Cloud-based solutions for distributed climate modeling

Nadya Vinogradova[1], Mark Shiffer[1], Gael Forget[2], and Chris Hill[2]
[1]Cambridge Climate Institute, MA
[2]Massachusetts Institute of Technology, MA

Climate models integrate our best knowledge of the climate system behavior, its governing principles and ongoing changes, providing unique tools for studying the Earth's past, present, and future states. In addition to their widespread use by the research community, model-based solutions offer crucial guidance for decision-makers in their efforts to anticipate and mitigate hazards associated with climate change. The success of both efforts is often tied to the ability of a user to interpret model results and reproduce solutions in order to build on previous achievements. However, modeling capabilities remain limited in their accessibility, as re-running simulations created by other groups can require expertise and manpower. Furthermore, potential users may face challenges associated with limited on premise computational and storage resources. These common impediments slow down the overall progress of model development, diminish the general openness of modeling activities, and make collaboration between various groups less efficient.

In this respect, cloud-based approaches open up promising new avenues for widely collaborative and distributed climate modeling. Today, running climate models in the cloud has become a practical alternative to the use of conventional on premise or government-sponsored computing facilities. Here, we present a framework that leverages existing cloud services and enables researchers to easily develop, archive, re-use, and share modeling tools (Fig. 1).

Cloud-based framework for dissemination of climate models

The main purpose of our framework, developed jointly at Cambridge Climate Institute (CCI) and Massachusetts Institute of Technology (MIT), is to replicate the computational environment and algorithms needed to reproduce a particular model simulation, and then provide these ready-to-use replicas to other researchers [*Shiffer et al.*, 2016; *Forget* 2017; and later updates]. These replicas can encompass the complete state of the operating system, as well as shared libraries, system codes, and model input. Instead of spending their time on the set up and configuration of a model that was created by another group, a user can thus simply launch pre-configured systems and run a model for diagnostic or experimental purposes.

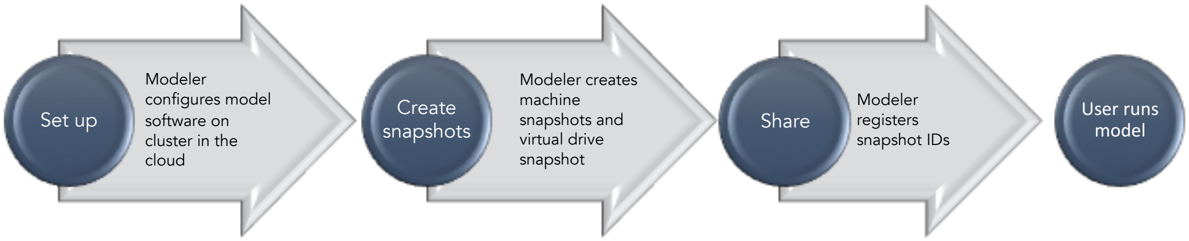
Set up
Modeler configures model software on cluster in the cloud
Create snapshots
Modeler creates machine snapshots and virtual drive snapshot
Share
Modeler registers snapshot IDs
User runs model

Fig. 1: Flowchart of the framework

To use a replica, a user only needs to create an account with a cloud provider, Amazon Web Services (AWS) in our prototype recipes, and then execute a series of commands that will (a) spawn a compute cluster ready to execute a model run, and (b) launch the simulation by following a few step-by-step instructions (*Shiffer et al.*, 2016; *Forget* 2017). Once done, the user returns computational resources to the cloud and thus pays only for the time that those resources were used and for the storage of output according to their needs. The virtualization and sharing of resources provided by cloud services allows immediate and as-needed access to exactly the required computing recourses. We provide both the tools to create and run the cluster as well as pre-configured machine images, which in combination allow a researcher to access models almost effortlessly, without incurring additional infrastructure or IT costs.

