Dataset Submission Guidance – Computational Models

Computational models have become a principal tool in modern science and are used to generate environmental forecasts or hindcasts of real-world conditions; to simulate processes observed in the lab or field; and to test model sensitivity. Generally, there are four components in a model dataset: (1) input data, (2) setup and configuration files, (3) model code, and (4) output data. What investigators submit to GRIIDC depends on the objective for sharing and the existing public availability of the dataset components. The sharing of model datasets should meet one or both of the following objectives: (1) enabling others to verify and reproduce results presented in publications or (2) allowing others to reuse elements of the model dataset, whether it be the model code, input, or output 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 and scope of archiving.

This document provides guidance regarding what data and information from modeling studies should be submitted to GRIIDC. Please contact griidc@gomri.org with any questions.

Model Components

1. Input data

- If input files are available online through inter/national data archives, these do not need to be submitted to GRIIDC. Examples of these archives include:
    - https://tidesandcurrents.noaa.gov/
    - https://waterdata.usgs.gov/nwis
    - https://www.ncei.noaa.gov/
    - https://data.nasa.gov/
    - https://www.copernicus.eu/en

  Documentation:
    - This situation should be described in the methods section of the descriptive information for the dataset.
    - A listing of specific data files used should be included with the dataset in a readme text file.
- If input files are not available online through inter/national data archives, they should be submitted to GRIIDC.
- In the case of fluid dynamic models, mesh files and mesh attributes are considered input files and are required to be submitted.
- In the case of some model sensitivity and process studies, the model may be run with specific values for each parameter.

  Documentation:

- This situation should be described in the methods section of the descriptive information for the dataset.
- A listing of specific data files used should be included with the dataset in a readme text file.

2. Setup and configuration files

- If the model requires distinct set up files specific to the research needed to re-execute the model, these should be submitted to GRIIDC as part of the dataset, or links to the external data sources should be provided.

  Documentation: This situation should be described in the methods section of the descriptive information for the dataset so that it is clear the dataset includes these files.
- If the model does not require separate setup files but another user would need additional information about model setup to understand the output or to re-run the model, the information should be provided.

  Documentation: Information should be described in the methods section of the descriptive information for the dataset and described in a readme text file with the dataset.

3. Model code

- If model code are publicly available on a long-term, static, version-controlled repository, the data do not need to be submitted to GRIIDC. GitHub is not included in this list as data are able to be removed from the repository and are not static. Examples of long-term archives include:
  - http://adcirc.org/
  - https://www.hycom.org/hycom/source-code
  - https://oss.deltares.nl/web/delft3d/source-code
  - https://biodiversityinformatics.amnh.org/open_source/maxent/
- If model code are not publicly available as described above, they should be submitted to GRIIDC.
- If the source code has been modified, it should be submitted to GRIIDC.
- Model code that is proprietary does not need to be submitted to GRIIDC.

  Documentation:
  - Model information (version number, description, restrictions on use and distribution to the code, etc.) should be provided in the methods section of the descriptive information of the dataset.

4. Output data

- Final model output files that are determined to be valuable for reuse should be submitted to GRIIDC.
- If the output is not valuable for reuse, submit only the portion of output data that supports the results of a publication so that others may assess the research.

- Intermediary output does not need to be submitted.

  Documentation: The data structures for all data types generated by the model should be well described in the descriptive information and the variables used defined appropriately.

File Formats

GRIIDC requires model data be submitted in the following file formats.

- Text/ASCII
- Matlab or other programming language data files
- NetCDF/HDF

Examples of different models and submission requirements

Example 1: Hydrodynamic Circulation Model

NOAA tidal data were used to force an ADCIRC circulation model of Texas bays and estuaries with the purpose of hindcasting water circulation and water levels. The researcher set the model to output NetCDF files of water level and current velocities every hour.

Note: The data used to force the model are found in a national repository and do not need to be submitted but need to be detailed in a readme file. However, the different files that make up the triangulated mesh and the properties needed to run the model are required: the mesh (fort.14), nodal attributes (fort.13), and model parameter (fort.15) files. Versions of the ADCIRC code used and links to the repository should be provided in a readme text file.

Dataset should include:

1. Links and details of the NOAA tide gauges used in the simulation
2. Mesh, nodal attributes, and model parameter files used to simulate circulation and water levels
3. Setup files used to run the simulation
4. Generated NetCDF files output files

Objective: Reuse of output data.

Objective	Input Data	Setup Files	Model Code	Output Data
Reuse of output	✗ ✓	✓	✗	✓

Example 2: Coastal Processes Mathematical Model

A researcher developed and coded a mathematical model of a process (e.g., marsh channel formation, plume formation, beach erosion, oil weathering). Data were developed to calibrate or validate the model.

Dataset should include:

1. Input data
2. Model code
3. Setup and configuration files and other model parameter specifications
4. Output data
5. Data used for model validation

Objective: 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.

Objective	Input Data	Setup Files	Model Code	Output Data
Verification and reproduction	✓	✓	✓	✓

Example 3: Hydrodynamic Model generated by the Navy

An NCOM model was run for the Gulf of Mexico by the U.S. Naval Research Laboratory that included assimilating all the relevant observational data the Navy collects. Some, but not all, of this is public, and some of the input data might even be classified. Because the data assimilation is done in real time, in the interim some input data may have been updated (better quality control) 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 commit computing resources to rerunning this complex of a model replicating its set-up and gathering all the input files.

Dataset should include:

1. Output data
2. The version number of the model should be included in the descriptive information for the dataset

Objective: Reuse of output data.

Objective	Input Data	Setup Files	Model Code	Output Data
Reuse of output	✗	✗	✗	✓

Example 4: Ecological Models

A researcher used Ecopath with EcoSim (EwE) as part of their research. The output was used in defining trophic models, ecological niches, and food web. The objective of the study was to determine sensitivity and resiliency of the ecosystem to anthropogenic disturbances, specifically oil spills. The software is maintained available for download via a web service.

In this case, the software is available via a website and need not be submitted, but it is important to provide the link to download the software and indicate the version used in the descriptive information. EwE may also use external datasets to supplement the confirmation parameters, and in this case, the links to the external datasets should also be listed in the descriptive information.

Dataset should include:

1. Input data
2. The configuration file used to initialize the model
3. Output data

Objective: Verification and reproduction of output data.

Objective	Input Data	Setup Files	Model Code	Output Data
Verification and reproduction	✓	✓	✗	✓

Example 5: LES or hypothetical simulation

A study produced multiple outputs of a simulation of a hypothetical volume of ocean. The output was used to study the character of various processes (e.g., mixing, turbulence, eddy formation). The output was produced with documented but proprietary computer code. 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, 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 text file, as appropriate. Because it is proprietary, 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.

Dataset should include:

1. Setup and configuration files
2. Version and information or a reference describing the model code
3. Output data

Objective: Reuse of output data.

Objective	Input Data	Setup Files	Model Code	Output Data
Reuse of output	✗	✓	✗	✓