Lawrence Livermore National Laboratory - University of California, San Diego Scientific Data Management Collaboration

Year 1 Progress Report

Tim Barnett
David W. Pierce
Climate Research Division
Scripps Institution of Oceanography

Mike Norman
Robert Harkness
Laboratory for Computational Astrophysics
University of California, San Diego

Randy Banks
Dan Reynolds
Mathematics Department
University of California, San Diego

Reagan W. Moore
Leesa Brieger
Arun Jagatheesan
San Diego Supercomputer Center

May 5, 2006

1. Introduction:

The goal of the joint LLNL/UCSD scientific data management project is to improve the conduct of science through the provision of scientific data management technology that enables the organization, manipulation, and analysis of observational and simulation data. This progress report covers two exemplary scientific applications: Global climate modeling to determine the impact of climate changes on water supply and Cosmology studies of the structure of the early universe. The Cosmology studies in turn are being used to simulate the images that will be seen by the Large-scale Synoptic Survey Telescope (LSST). The scientific applications drive the requirements for scientific data management by generating large simulation output files at LLNL, moving the data to storage systems at SDSC, and publishing derived data products in digital library technology for use by the broader research community. The data management tasks also include a design of a preservation environment for use by the LSST project.

2. Climate Simulations

In this part of the project, a detection and attribution analysis is used to answer the question: Can we detect a global warming signal in main hydrological features of the western United States? This will involve making runs of global climate and downscaling models that will be unprecedented in scope.

2.1 Summary of Progress on Global Climate Modeling

During the first year of the project, a long general circulation model (GCM) control run was executed at LLNL and the results transferred to SDSC. Researchers at SIO then accessed the

SDSC data store to pursue the scientific goals. The scientific data management aspect of the project worked reasonably well. Approximately 1 TB of daily data was transferred from the SDSC data store to UCSD in order to characterize the model and construct the daily forcing files needed for the statistical downscaling work (see below). This represents only a fraction of the total model output transferred from LLNL to SDSC, and was selected for detailed analysis based on the scientific goals of the project. Based on our experience moving data to UCSD for other projects, this amount of data was moved considerably faster and with perhaps an order of magnitude less effort than it would have required using traditional "old" tools such as ftp scripts.

The main issue that arose during this year was that we found the version of the global climate model we are using, CCSM3-FV, produces an equilibrium climate with substantial sea ice thickness and distribution errors in the Arctic Ocean, especially in the region of the East Greenland current. The modeling experts at LLNL and UCSD worked with the model developers at NCAR to evaluate the degree to which this is a problem for our particular application and how tuning of the model might reduce this problem. As of late April 2006, the evidence suggests that the western U.S. (our region of interest for this study) is less affected by this problem than locations closer to the problematic areas, such as the Northeastern U.S. Accordingly, the project is proceeding as planned, but the control run results are being evaluated to determine if it is reasonable to use MM5 to dynamically downscale the current version of the model to the high resolution local grid. As a strategy to continue progress towards the overall scientific goal while the feasibility of the dynamical downscaling is being examined, we are statistically downscaling the global model results to our area of interest. This process allows us to remove biases that the above-noted model deficiencies might impose on the climate.

2.2 Progress on project deliverables

The 1st year project deliverables relevant to the global climate change application, and their current status, are as follows:

a) Develop a standard scenario for climate modeling. We will define and complete a GCM control run. We will complete MM5 downscaling for selected geographic regions of the GCM control run.

Status: the standard scenario for climate modeling was developed, and the required output variables needed to accomplish the scientific goals for this application were defined (see Appendix A).

The GCM control run was defined and several hundred years of simulation, plenty to begin the scientific work and the statistical downscaling, was completed. Several hundred years of output from the control run was transferred to SDSC. Selected variables from the run were transferred to SIO and analyzed.

The MM5 downscaling is on hold pending our evaluation of the climate model errors described previously. In its place the statistical downscaling has been implemented and is proceeding over our region of interest, the western U.S.

b) Begin comparison of MM5 data with observations

Status: we have begun comparing the statistically downscaled data with the observations. We have focused on the distributions of daily runoff, temperature, and precipitation in the western U.S., as these are most relevant to our scientific objective (a detection and attribution study of hydrological features). The results look promising; the temperature and precipitation distributions are individually well simulated, although their joint distribution is skewed away from observed in some locations over the Rocky Mountains (i.e., storms there tend to be warmer than observed). We are investigating a correction to this. Our preliminary estimates of the signal to noise ratio, i.e., the observed change in hydrological parameters in the western U.S. compared to the model biases, suggests that the model biases are significantly smaller than the observed signal over parts of the western U.S. that supply a considerable amount of water for commercial, residential, agricultural, and hydropower needs (for example, the Sierra Nevada, Cascades, and Olympic mountains).

c) Develop forcing scenario for the GCM and a complete anthropogenic run

Status: this deliverable is affected by the equilibrium sea ice distribution problem with CCSM3-FV noted above, since climate sensitivity is partially determined by sea-ice albedo feedbacks. While we are evaluating whether it is sensible to perform an anthropogenic run with the CCSM3-FV, we are implementing a backup strategy that will enable us to achieve our scientific goals in the event that CCSM3-FV is found to be unsuitable for an anthropogenic run. Part of this involves generating VIC simulations from a downscaled anthropogenic run of a predecessor to CCSM3-FV, PCM (see Appendix B). Although this output is less desirable than that from a downscaled CCSM3-FV run, being coarser resolution, it nonetheless does cover our period and region of interest, and uses an earlier version of the same model (as opposed to a completely different model). It was also downscaled with the same version of VIC that we are using, so again there is as much consistency as possible. The other part involves evaluating the existing LLNL archive of other climate model runs to see if any are suitable for our use; this may not be possible, as most of the runs include little daily output, but if feasible this would have the advantage of providing a multi-model look at the problem.

d) Begin VIC simulations with downscaled data

We have begun the VIC simulations with the statistically downscaled data from CCSM3-FV, working on a 1/8th degree grid over the western U.S. We have so far completed a few decades of the multi-century integration. The bulk of this part of the effort will occur in year 2. Additionally, as noted above, we have generated VIC simulations from an anthropogenically forced run of PCM; this process is complete, although analysis will be in year 2.

e) Implement a data grid linking resources between LLNL and SDSC. The data grid will be used to manage the simulation output that is generated.

From the SIO researchers' point of view, this has been accomplished. We have been able to get the data we need in a timely fashion. This has made possible both the analysis of the control run and the statistical downscaling of the global model results to our area of interest. Currently, the VIC simulations being run at SIO are generating output that is stored locally at SIO; in year 2, as the volume of this data increases greatly, we plan to store this data at SDSC using the same mechanism that LLNL used.

2.3 Project sustainability

Researchers at both LLNL and UCSD have been eager to pursue collaborations that are made possible by the framework of this project. In the short time it has been active, two manuscripts have been written on efforts supported by this project. Both are centered around the core idea of this project, using large data sets to examine climate model simulations of global warming.

