Institute of Physics Publishing Journal of Physics: Conference Series **46** (2006) 550–555
doi:10.1088/1742-6596/46/1/076 SciDAC 2006

Multi-scale data visualization for computational astrophysics and climate dynamics at Oak Ridge National Laboratory

Sean Ahern [1], Jamison R. Daniel [1], Jinzhu Gao [1], George Ostrouchov [1], Ross J. Toedte [1], Chaoli Wang [1,2]

1 Center for Computational Sciences, Oak Ridge National Laboratory

2 Department of Computer Science and Engineering, Ohio State University

Abstract:

Computational astrophysics and climate dynamics are two principal application foci at the Center for Computational Sciences (CCS) at Oak Ridge National Laboratory (ORNL). We identify a dataset frontier that is shared by several SciDAC computational science domains and present an exploration of traditional production visualization techniques enhanced with new enabling research technologies such as advanced parallel occlusion culling and high resolution small multiples statistical analysis. In collaboration with our research partners, these techniques will allow the visual exploration of a new generation of peta-scale datasets that cross this data frontier along all axes.

Introduction:

A frontier is shared by several SciDAC computational science domains that generally correlates to aggregate dataset size. This frontier is a multidimensional collection of points each representing output from a particular simulation code. The loci of these points are determined by factors including spatial sampling, temporal sampling, field dynamic range, number of independent fields, and spatial coherence. It is this factorization that characterizes a simulation's data and determines the suitable visualization tools and techniques for exposing the underlying science. Many production visualization techniques begin to fail when considering data that cross this frontier. As such, this boundary marks the departure of visualization techniques that are production and requires specialized research efforts.

Choosing visualization techniques for a particular science domain is a process of stepwise refinement that must be verifiable, expository, and communicable with regard to the science. The visualization must bear a clear quantitative relationship to the data, deftly expose and support some element of scientific relevancy in the data, and easily facilitate a transfer of knowledge. Both climate and astrophysics simulations are generating data on the order of tens to hundreds of terabytes, and each dataset composition is predisposed to certain techniques. Science drivers and computational models are dynamic, however, and dataset compositions are often in a state of flux. Data analysis techniques must adapt to reflect these changes.

We are adapting these techniques for the changing needs of the science communities as they create datasets with increasing size, complexity, and fidelity. We present both production and research techniques that have demonstrated value with respect to scientific discovery within the respected domains. These techniques combine traditional visualization, parallel hardware exploitation, and coupled statistical analysis. It is only through this directed research toward large-scale data understanding that the realm of production visualization may be expanded to manage these datasets.

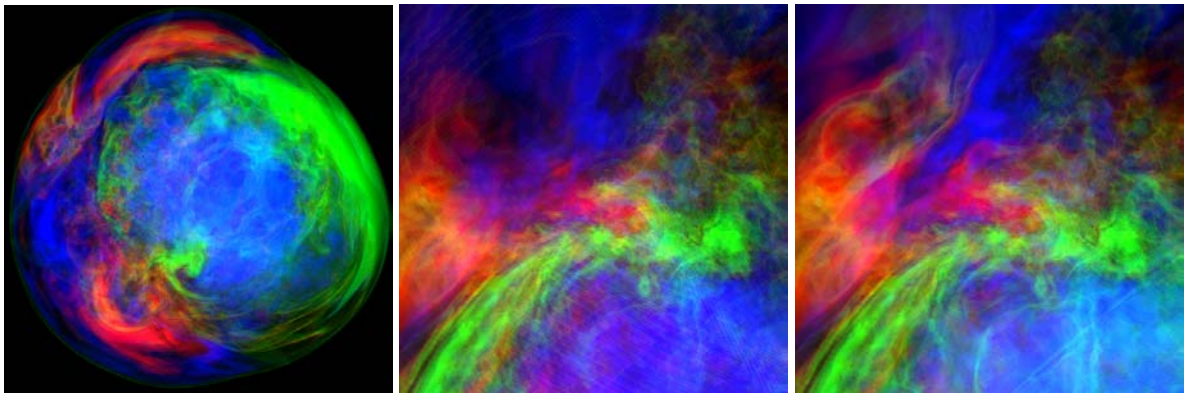
Computational Astrophysics Visualization:

