Title: A Scalable Framework for the Global Offline Community Land Model Ensemble Simulation

Dali Wang (Corresponding author)
Environmental Sciences Division
Climate Change Science Institute
PO Box 2008, MS 6301
Oak Ridge National Laboratory
Oak Ridge TN 37831, USA
Email: wangd@ornl.gov
Tel: 865-241-8679 Fax:865-574-4665

Jens Domke
Joint institute for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge TN 37831, USA
Email: domkej@ornl.gov
Tel: 865-241-6293

Jiafu Mao
Environmental Sciences Division
Climate Change Science Institute
PO Box 2008, MS 6301
Oak Ridge National Laboratory
Oak Ridge TN 37831, USA
Email: maoj@ornl.gov
Tel: 865-576-7815 Fax:865-574-4665

Xiaoying Shi
Environmental Sciences Division
Climate Change Science Institute
PO Box 2008, MS 6301
Oak Ridge National Laboratory
Oak Ridge TN 37831, USA
Email: shix@ornl.gov
Tel: 865-241-9199 Fax:865-574-4665

Daniel Ricciuto
Environmental Sciences Division
Climate Change Science Institute
PO Box 2008, MS 6301
Oak Ridge National Laboratory
Oak Ridge TN 37831, USA
Email: ricciutodm@ornl.gov
Tel: 865-574-7067 Fax:865-574-4665

Dali Wang is a staff scientist of Environmental Sciences Division, staff member of Climate Change Sciences Institute (CCSI) at Oak Ridge National Laboratory (ORNL), an adjunct professor of geography at the University of Tennessee, Knoxville, and a research associate of Global Monitoring Division at National Oceanic and Atmospheric Administration (NOAA). His research interests include environmental and climate modeling, environmental data sciences and systems, high-performance scientific computation and optimization, geographic information system, and large-scale system integration and simulation. Wang has a PhD in environmental engineering (scientific computation focus), MS in computer science and MSC in computational sciences and engineering, all from Rensselaer Polytechnic Institute (RPI), Troy, New York. After graduation from RPI, Wang became a research faculty of computer sciences at the University of Tennessee at Knoxville. He subsequently joined Southeastern Universities Research Association (SURA) at Washington, DC as an information technology infrastructure manager. Wang is the author and co-author of over 30 peer reviewed papers, over 30 presentations, several technical reports and software releases. He is a member of several societies, including Institute of Electrical and Electronics Engineers Computer Society, American Geophysical Union, American Association of Geographers.

Jiafu Mao is a research scientist with the Ecosystems Simulations Science group in the Environmental Sciences Division. He earned a combined MA-PhD in atmospheric sciences at the Institute of Atmospheric Physics, Chinese Academy of Sciences in 2007. After his work in China, he became a joint postdoctoral research fellow at the University of New South Whales and the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia (2008-2009). Mao was a Postdoctoral Research Associate with ORNL from August 2009 until his staff appointment in November of 2011. Mao is currently working on different projects such as the incorporation of the Global Land-Use Model/Global Climate Assessment Model (GLM/GCAM) systems in the Community Earth System Model (CESM) (as part of the iESM project), the replication and parameter improvement of the soil biogeochemical module for the Community Land Model (CLM) (as part of the Climate Science for a Sustainable Energy Future project), single-factor simulation and remote sensing evaluation of CLM and CESM. Systematic analysis and inter-comparisons of extreme climate products from station observations, reanalysis data, and Earth system models are some of Mao's other research interests.

Xiaoying Shi is a scientist for the Environmental Sciences Division at ORNL. Shi is an atmospheric scientist with training in atmospheric sciences during the periods of her BS, MS and PhD degrees' study. Since joining ORNL in 2009, her research has been focused on coupling Integrated Assess Model (GLM/GCAM) to the Community Earth System Model (CESM) to study on the global climatic changes and feedbacks by the introduction of human dimension in the complex earth system. Shi received her doctorate in atmospheric sciences from the Graduate University of the Chinese Academy of Sciences in 2007 and became an assistant researcher staff member at State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, where she worked on regional climate change. In 2008, she moved to CSIRO Marine and Atmospheric Research, where she calibrated CASACNP model (including carbon, nitrogen, and phosphorus cycles) and coupled it to Common Land Model, CoLM.

Jens Domke works as a research associate at the Joint Institute for Computational Sciences (JICS) at Oak Ridge National Laboratory (ORNL). The focus of his work is to support application developers with trace based performance analysis for their projects of porting/scaling applications for the next GPU-based HPC system at ORNL. His research interests are in the field of scientific computing and include, among others: high-performance computing, parallel algorithms, performance evaluation & optimization, HPC interconnects & routing schemes, and scheduling algorithms. Jens Domke received his Master's degree in mathematics (Dipl.-Math.) from the Technische Universität Dresden in 2010 for developing an optimized and deadlock-free routing algorithm for InfiniBand. He gained his 5 years of experience in HPC, performance analysis and routing algorithms while working at the Center for Information Services and High Performance Computing (ZIH), TU Dresden, as a student research assistant and as a research associate and while working at JICS, ORNL.

Daniel M. Ricciuto is a staff scientist in the Ecosystem Simulation Science group in the Environmental Sciences Division at ORNL. His main research interest is the application of data assimilation techniques that confront terrestrial carbon cycle models with observations to improve model parameterization and predictive skill at spatial scales ranging from site-level to global. Ricciuto received his PhD in meteorology from Pennsylvania State University in 2006. He then became a postdoctoral associate at ORNL under the supervision of Wilfred M. Post until transitioning to a staff position in the same group in March of 2010. Ricciuto is collaborating with experts in terrestrial carbon cycle modeling and computer sciences to develop a new data assimilation framework for terrestrial components of the Community Earth System Model that takes maximum advantage of ORNL's high-performance computing resources.

A Scalable Framework for Global Offline Community Land Model Ensemble Simulation

Abstract

Current earth system models have a large range of uncertainty, due to differences in the simulation of feedbacks and insufficient information to constrain model parameters. Parameter disturbance experiment provides a straightforward method to quantify the variation (uncertainty) outputs caused by model inputs. Due to the software complexity and computational intensity of earth system model, a large-scale simulation framework is needed to support ensemble simulation required by parameter disturbance experiment. This paper presents a parallel framework for the community land model ensemble simulation. After a software structure review of the community land model simulation, a single factor parameter disturbance experiment of a referenced computational experiment design is used to demonstrate the software design principles, computational characteristics of individual application, parallel ensemble simulation implementation, as well as the weak scalability of this simulation framework on a high-end computer. Finally, the paper discusses some preliminary diagnostic analysis result of the single factor parameter disturbance experiments. The framework design considerations and implementation details described in this paper can be beneficial to many other research programs involving large scale, legacy modeling system.