- *Variability of Ocean heat Uptake: Reconciling Observations and Models*. K. M. AchutaRao, B. D. Santer, P. J. Gleckler, K. E. Taylor, D. W. Pierce, T. P. Barnett, and T. M. L. Wigley. *Journal of Geophysical Research*, in press.

This study examines the temporal variability of ocean heat uptake in observations and in climate models. Previous work suggests that coupled Atmosphere-Ocean General Circulation Models (A-OGCMs) may have underestimated the observed natural variability of ocean heat content, particularly on decadal and longer timescales. To address this issue, we rely on observed estimates of heat content from the 2004 World Ocean Atlas (WOA-2004) compiled by Levitus et al. (2005). Given information about the distribution of observations in WOA-2004, we evaluate the effects of sparse observational coverage and the infilling that Levitus et al. use to produce the spatially-complete temperature fields required to compute heat content variations. We first show that in ocean basins with limited observational coverage, there are important differences between ocean temperature variability estimated from observed and infilled portions of the basin. We then employ data from control simulations performed with eight different A-OGCMs as a test-bed for studying the effects of sparse, space- and time-varying observational coverage. Subsampling model data with actual observational coverage has a large impact on the inferred temperature variability in the top 300 and 3000 meters of the ocean. This arises from changes in both sampling depth and in the geographical areas sampled. Our results illustrate that sub-sampling model data at the locations of available observations increases the variability, reducing the discrepancy between models and observations.

- *Three-dimensional tropospheric water vapor in coupled climate models compared with observations from the AIRS satellite system*. D. W. Pierce, T. P. Barnett, E. J. Fetzer, P. J. Gleckler. In preparation.

Water vapor is a major greenhouse gas, and how the distribution of water vapor changes in response to anthropogenic forcing will be a major factor in determining the warming the Earth experiences over the next century. It is therefore important to compare the observed distribution of atmospheric water vapor to that found in numerical climate models of the kind used to estimate future global warming. In this work the three-dimensional distribution of specific humidity in a number of global climate models is compared to observations from the AIRS satellite system. Our main finding is that the annual mean distribution of specific humidity in most models differs from observed in a consistent and systematic way. The large majority of models have a pattern of dryer than observed conditions (by 10-25%) in the tropics below 800 hPa, but overly moist conditions between 400 and 800 hPa, especially in the extra-tropics. The mean model atmosphere is 25-50% too moist in this region. At 200 hPa, the models tend to be substantially too moist, with the mean model showing 50% greater specific humidity than observed and a number of models showing errors > 100%. Analysis of the accuracy and sampling biases of the AIRS measurements suggests that these differences are, in fact, due to systematic model errors. If so, correcting them might affect the model-estimated range of climate warming anticipated over the next century.

3. ENZO Cosmology Runs

LUSciD – LLNL UCSD Scientific Data Project

Enzo Cosmology Simulation Data Grid (Cosmic Simulator) Year 1 Status Report
Prepared by Michael Norman, mlnorman@ucsd.edu
May 4, 2006

3.1 Summary

Enzo has been ported to LLNL Thunder and is in production mode. Two out of a planned suite of 16 very large adaptive mesh refinement (AMR) cosmological hydrodynamic simulations of galaxy clusters "on the lightcone" have been completed, and 4 TB of data has been successfully moved to SDSC. An SRB-managed data archive has been established at UCSD/SDSC. We are running our own SRB server at SDSC to catalog and preserve the data. A draft design for the metadata schema has been completed. Sky maps of unprecedented size (8192^2) have been produced and are being analyzed for publication. Version 2.0 of Enzo with performance and AMR file I/O enhancements is being prepared for public release. Enzo has been interfaced with our home-grown performance analysis package *jbPerf* and we are collecting performance data. A new isolating boundary Poisson solver has been installed. A detailed plan for incorporating radiation transfer/radiation hydrodynamics into Enzo has also been developed.

3.2 Internal Organization and Management

We have organized ourselves into four teams: production runs, data management, code development, and visibility. Table 1 lists the team leader, team members, and the functions of the teams. The PI M. Norman set year 1 goals for the teams. Team leaders are responsible developing plans to meet these goals, and then ensuring that the team carries them out. We meet weekly to hear progress reports from team leaders and set near-term goals. Progress reports and other documents are posted to the project web page at http://lca.ucsd.edu/projects/LLNL/.

Team	Description	Team Leader	Members
Production runs	Running production jobs on Thunder	Robert Harkness	James Bordner
Data management	Design and implementation of Cosmic Simulator data archive @ SDSC	Rick Wagner	Robert Harkness
Code development	Enhancements and support of Enzo software and tools	Dan Reynolds	James Bordner John Hayes Robert Harkness Alexei Kritsuk Pascal Paschos Rick Wagner
Visibility	Project webpage	Jake Streeter	James Bordner Rick Wagner

Table 1. Project management breakdown.

3.3 Science Application: Galaxy Clusters on the Lightcone

We have chosen an ambitious science driver for the early parts of this project. The application had to be new and scientifically significant, feasible on Thunder on a 1 year timescale, generate large amounts of data of different types, and produce large area sky maps that would be of interest to the LSST project. The project is to simulate the evolution of the galaxy cluster population over the redshift range $0 < z < 3$ on the light-cone covering 100 square degrees on the side (Fig. 1). Galaxy clusters are large concentrations of matter that gravitationally lens and distort the light from background galaxies. One of LSST's prime science missions is to use observations of gravitational lensing to map the large scale distribution of matter in the universe. Our light-cone simulations will be used to mock up LSST sky maps.

To simulate the lightcone, we subdivide it in redshift space into 16 sections, or tiles. Each tile will be a separate high resolution AMR simulation in a volume subtending the same solid angle as seen from the observer. The box sizes are chosen so as to achieve constant angular coverage and resolution in the observer frame, while the number of cells in the root grid is varied to achieve a constant mass resolution (Table 2).

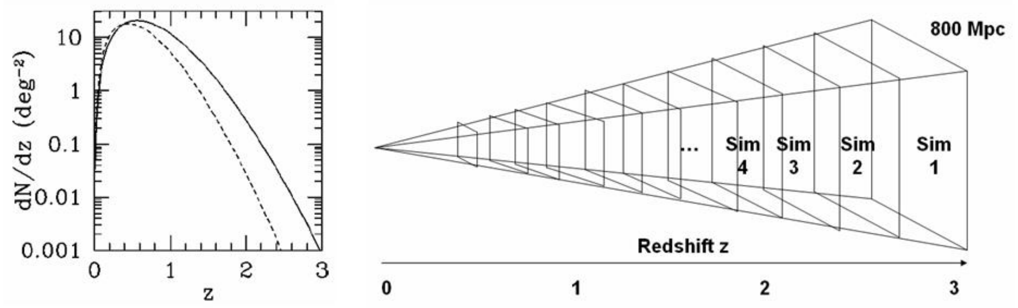

Figure 1 Simulating galaxy clusters on a light-cone. Left: differential number density of clusters above some mass threshold versus redshift for two values of the dark energy density Ω_Λ. Right: tiling strategy. Simulation volumes are varied so as to provide a constant angular resolution.