Replicating model simulations using snapshots of pre-configured machine images

To illustrate the usability of this new tool, we developed two prototype recipes. The first relies on a pre-installed, frozen copy of the numerical model [*Shiffer et al. 2016*], and the second downloads and compiles the latest model code on the fly [*Forget 2017*]. Our prototypes take a global ocean state simulation from Estimating the Circulation and Climate of the Ocean (ECCO; www.ecco-group.org) as a representative example. ECCO is a mature, global ocean modeling system that has gained recognition as one of the leading ocean state estimations. The system typically produces an estimate of the three-dimensional ocean state over the past decades that is consistent with most of the available ocean observations, estimated atmospheric forcing fields, and adjusted oceanic turbulent transport parameters [*Forget et al., 2015a,b; 2016*]. By porting ECCO to the cloud, we aim to further extend its user community by increasing resource access and facilitating the setup, reproduction, and modification of ECCO model runs.

An illustration of results using this approach is provided in Fig. 2. Following the recipes and step-by-step instructions [*Shiffer et al., 2016; Forget 2017*], one can replicate the ECCO solution and desired diagnostics, exactly as reported in the literature and system manuals.

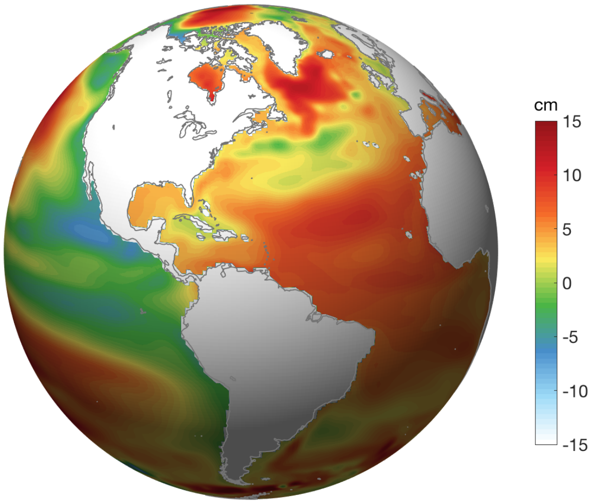

Fig. 2: Cloud-based modeling framework based on snapshots of pre-configured machine images facilitates transparent and broad use of climate models. Shown here is the change in sea surface height anomalies between 1992 and 2011 associated with ocean dynamics, heat uptake, and water discharge from land hydrology and melting ice. This example is obtained by replicating ECCO (v4 r2) simulations, exactly as reported in the literature. In a cloud-based framework, duplication of model runs becomes straightforward even for a user unfamiliar with climate modeling.

Anyone can thus explore what it means to be a climate modeler. In particular, first time model users may benefit the most from ready-to-use recipes and pre-configured systems. But experienced modelers, even those with access to dedicated computational and storage resources, may also find advantages to archiving, retrieving, sharing, and/or collaborating on model simulations via cloud services.

The accuracy of our ECCO v4 r2 replicas via the Amazon cloud is at the expected level. In terms of cost, to duplicate ECCO v4 r2, we were able to run the model with one-degree simulation over the 1992-2011 period using a virtual cluster of 6 nodes with 16 cores per node for 36 hours within the AWS cloud at a cost of about $40.

Expanding the user base through customized model simulations

In addition to making exact duplicates of model runs, one can also produce customized model outputs within the same modeling framework. In ECCO, many options are available at run-time to, e.g., adding perturbations to atmospheric forcing fields or internal ocean parameters (see *Forget et al., 2015a*, Table 7). When a researcher needs to have additional model diagnostics that are not provided by the production groups, he or she can specify desired variables and re-run the model in the exactly the same configuration, ensuring that the new diagnostics is consistent with the standard model solution.

