Datasheet for an Earth Science Dataset

Released:
Last updated:

Author One
Department Name
University Name
Location
email

Author Two
Department Name
University Name
Location
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 research group), 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 given dataset.

C. Was the author of the datasheet involved in creating the dataset? If so, how? If not, please describe your relation to this dataset.

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

D. What tasks 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 permanent reference (stable link, e.g., a DOI) here.

Motivation: Document use cases of the dataset within the scope of this datasheet.

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 this information is documented elsewhere you may simply provide a brief description and stable link in the relevant question(s).

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

Motivation: Basic information about data classification.

B. What is the data? (e.g., file format, dimensionality, variables and 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. What processing has been applied to this data? Is any code used to process the data available? If so, please provide a stable link or other access point.

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

D. Is the unprocessed data available in addition to the processed data? If so, please provide a stable link to the unprocessed data.

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

E. Is this dataset derived from another dataset? If so, please describe the process here or link to the relevant paper.?

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

F. Is any relevant information known to be missing from the dataset? If so, please provide an explanation.

Motivation: Document data missing or lost from the dataset.

G. Are there any sources of noise, redundancies, or errors in the dataset? If so, please provide a description.

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

H. Is the dataset self-contained, or does it rely on external resources? Please describe external resources and any associated restrictions, as well as relevant links or other access points.

Motivation: Explicitly track external dependencies that may otherwise go unacknowledged.

I. Any other comments?

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

3. Distribution and Maintenance

A. Is the dataset available to others? If not, why? If so, how will it be distributed (e.g., FTP, Earth System Grid, personal communication, etc.)? Is there a DOI or other stable link?

Motivation: Document availability and access points 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 this dataset.

C. Will the dataset be updated in the future (e.g., to add new data)? 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 is the dataset distributed under? 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 published document that describes an important error in this dataset (e.g., an erratum)? If so, please provide a link or other access point.

Motivation: Document any corrections to the dataset.

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

Motivation: Document stable access to the datasheet.

G. Any other comments?

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

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.

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. If the data has been assessed against some baseline(s) (e.g., an observational product or physical laws), please describe this assessment. If available, provide the relevant citation.

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

C. Please describe the model configuration and any modifications used to generate data within this dataset.

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

D. Describe relevant uncertainties associated with this data or provide citation(s). If no formal analysis of uncertainties has been completed, then please state this here.

Motivation: Provide information about known uncertainties within the scope of the project.

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

Motivation: Be transparent about important changes to instruments or methodology within the dataset.

F. 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.

G. Is this dataset an ensemble? If so, how many members are there? Are there differences in coverage 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, groupings, or labels within the data? If so, how are these determined?

Motivation: Document group definitions within the data.

I. Can users contribute to this dataset (e.g., citizen science or human labeling)? If so, please describe the process. Will these contributions be evaluated or verified? If so, please describe how. If not, why not?

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.

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 reader to think of common impacts from data. Please check all that apply, and include a brief text description with stable links to any references. Additionally, please document potential impacts relevant to the scope of the dataset that are not included on the checklist.

Motivation: Reflect on the potential impacts (direct or downstream) of the dataset on human systems.

Direct

- ☐ Does this dataset support reproducibility of a specific scientific finding or figure?
- ☐ Were there notable CO_2 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)

L. What biases were present in the construction or use of the dataset? The checklist below provides a non-exhaustive list of common examples in Earth science. Please check all that apply, and include a brief text description with stable links to any references. Additionally, please document any biases within the scope of the dataset that are not included in the checklist.

Motivation: Reflect on potential biases present in the dataset.

- ☐ 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 biases (e.g., seasonal cycle)
- ☐ Bias towards extreme or standard conditions (e.g., catchment error in high winds, failure to represent extremes)
- ☐ Unbalanced sampling (e.g., unequal classes)
- ☐ Adversarial impacts on data (e.g., fraudulent data in crowdsourcing)
- ☐ 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)

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 about the data.

REFERENCES