C^3F : Collaborative Container-based Model Coupling Framework

Jungha Woo
Lan Zhao
wooj@purdue.edu
lanzhao@purdue.edu
Research Computing at Purdue University
West Lafayette, Indiana, USA

Danielle S. Grogan
danielle.grogan@unh.edu
Institute for the Study of Earth, Oceans, and Space, University of New Hampshire
Durham, New Hampshire, USA

Iman Haqiqi
ihaqiqi@purdue.edu
Department of Agricultural Economics, Purdue University
West Lafayette, Indiana, USA

Richard Lammers
Richard.Lammers@unh.edu
Institute for the Study of Earth, Oceans, and Space, University of New Hampshire
Durham, New Hampshire, USA

Carol X. Song
carolxsong@purdue.edu
Research Computing at Purdue University
West Lafayette, Indiana, USA

ABSTRACT

Solving complex real-world grand challenge problems requires in-depth collaboration of researchers from multiple disciplines. Such collaboration often involves harnessing multiscale and multi-dimensional data and combining models from different fields to simulate systems. However, the progress on this front has been limited mainly due to significant gaps in domain knowledge and tools that are typically employed in silos of the domains. Researchers from different fields face considerable barriers to understanding and reusing each other's data/models in order to collaborate effectively. For example, in solving the global sustainability problems, researchers from hydrology, climate science, agriculture, and economics need to run their respective models to study different components of the global and local food, energy and water systems while, at the same time, need to interact with other researchers and integrate the results of one model with another. Developing this kind of model coupling workflow calls for (1) a large amount of data being processed and exchanged across domains and organizations, (2) identifying and processing the output of one model to make it ready for integration into another model, (3) controlling the workflow dynamically so that it runs until a certain convergence condition or other criteria is met, and (4) close collaboration among the modelers to explore, tune, and test the configuration and data transformation needed to link the models. We have developed C^3F, a flexible collaborative model coupling framework to help researchers accelerate their model integration and linking efforts by leveraging advanced cyberinfrastructure such as high-performance computing and virtual containers. In this paper, we describe our experience and lessons learned in developing this cyberinfrastructure solution to support the linking of Water Balance Model (WBM) and SIMPLE-G agricultural economic model in an NSF funded INFEWS project and a DOE-funded Program on Coupled Human and Earth Systems (PCHES) to study the implications of groundwater scarcity for food-energy-water systems. The C3F model coupling framework can be extended to facilitate other model linkages as well.

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3530298

CCS CONCEPTS

• **Computing methodologies** → **Simulation environments**.

KEYWORDS

Model coupling workflow, containerization, WBM, SIMPLE-G, Food-Energy-Water (FEW)

ACM Reference Format:
Jungha Woo, Lan Zhao, Danielle S. Grogan, Iman Haqiqi, Richard Lammers, and Carol X. Song. 2022. C^3F : Collaborative Container-based Model Coupling Framework. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491418.3530298

1 INTRODUCTION

The global Food-Energy-Water (FEW) system will face significant challenges in the next forty years driven by the growing population, rising per capita incomes, and climate impacts. These challenges are interconnected—both across systems and across scales—so that addressing one system or location will inevitably cascade into others, driven by socio-ecological feedback [2, 12]. Decision-makers without the capacity to factor in these interconnections risk inadvertently pursuing unsustainable FEW system solutions. Liu et al. [14] highlighted a lack of multi-scale, integrated systems approach in FEW science to address the role of spillover effects from one system to another, and argued for greater integrative analysis of human and natural systems across spatial scales. In both our NSF-funded INFEWS (Innovations at the Nexus of Food, Energy, and Water Systems) and DOE-funded PCHES (Program on Coupled Human and Earth Systems) projects, we aim to address this knowledge gap by applying an integrative framework for analysis of FEW solutions that highlights synergies and tradeoffs resulting

from multiple policy options and thereby allowing the development of more comprehensive sustainability solutions. In this paper, we present one of our FEW model linking frameworks designed to answer the research question: how do agricultural land and water use patterns respond to changing groundwater levels over time across the U.S.? Answering this question requires a hydrologic model (WBM) that can simulate groundwater systems and their response to agricultural activity, and an economic model (SIMPLE-G) [1] that can represent human decisions about land use and water use in response to information about groundwater depth, yield response to weather, regional market dynamics, and other changing economic factors.

Multiple significant challenges exist in model linking within the same domain and across multiple domains. First, from the technical perspective, existing models are very diverse, written in different programming languages and requiring different operating systems and different input/output file formats. Some models involve a large volume of input/output data, while others use proprietary data or software licenses. Second, from a conceptual perspective, models could come from fundamentally different paradigms (e.g., simulation models vs computable general/partial equilibrium models) and differences in terminology and vocabulary between disciplines can lead to different definitions of model variables or concepts. Third, from a collaborative perspective, researchers from different disciplines need to take the time to understand each other's models to overcome the conceptual challenges. They often need access to each other's computing systems, which can have institutional hurdles. All of this takes time and patience. Finally, as an intersection of all three (technical, conceptual, and collaborative), the development of a technical workflow requires researchers to know how to couple the models. Yet developing the coupling method may first require the researchers to test out different coupling options. Results of the model coupling tests often show researchers where they have misunderstood one another's models or disciplinary concepts, leading them to revise their workflow. This leads to a circular challenge: the technical implementation requires the conceptual and collaborative challenges to be solved yet solving those challenges may rely on already having solved the technical issues.

