Improvements in the Scalability of the NASA Goddard Multiscale Multicomponent Modeling Framework for Hurricane Climate Studies

Bo-Wen Shen[1], Wei-Kuo Tao[2], and Jiun-Dar Chern[3]
[1]UMCP/ESSIC, [2]NASA/GSFC, [3]UMBC/GEST
bo-wen.shen.1@gsfc.nasa.gov, wei-kuo.tao.1@gsfc.nasa.gov, jiun-dar.chern.1@gsfc.nasa.gov

1. Introduction:

The 2004 and 2005 Atlantic hurricane seasons were the most active in recorded history, but the 2006 season was not as active as predicted. Therefore, a challenging research topic is how to improve our understanding of hurricane inter-annual variability and the impact of climate change on hurricanes. To address it with numerical models, we need to improve hurricane simulations on a global scale. Paired with the substantial computing power of the NASA Columbia supercomputer, the newly-developed multi-scale modeling framework (MMF, Tao et al. 2007) at the NASA Goddard Space Flight Center shows potential for the related studies. The Goddard MMF consists of two NASA state-of-the-art modeling components: the finite-volume General Circulation Model (fvGCM, Lin et al. 2004) and the 2-D version of the Goddard Cloud Ensemble Model (GCE, Tao and Simpson 1993; Tao et al. 1993). While the fvGCM has shown remarkable capabilities in simulating large-scale flows and thus hurricane tracks (Atlas et al. 2005; Shen et al. 2006a,b), the GCE is well known for its superior performance in representing small cloud-scale motions and has been used to produce more than 90 referred journal papers (e.g., Lang et al. 2003; Tao et al. 2003). Preliminary results with the MMF are encouraging, showing a positive impact on simulations of large-scale flows via the feedback of resolved convection by the GCEs. Among them is the improved simulation of the Madden-Julian Oscillation, which could potentially improve long-term forecasts of tropical cyclones. Since a higher resolution (e.g., 1 degree) fvGCM and 3-D GCE are desired in the MMF for hurricane (long-term) climate studies, computational issues in the Goddard MMF (e.g., limited scalability) need to be addressed.

2. The Goddard MMF:

The Goddard MMF consists of the fvGCM at a 2^ox2.5^o resolution and 13,104 2D GCEs, each of which is “embedded” within one grid point of the fvGCM. Currently, only thermodynamic feedback between the fvGCM and the GCEs is implemented. While the time step for the individual GCE is ten seconds, the fvGCM-GCEs coupling interval is one hour at this resolution. Under this configuration, 95% or more of the total wall-time for running the MMF is spent on the GCEs. Thus, wall-time could be significantly reduced by efficiently distributing the large number of GCEs over a massive number of processors on a supercomputer.

During the past several years, an SPMD (single program multiple data) parallelism has been implemented in both the fvGCM and GCE with good parallel efficiency separately (Putman et al. 2005; Juang et al. 2007). While the fvGCM has a hybrid MPI-OpenMP parallelism, the GCE has a 2D domain decomposition using MPI-1. Since it would require a tremendous effort to implement an OpenMP parallelism into the GCE or extend the 1D domain decomposition to 2D in the fvGCM, the MMF only inherited the fvGCM’s 1D MPI parallelism. This limited the MMF’s scalability, and thereby posed a challenge for increasing the fvGCM’s resolution and/or extending the GCE’s dimensions from 2D to 3D.

3. A Revised Parallelism Implementation:

To overcome the aforementioned limitation, we propose a different strategic approach to coupling the GCEs to the fvGCM. From a computational perspective, we should completely forget about the concept of “embedded GCEs”, which restricts our view on the data parallelism of the fvGCM. Instead, we could view the 13,104 GCEs as a *meta global GCE* (mgGCE) in a *meta gridpoint system,* which includes 13,104 grid points (Figure 1). This grid system, which is not limited to any specific grid system, is assumed to be the same as the latitude-longitude grid structure in the fvGCM for convenience. With this concept in mind, either the fvGCM or mgGCE can have its own scaling properties. Thus, we could substantially reduce the execution wall-time by deploying a highly scalable mgGCE, and/or coupling the mgGCE with the fvGCM using an MPMD (multiple programs multiple data) parallelism.

Data parallelism in the mgGCE is indeed a task parallelism, which distributes 13,104 GCEs among processors. As a cyclic lateral boundary condition is used in each GCE, the mgGCE has no ghost region in the meta grid system, so the mgGCEs with a 2D domain decomposition can be scaled “embarrassingly”. The major overhead in the MMF occurs in data redistribution (or regridding) between the fvGCM and the mgGCE. Under this new concept, the grid inside each GCE becomes a *child grid* (or sub-grid) with respect to the parent (meta) grid. Since an individual GCE can still be executed with its native 2D-MPI implementation in the child grid-point space, this second level of parallelism can greatly expand the number of CPUs. Potentially, the final mgGCE and the coupled MMF as well could be scaled at a multiple of 13,104 CPUs. Another advantage of the mgGCEs is to permit the adoption of the idea of land-sea masks in a land model. For limited computing resources, we can create a cloud-mask file to specify limited regions where GCEs will be running, thereby possibly balancing computational loads. A sophisticated mgGCE implementation with the cloud-mask file will enable one to choose a variety of GCEs (2D vs. 3D, bulk vs. bin microphysics) depending on geographic location.