The SciDAC-sponsored TeraScale Supernova Initiative (TSI) is one of the principal users of CCS resources. TSI is a multidisciplinary, multi-institution effort involving the integrated efforts of astrophysicists, nuclear physicists, computer scientists, and mathematicians. Explorations undertaken by TSI focus on the birth, life, and death of stars eight to ten times more massive than our own sun. The extreme physical complexity of the supernova problem

includes the evolution of the mass distribution of the star as it marches toward incompressible density, the mechanics of the core "bounce" that starts the ejection of superheated material, the quantities and distribution of the elements of the periodic table that are expelled under explosive force, the distribution of the different "flavors" and energies of neutrinos and antineutrinos, and the imparting of spin to neutron stars that are the legacy of supernovae. The scales and complexities of the data into which these physics are translated illustrate the visualization challenges that exist today and the new challenges that are likely in the future. For example, prior simulations involving the VH-1 hydrodynamics code have yielded five scalar fields on dense meshes with over 800 grid points per spatial axis. A typical simulation run for 300 iterations at this spatial density yields over 3TB of data. On the other end of the data spectrum, TSI calculations of microscopic equations of state (EOS) have important relations to other TSI simulations in that they may have a significant impact on neutrino transport. These calculations produce data that has heretofore not exceeded 1MB in volume and has low parameterization. Clearly, TSI has a diversity of data in terms of variables, meshes, and aggregate data size

Because many of TSI's datasets are quite large and features of interest may be quite small (e.g. vortex cores), it is highly probable that, for volume rendered data, large portions of the mesh may be occluded. To avoid processing those invisible data regions, we proposed a visibility culling scheme that is highly scalable in a parallel and distributed environment. The scheme encodes the opacity of a volume block for all possible viewing angles around the block in a function called Plenoptic Opacity Function (POF) [4]. POFs are computed during preprocessing for all volume partitions. At run time, the opacity of each block can be quickly computed from its POFs. Based on the blocks' opacities, our visibility culling scheme can easily determine the visibility status for each block and only send visible blocks down to the visualization pipeline. In our experiments, only 30-40% of non-empty blocks are actually rendered after visibility culling is applied. About 81% parallel efficiency was observed when using 32 processors.

In another successful approach to rendering multiple terabyte TSI datasets, we developed an adaptive multi-resolution data selection scheme to use the lowest resolution possible for a region without sacrificing image quality. Leveraging lower resolution blocks, we reduce data movement and processing cost significantly. The selection of the resolution can depend on various factors. In [2], we select resolution based on raw data homogeneity. If the variance inside a block is lower than a user-specified error tolerance, the block will be rendered using a lower resolution. However, certain transfer functions may assign contrasting colors to voxel values within a small range or one single color to a block with a large variance. To further reduce the cost on regions classified as homogeneous in color and opacity, we introduce value histograms to our resolution selection scheme [5]. A value histogram captures the distribution of voxel values in a block over the possible range of values. Then, by comparing a block's value histogram with the transfer function, we can easily identify those homogenous regions and render them in lower resolution.



(a) a rendering of entropy field (b) 28 blocks rendered (c) 198 blocks rendered

Figure 1: Multiresolution volume rendering of the entropy field of the TSI data set. (a) Shows a rendering overview of the entropy field data. (b) and (c) are zoomed-in views using different error tolerances with the same viewing setting. The transfer function used in this work utilizes a rainbow color spectrum ranging from low entropy represented as blue to high entropy represented as red.

Figure 1 shows several results with different levels of detail for a TSI data set. The TSI data set has a spatial dimension of 800^3, and it consists of hundreds of iterations. Five variables (a density scalar, a pressure scalar, and a

velocity vector) are recorded at each iteration. The entropy field is derived from the density and pressure fields. For multiresolution volume rendering, data of lower resolution are used when the error tolerance becomes higher, resulting in a smaller number of blocks being rendered. It can be seen that, although more delicate details of the data are revealed when lowering the error tolerance, images of reasonable quality can still be obtained at lower resolutions. The use of wavelet-based compression also allows us to produce images of good visual quality with small space overhead.