Key words: Computational sciences, High-end computing, Earth system modeling, Performance profiling, Scalability

1. Introduction

The Community Earth System Model (CESM) administrated by the National Center for Atmospheric Research (NCAR) is one leading US earth system models. The newly released CESM contains five major community model components as well as associated data models to simulate earth systems: atmosphere, ocean, sea ice, land, and glacier. Like most earth system models, CESM has a large range of uncertainty, due to differences in the simulation of feedbacks and insufficient information to properly constrain model parameters. Advanced methods in non-linear dynamics and advanced statistics coupled with high end computing resources and integrated observational data sets can reduce these uncertainties. The whole system of CESM is reconfigurable, which provides a great flexibility to the community to design their own computational experiments. In our study, CESM has been configured into an offline global Community Land Model (CLM) simulation, which includes a data atmosphere model, active land model and stub models for ocean, ice and glacier. Scientifically, this configuration uses historical climate forcing to drive active land component simulation, and provides unique capability to verify and calibrate land modeling activities with observational datasets.

The overall uncertainty quantification (UQ) approach, from the perspective of land model evaluation and development, is first to characterize uncertainty in model parameters values, then to quantify the components of prediction uncertainty associated with parameter uncertainty, structural uncertainty, and uncertainty in model forcing (e.g. surface weather), and finally to propagate these multiple uncertainty components forward to generate an overall estimate of prediction uncertainty. This approach is envisioned as an iterative process, repeated as new observational datasets and new process model (structural and parametric) understanding is brought to bear. As the first step of the systematic UQ methodology is the single factor parameter disturbance experiment, which estimates uncertainty in

model outputs caused by single factor disturbances, and provides baseline information to guide subsequent observational campaigns towards improved estimates of those parameters of maximal impact on the uncertainty predictions. Due to the software complexity and computational intensity of CESM, a large-scale simulation framework is needed to support simulation required by parameter disturbance experiments.

This paper focuses on parallel framework design to support ensemble simulation of the global offline community land model. After a review of the current software system structure of the global offline Community Land Model (CLM), the authors explain the considerations to design a large simulation system to support ensemble simulation on high-end computers. After that, a single factor parameter disturbance experiment of the global offline CLM simulation is used to present the characteristics of the software system, and to demonstrate the scalability of this framework on a Cray XT5. Finally, authors discuss the diagnostics analysis of simulation results and explore future work.

2. Software Structure of The Global Offline CLM Simulation

The active land used in our CESM framework is the Community Land Model, version 4 (CLM4, (Oleson et al., 2010)). CLM4 includes a fully prognostic treatment of surface energy, water, carbon, and nitrogen fluxes and state variables for both vegetated and non-vegetated land surfaces. Recent model development includes improved surface energy partitioning and thus water cycling (Lawrence et al., 2007; Stöckli et al., 2008), and improved ability to reproduce contemporary global patterns of burned areas and fire emissions (Kloster et al., 2010). Sub-surface hydrology parameterizations have also been changed to improve prediction of permafrost dynamics. In our study, we will use the CLM model with coupled carbon and nitrogen cycle biogeochemistry functionality, wildfire, and a dynamic natural vegetation model (DGVM), which estimates states and fluxes of carbon and nitrogen for vegetation, litter, and soil organic matter, and associated exchange with the atmosphere (Thornton et al., 2009; Thornton et al., 2007).

The software system of the global offline CLM includes two groups: models and scripts. The model group includes physical earth system components, such as the CLM4, data atmosphere, stub ocean, stub ice and stub glacier. It contains a driver to configure the parallel computing environment and the whole simulation system (physical earth system components and flux mapping functions between those components). It also includes several shared software modules and utilities, such as a flux coupler and its APIs to individual earth system component, parallel IO library, performance profiling libraries. The schematic diagram of the CLM software structure is shown in Figure 1.

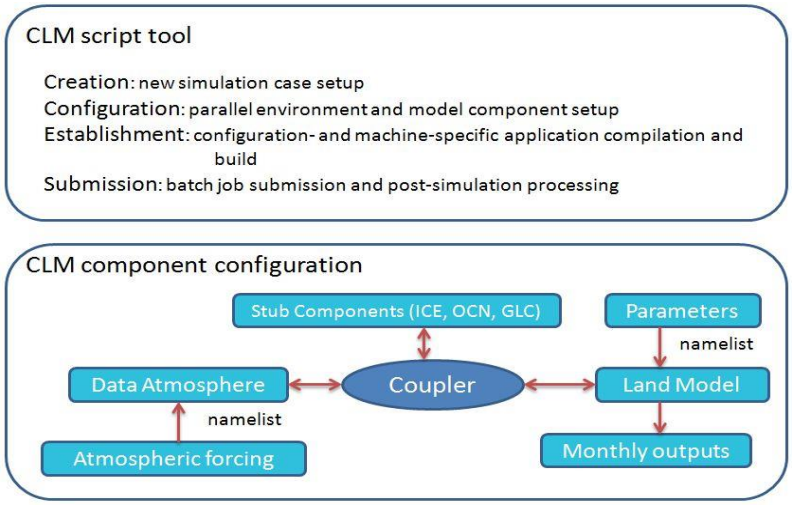

Figure 1: Software structure of the global offline CLM simulation

The CLM scripts contain tools and templates to create and run a CLM simulation case. Those scripts use pre-defined templates to create a simulation case and define CLM model components. Those scripts also use machine-specific computing environment settings, predefined make and configure templates, to generate software dependences, software libraries and the CLM executable, which are highly customized for each supported parallel machine. Specifically, there are 4 general steps:

1) Case creation

In this process, CLM scripts create a new case directory that contains simulation configurations (including case name, model component configuration, and grid resolution), environment variables, as well as case specific tools (such as environment settings for model preprocessing, compilation, build, and post-simulation processing). The list of environment variables is stored in XML format, and will be converted to session-only environment settings during the script execution.

2) Case configuration

The main purpose of this step is to configure parallel computing environment and to prepare the input datasets for each model component, that is, data atmosphere model, and land model and coupler in our case. After the configuration, case-specific scripts have been generated and copied into a new directory.

3) Case establishment

In this step, model source code and case-specific configuration information, machine-specific settings are used along with predefined templates to configure each model component, generate software dependencies, create libraries and then generate the executable with a list of input files for each model component.

4) Case submission and execution

In this step, a batch script is used to reserve computational resources, check case status, setup session-only environment variables, and other pre-simulation sanity checking. The scripts will also launch the parallel application, wait for the simulation to finish and then launch post simulation processing.

In the summary, the activities mentioned in those 4 steps involve several directory trees: including input datasets, case configuration, and working directories. Much of path information is hardwired into the executable.

