earth Interactions, 9(10): pp. 1-27, 2005.