Sky maps will be generated by casting geodesic rays through the simulations once the survey is completed. A mock sky map is shown in Fig. 2 by passing rays through the different redshift outputs of a single simulation run to z=0. The simulation had a root grid resolution of 512^3 cells and used 4 levels of refinement by a factor of 2 for a spatial dynamic range of 8192. A sky map utilizing the full dynamic range of the simulation would have a size of $(8192)^2 \sim$ 65 megapixels.

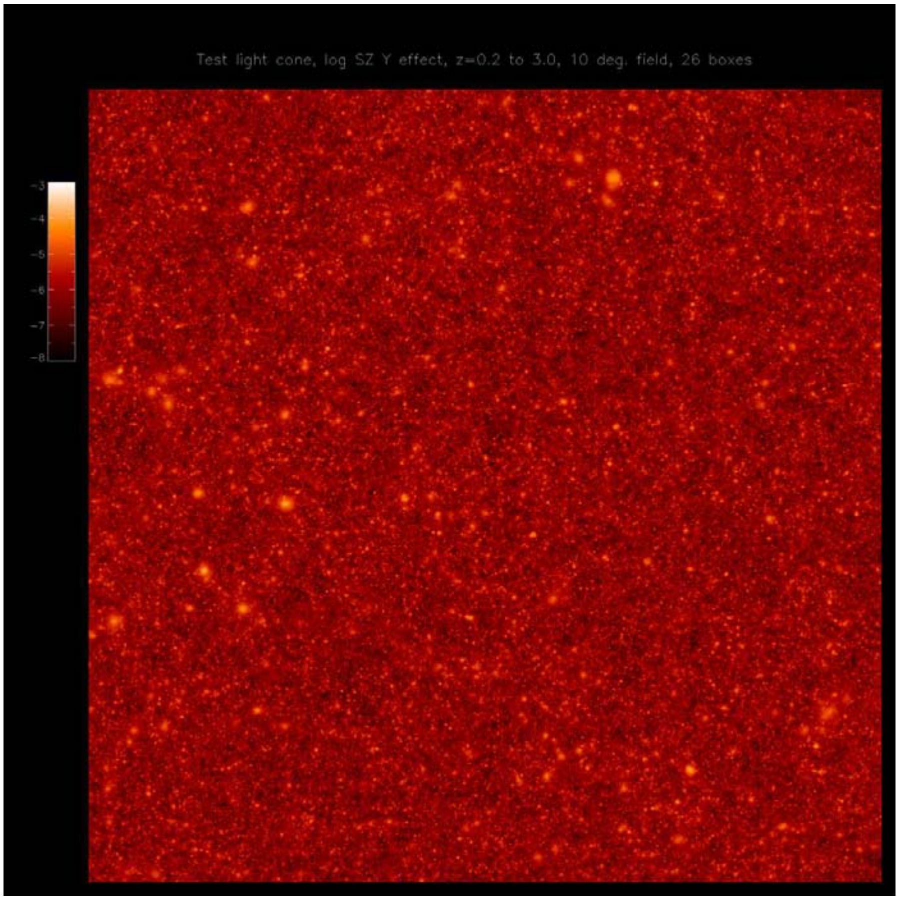

Figure 2. Sky map covering 100 square degrees on the sky derived from tile 7 of our lightcone simulation. The quantity plotted is the temperature decrement to the CMB background radiation as a result of the Sunyaev-Zeldovich effect.

Table 2. Light-cone simulations details and current status

Tile	**Redshift**[1]	**Box length**[2]	**N_root**[3]	**N_cpu**	**Status**
1	3	851.71	896	2744	completed
2	2.63	790.87	832	2197	not started
3	2.29	730.04	768	1728	not started
4	1.99	669.20	704	1331	not started
5	1.72	608.36	640	1000	not started
6	1.48	547.53	576	729	not started
7	1.26	486.69	512	512	completed
8	1.07	448.00	896	512	not started
9	0.91	384.00	832	512	not started
10	0.76	352	768	512	not started
11	0.63	288	704	512	not started
12	0.52	256	640	512	not started
13	0.42	256	576	512	not started
14	0.32	256	512	512	not started
15	0.22	256	896	512	not started
16	0.11	256	832	512	not started

[1]stopping redshift; [2] domain size (Mpc/h commoving); [3] number of cells per dimension in the AMR root grid.

3.4 Production Runs

Goal: Due to the delays getting user accounts on Thunder, the year 1 goal for production simulations was intentionally modest: to port Enzo to Thunder, to complete one of the 16 light cone simulations from Table 2, and transfer the data to SDSC by any means available.

Status. We have met this goal and exceeded it. Two simulations have been completed and ~ 4 TB of output has been transferred to SDSC. Many month delays were incurred at the outset of this project waiting for logins on Thunder and VPN accounts. These have now been secured for Harkness, Wagner, Bordner, Paschos, and Norman, but remain in limbo for Kritsuk—a Russian national. While it would be beneficial to the project if Kritsuk were provided access to Thunder, it is not essential for the success of the project. Further delays were incurred sorting out the data export policy at LLNL and identifying suitably high speed data transport mechanisms. Both of these issues are now resolved, and production runs began in earnest in February.

We have found Thunder to be an unstable platform for simulations of this scale. The most common failure mode is the file system which is critical for our data-intensive application. Reducing the number of nodes requested reduces our exposure to failures of this sort at the expense of increasing the wall-clock time required to complete a calculation. On the positive side, network speeds to SDSC have not become the bottleneck we feared. We achieve a sustained transfer rate of ~50 MB/s out of Thunder into the SDSC/HPSS, which is adequate at the present time given our slow rate of progress with the calculations themselves.

3.5 Data management and analysis pipeline

Goal. Our year 1 goal was to begin the development of the Cosmic Simulator data archive, produce a draft metadata schema, and begin building the analysis tools that will be incorporated into a pipeline managed by workflows. Another goal mentioned in the proposal--the incorporation of a more efficient file I/O method for large AMR data--was completed before this project began by Robert Harkness.

Status: We have met these goals. Details follow.

a) Data analysis tools

In order to produce the planar projections require for the initial analysis of the light cone, Rick Wagner parallelized an existing analysis tool, to enable the creation of large (i.e., up to 8192^2) projections of AMR datasets. Using data from an existing AMR cosmology simulation (with over 100,000 grids), Wagner created a series of 2048^2 projection of several fields, which were then used as the basis of prototype sky maps. This is an example of a terascale data pipeline; the input data was three-quarters of a terabyte, and the total projections were over 200 gigabytes. One of the major purposes of these projections was to "prime the pump" for the tiles of the light cone, whose scale will be at least an order of magnitude greater.