3. Computing Infrastructure

The computational platform used in this research is the Cray XT5 partition of Jaguar supercomputer at National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory (ORNL). The XT5 partition contained 18,688 compute nodes in addition to dedicated login/service nodes. Each compute node contains dual hex-core AMD Opteron 2435 (Istanbul) processors running at 2.6GHz, 16GB of DDR2-800 memory, and a SeaStar 2+ router. The resulting partition contains 224,256 processing cores, 300TB of memory, and a peak performance of 2.3 petaflop/s (2.3 quadrillion floating point operations per second). A center-wide Luster file system provides 5 PB of disk space for all NCCS computing resources. The massive longer term storage is provided via the NCCS High Performance Storage System (HPSS). The NCCS has implemented customized queue policies to enable large jobs to run in a timely fashion on the systems at ORNL. As a DOE Leadership Computing Facility (LCF), the ORNL has a metric that a large portion of the XT5's usage must come from jobs requesting at least 20% of the system (45,000 cores). Jobs requesting 45,000 or more cores are not subject to any over allocation penalties on the Cray XT5.

4. Single Factor Parameter Disturbance Experiment Design

4.1 A Reference Global Offline CLM Simulation

In our study, a CLM4 experiment was setup to simulate historical land surface conditions driven by a 111 year (1901-2010) observation-constrained half degree CRUNCEP dataset. The CRUNCEP is a combination of the CRU TS.2.1 0.5° monthly climatology covering the period 1901 to 2002 [Mitchell and Jones 2005] and the 2.5° NCEP2 reanalysis data beginning in 1948 and available in near real-time (see more details at http://dods.extra.cea.fr/data/p529viov/cruncep/). The simulation was spun up for 1850 conditions (atmospheric CO_2, nitrogen deposition, and land cover), driven by a repeating 20-year subset (1901-1920) of the meteorological forcing data. In this study for framework design purpose, the sensitive analysis simulation was initialized from the spin-up simulation, and the simulation period is 30 year driven by the climate forcing. Our production run will be the simulations for year of 1850-1900 used the same repeating 20 year climate forcing, but CO_2 concentration, nitrogen deposition, and land use and land cover change, for 1850–1900. Subsequently, the 1901-2009 meteorology was used in simulation experiments for 1901-2009. Annual land use change and harvest area for 1850-2005 were derived from the University of New Hampshire version 1 Land-Use History A (LUHa.v1) historical dataset based on that of Hurtt et al. [2006], and from IPCC RCP4.5 scenario for the period over 2006-2010. The transient atmospheric CO_2 concentration and nitrogen deposition are similar to the previous studies [Bonan and Levis 2010, Shi et al., 2011, Mao et al., 2012].

4.2 Reduction of Simulation Dependency

Since CESM/CLM has a large user community, our framework is designed with the considerations of community adaption and code coherence. The original scripts structure was changed to ensure the independency of each simulation session. Specifically, an individual simulation directory was created for each simulation. Six files have been moved into the simulation directory. Those files are used for runtime environment setup, environment variables retrieve, post-simulation processing (archive and resubmission), runtime simulation input generation, as well as machine-specific computing environment settings. Two data files (initial restart file and physiology parameter file) are also moved into this directory, furthermore, CLM is reconfigured to use those two data files for simulation, instead of default settings. The framework is also designed to separate the IO streams. Most of input data streams (except the restart files) are stored in a shared project directory, but output data are pushed into separate individual simulation directory. The original short-term and long-term archive capability has been disabled for individual simulation. After those modifications, each individual simulation can be

configured, established, submitted via the original scripts, and the executable can be moved away from the default execution directory without invalidating the CLM's sanity checking for runtime environment variable settings and simulation case.

4.3 Single factor parameter disturbance experiment

In CLM, many of these parameters appeared as fixed numerical values in various subroutines. Therefore, significant code modifications were made to place these parameters into an external input file to facilitate parameter manipulation. Specifically, 77 model parameters, controlling photosynthesis, energy balance, hydrologic cycling, autotrophic and heterotrophic respiration, decomposition, nitrogen cycling and phenology, were first identified and recorded in an external pft-physiology file (see Table 1, listed at the end of this manuscript). Also shown in Table 1, three values associated with each parameter are recorded, they are maximum, minimum and default values of those ecophysiological parameters. In our single factor parameter disturbance experiment, the value of each parameter in the new physiology file is recalculated as its default value plus 5% of parameter range (maximum value – minimum- value).

5. Characteristics of the Global Offline CLM Simulation

5.1 Strong scalability of the global offline CLM simulation

A one-year experiment is conducted in our study to investigate the strong scalability of CLM simulation under different parallel environment configuration. Five simulations have been conducted using 120, 516, 768, 1032 and 2040 CPU cores, respectively. Only the MPI communication mode has been enabled in our case. The data Atmosphere model, coupler and CLM use all the CPU cores, and a single CPU core has been assigned to stub components (Stub ocean, ice and glacier models). The strong scalability result of global offline CLM simulation is shown in Figure 2, which demonstrate a shorter simulation wall-time when more CPU cores are used. The one-year simulation using 768 CPU cores finishes in around 21 minutes, which is a good balance between simulation time and computational resource allocation. Therefore, the configuration of 768 cores is chosen for all the simulations in our following study.

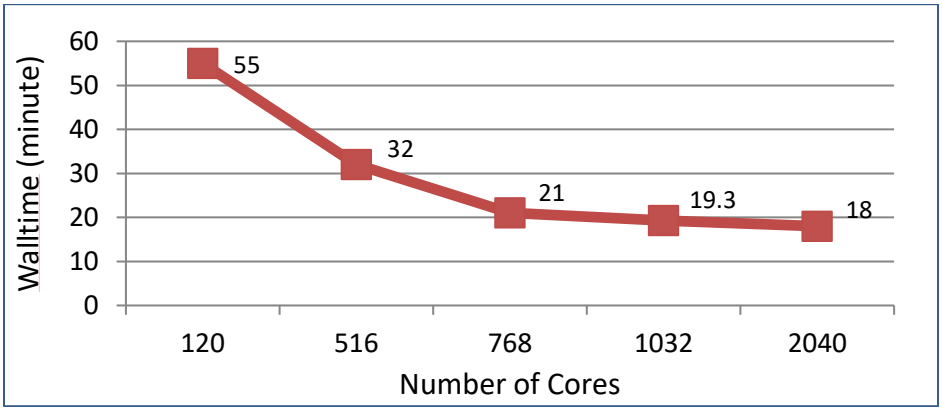

Figure 2: The strong scalability of the global offline CLM simulation

5.2 Computational characteristics

In order to get more detailed information on the global offline CLM simulation, the Vampir framework (Knüpfer et al., 2006, 2008) is used to gather information on simulation characteristics. The Vampir framework consists two components: VampirTrace and Vampir. VampirTrace is used to

instrument the source code of the program and to manage the recording of events while the instrumented program is running. Vampir is used to visualize the data/events gathered during the program execution. Within Vampir framework, the instrumentation could be done by methods via compiler flags or with TAU instrumentor (Shende et. al. 2006). A detailed explanation on how to configure the Vampir framework for the CLM global offline simulation can be found in another paper by Domke and Wang, 2012. The summarized simulation characteristics are shown in Figure 3.