Advanced data structures are critical to ensure low run-time cost on data selection. Since most large-scale data we encounter are time-varying, we adopt the skeleton of the Time-Space Partitioning (TSP) tree. We devise a multiresolution spatio-temporal hierarchy that encodes the input time-varying data set. The hierarchy is based upon the wavelet transform, where the data set is stored in a data structure called the wavelet-based TSP (WTSP) tree [1]. The WTSP tree is an octree (spatial hierarchy) of binary trees (temporal hierarchy). There is one octree skeleton, and at each octree node, there is a binary time tree. Each time tree spans the entire time sequence and combines data from multiple octrees. The WTSP tree exploits both spatial and temporal locality and coherence of the underlying time-varying data, thus allowing flexible spatio-temporal level-of-detail data selection and retrieval at run time. We also propose an enhanced TSP (ETSP) tree [4] to facilitate both visibility culling and multi-resolution data selection. Information required by those data selection techniques, such as POFs, spatial and temporal errors, and value histograms, can be stored in the tree structure. At run time, by traversing the tree, we can quickly select the suitable resolution for each visible (non-occluded) region.

In addition to performing simulations of supernovae that are on order of $4*10^8$ meters in radius, TSI also engages in studies of sub-nuclear proportions. Between these extremes exists a stunning span of 21 spatial orders of magnitude of simulation. Although the microphysics simulation code is still in the testing phase, the first results show clearly that, for the first time, we gain significantly new insight into properties of nuclear matter in three-dimensional space. Visualizations of "pasta phase" nuclear matter indicate that the previous models of neutron-heavy nuclei of exotic shapes, immersed in free neutron and electron gases, is only an approximation to the true distribution of nuclear density in a unit cell as demonstrated in Figure 2. The same volume visualization tool that incorporates the aforementioned adaptive multi-resolution data selection was employed for this work.

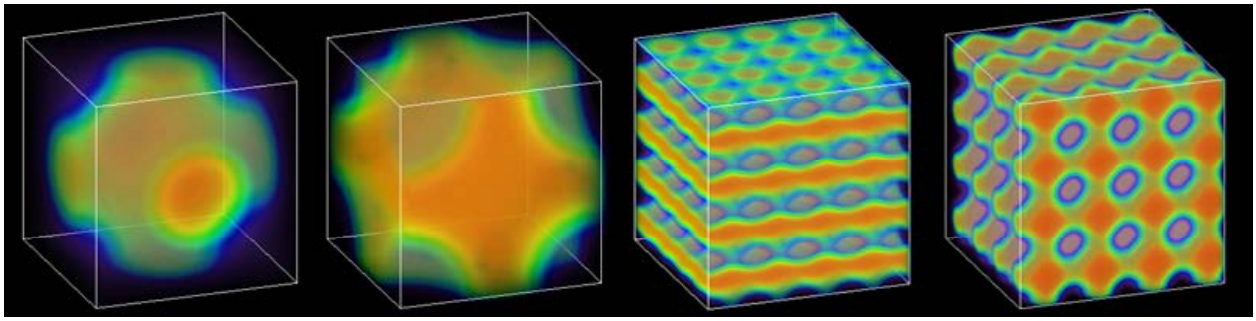


Figure 2: Volume visualizations of neutron density distributions from simulations of different temperature, density, and nucleon counts. Red colors indicate higher densities and blue colors represent lower densities. The two left hand visualizations are of single unit cells. The two right hand visualizations are of multiple unit cells. These visualizations support the idea that neutron matter densities near the cores of supernovae vary gradually, thus producing more complex structures than previously thought [8].

Computational Climate Dynamics Visualization:
The visualization challenges offered by the computational climate dynamics community are unique in both scale and complexity. The data representing the Earth's climate, generated by a set of simulations from the Community Climate System Model (CCSM), have high temporal resolution as well as a large number of degrees of freedom. The collective CCSM simulations are composed of several multi-disciplinary component models of the atmosphere (ocean, land surface, and sea ice) spanning a temporal space ranging from a pre-industrial phase, the 20^{th} century, and projecting into several future scenarios [10]. The Department of Energy, the National Science Foundation, and the National Aeronautics and Space Administration have all contributed to the CCSM climate simulation capabilities. The user base is composed of hundreds of climate scientists from academia and the national laboratories.

The CCSM community is responsible for several simulation contributions to the Intergovernmental Panel on Climate Change (IPCC) initiative. The IPCC was established to research climate change, the potential impacts, and to explore options for adaptation. The effort considers numerous domains including time-varying concentrations of greenhouse gases, prognostic sulfate with time-varying emissions of SO_2, time-varying concentrations of carbonaceous species scaled by population and SO_2 emissions, time-varying solar constant, time-varying concentrations of tropospheric and stratospheric ozone, and time-varying concentrations of stratospheric volcanic aerosols.