The steps in the data reduction were almost stereotypical: retrieving the datasets from an archival storage system; submit multiple batch jobs; verify the results and archiving the projections. This was a demonstration of the fact that as the simulations grow in memory, processor and storage requirements, the analysis of the simulations must follow the same pattern. The next steps will be to enhance the projection tool to reduce the memory requirements for creating larger images (i.e., 8192^2 and larger), and to continue modifying the other analysis to run in parallel.

b) Data Archive

With the aid of the SRB team at SDSC, we have set up an SRB domain for our lab, which uses SDSC's HPSS as its default resource. The true beauty of the SRB is its ability to associate an unlimited amount of additional data with the objects stored in it. To take advantage of this, we have chosen to use a new feature of the SRB: the extensible schema. With this, we will tailor the storage of our metadata to create a robust and scalable archive. The light cone, with its clearly defined series of simulations, and similarly well defined initial products (projections and sky maps), presents the cornerstone of our digital archive. I am using this data as a template to help me define our schema, or data hierarchy, and the metadata for both simulations and derived data.

Currently, we have the SRB domain, a draft schema, and data ready archival. The next steps will follow a practical route of expediency: the data will be ingested into the SRB in a logical structure that follows our new schema, and then, as we define our metadata, we will begin to associate this with the objects in the SRB. For now, we will take advantage of the several exiting SRB clients, which include command-line utilities, Java APIs, Windows clients, and a pre-packages web interface. However, our goal is to follow the example of other archives of large scientific datasets, and develop a website, or portal, which can provide our data for use to the community.

c) Sustainability

The data being generated for the simulated light cone is the first block of the digital archive being built by the LCA. The first use of the light cone data will be to create simulated sky maps, or virtual observations; but once in an archive, with associated metadata describing the features of the data, it will form a basis for future scientific questions. The obvious analogy here is to experimental research, where large amounts of effort are poured into data analysis. By building a repository of existing simulations, we are providing ourselves and others with the ability to continue to extract knowledge from our data, beyond just the initial goals.

These thoughts are not original, as seen by the prevalence of libraries in our society. However, we have a unique product to provide to the community, and the population of this archive will provide the community with several opportunities, such as the validation of our results, and original research based solely on our simulations.

3.6 Code development

Goals. In year 1 the code development team identified four goals: (1) place a baseline version of the Enzo code (Version 2) into a CVS repository and develop the policies and mechanisms for group code development; (2) incorporate *jbPerf* performance library [2] calls into Enzo V2.0 for performance analysis; (3) interface the multigrid elliptic solver package MGMPI [3] developed at UCSD into Enzo V2.0 and test it on an isolated Poisson problem. (4) Develop a staged plan for incorporating radiative transfer into Enzo for applications to cosmic reionization in years 2 and 3.

Status. All four of these goals have been met. Enzo V2.0 has been placed in a group-accessible CVS repository, and is basically the version obtained from Robert Harkness with the following enhancements relative to the public version [1]. (1) 64-bit integer arithmetic for indexing; (2) static processor map for root grid Poisson solver; (3) packed AMR file I/O based on HDF5; (4) fast sibling grid neighbor search based on chaining mesh; (5) generalized particle types. A code development plan enabling group development activities has been defined, and is described in Appendix C. Enzo V2.0 has been instrumented and has been used to establish baseline performance for Enzo as a part of DOE's Office of Science 2006 Joule metric. This baseline report is available at http://jbpc.ucsd.edu/~jbordner/nersc_baseline/. Enzo V2.0 has been interfaced with MGMPI, as described in Appendix D. The baseline runs were carried out on the NERSC bassi system. Performance enhancements made as a part of this exercise will benefit the LUSciD project.

Finally, the code development team developed a detailed plan for incorporating radiation diffusion effects within Enzo. This plan, included as an attached paper, describes the coupled system of radiation-hydrodynamics equations, along with the splitting approach for allowing implicitly-computed radiation effects to couple with explicitly-computed hydrodynamics evolution. Additionally, we examine the structure of the coupled nonlinear system of equations, and develop a systematic approach for efficiently solving the nonlinear and linear systems of equations. Dan Reynolds has begun code development on the data structures and nonlinear solver to be used throughout this approach within Enzo. Such code components will be applicable to all steps in the radiation plan, and should remain applicable to extensions for radiation transport computations in the future. The linear solver components have not yet been started, though the previously-described Enzo/MGMPI self-gravity interface can be easily extended to be used in the initial, non-AMR steps of the plan. Moreover, the simplified Enzo/MGMPI interface based on cell-centered variables will more readily extend to be used in radiation diffusion/transport effects.

3.7 Visibility

Goal: Re-implement the LCA (Laboratory for Computational Astrophysics) website using the PLONE open source content management system and create a LUSciD project page within it.

Status. Done. The home page for the project is http://lca.ucsd.edu/projects/LLNL/. PLONE [4] is a Python-based open source system for content management, providing much functionality needed for implementing intranets, extranets, including document management (see [4]). PLONE runs its own database for content management including access control. We use our project page to share progress reports and design documents. In the future, we will create an interface to our data archive within PLONE.

3.8 Status of code runs at LLNL Thunder (Robert Harkness)

The biggest advance since the last report is that now the first full-resolution tile of the LSST Light Cone simulation series on LLNL Thunder is being run. The first run was done with a 512^3 top mesh and 3 levels of refinement (512^3L3) to a redshift of Z = 3.

We re-designed the light cone series to maintain constant mass resolution. This necessitated a re-run with a much larger top mesh of 896^3 for the first tile. This run reached about Z = 3.5 when LLNL Thunder suffered a severe problem with the parallel file system. Apparently, 2.5 Million files were lost, including over 2,000 files from this simulation. Numerous attempts at re-starting the simulation have failed. This run is using 1372 processors and 343 nodes.

With help from Leesa Brieger, both HPSS and SRB pathways back to SDSC are operational. All of the 512^3L3 run was transferred to SDSC HPSS before the disk problems began, plus about 75% of the 896^3 L3 run. Unfortunately, some of the last few dumps may be corrupted

The 512^3L3 series produced 32 dumps of 11 GBytes each. The 896^3L3 series produced 25 dumps of 59 GBytes each. Clearly, it is now possible to move a reasonable volume of output back to SDSC for analysis.

Work continues to refine the Light Cone series. Unless we find a way to reduce or re-distribute the AMR hierarchy in memory we may not be able to follow the evolution of some tiles down to the desired redshift. Although Thunder has 8 GBytes/node this may not be adequate for the more extreme cases, at least with 4 processes per node. Experience with the NCSA Altix has shown that 4 GB per task is required for 512^3 L7, for example.

The code continues to evolve and of most relevance for Thunder, Harkness has added the Intel MKL FFTs as an option, as well as FFTE.

4. Mathematics for radiative transfer in ENZO

A paper has been written on the development of algorithms for incorporating radiation transfer effects within ENZO.