The plot A shows the general pattern of the one-year simulation on the first 21 processes of the total 768 processes. The green color represents the execution time recorded as application, the red color represents MPI activities, and the yellow color represents I/O activities, while the blue color represents the information of Vampir, like the markers for message bursts or ears with a large number of I/O events. There is a clear synchronization pattern during the simulation, caused by the I/O and flux communication between CLM and the coupler every 30 minutes. Plot B shows the computational intensity of CLM using PAPI (icl.cs.utk.edu/projects/papi) on process #10. The peak flop/s rate is around 150 Mflop/s and mostly comes from the flux calculation related to the coupler. The average flop/s rate is around 90 Mflop/s. Plot C categorizes all processes into ten groups and presents the statistically characteristic of each group. The first row is the number of processes in each group, followed by a graph of blue strips representing the distributions of processes in each group over the complete set of processes. The last is a histogram chart to show the statistical information on computation, communication as well as others (I/O and overhead of timer synchronization within VampirTrace). Plot D is a pie chart to demonstrate the ratio between computation (application), communication (MPI), others (I/O, and VampirTrace overhead). The chart shows that the computation takes around 68% of the time, and around 30 % of the time is spent in communication. The blue color represents less than 1% (percentage data not shown), and it covers I/O activities (0.08%) and the overhead of VampirTrace (0.01% of the simulation time). Plot E represents the breakdown of the communication time, in which the two major MPI events are MPI_Barrier and MPI_Bcast. The plot F represents the IO activities, which shows that most of I/O time (56%) is spent on opening and closing files (both fopen and fclose). The read and write operations take about 48 seconds, i.e., around 6% of the time for all I/O operations. All the I/O activities in CLM are synchronized events, and the majority of I/O activities are executed on process #1. Therefore, I/O activity on a single process can trigger significant communication wait time across the whole parallel environment. The MPI event shown on all processes except process #1 (in the Plot A) during the time period of 700–750 seconds is an excellent sample for that behavior. This section only describes the overall computational characteristics of the CLM simulation using 768 CPU cores. More detailed information on CLM functional call graphs, message communication pattern, I/O statistics, as well as the time information on specific CLM functions, could be found in our previous Vampir paper (Domke and Wang, 2012) showing the usage of the Vampir framework in combination with TAU instrumentor over a short-period simulation (2 days) on a small parallel environment (48 CPU cores)

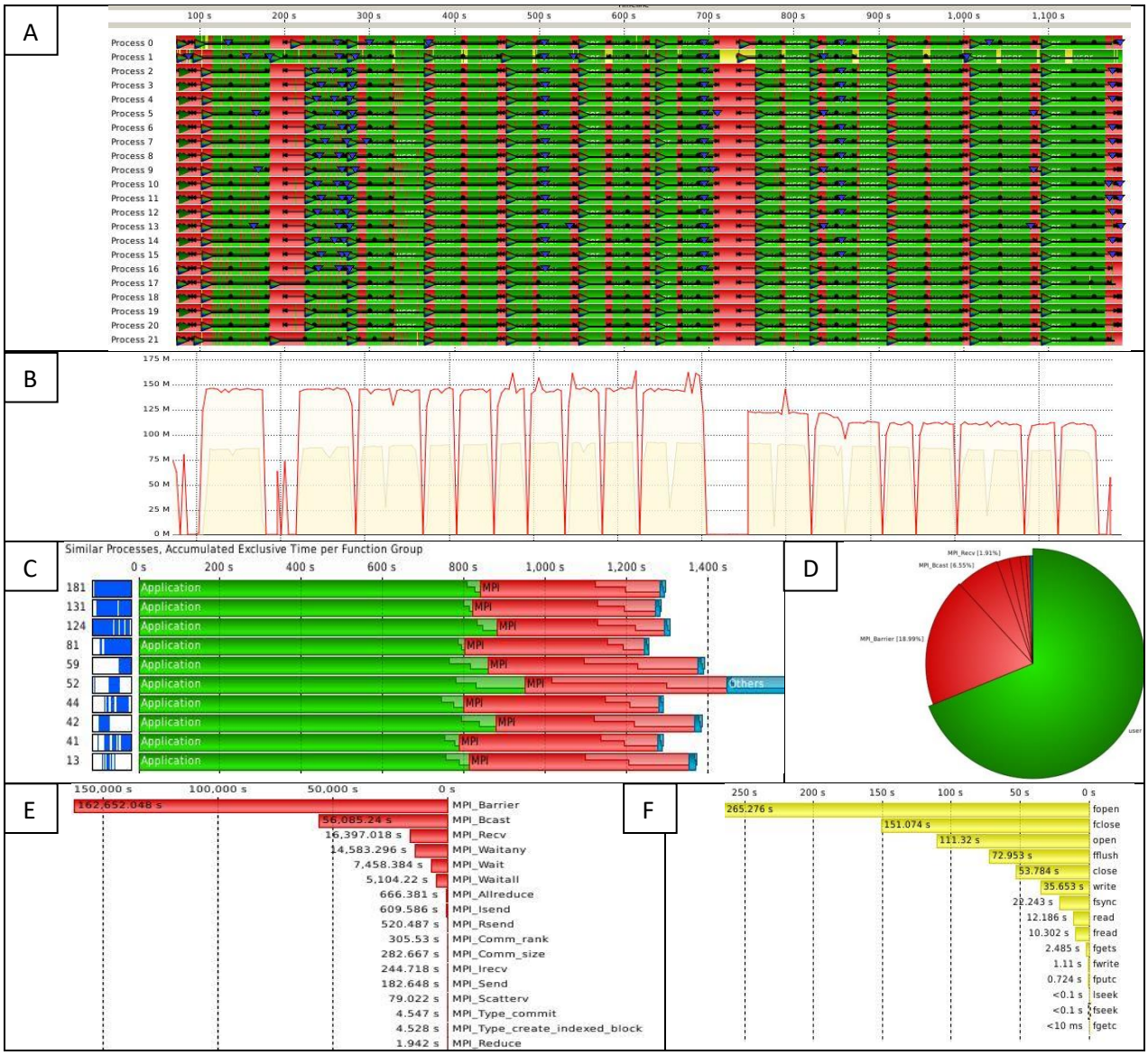

Figure 3: Computational characteristics of the global offline CLM simulation (A: general pattern of the one-year simulation on the first 21 processes; B: computational intensity of CLM using PAPI on one process (#10); C: statistical characteristics of all processes, grouped into ten groups; D: a pie chart demonstrating the ratio between computation, communication, others: E: communication time breakdown; F: IO activities)

