

Datasheet for an Earth Science Dataset

Released:
Last updated:

Author One
Affiliation
email

Author Two
Affiliation
email

1. PURPOSE

A. For what purpose was the dataset created?

Motivation: Describe the reason for the creation of the dataset (e.g., to provide insight on a knowledge gap, or to carry out some specific task).

B. Who created the dataset (e.g., which individual or team), on behalf of which entity (e.g., institution or company), and under what funding (e.g., grantor[s] and grant number[s])?

Motivation: Provide clarity about the authorship and funding source of the dataset.

C. Was the author of the datasheet involved in creating the dataset? If not, please describe their relation to the dataset.

Motivation: Document the authorship of the datasheet, which may be different than the creator of the dataset.

D. What task(s) has the dataset been used for? Please provide a description and/or citation(s); if there is a repository that archives uses of the dataset, provide a link.

Motivation: Document use cases of the dataset.

E. Any other comments?

Motivation: Space for any other relevant information about the creation of the dataset.

2. STRUCTURE AND PROCESSING

This section concerns technical aspects of the dataset. If documented elsewhere, provide a brief description and stable link (permanent reference, e.g., a DOI) in the relevant question(s).

A. What type(s) of data is/are contained in this dataset? (e.g., model output, observational data, reanalysis, etc.)

Motivation: Basic information about data classification.

B. What is the data? (e.g., file format, dimensionality, variables, metadata, spatiotemporal coverage). Is there important metadata in the data filenames? If so, document this here.

Motivation: Provide format and characteristics of the data.

C. Is this dataset derived from a preexisting dataset? (e.g., variable[s] drawn from a modeling experiment). If so, please describe the process or link to the relevant paper.

Motivation: Describe whether a dataset is drawn or derived from a preexisting dataset.

D. What processing, if any, has been applied to this data? Is any code used to process the data available? If so, please provide a stable link or other method of access.

Motivation: Minimal description of the process to obtain the data described by this datasheet from its unprocessed form.

E. Is any unprocessed data available? If so, please provide a stable link.

Motivation: Clarify the location of unprocessed data to facilitate reproducibility or unforeseen future uses, if possible.

F. Are there relevant known technical issues (e.g., redundancies, errors, missing data)? If so, please provide a description.

Motivation: Provide information about technical issues that affect all or portions of the dataset.

G. Are external resources required to access or use the dataset? If so, please describe them.

Motivation: Track resources required to access or use data.

H. Any other comments?

Motivation: Space for any other relevant information about the structure and processing of the dataset.

3. DISTRIBUTION AND MAINTENANCE

A. Where is the data made available (e.g., Earth System Grid Federation, FTP server, personal communication)? Is there a stable link? If the data is not available, why?

Motivation: Document availability and access to the dataset.

B. Who is/are the point(s) of contact for this dataset?

Motivation: Provide information about who is responsible for responding to inquiries about the dataset.

C. Will the dataset be updated in the future for any reason? If so, will older versions continue to be available?

Motivation: Clarify whether this version of the data is final.

D. What license or other terms of use apply to the dataset? Please link to any relevant licensing terms or terms of use (if in the public domain, simply state this).

Motivation: Provide information about what future uses of the data are permitted.

E. Is there a document that describes an important error in this dataset (e.g., an erratum)? If so, please provide a link or other method of access.

Motivation: Document any corrections to the dataset.

F. Who is hosting the datasheet? Will the datasheet be updated in the future?

Motivation: Document availability and access to the datasheet.

G. Any other comments?

Motivation: Space for any other relevant information about data distribution and maintenance.

A. How was the data generated or collected (e.g., model runs, reanalysis processes, observational measurements)? Please provide relevant citation(s); if none exist, describe why.

Motivation: Establish fundamental information about the methods used to generate or collect data in the dataset.

B. Has the data been assessed against some baseline(s) (e.g., an observational product or physical laws)? If so, describe how, and provide any relevant citations.

Motivation: Document evaluation of the data within the scope of this datasheet.

C. What sources of noise or uncertainty are relevant for this data? Has uncertainty quantification been carried out for this dataset? If so, describe how and provide citation(s).

