Datasheet for Marine Heatwaves under ARISE-SAI-1.5 (MHWs-ARISE-1.5)

Released:
Last updated: April 23, 2023

Daniel Hueholt
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
daniel.hueholt@colostate.edu

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). This dataset was created to investigate changes in marine heatwaves (MHWs) under scenarios of stratospheric aerosol injection (SAI) climate intervention and climate change. The conventional [1] definition of an MHW is used: an MHW is defined when sea surface temperature (SST) at a point exceeds the local 90th percentile for five or more days. The data is derived from daily-timestep SST data from the ARISE-SAI-1.5 [2] SAI modeling experiment using the [3] Python implementation of the [1] definition. ARISE-SAI-1.5 depicts one plausible scenario of climate intervention where SAI is used to maintain the 1.5°C target of the Paris Accords alongside moderate mitigation (SSP2-4.5) greenhouse gas forcing. MHWs have severe ecological impacts which are expected to worsen under climate change, and the impacts of SAI in the ocean are largely unknown.

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. This dataset was created by Daniel Hueholt at Colorado State University with support from the National Science Foundation Graduate Research Fellowship Program. The data was processed using the Colorado State University Walter Scott College of Engineering High-Performance Computing cluster ("ASHA"). This work was carried out for collaboration with Gouri Anil (Louisiana State University), Cheryl Harrison (LSU), Lala Kounta (Michigan State University), and Phoebe Zarnetske (MSU) under the Climate Intervention Biology Working Group. Multiple future analyses of this dataset by this team of people are currently planned and underway.

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. Yes–the datasheet was written by Daniel Hueholt, who also created 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. This dataset consists of marine heatwave properties derived from the ARISE-SAI-1.5 SAI experiment. The data consists of parallel ensembles of model output using the Community Earth System Model version 2 with Whole Atmosphere Community Climate Model version 6 (CESM2[WACCM6]). The ocean is represented using the Parallel Ocean Program version 2 (POP2), with interactive biogeochemistry through the Marine Biogeochemistry Library

(MARBL), and ocean surface waves from NOAA WaveWatch-III (version 3.14). One set of 10 ensemble members follows a no-SAI climate change trajectory (SSP2-4.5, moderate mitigation); a second set of 10 ensemble members uses SAI injected from multiple latitudes to maintain global mean temperatures at 1.5°C above pre-industrial as well as the pole-to-pole temperature gradient and pole-to-equator temperature gradient. See [2] for more information about ARISE-SAI-1.5.

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. All data files are in netCDF files. No-SAI (SSP2-4.5) files are identified by the token ***BWSSP245*** in the filename. Files where SAI is used are identified by the token ***SSP245-TSMLT-GAUSS-DEFAULT***.
The dataset contains two types of files: *frequency files* and *property files*.

- **Frequency files**
  - **Filename token:** *mhws*
  - **Dimensions:** time x lat x lon
  - **Description:** binary timeseries of MHW presence/absence (1/0) at each timestep at each gridpoint. Times are in serial datetime format (as required by the MHW calculation)–this can be converted to human-readable datetime using the `datetime` Python package.
- **Property files**
  - **Filename token:** *mhwProps*
  - **Dimensions:** event x lat x lon
  - **Description:** contains properties (start serial datetime, end serial datetime, mean intensity, duration, and numerical category [1]) for each MHW event that occurs at a given lat/lon point. Because the number of MHW events varies for each grid point, and Python xarray does not support jagged arrays, the event dimension has been padded to the same length for each point. Only the first n (for n=number of MHWs at a point) entries are meaningful; entries past this index are numpy NaN values.

In this dataset, we define MHWs relative to a fixed 2015-2024 climatology at each point. While a 10 year time period is normally too short to define a climatological baseline in the presence of internal variability, the ARISE-SAI-1.5 10-member ensemble size effectively increases the number of years available to calculate the climatology to 100 years (e.g., [4]). This is a sufficient period of time to calculate a meaningful climatology.