See the attached paper (pdf document):
Reynolds, D. R., P. Paschos, and J.C. Hayes, "Incorporating Radiation Transport Effects Within Enzo", March 2006.

5. Digital Library and Data Management Tasks

The data management tasks have four components:

- Data management support for the Global Climate Simulations. Explicit tasks include providing tools to support data transport, integrating the Netcdf data file manipulation utilities on top of the SRB data grid, building a digital library to support access to derived data products, and helping with visualizations.
- Data management support for the Cosmology Simulations. The tasks include helping transport data from LLNL to SDSC, building a digital library for derived data products, and helping with visualizations.
- Preservation assessment for the Large Scale Synoptic Survey Telescope. A revised set of deliverables have been coordinated with LLNL in support of the LSST team.
- Collaboration with LLNL on additional uses of data grid technology for managing simulation output.

5.1 Supporting data transfer to SDSC

A major requirement has been the establishment of LLNL data transfer policies under which data may be moved from LLNL resources to SDSC. Originally, all data movement needed to be done by LLNL staff. Jim McGraw provided the required policy, listed below:

> **Data Transfer Policy**
>
> The bottom line is that UCSD staff need to make sure that all work they do on LLNL machines is within the scope of the project, as defined in the proposal. All LLNL staff have the responsibility to make sure all data they provide, in whatever fashion, is appropriately R&R'd prior to giving it to a partner. The conclusion is that UCSD folks can move data off of our systems without any further R&R.

Two data movement environments were established at LLNL, in collaboration with Bala Govindasamy. Initially the HSI utility was installed for direct access to HPSS at SDSC. Bala used the system to move 6 TBs of data to SDSC. Data rates of 7-9 MB/sec were achieved using a single data stream.

In collaboration with Brent Gorda, a Storage Resource Broker data grid server is being installed on the Green Data Oasis disk server. The SRB version is being upgraded to the latest release of the software, version 3.4.1. This SRB version was released on April 28, 2006.

5.2 LSST Preservation Assessment Support

SDSC is collaborating with LLNL on three preservation activities:

- Participation in LSST data management planning meetings
- Demonstration of preservation approaches based on the existing 2MASS sky survey
- Participation in International Virtual Observatory Alliance preservation standards efforts, and in Global Grid Forum preservation standards efforts.

The set of activities desired by the LSST data management group is listed below:

> To: LSST Data Management Mailing List
> Subject: [LSST-data] DMGeneral - Notes from DM PM Telecon 20060421
>
> Attendees:
> Ray Plante NCSA
> Deborah Levine IPAC
> Arun Jagatheesan SDSC
> Jeff Kantor LSSTC
>
> SDSC Involvement – Based on past meetings, we want SDSC to focus on:
> * Long-term preservation – e.g. Establish LSST as Strategic Scientific Database, create UML use cases for long-term storage, post-survey access, technology migration, disaster recovery
> * Community Data Access – model distribution of data out from Archive Center to Data Access Centers and Tiered End User sites, coordinate with Data Access Working Group and NCSA to understand Archive Data organization and with Data Products Working Group to understand community access requirements. Arun Jagatheesan to keep Jeff Kantor and Don Dossa informed of activities
>
> ____
> LSST-data mailing list
> LSST-data@lsstmail.org
> http://www.lsstmail.org/mailman/listinfo/lsst-data

The preservation approach being taken by SDSC builds upon current research projects with NSF Information Technology Research initiatives (including the National Virtual Observatory), NARA, Library of Congress, State Archives, and the University of California. The current state-of-the-art uses data grid technology to build preservation systems that provide:

- Authenticity mechanisms, the assertion that information identifying the source of the data remains linked to the data
- Integrity mechanisms, the assertion that the data bits remain uncorrupted, and that the chain of custody can be tracked, including management of access control. This includes maintenance of the links between data objects, such as a link from a publication to the referenced data set and algorithms to enable reproduction of results.

- Infrastructure independence, the assertion that the data collection does not incorporate any proprietary mechanisms that restrict the migration of the data collection to new technology.

The generation of the preservation metadata (authenticity and integrity metadata) requires the application of archival processes for

- Appraisal – selection of data to be archived
- Accession – registration and loading of data through a submission pipeline
- Arrangement – the organization of the material for browsing
- Description – the creation of metadata attributes used in discovery
- Preservation – the packaging for storage, including replication
- Access – the creation of products that can be viewed

The types of technologies that are used in preservation environments include:

- Standard archival processes
- Workflow system for applying the archival processes
- Data grid system for managing the data while maintaining authenticity, integrity, and infrastructure independence
- Database system for storing preservation metadata
- Storage systems for storing data
- Networks for moving, replicating, and accessing the data

These technologies are also used in the management of the original images and generation of the derived data products. One goal of the preservation environment is to show that generic data management infrastructure can be used for both the original data processing pipelines, and also for the preservation environment. Towards this end, the initial creation of a viable preservation environment can be accomplished by demonstrating that the data management system is able to support multiple types of processing pipelines. The migration of a preservation environment to new technology constitutes one of the types of processing pipelines that need to be supported by the data management system.

LLNL and SDSC have access to the following data processing pipelines:

- 2MASS – (NVO, TeraGrid) creation of a hyperatlas of re-projected mosaics to create standard plates. The process involves extraction of object locations from the standard plates, and the manipulation of plates from multiple surveys
- Quest – (NVO, TeraGrid) application of standard processing procedures to synoptic data
- SuperMacho – (LLNL) application of standard processing procedures to synoptic data
- ENZO – (LLNL, TeraGrid) generation of synthetic views of the early universe, and processing of the synthetic views through LSST processing
- MOPS - LLNL does not have access to MOPS, the Moving Object Processing System, at the time being. However, we are interested in collaborating with the pipeline developers.

We can demonstrate a preservation environment by asserting that the same data management infrastructure is capable of handling the output from multiple generations of processing pipelines. The explicit tasks that SDSC believes are important for preservation assessment include:

1. Demonstration of processing pipeline that analyzes data from a generic data management environment (we are using the Storage Resource Broker data grid as it supports authenticity, integrity, and infrastructure independence for multiple existing preservation projects). To do this, we are creating a processing pipeline that provides the required scalability, manages process provenance information (processing pipeline state

information), and preserves derived data products. We are developing a version of the Matrix workflow language to support massive collections. This is one of the essential development activities needed to handle the scale of data that will be generated by LSST.