Motivation: Provide information about known uncertainties.

D. Did the method of generation or collection of the data change within the scope of the dataset?

Motivation: Describe important changes to instruments or methodology within the dataset.

E. Are there any unexplained but relevant numerical values ("magic numbers") that go into the data generation, collection, or processing? (e.g., calibration constants, hyperparameters)

Motivation: Define unique numerical values that exist within or impact this data, but may not be documented elsewhere.

F. Was a model used to generate data in this dataset? If so, please describe the configuration and any modifications.

Motivation: Record the exact model setup used to create data.

G. Is this dataset an ensemble? If so, how many members are there? Are there any differences between members? Describe the perturbation of the members, and any relevant sampling limitations (e.g., ocean states).

Motivation: Describe the sampling, construction, and any important limitations of the ensemble.

H. Are there relevant categories or labels within the data? If so, how are these determined?

Motivation: Define relevant subsets within the data.

I. Can users contribute to this dataset (e.g., citizen science or human labeling)? If so, please describe the process including evaluation or verification.

Motivation: Describe if the data includes user contributions.

J. Are there specific tasks for which the dataset should not be used? If so, please provide a description.

4. DATA-DEPENDENT QUESTIONS

Responses in this section will depend on the type(s) of data within the dataset. Questions that do not apply can be left blank.

Motivation: Address relevant gaps or inadequacies of the data for specific use cases.

K. What are the direct or downstream impacts on humans from this dataset?

The non-comprehensive checklist below is intended to prompt the author to think of common impacts from data. Please check all that apply.

Direct

- ☐ Does this dataset support reproducibility of a specific scientific finding or figure?
- ☐ Were there notable CO₂ emissions in creating this dataset? (e.g., from large machine learning models)
- ☐ Were there notable land use impacts from equipment? (e.g., in situ instruments during a field experiment)
- ☐ Was this dataset created through co-production of research? (e.g., for fieldwork in vulnerable communities)
- ☐ Does this dataset include identifying information? (e.g., community-level data, social information)

Downstream

- ☐ Is this dataset intended for development of a research tool? (e.g., model improvement, sensor design)
- ☐ Does this dataset support further use for novel research? (e.g., unrelated scientific studies)
- ☐ Would analysis of this dataset be policy relevant? (e.g., climate, environmental, public health issues)
- ☐ Would this dataset be considered actionable science? (e.g., completed with use by a specific stakeholder in mind)
- ☐ Could this dataset inspire behavioral changes? (e.g., change agricultural practices, city planning)
- ☐ Could this dataset affect operational forecasting? (e.g., improve models, forecasting, predictability)

Briefly describe your decision-making process for this question with stable links to any references, as available. Additionally, please document any impacts relevant to the scope of the dataset that are not included in the checklist.

L. What biases were present in the construction or use of the dataset?

The non-comprehensive checklist below is intended to prompt the author to think of common biases in data. Please check all that apply.

- ☐ Geographic bias (e.g., restricted or weighted to specific regions)
- ☐ Model bias (e.g., error relative to evaluation product)
- ☐ Sensor bias (e.g., calibration)
- ☐ Temporal bias (e.g., diurnal cycle, restrictions on detection)
- ☐ Seasonal bias (e.g., seasonal cycle)
- ☐ Bias towards extreme or standard conditions (e.g., catchment error, failure to represent extremes)

- ☐ Unbalanced sampling (e.g., unequal classes)
- ☐ Adversarial impacts (e.g., fraudulent samples in crowd-sourced data)
- ☐ Label bias (e.g., subjective labeling)
- ☐ Threshold sensitivity (e.g., for an extreme index)
- ☐ Regime dependence (e.g., convective structure, mode of variability)
- ☐ Selection bias (e.g., case studies, survivorship effects, loss of historical data over time)

Briefly describe your decision-making process for this question with stable links to any references, as available. Additionally, please document any biases within the scope of the dataset that are not included in the checklist.

M. Any other comments? Are there any other citations necessary to document some important aspect of the data? If so, provide the citation(s) and describe their purpose.

Motivation: Space for any other relevant information.

REFERENCES