6. Parallel Simulation for Single Factor Parameter Disturbance Experiment

6.1 Ensemble simulation preparation

All the members in our ensemble simulation share the same case configuration and same execution images, as developed in section 5. The workflow of the ensemble simulation preparation and parameter disturbance can be described as following:

1) *Create a simulation directory for each target parameter (see Table 1 at the end of manuscript)*
2) *Setup an independent simulation case within each simulation directory (copy the executable image from the default run directory (see section 5.1), copy a restart file and the default physiology file)*
3) *Increase the value of the target parameter in the physiology file by 5%.*
4) *Configure the runtime environment variables for each independent ensemble member run.*
5) *Create a new post-simulation processing script for the ensemble run (short-term and long-term archive, and individual post-simulation sanity check).*

6.2 Parallel execution

The ensemble simulation starts with a block request of computational resources through the default batch queue system. The number of CPU cores requested equals to the product of the number of ensemble runs and CPU core requested in each simulation (i.e., 768 CPU cores in our case). Then, a backend computing session is created for each ensemble member simulation (using 768 CPU cores) with independent environment variable settings, which is required to execute CLM offline simulation. The workload balance between the simulations of all ensemble members is achieved by the underlying OS, which guarantees that all the simulations do not share nodes if the number of available cores is large than the requested number of CPU cores. To prepare for the unexpected performance variations of the long-term ensemble simulation on a significant port of a high-end computer, each ensemble simulation was configured to generate restart files at end of each simulation year. At the end of ensemble simulation, a script is developed to check the simulation progress. If some individual simulations fails, or do not finish in time, the script will record those ensemble members and resubmit a continuous restart simulation for those ensemble members if necessary.

6.3 Weak scalability demonstration

As mentioned before, each ensemble member of CLM simulation is assigned with 768 CPU cores in our study. Weak scalability is used here to demonstrate how the execution time varies with the number of processors, given for a fixed number of CPU core (786) is used for each individual CLM simulation. Four tests have been conducted using 768 (one ensemble member), 7680 (10 members), 23040 (30 members) and 59136 CPU cores (77 members), respectively, to demonstrate the weak scalability of our simulation framework. The result is shown in Figure 4. It is exciting to see the excellent weak scalability of our simulation framework, that is, the execution time of 77 ensemble simulations using 59136 CPU cores is close to the execution time of 1 simulation using 768 CPU cores. This weak scalability is achieved via both our software dependency reductions, which eliminate necessary communications between ensemble member, and the IO operation capability from the Lustre file system, which is capable of handling large number IO operations simultaneously.

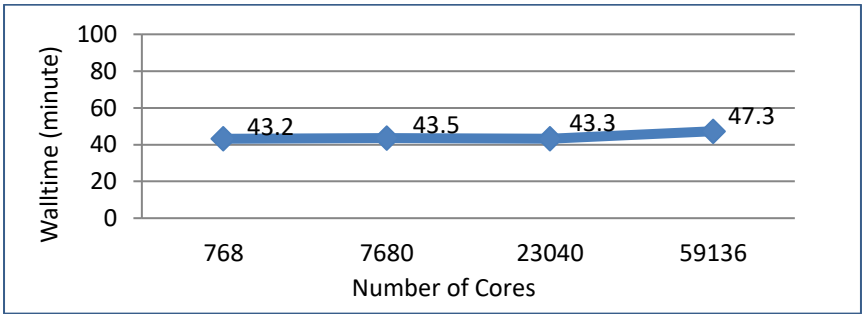

Figure 4: Weak scalability of the parallel simulation framework

7. Preliminary Simulation Results and Diagnostic Analysis

CESM land model working group has developed a CLM diagnostic analysis package (www.cgd.ucar.edu/tss/clm/diagnostics/). The diagnostics package consists of scripts that provide job control over plotting programs written in the NCAR Command Language (NCL), Perl and the NetCDF operators (NCO). The land model diagnostics package can be used to compare two model simulations or compare one model simulation to observational data. The diagnostics package produces postscript plots of long-term trends and seasonal means, in regional, global, and globally averaged formats from monthly NetCDF history files. However, due to the performance and the differences of the design purpose, the diagnostic package is not suitable for our ensemble sensitivity analysis. In this study, the total output is around 1.4 TB, it would take weeks to process those data using exist CLM diagnostic analysis package; therefore, an analysis toolkit has been developed to investigate the parameter disturbance impact on the model simulation results. Two types of plots are presented here to demonstrate our preliminary sensitivity analysis. They are range plot and spatial-explicit impact rank plot.