WBM researchers have engaged in numerous model-linking projects over the past decade [4, 5, 11, 13, 16, 17, 20]. Despite many of these exercises beginning with the goal of producing 2-way, fully automated coupled-models, all but one [16] stopped at the phase of one-way, offline coupling due to the challenges discussed above. The work presented here builds on such prior experiences [5, 13].

In this effort, we developed a flexible Collaborative Container-based model Coupling Framework (C^3F) to help the FEW researchers accelerate their model integration and linking efforts. It addresses some of the interconnected technical, conceptual, and collaborative challenges by leveraging advanced technologies and cyberinfrastructure (CI) such as high-performance computing (HPC), containers, Open OnDemand (OOD), and XSEDE. Using this framework WBM and SIMPLE-G researchers can independently package their models and data processing code into Singularity containers and collaboratively explore, create, and execute the coupled modeling workflows in the XSEDE HPC environment. In this paper, we will first introduce the WBM and SIMPLE-G models and the model coupling workflow in section 2. We will then describe the two phases of our model integration and coupling CI solutions and present preliminary results in sections 3 and 4. We will discuss our experience and lessons learned in section 5 and then conclude the paper in section 6.

2 WBM-SIMPLE-G MODELING WORKFLOW

Our FEW projects aim to link several internationally vetted open-source models from different scientific disciplines including hydrology, climate, and economics. As an initial effort, we focused on developing a CI solution to improve the productivity of the WBM and SIMPLE-G researchers in linking their models to understand how economically driven changes in agricultural production may impact on sustainable water use.

2.1 WBM Model

The University of New Hampshire Water Balance Model (WBM) [3, 19] is a process-based, gridded hydrologic model that simulates spatially and temporally varying water volumes and quality. It was one of the first Global Hydrologic Models (GHMs) developed [18]. WBM represents all major land surface components of the hydrologic cycle and tracks fluxes and balances between the atmosphere, above-ground water storage, soil, vegetation, groundwater, and runoff. A digitized river network connects each grid cell to the next, enabling the simulation of flow through river systems. The model's representation of these natural processes is based on well-established principles from the fields of physics and hydrology; the daily simulation of natural water fluxes achieves high fidelity when compared to historical observational data.

WBM also includes modules to represent human interactions with the water cycle, such as estimates of domestic, industrial, and agricultural water requirements and use, as well as hydro-infrastructure (dams, canals, and inter-basin transfers). These human processes are not governed by physics, but rather are functions of complex decision-making processes. WBM, like all GHMs that attempt to include these human water uses, relies on input data about key human activities. While this data may be available for historical periods, simulating future trajectories requires the expertise of other fields like economics or policy analysis to devise reasonable scenarios or models of human activity relevant to water use.

WBM is written in Perl with heavy use of Perl Data Language. It relies on a custom module called RIMS and has several open-source software dependencies including GDAL. The model runs in Unix and can be launched by a command or a submission script. The data involved in WBM is large and complex. Most of its input files are geospatial data (e.g., GeoTIFF, NetCDF, or shapefiles), and its output files are multi-dimensional NetCDF files with a time dimension (daily, monthly, or yearly), latitude and longitude spatial dimensions, and also multiple (ranging from 1 to over 500) individual output variables. WBM simulates the processes in daily time steps determined by input data and the model algorithms. Unlike the types of models typically employed in economics, the model is not solving an optimization or linear programming problem.

2.2 SIMPLE-G Model

SIMPLE-G is one of the first global gridded economic models that looks into economic decisions and market forces at a higher resolution [1]. It is the gridded version of SIMPLE (Simplified International Model of agricultural Prices, Land use and the Environment) developed by the Purdue University GLASS team (Global to Local Analysis of System Sustainability). It is a validated multi-scale economic model that is designed to better understand the local economic decisions about land use, water use, and agricultural production while taking into account the competing forces of the food system at the global, regional, and subregional levels. The simple yet powerful economic structure connects agricultural economic systems with biophysical systems at a gridded and high-resolution scale. The model framework includes economic supply and demand modeling of the food system. Crop production is a function of land, water, fertilizer, climate-driven productivity, and technology.

SIMPLE-G modeling solves a system of equations in GEMPACK (General Equilibrium Modelling PACKage) [6] as a suite of economic modeling software. GEMPACK is especially suitable for large computable equilibrium models. SIMPLE-G can involve half a billion endogenous variables depending on the resolution and the spatial scope of the study. Here, GEMPACK provides condensation techniques to speed up the calculations for this large system. However, it runs on a Windows-based system and requires a license. A set of files are required to be compiled to create an EXE file. The input files are in HAR format. The standard outputs are also in HAR or SL4 formats readable to GEMPACK and need to be converted to TXT to communicate with other models.

