Dataset Submission Guidance Computational Models

GoMRI investigators use computational models to generate environmental forecasts or hindcasts of real-world conditions, to simulate particular processes observed in the lab or field; and to test model sensitivity. Model datasets can range in size from a few kilobytes to hundreds of terabytes, depending on what information is included in the dataset. This document provides guidance to researchers regarding what data and information from modeling studies should be shared to meet the goals of the GoMRI data policy. Given the variety in purpose and approach of modeling studies, however, it is expected there will be a continuing need for considering specific circumstances for particular datasets.

Generally, there are four components in a model dataset (1) input data, (2) set up information, (3) computer code, and (4) output data. What investigators should share depends on the objective for sharing, and the existing public availability of the dataset components. Given the large size of some modeling datasets, it is important that resources are not spent on archiving data that will not be used for verification or reuse in other studies. In this light, consultation with the researcher producing a model dataset is often needed to help determine the value of archiving.

The sharing of model datasets should meet one or both of the following objectives: (1) enabling others to verify results presented in publications or (2) allowing others to reuse elements of the model dataset, whether it be the computer code, input, or output portions.

General requirements by objective:

Objective	Input Data	Setup Files	Computer Code	Output Data
Verification	Yes, If not publicly available	Yes, if computer code is available	Yes, if not proprietary or not publicly available	Yes, only portion used in publication or required to verify analysis
Reuse of Output	Yes, if acquired with GoMRI funds	No	Yes, if code development is part of GoMRI-funded work	Yes, all useable output

The sharing objective(s) is largely determined by the purpose and publication status of the modeling work. Furthermore, the components that the researcher needs to provide will also depend on what is already publicly available. Following is further guidance for each model component including documentation needs. At the end of the document are a few examples of model dataset sharing scenarios.

Model Dataset Components and Sharing Requirements

1) Input data (required to meet verification objective)
   - If input files are available online through national data archives, these do not need to be submitted to GRIIDC. This situation should be described in the methods section of the descriptive information for the dataset and a listing of specific data files used should be included with the dataset in a ReadMe file.
   - In the case of some model sensitivity and process studies, the model may be run with specific values for each parameter. This situation should be described in the methods section of the descriptive information for the dataset and a listing of specific data files used should be included with the dataset in a ReadMe file.
   - If input data are not available through national data archives, it should be submitted to GRIIDC as part of the dataset file. Alternatively, input data may be submitted as a unique dataset in the GRIIDC system.

2) Set up data and information (required to meet verification objective)
   - If the modeling requires distinct set up files specific to the research, these should be submitted to GRIIDC as part of the dataset. This situation should be described in the methods section of descriptive information for the dataset so that it is clear the dataset includes these files.
   - If the model set up files and information are available online with the computer code then there is no need to submit them to GRIIDC.
   - If the model does not require separate set up files but another user would need additional information about model set up to understand the output or to re-run the model, this information should be described in the methods section of the descriptive information for the dataset and described in a ReadMe file with the dataset, if needed.

3) Computer Code (may be required to meet verification or reuse objectives)
   - Community based code that has broad acceptance and use within the particular modeling community and is publicly available online does not need to be submitted to GRIIDC. The model information, including version number, should be provided in the methods section of the descriptive information for the dataset.
   - Submit model computer code developed using GoMRI funding. This is particularly true if developing the code is an objective of the work or is the subject of a publication.
   - Do not submit model code if it is proprietary and not developed using GoMRI funding. In the methods section of the descriptive information; however, document the version and describe or provide references describing the model code.

4) Output data (required to meet verification and reuse objectives)
   - Model output intervals that are used for computational stability during processing should not be submitted.
   - Submit all model steps and final model output files that are determined to be valuable for reuse.

- If the output is not generally reusable, submit only the portion of output data that supports the results of a publication so that others may assess the research.

Examples

Example 1

An open-source community ocean/coastal model (e.g., NCOM, HYCOM, ADCIRC, FVCOM) is implemented in a GoMRI study to forecast currents. Some input data were gathered from public sources (e.g., National Weather Service, NOAA tide observations, other archived model outputs) while some were acquired in the field or specially developed by the GoMRI project. A portion of the output is used in a publication, and the entire output has potential to be used for other research.

In this case, there is no need to submit the input data gathered from the public source, but do document the source in the model dataset descriptive information and include a listing of the specific data files used with the dataset. The portion of the input data acquired or developed specifically for the GoMRI project, however, should be provide as part of the modeling dataset or submitted as a separate dataset and referenced in the modeling dataset descriptive information. Submit any setup or auxiliary files or information (could be a ReadMe file) in the model dataset if they are needed to explain the output. There is no need to submit the computer code because it is publicly available, but include the version of the code in the descriptive information for the dataset.

Because the output has potential for reuse, submit all the output, not just the portion used in the publication. It may be logical to submit the publication dataset and the remainder of the dataset, separately. The objective to be met in this example is for both verification and reuse, but in some cases, this is not achievable as described in example 2.

Example 2:

An NCOM model was run for the Gulf of Mexico by NRL that included assimilating data of all the relevant observational data the Navy collects. Some of this is public, but not all of it. Some of the input data might even be classified. The data assimilation is done in real time, therefore since the model run, some input data may have been updated (better QC) and the original data files overwritten. Most of the computer code is public, but parts are proprietary and cannot be shared. It is very unlikely that anyone would want to commit computing resources to rerunning this complex of a model, replicating exactly its set-up, and gathering all the myriad input files. Especially not, if the output is available. The objective to be met therefore is reuse, and only the output data needs to be submitted.

Example 3:

A GoMRI study developed a computer model of a process (e.g., marsh channel formation, plume formation, beach erosion, oil weathering). Data were developed to calibrate or validate the model, but there are no input datasets - just sets of parameter values used in the model.

In this case, the code and setup information or parameters developed under the GoMRI project should be submitted and documented in the descriptive information for the dataset. One should also submit examples of the output along with the model parameter values used. In addition, submit data used for model validation. The objective being met is verification, which enables other researchers to use the code. Some of these studies may also include a full output dataset that is valuable for reuse as well.

Example 4:

A GoMRI study produced multiple outputs of a simulation of a hypothetical volume of the ocean. The output is used to study the character of various processes (e.g., mixing, turbulence, eddy formation). The output was produced with documented but proprietary computer code that was not developed with GoMRI funding. The researchers have published using the output, but they recognize that other studies could use the output, which required considerable computing resources to generate.

In this case, input data are not pertinent, but any setup or parameter information regarding how the model was initiated and output generated should be submitted in the descriptive information for the dataset or a ReadMe file, as appropriate. Because it is proprietary and not created with GoMRI support, the computer code should not be submitted, but the version and information or a reference describing it should be provided in the descriptive information for the dataset. The objective to be met in this case is reuse; therefore, all potentially reusable output should be submitted.

Investigators should contact GRIIDC (griidc@gomri.org) if they have questions about what data to submit.