2. Demonstration of data grid technology for preserving collections. This includes the demonstration of preservation processes on an existing 10-TB sky survey (2-MASS). We have migrated the 2MASS image survey to new storage technology, validated the integrity of the collection using MD5 checksums, replicated the collection, and migrated the preservation metadata to new database technology versions (Oracle 9 to Oracle 10).
3. Demonstration of management of derived data products. We will work with ENZO simulation data to manage synthetic views of the universe that will be stored in the data grid, along with the derived products (object locations) as they are created. We have applied NVO Mosaic technology to the 2MASS 10-TB sky survey to demonstrate star extraction, standard projection onto reference plates, and ingestion of the plates into a digital library. This process demonstrates the formation of standard plates against which future observations can be directly composited.
4. Demonstration of database preservation. A future task is to collaborate with LLNL on the demonstration of the export of database tables into the preservation environment, and the import of archived database tables into a database. The goal is to show that selected sky survey catalogs can be dynamically loaded into a large parallel database, and that queries can be done across the resulting database, without loss of authenticity.
5. Architecture design. SDSC is collaborating with the LSST data management team on an assessment of the ability of data grid technology to meet the design requirements for LSST infrastructure. SDSC is participating in the LSST middleware working group, and mapping their data management requirements onto the capabilities provided by the Storage Resource Broker data grid
6. Storage. SDSC expects to provide up to 10 TBs of storage space under SRB collection management control to hold LSST data products.
7. Data distribution. SDSC has demonstrated use of data grids to distribute and replicate data across multiple storage systems at multiple institutions. A similar system is in use by NOAO to manage federation of data grids between Chile, Tucson, and NCSA. SDSC is collaborating with NVO on use of NVO services to access the distributed data. SDSC has also demonstrated the bulk data transport capabilities provided by data grids.
8. Data grid federation. SDSC is tracking evolution of Global Grid Forum standards for the federation of data grids. A demonstration was given at the 17th Global Grid Forum meeting on the federation of 14 data grids that are based on the SRB technology, including the NOAO data grid. The participating data grids span institutions in Europe, the US, South America, Australia, and the Far East.
9. Risk mitigation. SDSC has demonstrated technologies that minimize the risk of data loss. These include use of checksums, file replication, validation of checksums, synchronization of replicas, and federation to replicate metadata catalogs. A paper was written summarizing the capabilities.
10. Distributed pipeline processing. SDSC plans to demonstrate distributed workflow management on top of the preservation environment.

5.3 LSST Data Preservation Assessment using 2MASS

The addition of 2MASS to the NVO Hyperatlas has provided a testbed for SDSC to begin assessing the data preservation needs of the LSST project. The raw data of the 2MASS survey was transferred to SDSC, checked for integrity, and processed to generate the data products that will go into the Hyperatlas. These consist of standard plates that can be used to support compositing of future observations. The standard plates are now being validated for registration into a Hyperatlas and placement into on-line storage from which it will be available for on-demand access by the astronomical community.

The 2MASS archive is on the order of the 10 TB and is therefore comparable to an LSST single-day acquisition. The challenges of large-scale data transfer and data integrity checking became clear when it became necessary to move this data set to SDSC for processing. Data grid software that automates integrity checking and the restart of transfers when necessary will be of utmost importance for the daily transfers on this scale for LSST.

The data products of the processing pipeline were 6-degree mosaics, in each of the three infrared bands, calculated to match the Hyperatlas format. The processing of 5200 mosaics was carried out on Teragrid resources, organized in parallel MPI jobs of 128 or 256 CPUs at a time, calculating 128 or 256 mosaics simultaneously. Significant inefficiencies in the calculations were experienced due to long-undetected hardware problems in the compute resources. It is very interesting to note that problems similar to those experienced by ENZO on Thunder were encountered, in which the file system lost data. The problem on the Teragrid was traced to hardware. The importance of the integrity of the compute resources is another important factor, not to be taken for granted, in processing data on the scale of LSST. Ultimately, we utilized over 100,000 CPU-hours to apply the NVO Montage service, developed at IPAC, to 4,121,440 images to generate 5,196 standard plates.

The raw data was stored as 32-bit floating point data but was converted to 64-bit double precision for the calculations. Output was written in the 64-bit format, resulting in an inflation of the output data, which came to 20 TB. The data has subsequently been converted back to 32-bit precision. Additionally, it has been decided that additional reduced images in jpeg format are needed for browsing.

Assessment of the output will now take place through a validation program that extracts the location of 1000 stars for a given plate from the 2MASS catalog, evaluates photometry for each star and evaluates astrometry for each star. Such validation after processing must be verified in the LSST preservation environment, as part of the authentication of output data. FITS header information is being updated for the collection to contain assertions about the image creation process. This serves as metadata for each plate of output data; having the right information stored as metadata is crucial to being able to verify the authenticity of each data product.

Zero-point corrections will be incorporated into the calculation of radiance so as to correct even further for background seeing conditions, and the data processing pipeline will be put into action again: calculations, validation, generation of jpeg archive, storage. The data preservation environment for LSST will have to guarantee the robustness of each of these steps. We are building up the experience to be able to deliver that with our work with the 2MASS archive.

6. Budget Summary

The first year of funding for the project is for the time period July 1, 2005 through June 30, 2006. While the research activities proceeded immediately on award, the deployment of the application codes and data management technology to LLNL was substantially delayed. Several important data access issues were clarified by LLNL, the network bandwidth from LLNL to UCSD was upgraded, and LLNL policies were followed for the award of user accounts and installation of software. These policy issues were resolved on the following dates:

- July 18, 2005 – data transmission rates of 50 MB/sec were measured from a computer outside of the LLNL firewall to SDSC.

- November 2005 – user accounts were established for all but one UCSD person.
- March 14, 2006 – the data distribution requirements for moving data from LLNL to UCSD were defined. Prior to this date, LLNL staff initiated data movement from LLNL to UCSD.
- May, 2006 – the SRB data grid software is being installed on the Blue Oasis disk cache at LLNL

The hiring of staff was delayed at SIO and SDSC as the above issues were resolved. The delayed hiring and the late start of application execution at LLNL have resulted in an under-run of the proposed budget. The under-run at SDSC was caused by a smaller use of storage than originally planned. The total amount of data stored at SDSC was ten times smaller than originally budgeted. This decreased the storage costs by a factor of ten at SDSC for the first year, and resulted in smaller storage charges by about $80,000. As the project simulations complete, the amount of data stored will grow to the targeted amount and the SDSC costs will match the budget. The under-run at SIO was caused by the inability to hire a person willing to work part time at UCSD and part time at LLNL. The approach now being followed is to hire a person that will work full time at either LLNL or UCSD. In the interim, staff at SIO are supporting the project.

The budget figures for July 2005 through April 2006 are provided below. Note that an additional two months of expenses will be accrued on the first year of the project for the months of May and June.

Table 1. Cumulative expenses through April 2006.

