Datasheet for an Earth Science Dataset

Released: September 18 2024
Last updated: October 8 2024

Marybeth C. Arcodia
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
marcodia@colostate.edu

1. Purpose

A. For what purpose was the dataset created?

This datasheet describes the data used in the following publication: Arcodia, Marybeth C., Elizabeth A. Barnes, Paul J. Durack, Patrick W. Keys, Juliette Rocha, 2024. "Sea Surface Salinity Provides Subseasonal Predictability for Forecasts of Opportunity of U.S. Summertime Precipitation".

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])?

Two datasets were used in this study. 1) The Community Earth System Model version 2 Large Ensemble (CESM2-LENS). The data are available freely to the public at https://www.cesm.ucar.edu/community-projects/lens2. 2) ECMWF Reanalysis version5 (ERA5) data are available freely at https://cds.climate.copernicus.eu/.

C. Was the author of the datasheet involved in creating the dataset? If so, how?

All data was downloaded from the above websites.

D. Any other comments?

n/a

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 (e.g., digital object identifier [DOI]) 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.?)

The CESM2 data contain model ouput from 10 ensemble members from the large ensemble dataset. The ERA5 data are reanalysis data.

B. What is the data? (e.g., file format, dimensionality, variables and metadata, spatiotemporal coverage)

C. What processing has been applied to this data?

For the CESM2 data, daily data are interpolated to 2.5 by 2.5 degree resolution via bilinear interpolation and selected from May-August from 1850-1949 for CESM2. Sea surface salinity anomalies (SSSAs) are computed via subtraction of the linear trend at each grid point of the ensemble mean for each calendar-day of the year across the 100 years of simulation to remove the forced response. The data are normalized by subtracting the mean and dividing by the standard deviation at each grid point. The moisture-tracking model uses European Centre for Medium-Range Weather Forecasts v5 (ERA5; Hersbach et al. 2020) reanalysis data, including hourly, 2-dimensional surface pressure, evaporation and precipitation, and hourly 3-dimensional specific humidity, and zonal and meridional winds with 0.25deg resolution. We use data from 2008-2021 for this analysis.

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

The unprocessed data is available via the links above.

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

The processed data can be computed via the Python scripts found at https://github.com/mbarcodia/paper_salinity_s2s_predictor/.

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

n/a

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

We use climate model and reanalysis data so there are no known missing data points.

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

There are no known sources of errors within the data used. However, any data-specific noise, redundancies, or errors can be found within the dataset documentation (links provided above).

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.

No external resources are needed for this data.

J. Any other comments?

n/a

3. Uses

A. What tasks has the dataset been used for?

Artificial neural networks are trained to ingest maps of sea surface salinity anomalies to classify precipitation into light or heavy events over the U.S. Midwest at leads of 0-56 days.

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

n/a

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

n/a

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.

The datasets used are representations of the climate system, but do not represent the true climate system perfectly. Therefore, conclusions for human systems should be taken within that context.

E. Any other comments?

n/a

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?

Two datasets were used in this study. 1) The Community Earth System Model version 2 Large Ensemble (CESM2-LENS). The data are available freely to the public at https://www.cesm.ucar.edu/community-projects/lens2. 2) ECMWF Reanalysis version5 (ERA5) data are available freely at https://cds.climate.copernicus.eu/.

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

Information on the current point(s) of contact for both datasets are included in the links above.

C. Is the dataset complete or will it be updated in the future?

The data processed for this study is final. However, there will be continued updates to both raw datasets in the future.

D. Is the dataset receiving ongoing maintenance? If so, please provide one or more point(s) of contact and describe the method (if any) by which updates would be communicated to users.

The data processed for this study is final. However, there will be continued updates to both raw datasets in the future.

E. 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).

The data and code provided on the Github link operate under a MIT license.

F. Is there an erratum? If so, please provide a link or other access point.

Any errata for the data can be found in the links to the raw datasets provided above.

G. Will older versions of the dataset continue to be available? If so, please describe where.

All data used in this study are final.

H. Who is hosting the datasheet? Is the datasheet receiving ongoing maintenance?

This datasheet can be found at https://github.com/mbarcodia/paper_salinity_s2s_predictor/. The datasheet may be updated after acceptance for publication, but it will not receive routine updates

I. Any other comments?

n/a

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.

Two datasets were used in this study. 1) The Community Earth System Model version 2 Large Ensemble (CESM2-LENS). The data are available freely to the public at https://www.cesm.ucar.edu/community-projects/lens2. This climate model uses a large ensemble approach to produce the model output. More information can be found at the documentation link: https://www.cesm.ucar.edu/community-projects/lens2. 2) ECMWF Reanalysis version5 (ERA5) data are available freely at https://cds.climate.copernicus.eu/. The reanalysis uses data assimilation and is produced on the ECMWF high-performance computing facility. Additional information can be found at https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3803.

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 citation for the article on CESM2 can be found at https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001916 [1]. The citation for the article on ERA5 can be found at https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3803 [2].

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.

Any known biases can be found in the articles cited in the answer to question 5B.

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.

Any known configurations or modifications can be found in the articles cited in the answer to question 5B.

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?

n/a

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.

Any known uncertainties and their quantifications can be found in the articles cited in the answer to question 5B.

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

No.

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)

The architecture is identical for networks trained for predictions from leads of 0-35 days and then a slightly different architecture was used for leads of 42-56 days. Hyperparameter tuning was performed using the KerasTuner to find the optimal set of parameters determined via validation accuracy. For the shorter lead forecasts, the network architecture consists of 1 hidden layer with 128 nodes with a rectified linear activation function applied (ReLU), a dropout rate of 50% and ridge regression coefficient of 0.1 to reduce overfitting, batch size of 32 samples, and a learning rate of 1.618e-5. For the longer lead forecasts beyond 35 days, the network architecture consists of 2 hidden layers with 160 and 192 nodes with a rectified linear activation function applied (ReLU) to each, a dropout rate of 80% and ridge regression coefficient of 0.01 to reduce overfitting, batch size of 32 samples, and a learning rate of 2.886e-6. All networks have a set global seed of 147483648 and are initialized with the following random seeds: 6, 26, 19, 54, 68. Networks are trained using the categorical cross-entropy loss function. Networks are trained with early stopping when the validation loss does not decrease after 25 epochs. We note that for the lead of 7 days, the network architecture with the highest validation accuracy was slightly different than the one used here. However, the same architecture which resulted in

the highest validation accuracy for leads 0, 14, 21, 28, and 35 resulted in a validation accuracy on the order of 0.001 less than the highest performing architecture. Therefore, we used the same architecture for all leads 0-35 days for simplicity. A table with the hyperparameter search can be found in the Supplemental Material.

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?

The CESM2 Large Ensemble (LENS2) consists of 100 members at 1-degree spatial resolution covering the period 1850-2100 under CMIP6 historical. Only 10 ensemble members are used in this study (more information on the ensemble members chosen for the study can be found in the manuscript). Additional documentation on the perturbations to the ensemble members can be found at https://www.cesm.ucar.edu/community-projects/lens2. The exact ensemble members used can be found in a table in the Supplemental Material.

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

Documentation on the categories and groupings of ensemble members can be found at https://www.cesm.ucar.edu/community-projects/lens2.

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?

Contributions by users to the datasets described would need to be coordinated with the individual data providers.

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.

n/a

REFERENCES

[1] Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.

[2] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.