2.3 Linked WBM-SIMPLE-G Models

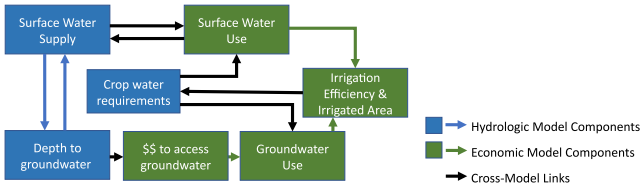

Figure 1: Model components and links used by the coupled model framework

The research questions posed by our FEW systems projects require information from both hydrology and economics. Each of the two models described above represents part of the FEW system, but neither includes all the dynamics of interest. As shown in Figure 1, a more complete system representation can be achieved by linking the two models.

Developing such model linking requires close collaboration between the modelers of SIMPLE-G and WBM and computer scientists. Before the final workflow may be developed, there is a significant amount of time spent in an exploratory step where the modelers and CI experts from the two institutions need to achieve a common understanding of details such as how their models run, what input/output is needed and created, how to convert the output from one model to the right input format for the other model, etc. Key challenges include harmonizing the time steps represented by the two models, developing variable consistency between model concepts, and generating linking functions that convert data to the right format. For example, WBM runs at daily time steps, while SIMPLE-G runs at 5 to 10-year steps. Conceptually, the two models do not have the same theory of efficiency; hydrologic irrigation efficiency refers to physical water systems, while economic irrigation efficiency is based on balancing supply and demand for water. Lastly, the conversion of model outputs to model inputs requires not only time-step and file format conversions, but more importantly data-based transfers of information of one type to another. For example, a change in depth to groundwater as simulated by WBM must be converted to a change in the cost to access groundwater to become an input to SIMPLE-G. Similarly, a change in water use per acre in SIMPLE-G must be converted to a change in physical irrigation efficiency to be used as an input to WBM. These data conversions are effectively a second set of two models; simpler models than WBM and SIMPLE-G, yet still requiring their own place in the workflow, making the model linking activity include four models, not just two.

3 A DISTRIBUTED LOOSELY COUPLED MODELING SYSTEM (PHASE I)

We first developed a distributed model coupling solution based on the existing infrastructure on MyGeoHub [9]. The main goal is to establish an initial common workflow for data organization and sharing, and identify the pain points of the model coupling process while enabling the research groups to work out the conceptual and coupling steps. Built on the HUBzero platform [15], MyGeoHub is a science gateway that supports geospatial data management, processing, and simulations. Funded by the NSF DIBBS and CSSI programs, the Geospatial Data Analysis Building Blocks (GABBs) [8, 21] and Extensible Geospatial Data Framework (GeoEDF) [10] projects developed and deployed reusable software modules and libraries on MyGeoHub to create a powerful web-based system that allows researchers worldwide to easily manage, share, analyze, and visualize geospatial data. MyGeoHub provides the infrastructure and a well-defined procedure that enables developers and domain scientists to create and publish their models and data processing tools online using popular programming languages such as R and Jupyter Notebook. It provides a "submit" middleware that allows these online tools to run simulations on HPC resources at the backend.

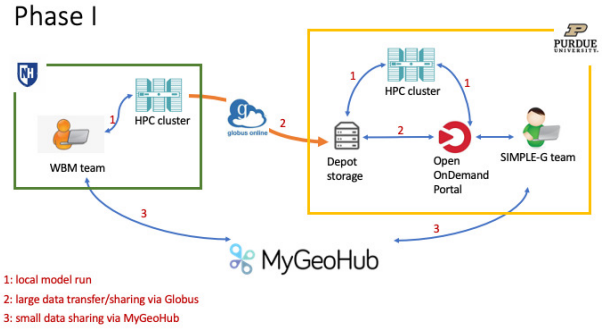

Figure 2: A distributed model coupling system

While some models are available to run online on MyGeoHub or other connected CIs, most models either run on a researcher's local computers or a campus HPC resource, depending on the individual's preference, different stages of the research, and flexibility and OS requirements of the models. In this initial solution, MyGeoHub served as a bridge for loosely linking the distributed modeling workflows. As shown in Figure 2, researchers at different locations download the data from MyGeoHub each time it is shared by collaborators, preprocess the data, feed it into their models (for example, SIMPLE-G researchers could run the model on their laptops at Purdue University or on MyGeoHub which submits the model runs to a campus cluster, while the WBM team runs their model on an HPC resource in the University of New Hampshire), postprocess the output, and share the result by uploading them back to the project space on MyGeoHub.