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. The processing workflow uses code written by DMH (wrapping the [3] marineHeatWaves package by Oliver et al.) stored in the mhws-esm repository on Github: github.com/dmhuehol/mhws-esm.

1) Obtain daily-timestep SST data (see Question 2D)
2) Use `remap_ocean_data` to remap data from POP Arakawa B-grid to lat/lon coordinates
3) Use `run_selyear` to separate out files during the 2015-2024 climatology reference period
4) Run `shatter` with the `calcRlzMn` flag set to True and `calcEachRlz` flag set to False to break the reference files into individual ensemble mean latitude shards used for the climatology reference
5) Run `shatter` with the `calcRlzMn` flag set to True and `calcEachRlz` flag set to False to break the reference files into individual latitude shards for each realization used to calculate MHWs
6) Run `calc_shard_mhws` to make shard mhw files for each realization
7) Run `mending` to combine the shard mhw files into global mhw files

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. Unprocessed SST data from ARISE-SAI-1.5 can be obtained from the NCAR Climate Data Gateway doi.org/10.5065/9kcn-9y79 (all SAI, 5 no-SAI members) and doi.org/10.26024/0cs0-ev98 (remaining 5 no-SAI members); or Amazon Web Services registry.opendata.aws/ncar-cesm2-arise/.

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. The code used to process the data can be found at the mhws-esm repository on Github: github.com/dmhuehol/mhws-esm. This package itself relies on the marineHeatWaves implementation of the [1] MHW definition [3]: github.com/ecjoliver/marineHeatWaves. Note that even with parallel processing, the computational burden to calculate MHWs on the entire ARISE-SAI-1.5 dataset is high and is very time-consuming (days) without access to an HPC environment.

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). This dataset is derived from daily-timestep SST data in the ARISE-SAI-1.5 SAI experiment [2].

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. No data known to be missing.

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. None known to exist from first principles, but the dataset has yet to be analyzed. If errors are found during the research process, they will be documented in this datasheet.

Note sources of noise, redundancies, or errors may exist elsewhere in the greater ARISE-SAI-1.5 dataset–this datasheet documents only the specific data in MHWs-ARISE-1.5.

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. The dataset is self-contained. The marineHeatWaves package [3] used to calculate MHWs is not maintained by the authors, and they make no claim to its upkeep.

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. This dataset was created by DMH for multiple planned analyses by the Climate Intervention Biology Working Group. The dataset will be made public upon publication of these works.

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. The dataset only contains a subset of possible properties about MHWs, as documented in Question 2B. Analysis beyond these variables would require further data creation. Additionally, this dataset strictly contains data about MHW frequency and properties–further data will be needed to explore reasons for and impacts of changes to MHWs shown in this dataset.

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. ARISE-SAI-1.5 depicts only one scenario of SAI deployment. Since the response to SAI deployment in models is known to be dependent on the design of the scenario, results from this dataset must be considered in context of this specific scenario–they cannot be assumed to be true of any general SAI deployment.

The [1] definition of MHWs does not consider the effects of sea ice, and should be used with extreme caution where sea ice is present.

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. This dataset was created to explore changes to MHWs under one scenario of stratospheric aerosol injection. Following the 2021 U.S. National Academies of Sciences report [5] we explicitly distinguish this goal from research on the practical deployment of SAI, about which critical ethical and governance concerns exist.

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. TBD upon publication of data

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. TBD upon publication of data

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. TBD upon publication of data

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. TBD upon publication of data

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. None

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

Motivation: Document stable access to the datasheet. TBD upon publication of data

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. The data is derived from ARISE-SAI-1.5, consisting of model output generated by the Community Earth System Model version 2 with Whole Atmosphere Community Climate Model version 6 (CESM2[WACCM6]). See Question 2A or [2] for further details.

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. While CESM2(WACCM6) does simulate MHWs, the model representation has considerable errors relative to observations [6]. On average, modeled MHWs tend to be overly long-lasting and too weak.