At ORNL, we were interested in visually exploring twelve of the IPCC fields over a temporal space surrounding the introduction of several tons of sulfur dioxide into the global atmosphere resulting from a historic volcanic eruption. While each discrete domain was characterized by a relatively low spatial resolution, we needed to visually correlate the numerous degrees of freedom; including total aerosol optical depth, volcanic aerosol depth, carbon optical depth, vertically integrated high cloud, vertically integrated low cloud, net longwave flux at the surface, net longwave flux over open ocean and ice, net solar flux at surface, net solar flux at the top of the atmosphere, surface latent heat flux, large-scale stable precipitation rate, and shallow convection precipitation rate. In order to demonstrate the field correlation over a temporal space of five years while maintaining the clarity of context, we decided to pursue an alternative small multiples visualization technique.

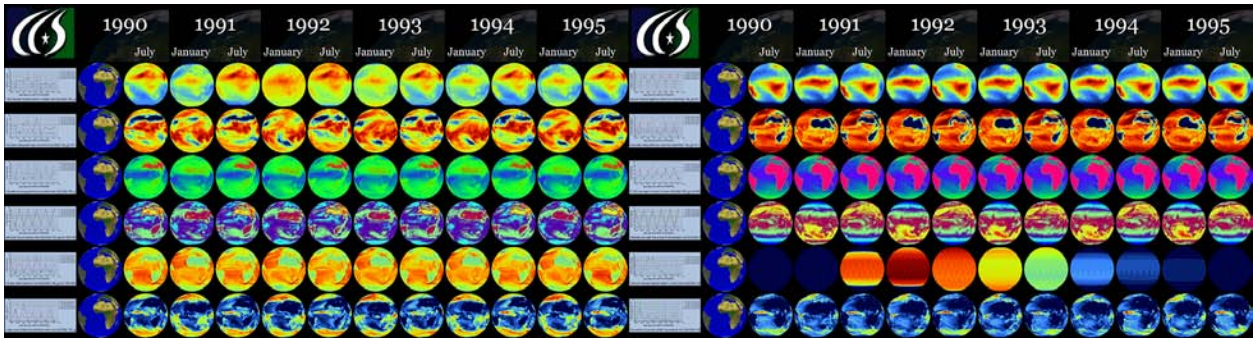


Figure 3: High-resolution small multiples visualization exploring the modification of the global radiation budget as a result of changes in stratospheric aerosols and polar stratospheric clouds. The visualization reveals the large amount of SO_2 injected into the atmosphere by the June 15, 1991 eruption of Mount Pinatubo. This eruption was a significant source of stratospheric aerosols, which plays a critical role in determining the temperature distribution of the Earth's atmosphere.