While it is feasible to share and exchange small datasets (up to a few GB) via HTTP upload to MyGeoHub, it becomes insufficient when a large amount of modeling data, such as the WBM output, needs to be exchanged across the teams and institutional CIs. Although there are ways to move data in and out of each CI or HPC system, it remains a challenge to provide a seamless experience to the end users in a trackable manner. The challenge of data sharing across the networks is further exacerbated by the greater variability of local platforms, tools, and researcher expertise and hence calls for a solution to enable seamless data-driven collaboration. In phase I, large dataset transfers were carried out via Globus endpoints between the HPC storage systems of different institutions which involved coordination and assistance of system admins. These data transfer and sharing steps are mostly manual and time consuming, causing significant delays as researchers need to share data to explore how to couple the models and repeat the workflow many times to tune the parameters and test model output. It is evident that research productivity and reproducibility would be significantly improved with a more flexible and seamless solution for exchanging model data and automating the model linking workflow with FAIR (Findable, Accessible, Interoperable, and Reproducible) properties automatically enabled.

4 C^3F– COLLABORATIVE CONTAINER-BASED MODEL COUPLING FRAMEWORK (PHASE II)

To overcome the shortcomings of the Phase I solution, we created C^3F, a modeling environment to provide more seamless model coupling support. As shown in Figure 3, in this system, researchers package their modeling code and data processing scripts into portable Singularity containers that are secure environments suitable to run on HPC resources. The models run on XSEDE's Anvil system which provides a common HPC environment researchers from different institutions can access simultaneously. Anvil's OOD portal offers a web user interface to access the HPC resources for users who are not familiar with Linux terminals and commands. The model coupling workflow is controlled by a shell script. The input and output data are stored in a project space on Anvil's project storage and can be accessed by all researchers working on the same project. The model results may be published with a DOI using MyGeoHub's automated data publishing function. It may also be accessed from MyGeoHub for online data analysis or visualization. The Singularity container recipe files and data conversion scripts are managed in a GitHub repository.

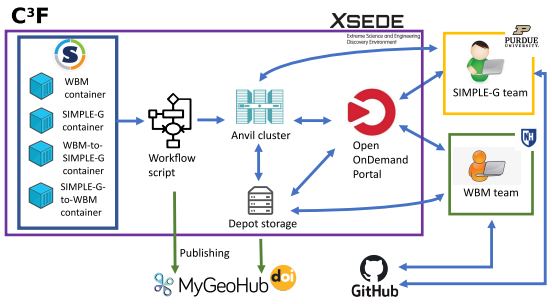

Figure 3: C^3F- A container-based model linking system

At the initial exploration stage, researchers need to understand each model, access each model's output data, and test their scripts that convert one model's output to the right format the other model can read. They may want to customize model input by adjusting the parameters. The containers are separated from the model input files, which makes them easy to customize. Using a shared XSEDE project space, neither data transfer nor waiting for the administrator's assistance/other researcher's response is necessary. Later during the research stage, both modelers can update their modeling code/data processing code and the containers, change the script that controls how the models are linked, and access modeling results. Finally, at the production stage, researchers can easily launch multiple workflows simultaneously either from the command line or using the OOD portal and access results immediately when a workflow completes. They can also use the Jupyter Notebook and R Studio environment directly from the OOD portal to analyze the data and visualize results.

4.1 Workflow Implementation

The workflow implementation consists of four Singularity containers – two model containers (i.e., WBM and SIMPLE-G), and two data conversion containers as shown in Figure 4a. The models communicate by passing data through mounted directories. The workflow execution is controlled by a bash script. The WBM and SIMPLE-G models are executed sequentially and iteratively until predetermined criteria are met. Each iteration of the workflow consists of four steps. Step 1 executes the WBM model container and outputs groundwater levels, surface water supply, and crop water requirements for each grid cell in the simulation domain in NetCDF format. In step 2 the WBM-to-SIMPLE-G container calculates the change rates of WBM output and converts it into a SIMPLE-G readable format. Step 3 runs the SIMPLE-G model container and produces new irrigated areas. Step 4 runs the SIMPLE-G-to-WBM container that converts the updated change in water use information to change in physical irrigation efficiency and passes it to the WBM model.

After exploring different approaches to linking the two models, the teams developed two workflow options based on two alternative

assumptions regarding the economic expectations and behaviors of the farmers in the SIMPLE-G model. Workflow option 1 (Figure 4b) assumes that farmers have perfect foresight, i.e., they have some information about the near future weather and markets. This workflow requires that the two models simulate over the same period iteratively until the output of at least one model reaches convergence, i.e., when there is no (or extremely little) change in the results from one iteration to the next over a period of T known as a lookback window. For the first iteration, WBM runs for T years with the given climate data and produces hydrological variables, followed by running a WBM-to-SIMPLE-G R script that converts the selected WBM output variables to percentage change over T to be further processed through the transfer functions. Then SIMPLE-G takes these changes and calculates the new equilibrium variables. A SIMPLE-G-to-WBM R script then sends the new SIMPLE-G output variables to WBM for the next iteration of the same period. The runs will continue until the changes in SIMPLE-G outputs are converged. After reaching convergence for the first T years, this process will be repeated for each T year until the end of the study period.