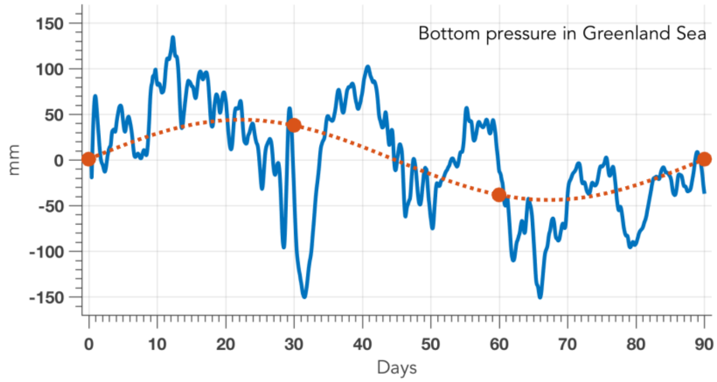

Fig. 3 In addition to standard monthly diagnostics, one can output model variables at higher temporal resolution. Shown here is an example of hourly (blue curve) time series of ocean bottom pressure anomalies in Greenland Sea (-10°W, 70°N), which are obtained by re-running a model in its standard configuration (ECCO v4 r2) but with customized model output settings. Corresponding monthly-averaged values, based on the output as distributed by ecco-group.org, are shown as red circles.

For example, Fig. 3 illustrates how a customized model run generated hourly fields of ocean bottom pressure to complement monthly model output which are already available online. This new, hourly-varying time series was obtained by re-running the model in its standard configuration (Fig.3) and specifying a high-frequency output request as explained in the model documentation[1]. The ability to produce customized outputs makes model solutions applicable for a broader range of research problems. It also facilitates model development by expanding metrics of model validation and by providing the ability to perform sensitivity runs and exploratory experiments.

[1] http://mitgcm.org/public/r2_manual/latest/

Increasing transparency of modeling tools

Much like the importance of sharing data, enabling transparency and the easy use of modeling tools is essential to the integrity of scientific results. In addition to traditional archiving approaches, the cloud-based framework allows research groups to store and give access to all configuration files associated with a particular model run. Such an approach not only simplifies duplication of previous results as discussed above, but also pushes model archiving and sharing practices towards higher standards such that each model experiment can be referred to and then viewed by an independent researcher, including reviewers of scientific papers. Such practice alone can eliminate common frustration associated with "black-box" model results by allowing other scientists to evaluate and extend previous efforts in a traceable fashion. It has the potential to bring modeling results to the same standard of transparency championed by the AGU community for observations.

Facilitating collaboration via cloud-based solutions

The examples discussed here, with pre-configured images of the ECCO solutions, illustrate how access to near-limitless computational power can be leveraged via cloud-based cluster technology to allow effortless user access to complex models. These examples can also be used as a template for running and sharing other climate models and systems. As shown here, porting a simulation to the cloud just requires replicating the original computational environment, model codes, and model inputs. We therefore expect that sharing model simulations via replicas in the cloud will rapidly become common practice, foster widely distributed modeling activities, and increase transparency within climate research and beyond.

For more information, please contact cloud-support@camclimate.org at Cambridge Climate Institute (attn: Mark Shiffer), or mitgcm-support@mitgcm.org at MIT (attn: Gael Forget)

References:

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch (2015a). ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation, *Geoscientific Model Development*, 8(5), 3071–3104, doi:10.5194/gmd-8-3071-2015.

Forget, G., Ferreira, D., & Liang, X. (2015b). On the observability of turbulent transport rates by Argo: supporting evidence from an inversion experiment. *Ocean Science*, *11*(5), 839-853.

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch (2016), ECCO Version 4: Second Release, http://hdl.handle.net/1721.1/102062

Forget, G. (2017). gaelforget/ECCO_v4_r2: Updates and improved documentation of on-premise and cloud recipes. *Zenodo*. http://doi.org/10.5281/zenodo.834082

Shiffer, M., G. Forget, and N. T. Vinogradova (2016). ECCO in the cloud. *Zenodo*, Geneva, Switzerland, doi: 10.5281/zenodo.199307.