Datasheet for an Earth Science Dataset

Released: Nov. 08, 2023
Last updated: Nov. 08, 2023

Jamin Rader
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
jamin.rader@colostate.edu

1. PURPOSE

A. For what purpose was the dataset created?

These weighted masks are intended to be used to identify best analogs for analog forecasting. The weighted mask can be multiplied by a state of interest (a state for which you are making a forecast) and a potential analog (a climate state that may evolve similarly to the state of interest). A potential analog with the lowest weighted MSE is msot likely to evolve similarly to the state of interest. These weighted masks were trained using the Max Planck Institute for Meteorology Grand Ensemble (MPI-GE [1]) historical runs. Find details in Rader and Barnes (2023) [**?**] for a description of the data-driven prediction method, training process, and use examples. Two weighted masks are available: exp-Niño(.npz) and exp-NorAtl(.npz). Briefly, exp-Niño is the mask for predicting the SST anomaly in the Niño3.4 region in winter (November-March) given a global map of wintertime SST anomalies the winter prior. Exp-NortAtl is the mask for predicting the 5-year SST anomaly in the North Atlantic given a global map of SST anomalies from the 5 years prior. More information on these experiments can be found in the main text of Rader and Barnes (2023) and Section S1 of the supporting information.

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])?

Jamin K. Rader and Elizabeth A. Barnes, Colorado State University, supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0020347, and by grant AGS-2210068 from the National Science Foundation.

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

Yes, JKR trained the weighted masks and wrote this datasheet.

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.

The weighted masks have been used to identify analogs within MPI-GE. These have been applied to test samples from the MPI-GE itself (Sections 4-5 of Rader et al.) as well as observations (Section S3 in the supporting information for Rader and Barnes, 2023).

E. Any other comments?

N/A

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.?)

Matrices of weights (weighted masks) which can be used for identifying best analogs for the prediction tasks they were trained on.

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.

The data is a matrix of weights (the weighted mask), an array of latitude and an array of longitude. These data are zipped together and saved as an .npz file—NumPy's file format for saving multiple arrays ([2]). The data can be opened in Python using the numpy.load() function and the keys 'mask', 'lat' and 'lon'. The mask has a resolution of 1.875 degrees (latitude and longitude) over the entire globe. The weighted mask is not state-dependent, thus this weighted mask is made to be applied to any initial climate state (over the same period/variable as the weighted mask was trained for, see Section 1.A) to find an analog climate state within the MPI-GE.

C. What processing has been applied to this data?

The weighted mask, as originally learned by the interpretable neural network, was normalized such that the weights sum to 12210—the number of ocean grid points in the mask. All land grid points are set to zero.

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

No.

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

DOI: XXX

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

The weighted mask is trained using the historical run of the MPI-GE with the forced response removed.

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

N/A

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

Noise in the weighted mask, a product of the neural network training, can be removed manually as in Section 5 of Rader and Barnes (2023).

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.

The weighted mask has been designed to identify best analogs within the MPI-GE historical run with the forced response removed. Data for MPI-GE can be located at https://esgf-data.dkrz.de/projects/mpi-ge/.

J. Any other comments?

N/A

3. 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?

Zenodo DOI: XXX.

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

Jamin K. Rader, jamin.rader@colostate.edu

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?

This dataset is complete. Further iterations of neural network-learned weighted masks may arise in future publications, but those masks will not be kept here.

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).

Public use.

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.

N/A

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

Zenodo DOI: XXX. This datasheet will not be updated.

G. Any other comments?

N/A

4. 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.

An interpretable neural network was trained to identify the important regions for determining whether two climate states within the internal variability of MPI-GE historical run would evolve similarly. The weighted mask is an output of this interpretable neural network, and can then be used for analog forecasting as in Rader and Barnes (2023)

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.

The weighted mask has been shown to provide better analog forecasts for 1-year predictions of the tropical Pacific and 5-year predictions of the North Atlantic than traditional analog approaches. In addition, when applied to observations, it has comparable skill to initialized decadal prediction systems, as in Section S3 of Rader and Barnes (2023).

C. 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.

N/A

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.

While these weighted masks have been shown to improve analog forecasting, we have not quantified the uncertainty in these weights.

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

N/A

F. 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)

Yes, the specific code for creating these weighted masks can be found in the Zenodo repository. The saved weighted masks used a random seed of 0.

G. 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?

N/A

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

N/A

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?

No. These weighted masks are complete.

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

These weighted masks were trained within the simulated climate of the MPI-GE, and thus they may not represent the precursor patterns seen in the true Earth System. Please refer to Section 6 of Rader and Barnes (2022).

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.

The weighted mask methodology can be used to improve operational analog forecasts by identifying analogs using the important precursor regions rather than the similarity between two maps over the entire globe or a single predefined region of the globe.

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.

The weighted mask is created within the space of the MPI-GE. Model biases within MPI-GE may cause this mask to be not optimal for the observed Earth System.

☐ Geographic bias (e.g., restricted or weighted to specific regions)
✓ Model bias (e.g., error relative to observations or other ground truth product)
☐ Sensor bias (e.g., calibration)
☐ Day/night bias (e.g., diurnal cycle, restrictions)
☐ 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., incorrect or incomplete labeling)
☐ Threshold sensitivity (e.g., for an extreme index)
☐ Regime dependence (e.g., convective structure, mode of variability)

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.

N/A

REFERENCES

[1] N Maher, S Milinski, L Suarez-Gutierrez, and others. The max planck institute grand ensemble: enabling the exploration of climate system variability. *Journal of Advances*, 2019.

[2] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández Del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.