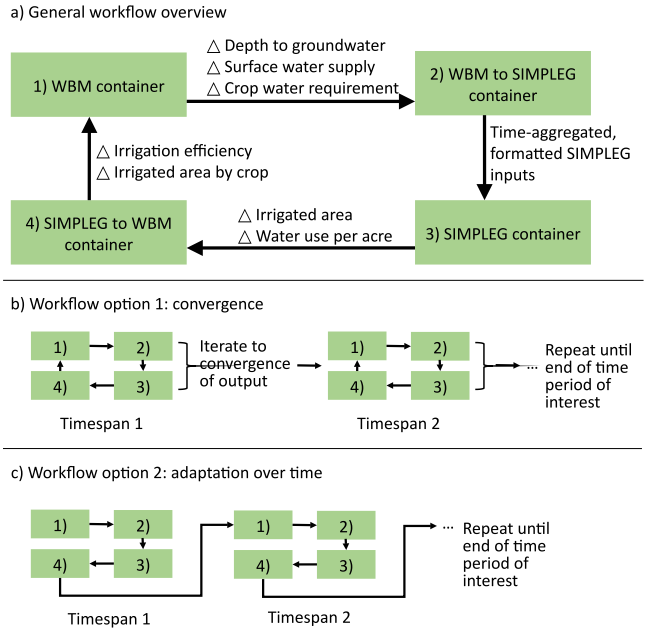

Figure 4: WBM-SIMPLE-G linkage workflow

Workflow option 2 (Figure 4c) is based on adaptive expectations that assume individual farmers form their expectations about future hydroclimatic conditions based on what happened in the past. This method does not check for convergence but requires consecutive runs of WBM and SIMPLE-G. Here, WBM takes SIMPLE-G economic outcome variables based on the previous years, while SIMPLE-G takes the WBM inputs based on changes in the moving average of hydroclimatic variables in the latest T years. In this workflow, WBM runs for the T-year and generates T yearly water data.

Having two workflows allows the possibility of future sensitivity analysis to the background assumptions. It is also beneficial for addressing different economic questions that support one of the alternative assumptions. Importantly, the domain experts found that it was only possible to develop and evaluate these two workflows once the CI was in place, enabling multiple reproducible simulations for analysis. This is one of the key reasons CI experts must be included in the collaboration from an early stage in the project; without them, workflow option 2 (which has been chosen for the main analysis in the FEWS project) would not have been developed in a timely manner.

4.2 Containerization of Models and Data Processing Scripts

As described in sections 2.1 and 2.2, the WBM and SIMPLE-G models were written in different programming languages for different operating systems, and subject to software license requirements. Although the data conversion code used in steps 2 and 4 are both written in R, they use different sets of libraries. To provide an isolated and easily reproducible environment, the computation code in each step of Figure 4a is packaged into an individual Singularity container. All four Singularity images are accessible from the project storage by both modeling teams. The SIMPLE-G model runs in a single thread while the WBM model runs with multiple threads which are determined based on the domain grid size.

To migrate the SIMPLE-G model to run in a Linux environment, a license for the source code of GEMPACK Release 11.1 for the Linux platform was obtained. A GEMPACK docker container that packages all the software needed to make a model executable was created. Next, a SIMPLE-G model Singularity container was built incrementally on top of the GEMPACK container image. It uses the docker bootstrap agent to pull the GEMPACK container from a private Docker Hub repository as it is a proprietary software requiring a license file. In the recipe file, the SIMPLE-G model code is added to the base GEMPACK docker image layer. The executables are generated in the post section. In the runscript section, the SIMPLE-G model is executed with user-supplied input and writes output to the mounted directories. The WBM team containerized their model into Docker and Singularity containers. WBM model uses initialization files that list all input metadata files and parameters files. The input to the WBM model is quite large as it involves a time series of gridded weather data. Both the WBM and SIMPLE-G Singularity containers set up some static directories for input, output, and command files in the container which are mounted at run time, and the data can be changed without impacting the container images. Only when the model code is updated does the container image have to be rebuilt. The R scripts to aggregate and convert data to the right format were packaged into separate containers based on the rocker/r-ver:4.1 (https://hub.docker.com/r/rocker/r-ver) image.

4.3 Workflow Execution

A linked model workflow implementing the adaptive option in Figure 4c was tested on Purdue's Brown cluster. A Slurm job submission script written in Bash is used to control the execution of the workflow. The number of iterations is calculated based on the start year, the end year, and the lookback window size which is set

Table 1: Breakdown of the run time for each step of the WBM-SIMPLE-G linked workflow for 2000-2007

Step	Container	Total runtime for three iterations (min)	Percentage
1	WBM	192	17%
2	WBM-to-SIMPLE-G	900	80%
3	SIMPLE-G	18	1.5%
4	SIMPLE-G-to-WBM	18	1.5%

up as environment variables at the beginning of the script. Suppose we run a coupled model for 2000-2007 with a five-year lookback window, the workflow will iterate three times. At the first iteration, the WBM container is set to simulate for years 2000-2005. Step 2 calculates the change ratio for 2005 and outputs a new irrigation area for 2006. In the second iteration, the WBM container simulates 2001-2006. Since the years 2001-2005 have been simulated in iteration 1, the model only simulates the year 2006 in this round reusing the output from iteration 1. Similarly in iteration 3, the WBM container produces results for 2002-2007, and the WBM-to-SIMPLE-G container calculates the percent changes of shocks for 2007 which is fed into the SIMPLE-G container to produce predictions of irrigation area for 2008. Researchers submit a coupling job from the terminal at this moment. They customize parameters such as simulation period and the lookback window size by editing the job submission script. They can also access the data and edit the file using the graphical interface provided by the OOD portal. Customized interactive applications may be developed using Jupyter Notebook or OOD's extensible app development framework.

