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. Any other comments?

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

2. COMPOSITION

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 stable link is a permanent reference for an object (e.g., a digital object identifier [DOI]).

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

Motivation: Basic information about the fundamental classification of your data.

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

Motivation: Provide format and characteristics of the data.

C. What processing has been applied to this data?

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 the code used to process the data available? If so, please provide a stable link or other access point.

Motivation: Share processing methodology to facilitate reproducibility, if possible.

F. Is this dataset derived from another dataset? If so, how?

Motivation: Describe whether a dataset is drawn or derived from another preexisting dataset (e.g., field campaign, model intercomparison).

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

Motivation: Describe missing data and be transparent about causes of missing data within the dataset.

H. 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.

I. 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.

J. Any other comments?

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

3. USES

A. 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 the stable link here.

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

B. Is there anything about the construction of the dataset that might impact future uses?

Motivation: Be transparent about how the composition or processing of the dataset could affect future uses.

C. 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.

D. What are the potential impacts of this dataset on humans? Please provide a description as well as a stable link to any supporting documentation.

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

E. Any other comments?

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

4. DISTRIBUTION AND MAINTENANCE

A. How will the dataset be distributed (e.g., FTP server, Earth System Grid, Amazon Web Services, etc.)? Is there a DOI or other stable link?

Motivation: Document stable 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 this dataset.

C. Is the dataset complete or will it be updated in the future (e.g., to add new data, or make corrections)? 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.

5. DATA-DEPENDENT QUESTIONS

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

A. How was the data generated or collected? (e.g., a model used to produce output, reanalysis estimation of conditions, observations using remote sensing methods or in situ sensors) 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 evaluated against some baseline(s) (e.g., an observational product or fundamental physical laws), please describe its evaluation against that baseline(s). If available, simply provide the relevant citation.

Motivation: Document adequacy of the method (e.g., model, remote sensing retrieval) within the scope of this datasheet.

C. Please provide relevant known biases in the generation or collection method of this data and citations as available. This list does not need to be exhaustive, but should include any known biases relevant to the scope of the project the data was created for.

Motivation: Document known biases that pertain specifically to the scope of the project at hand.

D. Please note configurations or modifications made to any model used to complete runs in this dataset (e.g. changes to seasonality, changes to coupling, nudging), or provide relevant startup files.

Motivation: Be transparent about the exact setup of the model to create the data at hand.

E. If this data is restricted to a single point or region, why was this location or region chosen? What are some potential implications of this choice of location on the interpretation of the data?

Motivation: Describe the reasoning for and any relevant impacts of the selection of this location.

F. Describe relevant uncertainties associated with this data or provide relevant 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.

G. Did the method of generation or collection of the data change within the extent of the dataset?

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

H. Are there any relevant unexplained but important numerical values ("magic numbers") that go into the generation, collection, or processing of this data? (e.g., model tuning values, calibration constants, machine learning hyperparameters)

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

I. Is this dataset an ensemble? If so, how many members are there? Describe how the ensemble is perturbed, and whether there are relevant forms of variability that are not dispersed. Are there differences in coverage between the ensemble members?

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

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

Motivation: Be transparent about the processes used to define groups within the data.

K. Can users contribute to this dataset? 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 user contributions make up part of the dataset (e.g., citizen science or human labeling).

L. 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. Can include specific useful citations that do not fall naturally into any other question.

REFERENCES