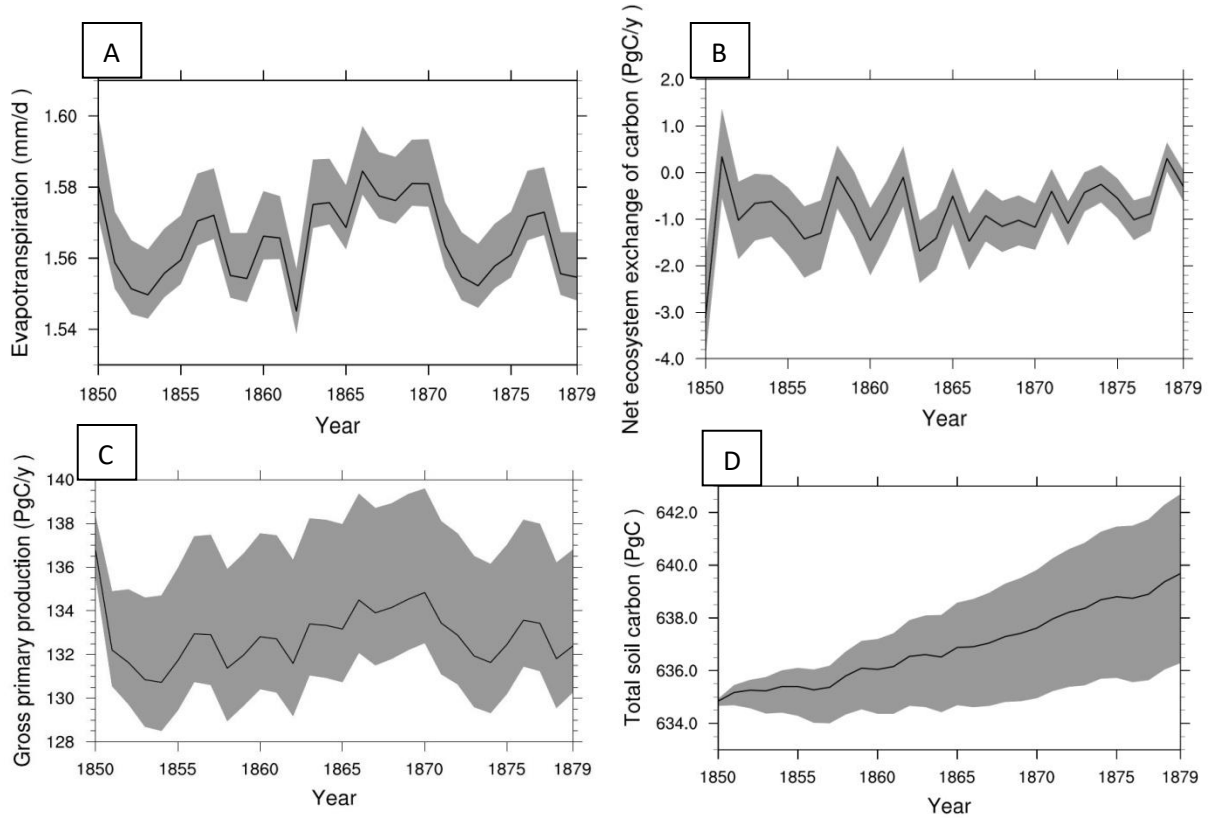

Figure 5: Uncertainty range plot of ET(A), NEE(B), GPP(C) and Total Soil Carbon(D)

8.1 Range plots of major water-carbon-flux states

The range plots give an overview of the range of uncertainty on specific variables, caused by the 5% parameter disturbance. At the end of each year, the maximum, minimum and mean value of each state are recorded, and then the range values of whole simulation (30 years) are plotted to present the trend line of the impact of parameter disturbance on a specific water-energy-flux state. Figure 5 shows the uncertainty range of Evapotranspiration (ET), Net Ecosystem Change of Carbon (NEE), Gross Primary

Production (GPP), and Total Soil Carbon (TSC), caused by the parameter disturbances. Plot A, B and C show that the uncertainty of ET, NEE, and GPP are bounded by the ecosystem processes through the simulation period. Plot C shows that the uncertainty of the NEE estimation decreases with the decadal simulation. And it is obvious in plot D that the uncertainty of the total soil carbon pool is accumulative.

8.2 Spatially-explicit rank plot of parameter impact on major water-carbon-flux states

In this case, the primary impact factor of each of the major water-carbon-flux states (such as the GPP, NEE, ET and Total Soil Carbon) on each grid cell is identified. The analysis results for the most significant factor of the model uncertainty related to the increase of GPP are presented in Figure 6.

Since GPP is considered, the bare ground areas (for example, desert and glacier area) have been masked out for the GPP plot. As shown in Figure 6, the disturbance of parameter #31 (fraction of allocation to currently displayed growth) is the most significant impact factor on the tropical areas (Amazon region) and boreal needleleaf forest (60 degree north latitude zone), see plot C. The parameter #50 (respiration fraction when SOM3 is converted to SOM4) has the most significant impact on increasing the GPP value on large areas (plot E). Similar plots can be generated for all the parameter disturbances which decrease the GPP value. This analysis can be extended to many other major carbon-water-flux states, and provides a new method to rank the top significant impact factors on the carbon-water-flux state in a spatially-explicit way, which in turn, provide potential baseline reference to site experiment establishment as well as regional field campaigns.

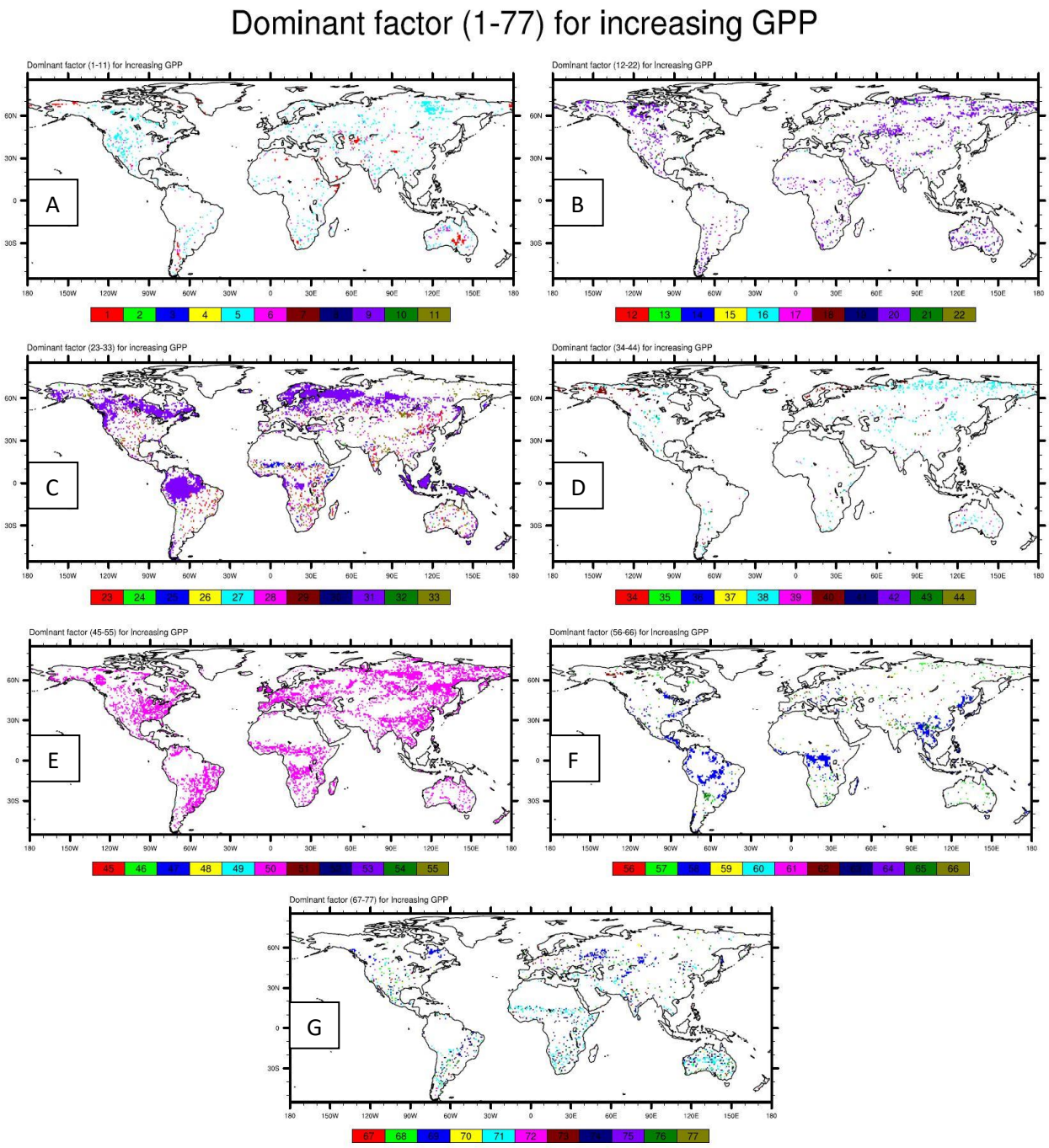

Figure 6: Spatially-explicit rank plot of parameter disturbance impact on the increase of GPP. The plots demonstrate the strong spatial patterns of the influences of ecophysiological factor on GPP.