SDSC	Budget	Expense	Balance
Labor & benefits	202,235	154,861	47,373
Supplies & expenses	90,142	7,049	83,092
Equipment	4,000	0	4,000
Travel	4,600	2,083	2,517
Subtotal	300,977	163,994	136,983
Lab. for Comp. Astrophysics			
Labor & benefits	223,541	205,597	17,945
Supplies & expenses	14,482	12,587	1,885
Equipment			
Travel	2,000	1,554	446
Subtotal	240,023	219,747	20,276
SIO			
Labor & benefits	168,822	41,595	127,227
Supplies & expenses	12,938	6,515	6,423
Equipment			
Travel	18,240	2,164	16,076
Subtotal	200,000	50,274	149,726
Total	741,000	434,015	306,985

The projected expenses for May and June are

- SDSC - $50,000
- Center for Astrophysics - $38,000
- SIO - $45,000

The under-run will be applied to the following tasks:

- SDSC - $86,983 from under-utilization of storage will be applied to increased visualization support by Steven Cutchin, and used to support increased storage in the second two year

- SIO - $105,000 from delayed hiring of a full time postdoc will be used to hire a full time person at either LLNL or UCSD

A budget is being prepared for the second year that will total $750,000. The allocation to the Laboratory for Computational Astrophysics will be increased to $249,023, to help cover the costs for supporting Dan Reynolds on the project. The revised budget will also include an allocation from the SDSC under-run to support travel costs for participation in the Global Grid Forum on grid standards.

7. References

[1] Enzo code website: http://lca.ucsd.edu/codes/currentcodes/enzo
[2] jbPerf website: http://lca.ucsd.edu/codes/currencodes/jbPerf
[3] MGMPI website: http://lca.ucsd.edu/codes/currentcodes/MGMPI
[4] PLONE website: http://plone.org

Appendix A. Definition of variables to be saved from the global climate model runs (David Pierce)

There will be the following runs:

> 1) A global run of CCSM3 with the finite volume dynamical core, at a resolution of 1.25 deg longitude by 1.0 deg latitude. **** CORRECTED ****
> 2) A downscaling run using MM5, which will be forced by the CCSM3 results. I don't know what the resolution of this run will be. In the estimates below, I assume it will be 1/8 degree by 1/8 degree over the domain 130 West to 70 West, 25 North to 50 North (basically, continental U.S. at what I believe is VIC standard resolution). If someone knows otherwise, please correct me.
> 3) SIO will then produce VIC runs using the saved MM5 data. The VIC model output will be saved at SDSC, and so is included in the storage size calculations below.

Note that this list does NOT include whatever it is the LLNL folks will need to run MM5. I am assuming their MM5 people will specify what needs to be saved for them, how frequently, and where it will be stored. The following variables need to be stored:

39 2-D fields:

- PBL (planetary boundary layer depth)
- TREFHT (reference height air temperature)
- TS (surface temperature)
- Convective Precipitation rate
- Large Scale precipition rate
- Snowfall rate
- Snow depth
- Tau-X, Tau-Y
- UBOT, VBOT
- Latent, Sensible, Longwave, Shortwave net fluxes at surface
- FSDS (downwelling solar flux at surface)
- TREFHTmax, TREFHTmin for the day
- QREFHT (reference height specific humidity)
- CLDTOT (vertical integrated total cloud)
- SLP (sea level pressure)
- PS (surface pressure)
- ICEFRAC (sea ice fractional coverage; global model only)
- PRW (total columnar water vapor)
- The following atmospheric variables to be saved at 850, 500, and 250 hPa:
- U; V; T; Q; Geopotential height (NOTE: requires postprocessing)

Pointwise values extracted from a 2-D field:

- Values of the river transport model (variable QCHANR) at 28 points to be supplied by SIO (NOTE: requires postprocessing)

5-DAY FIELDS (3-D, from global model only):

- Q (specific humidity)

MONTHLY FIELDS (2-D):

- Soil moisture on top 2 levels

MONTHLY FIELDS FROM OCEAN (2-D):

- OCEAN SSH (sea surface height)
- OCEAN Utop, Vtop
- OCEAN U100m, V100m

MONTHLY FIELDS (3-D, from global model only):

- AIR TEMP (from atmos model)

MONTHLY FIELDS FROM OCEAN (3-D):

- OCEAN TEMP (from ocean model)
- SALIN (from ocean model)

Data storage requirements

For the global atmospheric model: **** VALUES IN THIS SECTION CORRECTED ****

1.25x1 deg = 288x181 = 52.1k values per 2D field. *4 bytes/value = 209 KB/2D field. *26 levels (is this what is planned?) = 5.42 MB/3D field

The 39 2D fields would then take 40*173KB = 8.15 MB

If we stored daily averaged 2D fields, water vapor averaged over 5 days, and monthly averaged 3-D temp, this would give about:

30*8.15MB + 6*5.42MB + 1*5.42MB = 282 MB/month = 3.39 GB/model year (global atmosphere)

A 500-yr run would then produce 1.69 TB (global atmosphere)

For the MM5 model:
Assume 1/8 x 1/8 deg over continental US ~ 480x240 = 115k values per field.
*4 bytes/value = 460KB/2D field.

I have not heard anyone request any 3-D fields from the MM5 run (note again, this estimate does not include whatever is needed to run MM5; I'm assuming the MM5 folks will worry about that part).

If we stored the 39 2D fields 4 times/day, this would give about:

39*460KB = 17.9MB/slice *4slice/day = 71.8MB/day *30 days = 2.15GB/month = 25.8 GB/model year (MM5)

A 500-yr run would then produce 12.9 TB (MM5)

For the ocean model: **** VALUES IN THIS SECTION CORRECTED ****
The ocean model will be run at a resolution of 320x384x40

320*384 = 123k values per 2D field. *4 bytes/value = 492 KB/2D field. *40 levels = 19.7 MB/3D field

The 5 2-D fields would then require 5*492 KB = 2.46 MB/month = 29.5 MB/model year
The 2 3-D fields would then require 2*19.7 MB = 39.4MB/month = 49 MB/model year
The total per model year would then be (29.5 MB + 493 MB) = 523 MB/model year (ocean)

A 500-yr run would then produce 500*523 = 261 GB (ocean)

For all models combined: **** VALUES IN THIS SECTION CORRECTED ****

3.39 GB/yr (global atmo) + 25.8 GB/year (MM5) + 523 MB/yr (ocean) = 29.7GB/yr

So a terabyte of disk could hold 33 years of the model run.

VIC Model output

The following variables are those suggested to be saved to SDSC from the VIC model output to be generated by the SIO folks (i.e., this is NOT a list of variables to be saved by the LLNL folks). It is included here for documentation and to calculate the data size. It is taken from a list compiled by Hugo Hidalgo.

- day
- month
- year
- evap
- runoff
- baseflow
- soil moisture top layer
- soil moisture middle layer
- soil moisture bottom layer
- snowpack water equivalence
- net shortwave radiation
- incomming longwave radiation
- latent heat flux
- sensible heat flux
- net surface radiation
- ground heat flux
- surface albedo
- surface temperature
- relative humidity
- soil ice content
- transpiration
- snow sublimation
- bare soil evap
- interception
- snowpack depth

Hugo has calculated that this will take 4.1GB/model year. Thus a 500-yr run will take 2.1TB.