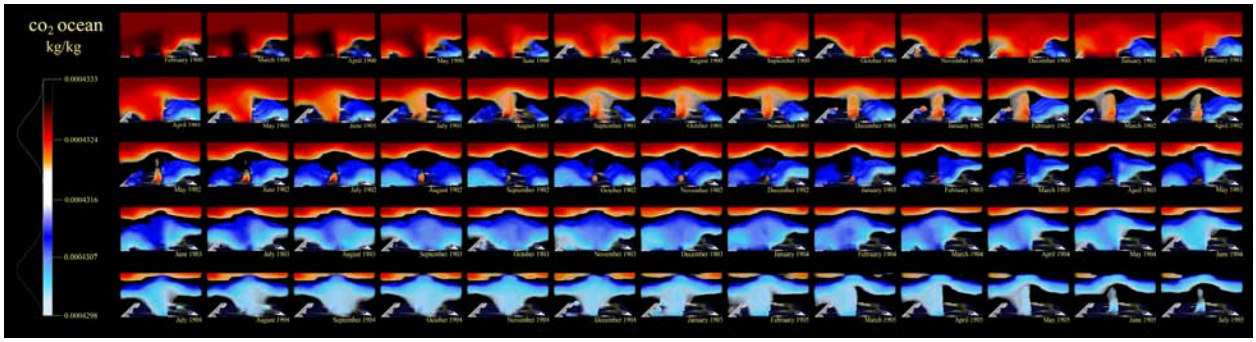
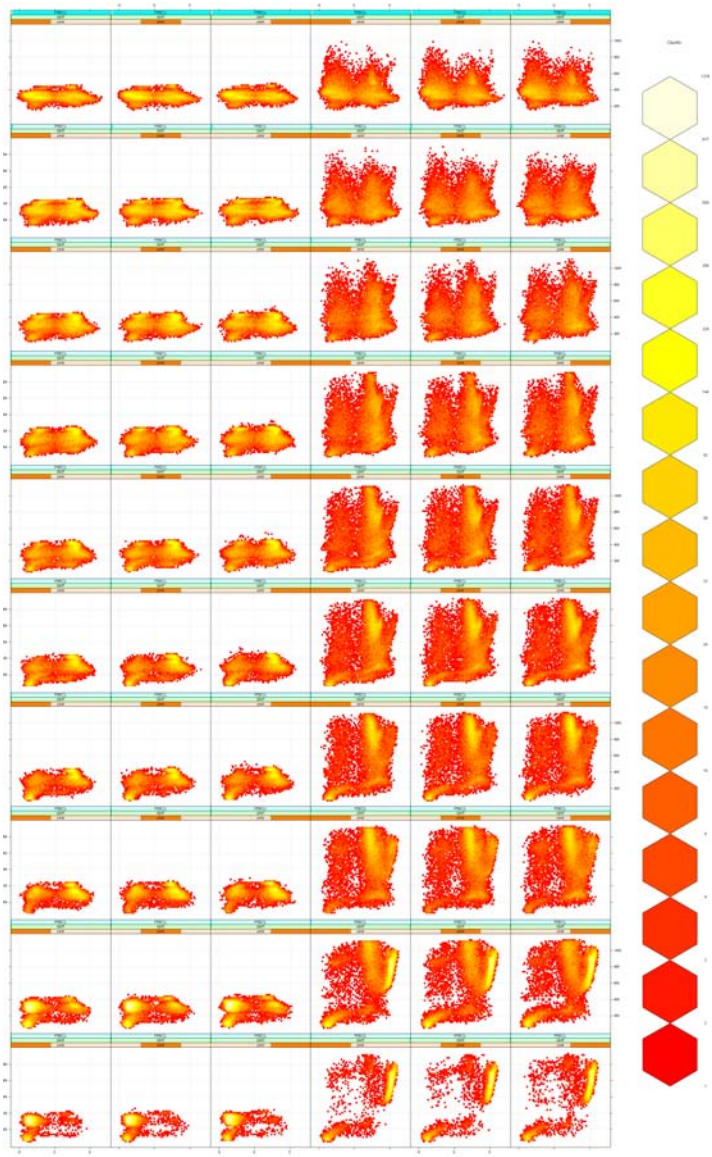


Figure 4: High-resolution small multiple visualization of the simulated time evolution of the component of atmospheric carbon dioxide concentration originating from the ocean surface. The underlying simulation is one of a number of transient runs performed for Phase 1 of the CCSM on the Leadership Computing Facility (LCF) Computational Climate Science End Station (CCSES). These CCSM simulations were run on the Cray X1E at ORNL.

The technique of small multiples is motivated by the concept of data density, defined as the amount of data divided

by the area of the resulting graphic. The human eye has the capacity to process 3,500 distinctions per square centimeter [9]. Thus, a high resolution powerwall can allow quick visual assimilation of numerous domains arranged within a single eye span. Each graphic can range from a conditional scatter plot to a more traditional scientific visualization rendering. Once the enforced standard is understood by the observer, multivariate differences can be perceived through visual comparison of the individual graphics.



New visualization challenges will arise as the complexity of the CCSM models increase. Future models will include both chemical and biological processes that govern the absorption and emission of greenhouse gases, a comprehensive atmospheric chemistry formation, and both land and ocean ecosystems. In order to resolve critical mesoscale eddies, simulations of the Parallel Ocean Program (POP), the ocean component of the CCSM, are anticipated at 1/10th degree spatial resolutions. These resolutions have previously been computationally impossible or achievable only in limited domains of the ocean. Coupled with the availability of new computational resources, peta-scale computing will challenge the scientific community to develop effective new visualization techniques.

Figure 5: This set of 60 scatter plots provides access to a five-dimensional relationship between radiative heat flux and sensible heat flux under various conditions: morning versus evening, early - mid - and late 21st century, and ten different precipitation levels. Overplotting is handled by hexagonal binning with a log-based colorscale. The display is generated by the lattice package of the R statistical computing environment [7].