8. Conclusion and discussions

This study presents a parallel framework for the global offline community land model ensemble simulation, which take the advantage of a high-end computer to explore new avenue to reduce the model uncertainty caused by the disturbances of model parameters. Based on the analysis of the software structure and the simulation runtime characteristics of the global offline CLM simulation, the

authors developed a simulation framework, which shows a very good weak scalability for large-scale ensemble simulations. Diagnostic analysis functionality is described to demonstrate the potentials enabled by the single-factor parameter disturbance experiments. From the software design perspective, the methods to reduce the simulation dependency, the investigation of simulation characteristics, the approach of parallel ensemble execution, as well as the resubmission strategies for the automatic simulation, are applicable to other ensemble simulation of legacy applications on high performance computing systems.

Future work could focus on further sensitivity analysis of CLM beyond the single factor parameter disturbance experiments and the data analysis for large-scale spatial-temporal simulation results. This study creates opportunities for a sensitivity analysis of other CESM model components (such as the atmosphere, ocean, sea ice and glacier model), as well as a fully-coupled CESM simulation on a high-end computer. The range plot and spatially-explicit rank plot, presented in this study, are just two examples of spatial and temporal multivariate analysis, which can be further developed on high performance computers to generate a comprehensive knowledge database to reduce the earth system model uncertainty, as well as to provide the baseline information for regional observation campaign and site experiment establishment.

Acknowledgement

This research was funded by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research (BER). This research used resources of the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725. Oak Ridge National Laboratory is managed by UT-Battelle LLC for the Department of Energy under contract DE-AC05-00OR22725.

References

Bonan, G., and S. Levis (2010), Quantifying carbon-nitrogen feedbacks in the Community Land Model (CLM4), *Geophysical Research Letters*, *37*(7), L07401.

Domke, J., D. Wang,(2012), Runtime Tracing of the CESM: Feasibility Study and Benefits, in press, International Conference on Computational Sciences, 2012.

Hurtt, G., S. Frolking, M. Fearon, B. Moore, E. Shevliakova, S. Malyshev, S. Pacala, and R. Houghton (2006), The underpinnings of land use history: three centuries of global gridded land use transitions, wood harvest activity, and resulting secondary lands, *Global Change Biology*, *12*(7), 1208-1229.

Knüpfer, A., R. Brendel, H. Brunst, H. Mix, W. E. Nagel, (2006), Introducing the Open Trace Format (OTF), in: V. N. Alexandrov, G. D. van Albada, P. M. A. Sloot, J. Dongarra (Eds.), International Conference on Computational Science (2), Vol. 3992 of Lecture Notes in Computer Science, Springer, 2006, pp. 526–533

Knüpfer, A., H. Brunst, J. Doleschal, M. Jurenz, M. Lieber, H. Mickler, M. S. Müller, W. E. Nagel, (2008), The Vampir Performance Analysis Tool-Set, in: M. Resch, R. Keller, V. Himmler, B. Krammer, A. Schulz (Eds.), Tools for High Performance Computing, Springer Berlin Heidelberg, 2008, pp. 139–155

Lawrence, D.M., P.E. Thornton, K.W. Oleson, and G.B. Bonan (2007), The partitioning of evapotranspiration into transpiration, soil evaporation, and canopy evaporation in a GCM: Impacts on land-atmosphere interaction.. *J. Hydromet.*, **8**, 862-880.

Mao, J.F., X.Y. Shi, P.E. Thornton, S.L. Piao, and X.H. Wang (2012), Causes of spring vegetation growth trends in the northern mid-high latitudes from 1982 to 2004. *Environmental Research Letters*, 7, doi:10.1088/1748-9326/7/1/014010.

Mitchell, T. D., and P. D. Jones (2005), An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, **25,** 693–712.

Oleson, K., D. Lawrence, B. Gordon, M. Flanner, E. Kluzek, J. Peter, S. Levis, S. Swenson, E. Thornton, and J. Feddema (2010), Technical description of version 4.0 of the Community Land Model (CLM).

Shende, S.S., A. D. Malony, (2006), The Tau Parallel Performance System, The International Journal of High Performance Computing Applications 20 (2006) 287–331.

Shi, X., J. Mao, P. E. Thornton, F. M. Hoffman, and W. M. Post (2011), The impact of climate, CO2, nitrogen deposition and land use change on simulated contemporary global river f low, *Geophysical Research Letters* , 38, L08704, doi:10.1029/2011GL046773.

Stöckli, R., D. Lawrence, G. Niu, K. Oleson, P. Thornton, Z. Yang, G. Bonan, A. Denning, and S. Running (2008), Use of FLUXNET in the Community Land Model development, *J. Geophys. Res*, *113*, G01025.

Thornton, P., and N. Zimmermann (2007), An improved canopy integration scheme for a land surface model with prognostic canopy structure, *Journal of Climate*, *20*(15), 3902-3923.

Thornton, P., S. C. Doney, K. Lindsay, J. K. Moore, N. Mahowald, J. T. Randerson, I. Fung, J. F. Lamarque, J. J. Feddema, and Y. H. Lee (2009), Carbon-nitrogen interactions regulate climate-carbon cycle feedbacks: results from an atmosphere-ocean general circulation model, Biogeosciences, 6(10), 2099-2120.