4.4 Preliminary Result

The team is conducting workflow runs to gather runtime statistics. Here we present the preliminary data obtained from an early run for illustration: a WBM-SIMPLE-G linked workflow for the years 2000-2007 with a 5-year lookback window ran on Purdue's Brown cluster which consists of Dell compute nodes with two 12-core Intel Xeon Gold "Sky Lake" processors (24 cores per node) and 96 GB of memory. The workflow ran three iterations and took approximately 19 hours. Table 1 shows the run time for each iteration step. Note that the runtime for step 1 is iteration-dependent as in the first iteration step 1 simulates five years (the lookback window), while in the subsequent iterations it simulates a single year using cached results. The runtime for step 2 varies slightly among iterations, and the runtime for steps 3 and 4 stays the same for each iteration.

Figure 5 shows how the percentage change in surface water storage evolves over three iterations due to a combination of weather (an input to the coupled model system) and irrigation water use (a result of the coupled model system). This is a key research output from the WBM model within the linked framework. In the 2000-2005 period, California, Nevada, and Arizona showed an increase in surface water storage. West north central and west south central areas showed mild decreases for 2001-2006, but their surface water storage turned back to increases for 2002-2007. Figure 6 demonstrates the computed changes in irrigated crop production, output from SIMPLE-G, from three iterations of the coupled system at each 5 arc-min grid cell for the cultivated US. This exemplifies the equilibrium economic responses to changes in water availability

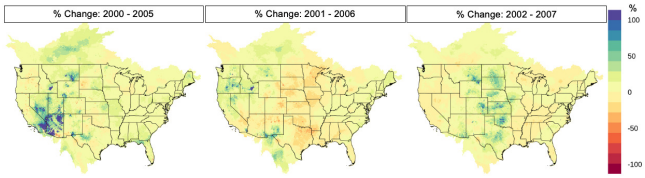

Figure 5: Percent change in irrigated crop production for three iterations in WBM-SIMPLE-G. The green color illustrates the percentage change increase in irrigated crop production and the red color shows the percentage change decline in irrigated crop production due to changes in water availability and economic market responses. White color means no crop production.

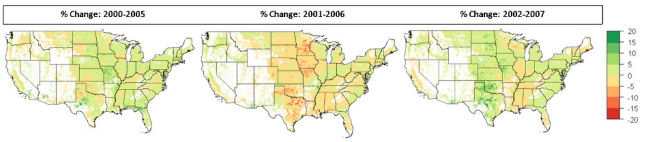

Figure 6: Percent change in irrigated crop production for three iterations in WBM-SIMPLE-G. The green color illustrates the percentage change increase in irrigated crop production and the red color shows the percentage change decline in irrigated crop production due to changes in water availability and economic market responses. White color means no crop production.

and the overall market interactions. Compared with figure 5, the pattern of change in irrigated production follows the pattern of change in surface water storage. However, the magnitude of the change in crop production is smaller due to economic responses and the economic reallocation of production resources. Although the maps demonstrate only one type of output from the coupled system, it shows successful interactions between WBM and SIMPLE-G models. Neither Figure 5 nor Figure 6 results would have been possible without the fully coupled system, as these results rely on the passing of information between the two domain models.

5 DISCUSSION

Based on our experimental runs, the most time-consuming part of the workflow is step 2 as shown in Table 1. It involves I/O operations of large raster files which computes the average groundwater level for each pixel throughout the given period. The performance of this step may be improved by caching some intermediate yearly average results for subsequent iterations. Furthermore, although

WBM computes daily, monthly, and yearly average outputs, the monthly and yearly files are calculated as a background process and sometimes fail to complete due to a timing issue. We will adjust it to write yearly output instead of daily, which could significantly speed up step 2. Additionally, the WBM input directory is around 60 GB and its output directory is even bigger depending on the number of years to simulate. We need to improve the management of model copies to enable larger-scale runs.

Cyber training of domain scientists is an important aspect of this work. In the exploration and research stages, the WBM-SIMPLE-G models and the data conversion scripts need to be tuned and tested by the modelers. As a result, they need to learn the fundamentals of working with a Singularity container and be able to use GitHub to clone/modify/build/test the code. While the learning curve is non-trivial, it is not insurmountable. The CI experts on the team held two hands-on tutorial sessions on topics such as installing Singularity, cloning projects from GitHub repositories, setting up environments for linked workflow submission, and editing Slurm submission scripts. While logging in directly on the cluster and running everything through the command line provides maximum flexibility for researchers, some may benefit from using the OOD portal's GUI interface to access the data, modify the Slurm script, and invoke the workflow, etc., to quickly ramp up on the HPC system and use the CI resources in teaching and training themselves.