Future Work:

The future of scientific visualization within the SciDAC program is a challenging one. Dataset sizes are expected to grow by two to three orders of magnitude over the next five years, greatly stressing existing analysis techniques. In the short term, increases in dataset sizes can be met by increasing the scale at which parallel visualization is performed. However, advances in the simulation domains of both computational astrophysics and climate research will require fundamental changes in how analysis is performed.

3D datasets generated by simulation codes developed by astrophysics simulation efforts are undergoing radical changes. In addition to changes to mesh, simulation codes will be tracking many more variables that need to be correlated. Hydrodynamics will be augmented to full magneto hydrodynamics. Most disruptive will be the inclusion of radiation transport in the form of anisotropic, multi-frequency neutrino flux, causing these datasets to enter the realm of high-dimensional analysis. Traditional visualization techniques are inadequate for performing data

understanding on datasets of this type, and new methods are required.

In the area of computational climate research, the push is toward greater spatial resolution. Considering that climate simulation codes already push the bounds in degrees of freedom, as well as temporal resolution, the addition of greatly increased spatial resolution further obscures data understanding. Computational climate research is complicated by the distributed nature of the simulation efforts. Oftentimes, researchers wish to do analysis of data that sits at a laboratory that is remote to the researchers. This requires solutions that encompass remote data access or remote visualization. While remote visualization solutions have been successfully deployed for scientific domains, the increase in dataset size will complicate the selection of remote datasets.

Statistical visualization with small multiples is very advanced in the specification and manipulation of high dimensional relationships, though it lacks parallel computing support. This currently forces severe (but careful) downsampling to achieve effective visualization of large high-dimensional data sets. On the other hand, scientific visualization is very advanced in parallel computing support and rendering techniques. We are exploring the unification of high-dimensional statistical analysis with traditional parallel visualization and data management techniques. Utilizing the fundamental data-parallel nature of large-scale scientific visualization software, we expect to bring this model to statistical visualization with small multiples and combine it with advanced visualization techniques.

References:

[1] Chaoli Wang, Jinzhu Gao, Liya Li, and Han-Wei Shen, A Multiresolution Volume Rendering Framework for Large-Scale Time-Varying Data Visualization, In *Proceedings of International Workshop on Volume Graphics 2005*, Stony Brook, New York, pp. 11-19, June 2005.

[2] Chaoli Wang, Jinzhu Gao, and Han-Wei Shen, Parallel Multiresolution Volume Rendering of Large Data Sets with Error-Guided Load Balancing, In *Proceedings of Eurographics Symposium on Parallel Graphics and Visualization 2004*, Grenoble, France, pp. 23-30, June 2004.

[3] Hank Childs and Mark Miller, Beyond Meat Grinders: An Analysis Framework Addressing the Scale and Complexity of Large Data Sets, In *SpringSim High Performance Computing Symposium (HPC 2006)*, pp 181-186, 2006.

[4] Jinzhu Gao, Jian Huang, C. Ryan Johnson, Scott Atchley, and Jim Kohl, Distributed Data Management for Large Volume Visualization, In *Proceedings of IEEE Visualization 2005*, pp. 183-189, Minneapolis, MN, USA, October 2005.

[5] Jinzhu Gao, Jian Huang, Han-Wei Shen, and Jim Kohl, Visibility Culling Using Plenoptic Opacity Function for Large Scale Volume Visualization, In *Proceedings of IEEE Visualization 2003,* pp. 341-348, Seattle, Washington, USA. October 2003.

[6] R. J. Gurney, J. L. Foster, C. L. Parkinson, Atlas of Satellite Observations related to Global Change, Cambridge University Press, 1993.

[7] R. Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (http://www.R-project.org).

[8] J. R. Stone, W.G. Newton, A. Mezzacappa, From Microscales to Macroscales in 3D: Selfconsistent Equation of State for Supernova and Neutron Star Models, poster, In *Proceedings of DOE SciDAC 2006 Conference*, Denver, CO, USA, June 2006.

[9] E. R. Tufte, The Visual Display of Quantitative Information, Graphics Press, 2001.

[10] W. M. Washington, The Computational Future for Climate Change Research, In *Proceedings of DOE SciDAC 2005 Conference (Journal of Physics: Conference Series)*, vol. 16, 2005.