Table 1. Total corrected values for SDSC storage, calculated for a 500-yr run of the global model, MM5, and VIC:

Global Atmosphere	1.7 TB
MM5	12.9 TB
VIC	2.1 TB
Ocean	0.3 TB
Total	**17.0 TB**

Table 2. Locations of Points to Save from River Runoff Model (QCHANR)

32	306
92	317
113	273
178	241
222	277
130	249
117	257
119	255
257	179
237	199
245	111
373	190
386	169
432	144
423	242
453	279
371	284
370	268
442	309
503	314
615	325
528	321
640	284
593	253
602	245
556	215
537	229
497	230

Appendix B: Description of the PCM downscaling runs

Dr. Andy Wood, Res. Asst. Prof., UW Dept. of Civil Engineering
Mr. Niklas Christensen, MSE

Background

As part of the Dept. of Energy Accelerated Climate Prediction Initiative (ACPI), the University of Washington (UW) bias-corrected and downscaled four Parallel Climate Model (PCM) integrations, and used the resulting temperature and precipitation scenarios to drive simulations of the UW Variable Infiltration Capacity (VIC, version 4.0.3 rev. 3) hydrology model, applied to a domain comprising the Pacific Northwest (at 1/4 degree spatial resolution), California and the Columbia River basin (both at 1/8 degree spatial resolution). The four climate scenarios were a control climate run (b0645: 1999-2048 with static 1995 greenhouse gas concentrations) and three future climate runs (b0644, b0646, b0647: 1999-2098 with evolving greenhouse gas concentrations). Output from a fifth PCM integration (b0628: 1870-1999 with historical greenhouse gas concentrations) was also archived at UW and used in bias-correcting the other four runs, but it was never downscaled or processed into hydrology model forcings, nor was an associated hydrologic simulation performed. Output from a sixth PCM integration (b0622: 1870-1999 with historical greenhouse gas concentrations) was also archived, and a 20-year subset (1975-95) was downscaled and used to force the VIC model as part of a downscaling methods analysis. The four primary runs, however, formed the basis for four papers appearing in a special journal issue devoted to the ACPI project, *Climatic Change* Vol. 62, Issue 1-3 (January, 2004), in which relevant methods are described.

Objective of Proposed Work

As part of an analysis of climate trends and variability and associated hydrologic sensitivities, Scripps researchers intend to evaluate more fully two PCM integration: (a) an historical run for the full period 1870-1999; and (b) the control run, nominally 1999-2048. UW researchers will downscale and process the runs using consistent methods and model versions to those used for the ACPI project, and provide the resulting daily hydrologic output datasets to Scripps. UW will use a current version of VIC source code (4.0.5) and current basin hydrology model parameters sets to do so.

Deliverables

- Daily time-step outputs for the entire PCM b0628 historical scenario and the PCM b0645 control scenario, both used previously as part of the DOE ACPI project, over the ACPI study regions (the Colorado and Columbia River basins and California at 1/8 degree spatial resolution). The hydrologic output dataset includes precipitation, evaporation, surface runoff, baseflow, soil moisture (in 3 layers), and snow water equivalence, all at a daily time resolution.
- Monthly time-step averages or totals of the daily outputs.
- Monthly time-step streamflow at the locations used in the earlier ACPI work (35-40 locations total).
- Summary diagnostic graphics as time permits, and a brief summary of relevant methods.

Appendix C. Enzo Code Development Plan (Dan Reynolds)

Enzo is currently at somewhat of a crossroads, in that it will soon be used by a large portion of the scientific computing community for both scientific and benchmark computing, the number of core Enzo developers is growing, we have one current code version on which all future development will be based, and a number of new additions will soon be made to Enzo's physical and algorithmic functionality. As a result, we wish to decide on a development strategy for current and future Enzo development that attains the following goals:

1. allows for simplified integration of new physics modules
2. minimizes inefficiencies due to different code versions (playing 'catch up' to new versions)
3. allows for simplified merging of contributions from the multiple Enzo developers
4. incorporates regular regression testing, to ensure that code additions do not destroy Enzo functionality

We plan to address these in the manner outlined as follows:

1. Document the existing Enzo code infrastructure, including current algorithms and program flow. This should build off of the documentation template that Dave Collins and Brian O'Shea have sent out, and should serve as the de-facto description of Enzo's capabilities, usage, and structure. We all plan to help out with this phase, as no single one of us understands all of Enzo, but our combined knowledge should span the large majority of the code.
2. Set up regular regression testing. To do this, James Bordner will extend the current build system using the approach that he developed for the previous Enzo version. Additionally, Pascal Paschos has proposed to develop one (or more) tests that will exercise Enzo's physical functionality, which will be used as the example code(s) for regression testing. John Hayes will also help out in this area. Once these tests are in place, we will run some base tests and save their output. Regression tests will then compare results against these 'trusted' results. James Bordner has proposed to set up the automated regression testing system in the same manner as what he has previously done for MGMPI, in which regression test results are automatically uploaded to a web page.
3. Compartmentalize Enzo based on functionality. Since much of the code inside Enzo overlaps between different routines and serves multiple purposes, development may be difficult for a large team, as any changes by one author will almost certainly interfere with changes by the others. Therefore, we plan to re-work much of the 'glue code' within Enzo so that different functional modules are each accessed through clean interfaces, so that , for instance, development on an I/O module does not affect functionality of various physics modules. To this end, we plan to each look through Enzo to consider where these clean interfaces should lie, and how they should interact. We then plan to meet during the first week of April to discuss this compartmentalization, and decide on a plan of action from there.
4. Set up regularly-scheduled meetings between core Enzo developers. We plan to meet every two weeks or so to discuss what we are currently working on and how best to interact on software development. It will be at these meetings where we will determine when and how to perform large-scale Enzo enhancements (e.g. incorporation of MHD or radiation, etc.).
5. Use CVS regularly to both update personal code to current code base, and commit contributions to the repository for regression testing and use by others. This should serve

to minimize parallel development, in which multiple developers are working on the same bugs and/or bug-fixes by one developer can be propagated to the others and to scientific users.

6. Update the documentation regularly as new functionality and/or algorithms are added to Enzo.

Appendix D**. Enzo/MGMPI interface (Dan Reynolds)**

In anticipation of incorporating implicit nonlinear and linear solvers into Enzo for solution of coupled radiation-hydrodynamics systems, I have developed an Enzo/MGMPI interface for solving self-gravity effects based on non-periodic domains. This interface enables the use of geometric multigrid and Krylov linear solvers from the MGMPI package within Enzo, allowing problems based on a combination of Dirichlet (isolating) and periodic boundary conditions in each of the three spatial dimensions. The resulting solver may currently be used by either fresh or restarted Enzo runs; however, parallel I/O of the potential field boundary conditions is still under development (nearly complete). Moreover, in discussions with Alexei Kritsuk and James Bordner, I believe that the interface may be simplified considerably so that these solutions may be performed on Enzo's cell-centered hydrodynamics mesh (as opposed to interpolations to/from a nodal mesh) -- allowing for increased accuracy and efficiency in the resulting potential field solutions. Development on these simplifications is also currently under way, and should be completed shortly.