Our experience has shown that fully coupling models in an automated framework often requires a greater period than is provided by a single grant-funded project (which is typically 3-4 years), due to the conceptual and collaborative challenges and the important intersection challenges that require both the technical and conceptual challenges to be solved. We have also found that it is only with the inclusion of computing experts and software engineers that the technical challenges are solved well enough and early enough in the research process to enable fully automated model coupling. These collaborators must be involved in the model coupling project from the beginning, as they are needed in working through the intersection challenges.

In addition to improving the performance of the WBM-SIMPLE-G workflow, we plan to deploy the workflow on the newly available XSEDE Anvil cluster where the research team will conduct "production" runs of their fully linked modeling experiments. We also plan to automate container creation, testing, and deployment using GitHub's CI/CD capabilities. In the future, we plan to investigate how to extend this system to support more model coupling use cases and incorporate community standards where appropriate. For example, in another NSF-funded project, researchers from the Natural Capital Project (NatCap) and the Global Trade Analysis Project (GTAP) are looking into linking the InVEST ecosystem services model and GTAP economic model to study global sustainability challenges and make informed environmental decision making [7]. Many complex factors impact how models may be linked. We believe the lessons and experience learned here will help guide our future work in coupling other models critical in multi-disciplinary research.

6 CONCLUSION

The linking of WBM and SIMPLE-G allows members of these two different disciplinary communities (hydrology and economics) to collaborate on the analysis of land and water resource issues. The software modules, data processing and modeling tools, and data services developed in C^3F are mostly generic and portable using containers, which can be applied to other cyberinfrastructure platforms. The pioneering work on WBM-SIMPLE-G model coupling workflow lays the foundation for a broader exploration of the use cases and solutions for cross-disciplinary research involving multiple models. All the infrastructure developed and deployed follows and promotes the FAIR principles. It will help researchers make their research software and tools more easily reusable and interoperable.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grants #1855937, #2020635, and the U.S. Department of Energy grants DE-SC0016162 and DE-SC0022141.

REFERENCES

[1] U. L. C. Baldos, I. Haqiqi, T. W. Hertel, M. Horridge, and J. Liu. 2020. SIMPLE-G: A multiscale framework for integration of economic and biophysical determinants of sustainability. *Environmental Modelling and Software* 133 (11 2020). https://doi.org/10.1016/j.envsoft.2020.104805

[2] Jonathan F Donges, Ricarda Winkelmann, Wolfgang Lucht, Sarah E Cornell, James G Dyke, Johan Rockström, Jobst Heitzig, and Hans Joachim Schellnhuber. 2017. Closing the loop: Reconnecting human dynamics to Earth System science. *The Anthropocene Review* 4, 2 (2017), 151–157. https://doi.org/10.1177/2053019617725537 arXiv:https://doi.org/10.1177/2053019617725537

[3] Danielle Sarah Grogan. 2016. *Global and regional assessments of unsustainable groundwater use in irrigated agriculture.* Ph. D. Dissertation. University of New Hampshire, Durham, NH. https://scholars.unh.edu/dissertation/2.

[4] Danielle S Grogan, Fan Zhang, Alexander Prusevich, Richard B Lammers, Dominik Wisser, Stanley Glidden, Changsheng Li, and Steve Frolking. 2015. Quantifying the link between crop production and mined groundwater irrigation in China. *The Science of the total environment* 511 (April 2015), 161—175. https://doi.org/10.1016/j.scitotenv.2014.11.076

[5] I. Haqiqi, D. S. Grogan, T. W. Hertel, and W. Schlenker. 2021. Quantifying the impacts of compound extremes on agriculture. *Hydrology and Earth System Sciences* 25, 2 (2021), 551–564. https://doi.org/10.5194/hess-25-551-2021

[6] Mark Horridge, Michael Jerie, Dean Mustakinov, and Florian Schiffmann. 2018. *GEMPACK Manual.* The Centre of Policy Studies., Victoria, Australia. https://www.copsmodels.com/

[7] Justin Andrew Johnson, Giovanni Ruta, Uris Baldos, Raffaello Cervigni, Shun Chonabayashi, Erwin Corong, Olga Gavryliuk, James Gerber, Thomas Hertel, Christopher Nootenboom, and Stephen Polasky. 2021. The Economic Case for Nature : A Global Earth-Economy Model to Assess Development Policy Pathways. *World Bank, Washington, DC.* (2021). https://openknowledge.worldbank.org/handle/10986/35882

[8] R. Kalyanam, R. Campbell, D. Kearney, L. Delgass, L. Biehl, L.and Zhao, C. Ellis, and C. X Song. 2017. Cloud-enabling a Collaborative Research Platform: The GABBs Story. *Practice and Experience in Advanced Research Computing (PEARC17)* (July 2017).