Currently, a prototype MMF with the mgGCE idea has been successfully implemented. The technical approaches are briefly summarized at follows: (1) a master process allocates a shared memory arena for data redistribution between the fvGCM and the mgGCE by calling the Unix *mmap* function; (2) the master process spawns multiple (parent) processes with a 1-D domain decomposition in the y direction by a series of Unix *fork* system calls; (3) each of these parent processes then forks several child processes with another 1-D domain decomposition along the x direction; (4) data gathering in the mgGCE is done via data communication along the x and then y directions; (5) synchronization is implemented with the atomic *__sync_add_and_fetch* function call on the Columbia supercomputer. While steps (1), (2), and (5) were previously used in single-component models by Taft (2001), we extend this methodology to our multicomponent modeling system. Figure 2 shows very promising scalability up to 364 CPUs, giving a super-linear speedup as measured by the production run with 30 CPUs.

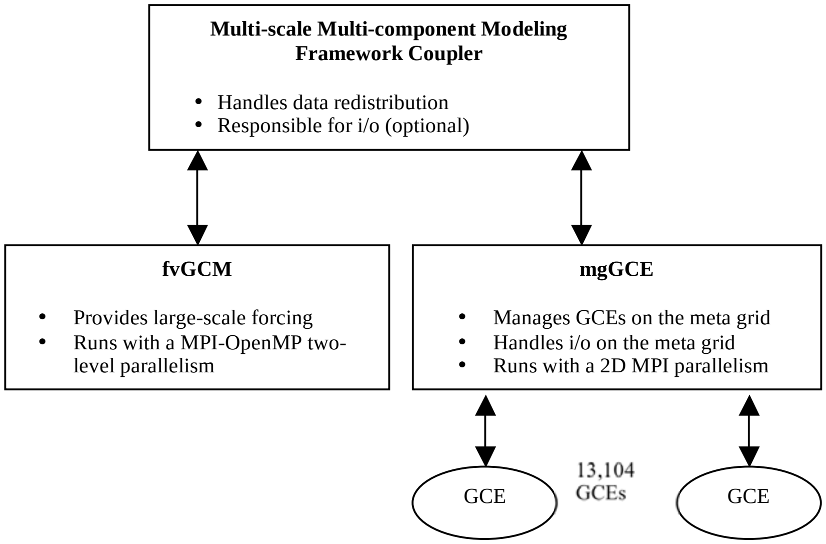
4. Concluding Remarks:

Improving our understanding of hurricane inter-annual variability and the impact of climate change (e.g., doubling CO2 and/or global warming) on hurricanes brings both scientific and computational challenges to researchers. The newly-developed MMF (Tao et al. 2007) and the substantial computing power of the NASA Columbia supercomputer show promise in studying this topic. In this article, we discuss the computational issues in hurricane climate studies with the MMF, and propose a revised methodology to improve the MMF's performance and scalability. A prototype of a revised MMF, which allows data redistribution between the fvGCM grid space and the mgGCE meta grid space, is being implemented with remarkable scalability. This proof-of-concept approach encourages us to implement a more sophisticated modeling coupler to solve complex problems with advanced computing power.

As the meta grid system in the mgGCE is no longer bound to the fvGCM's grid system, we could avoid the performance issues of a latitude-longitude grid system by implementing a quasi-uniform grid system (such as a cube grid or geodesic grid) in the mgGCE. Finally, as the MMF's major computing is done in the mgGCE, which has no ghost points, we envision the next version of the MMF with the mgGCE to be a good candidate for the meta- (grid-) computing like the SETI@home project. Namely computations in the mgGCE could be distributed among available (personal) computers connected by the Internet.

References:

Atlas et al. 2005, *GRL*, **32**, L03801.
Juang et al. 2007, *TAO*, in press.
Lang et al. 2003, *J. Appl. Meteor.* **42**, 505-527.
Lin et al 2004, *CISE*, **6(1)**, 29-35.
Putman et al. 2005, *IJHPCA*, **19**, 213-223.
Shen et al 2006a: *GRL*, **33**, L05801.
Shen et al 2006b: *GRL*, **33**, L13813.
Taft, J. R. 2001, *P.C.*, 27 (4), 521-536
Tao and Simpson, 1993, *TAO*, **4**, 19-54.
Tao et al. 1993, *JAS*, **50**, 673-690.
Tao et al. 2003, *JAS*, **60**, 2929-2956.
Tao et al. 2007, *JGR*, submitted.


Multi-scale Multi-component Modeling Framework Coupler
• Handles data redistribution
• Responsible for i/o (optional)
fvGCM
• Provides large-scale forcing
• Runs with a MPI-OpenMP two-level parallelism
mgGCE
• Manages GCEs on the meta grid
• Handles i/o on the meta grid
• Runs with a 2D MPI parallelism
GCE
13,104 GCEs
GCE

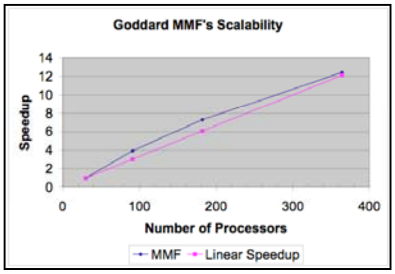


Figure 1 (top): Schematic diagram of the meta-global GCE and the improved MMF coupler.
Figure 2 (right): Scalability of the Goddard MMF.