Table 1: List of parameter names, values, units and descriptions

#	Min	Max	Default	Units	Descriptions
1	1.00E-02	1.00E-01	5.50E-02	m	roughness length:canopy top
2	1.00E-01	1.00E+00	6.70E-01	m	displacement length: canopy top
3	1.00E-02	1.00E-01	4.00E-02	m	characteristic leaf dimension
4	1.00E+00	2.00E+02	5.10E+01	umolC/m2/s	Vcmax at 25C
5	3.00E+00	2.00E+01	6.00E+00	none	slope of conductance to photosynthesis
6	1.00E-02	5.00E-01	6.00E-02	umol C/ umol phot	Quantum efficiency
7	1.00E-02	1.00E+00	7.00E-02	none	leaf reflectance (vis)
8	1.00E-02	1.00E+00	3.50E-01	none	leaf reflectance (nir)
9	1.00E-02	1.00E+00	1.60E-01	none	stem relectance (vis)
10	1.00E-02	1.00E+00	3.90E-01	none	stem reflectance(nir)
11	1.00E-02	1.00E+00	5.00E-02	none	leaf transmittance (vis)
12	1.00E-02	1.00E+00	1.00E-01	none	leaf transmittance (nir)
13	1.00E-04	1.00E-02	1.00E-03	none	stem transmittance (vis)
14	1.00E-04	1.00E-02	1.00E-03	none	stem transmittance (nir)
15	1.00E-02	1.00E+00	1.00E-02	none	leaf/stem orientation index
16	1.00E+00	2.00E+01	7.00E+00	m-1	rooting distribution parameter
17	5.00E-01	1.00E+01	2.00E+00	m-1	rooting distribution parameter
18	1.00E-02	1.00E-01	1.00E-02	m2/gC	SLA at top of canopy
19	0.00E+00	1.00E-02	1.20E-03	m2/gC/LAI	SLA/dLAI
20	5.00E+00	4.00E+01	3.50E+01	gC/gN	leaf C to N ratio #max was 70
21	2.00E-02	1.50E-01	5.00E-02	none	frac of leaf N in Rubisco #was 0.25
22	-2.00E+05	-5.00E+03	-6.60E+04	mm	soil water pot. at full opening
23	-8.00E+05	-5.00E+03	-2.55E+05	mm	soil water pot. at closure
24	5.00E+00	2.00E+02	7.00E+01	gC/gN	leaf litter C:N
25	2.00E+01	1.00E+02	4.20E+01	gC/gN	fine root C:N
26	5.00E+00	5.00E+02	5.00E+01	gC/gN	live wood C:N
27	5.00E+01	5.00E+03	5.00E+02	gC/gN	dead wood C:N
28	0.00E+00	2.00E+00	1.00E+00	gC/gC	new fine root alloc C /leaf C #was 4.0
29	5.00E-02	1.00E+00	3.00E-01	gC/gC	new croot alloc C per stem C
30	0.00E+00	1.00E+00	1.00E-01	none	fraction of new wood that is live
31	0.00E+00	1.00E+00	1.00E+00	none	frac of alloc. currently disp grwoth
32	2.00E-01	1.00E+01	1.00E+00	yr	leaf longevity
33	1.00E-02	9.00E-01	1.20E-01	none	fire resistance index
34	1.00E-01	5.00E-01	3.00E-01	none	growth respiration factor 1
35	1.00E+00	1.00E+00	1.00E+00	none	growth respiration factor 2
36	0.00E+00	6.00E-01	2.50E-01	(1/s)	bulk denitrification rate (CNAllocationMod)
37	1.00E+00	9.00E+02	3.00E+02	days	days to recover negative cpool
38	5.00E-01	5.00E+01	1.00E+02	none	resistance for uptake from plant npool
39	4.00E-07	1.00E-05	2.53E-06	gC/gN/s	base rate for maintenance respiration
40	1.00E+00	4.50E+00	1.50E+00	none	q10 for maintenance respiration

41	3.00E+00	3.00E+01	1.20E+01	gC/gN	carbon:nitrogen for SOM 1
42	3.00E+00	3.00E+01	1.20E+01	gC/gN	carbon:nitrogen for SOM 2
43	3.00E+00	3.00E+01	1.00E+01	gC/gN	carbon:nitrogen for SOM 3
44	3.00E+00	3.00E+01	1.00E+01	gC/gN	carbon:nitrogen for SOM 4
45	5.00E-02	9.00E-01	3.90E-01	none	resp. fraction for litter 1 ->SOM1
46	5.00E-02	9.00E-01	5.50E-01	none	resp. fraction for litter 2 ->SOM2
47	5.00E-02	9.00E-01	2.90E-01	none	resp. fraction for litter 3 ->SOM3
48	5.00E-02	9.00E-01	2.80E-01	none	resp. fraction for SOM 1 -> SOM2
49	5.00E-02	9.00E-01	4.60E-01	none	resp. fraction for SOM 2 -> SOM3
50	5.00E-02	9.00E-01	5.50E-01	none	resp. fraction for SOM 3 -> SOM4
51	5.00E-02	9.00E-01	1.20E+00	1/day	decomp rate for litter 1
52	5.00E-02	9.00E-01	7.26E-02	1/day	decomp rate for litter 2
53	5.00E-02	9.00E-01	1.41E-02	1/day	decomp rate for litter 3
54	5.00E-02	9.00E-01	7.26E-02	1/day	decomp rate for SOM 1
55	5.00E-02	9.00E-01	1.41E-02	1/day	decomp rate for SOM 2
56	5.00E-02	9.00E-01	1.40E-03	1/day	decomp rate for SOM 3
57	5.00E-02	9.00E-01	1.00E-04	1/day	decomp rate for SOM 4
58	5.00E-02	9.00E-01	1.00E-03	1/day	fragmentation rate for CWD
59	1.00E-02	9.90E-01	7.69E-01	none	fraction of cellulose in CWD
60	1.00E-03	1.00E-01	1.00E-02	none	denitrification proportion
61	-1.50E+02	-2.00E+01	-1.00E+02	MPa	minimum psi for heterotrophic resp
62	1.00E+00	4.50E+00	1.50E+00	none	q10 for heterotrophic respiration
63	2.00E-03	2.00E-01	2.00E-02	1/year	mortality rate
64	1.00E-02	5.00E-01	1.00E-01	none	solulble fraction of mineral N
65	2.00E+04	5.00E+04	3.93E+04	seconds	critical daylength for senescence onset
66	5.00E+00	6.00E+02	3.00E+02	days	no. of days to complete leaf onset
67	5.00E+00	4.00E+02	1.50E+02	days	no. of days to complete leaf offset
68	1.00E-01	8.00E-01	5.00E-01	none	fraction of strage to move to transer
69	2.00E+00	4.00E+02	1.50E+02	days	no. of freezing days to set GDD counter
70	2.00E+00	4.00E+02	1.50E+02	days	no. of water stress-free days for leaf onset
71	-5.00E+00	-7.50E-01	-2.00E+00	MPa	critical soil water potential for leaf onset
72	2.00E+00	4.00E+02	1.50E+02	days	no. of freezing days for leaf offset
73	2.00E+00	4.00E+02	1.50E+02	days	no. of water stress days for leaf offset
74	-5.00E+00	-7.50E-01	-2.00E+00	MPa	critical soil water potential for leaf offset
75	2.00E-01	9.00E-01	7.00E-01	1/year	live wood turnover proportion
76	3.00E+00	7.00E+00	4.80E+00	none	gdd threshold parameter 1
77	5.00E-02	3.00E-01	1.30E-01	none	gdd threshold parameter 2