[9] Rajesh Kalyanam, Lan Zhao, Carol Song, Larry Biehl, Derrick Kearney, I. Luk Kim, Jaewoo Shin, Nelson Villoria, and Venkatesh Merwade. 2019. MyGeoHub—A sustainable and evolving geospatial science gateway. *Future Generation Computer Systems* 94 (2019), 820–832. https://doi.org/10.1016/j.future.2018.02.005

[10] Rajesh Kalyanam, Lan Zhao, X. Carol Song, Venkatesh Merwade, Jian Jin, Uris Baldos, and Jack Smith. 2020. *GeoEDF: An Extensible Geospatial Data Framework for FAIR Science.* Association for Computing Machinery, New York, NY, USA, 207–214. https://doi.org/10.1145/3311790.3396631

[11] R.B. Lammers, A. Bliss, R. Hock, A.A. Proussevitch, D.S. Grogan, S. Glidden, S. Frolking, and V. Radic. 2013. Contributions of the world's glaciers to the hydrological cycle in the 21st Century. *Abstract GC21E-03 presented at 2013 Fall Meeting, AGU* GC21E-03 (December 2013).

[12] Jianguo Liu, Thomas Dietz, Stephen R. Carpenter, Marina Alberti, Carl Folke, Emilio Moran, Alice N. Pell, Peter Deadman, Timothy Kratz, Jane Lubchenco, Elinor Ostrom, Zhiyun Ouyang, William Provencher, Charles Redman, Stephen H. Schneider, and William W. Taylor. 2007. Complexity of coupled human and natural systems. *Science* 317, 5844 (14 Sept. 2007), 1513–1516. https://doi.org/10.1126/science.1144004

[13] Jing Liu, Thomas W Hertel, Richard B Lammers, Alexander Prusevich, Uris Lantz C Baldos, Danielle S Grogan, and Steve Frolking. 2017. Achieving sustainable irrigation water withdrawals: global impacts on food security and land use. *Environmental Research Letters* 12, 10 (2017), 104009.

[14] Jianguo Liu, Harold Mooney, Vanessa Hull, Steven J Davis, Joanne Gaskell, Thomas Hertel, Jane Lubchenco, Karen C Seto, Peter Gleick, Claire Kremen, et al. 2015. Systems integration for global sustainability. *Science* 347, 6225 (2015), 1258832.

[15] Michael McLennan and Rick Kennell. 2010. HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Computing in Science Engineering* 12, 2 (2010), 48–53. https://doi.org/10.1109/MCSE.2010.41

[16] John T. Murphy, Jonathan Ozik, Nicholson Collier, Mark Altaweel, Richard B. Lammers, Alexander A. Prusevich, Andrew Kliskey, and Lilian Alessa. 2015. Simulating regional hydrology and water management: An integrated agent-based approach. *Proceedings of the 2014 Winter Simulation Conference* (2015), 3913–3924.

[17] Nihar R. Samal, Wilfred M. Wollheim, Shan Zuidema, Robert J. Stewart, Zaixing Zhou, Madeleine M. Mineau, Mark E. Borsuk, Kevin H. Gardner, Stanley Glidden, Tao Huang, David A. Lutz, Georgia Mavrommati, Alexandra M. Thorn, Cameron P. Wake, and Matthew Huber. 2017. A coupled terrestrial and aquatic biogeophysical model of the Upper Merrimack River watershed, New Hampshire, to inform ecosystem services evaluation and management under climate and land-cover change. *Ecology and Society* (2017). https://doi.org/10.5751/ES-09662-220418

[18] Charles J. Vörösmarty, Berrien Moore III, Annette L. Grace, M. Patricia Gildea, Jerry M. Melillo, Bruce J. Peterson, Edward B. Rastetter, and Paul A. Steudler. 1989. Continental scale models of water balance and fluvial transport: An application to South America. *Global Biogeochemical Cycles* 3, 3 (1989), 241–265. https://doi.org/10.1029/GB003i003p00241

[19] D. Wisser, B. M. Fekete, C. J. Vörösmarty, and A. H. Schumann. 2010. Reconstructing 20th century global hydrography: a contribution to the Global Terrestrial Network- Hydrology (GTN-H). *Hydrology and Earth System Sciences* 14, 1 (2010), 1–24. https://doi.org/10.5194/hess-14-1-2010

[20] Esha Zaveri, Danielle S. Grogan, Karen Fisher-Vanden, Steve Frolking, Richard B. Lammers, Douglas H. Wrenn, Alexander Prusevich, and Robert E. Nicholas. 2016. Invisible water, visible impact: Groundwater use and Indian agriculture under climate change. *Environmental Research Letters* 11, 8 (3 Aug. 2016). https://doi.org/10.1088/1748-9326/11/8/084005 Publisher Copyright: © 2016 IOP Publishing Ltd..

[21] Lan Zhao, Carol X. Song, Rajesh Kalyanam, Larry Biehl, Robert Campbell, Leif Delgass, Derrick Kearney, Wei Wan, Jaewoo Shin, I Luk Kim, and Carolyn Ellis. 2017. GABBs - Reusable Geospatial Data Analysis Building Blocks for Science Gateways. *9th International Workshop on Science Gateways (IWSG 2017)* (June 2017).