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. In general, Earth system model output is most reliable where long-term observational datasets are present to evaluate the model behavior. SST data is sparse over large sections of the ocean, particularly the poles and the Southern Ocean. Additionally, SST representation in coarse-grid Earth system models tends to be less reliable in regions where small-scale features are important to local dynamics (e.g., near coastlines and around mesoscale currents or eddies).

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. Data was derived from the model runs in [2]; no original model output was created as part of this dataset.

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. Data in MHWs-ARISE-1.5 has global coverage.

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. As ARISE-SAI-1.5 is a single-model initial condition large ensemble, the spread of data among different realizations represents a range of inter-

nal climate variability. General uncertainty quantification for CESM2(WACCM6) is carried out in [7].

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. No known changes to the method of generation within the extent of 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. The length of the "event" dimension used for the mhwProps data (start time, end time, duration, mean intensity, category) is the length of the time dimension divided by 7. This pads the "events" to include the theoretical maximum number of MHWs at a point (as individual MHWs are 5-day events separated by at least 2 days).

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. ARISE-SAI-1.5 is a single model initial condition large ensemble with 10 SAI members (6 members 2035-2069, 4 members 2035-2070) and 10 no-SAI members (4 members 2015-2069, 1 member 2015-2070, 5 members 2015-2100). The ensemble is perturbed both by microperturbations to the atmospheric initial conditions, and by choosing different ocean state initializations. Thus, the data samples internal climate variability in both the atmosphere and ocean.

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. Beyond the natural division of SAI and no-SAI ensemble members, there are no relevant categories, groupings, or labels 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). User contributions are not a part of this dataset.

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

[1] Alistair J. Hobday, Lisa V. Alexander, Sarah E. Perkins, Dan A. Smale, Sandra C. Straub, Eric C.J. Oliver, Jessica A. Benthuysen, Michael T. Burrows, Markus G. Donat, Ming Feng, Neil J. Holbrook, Pippa J. Moore, Hillary A. Scannell, Alex Sen Gupta, and Thomas Wernberg. A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, 141:227–238, February 2016.

[2] Jadwiga H. Richter, Daniele Visioni, Douglas G. MacMartin, David A. Bailey, Nan Rosenbloom, Brian Dobbins, Walker R. Lee, Mari Tye, and Jean-Francois Lamarque. Assessing Responses and Impacts of Solar climate intervention on the Earth system with stratospheric aerosol injection (ARISE-SAI): protocol and initial results from the first simulations. *Geoscientific Model Development*, 15(22):8221–8243, November 2022. Publisher: Copernicus GmbH.

[3] Eric Oliver. Marine Heatwaves detection code, May 2022. original-date: 2015-04-28T07:33:04Z.

[4] Claudia Tebaldi, Kalyn Dorheim, Michael Wehner, and Ruby Leung. Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates. *Earth System Dynamics*, 12(4):1427–1501, December 2021. Publisher: Copernicus GmbH.

[5] NASEM. Reflecting Sunlight: Recommendations for Solar Geoengineering Research and Research Governance. Technical report, National Academies of Science, Engineering, and Medicine, 2021.

[6] Zijian Qiu, Fangli Qiao, Chan Joo Jang, Lujun Zhang, and Zhenya Song. Evaluation and projection of global marine heatwaves based on CMIP6 models. *Deep Sea Research Part II: Topical Studies in Oceanography*, 194:104998, December 2021.

[7] G. Danabasoglu, J.-F. Lamarque, J. Bacmeister, D. A. Bailey, A. K. DuVivier, J. Edwards, L. K. Emmons, J. Fasullo, R. Garcia, A. Gettelman, C. Hannay, M. M. Holland, W. G. Large, P. H. Lauritzen, D. M. Lawrence, J. T. M. Lenaerts, K. Lindsay, W. H. Lipscomb, M. J. Mills, R. Neale, K. W. Oleson, B. Otto-Bliesner, A. S. Phillips, W. Sacks, S. Tilmes, L. van Kampenhout, M. Vertenstein, A. Bertini, J. Dennis, C. Deser, C. Fischer, B. Fox-Kemper, J. E. Kay, D. Kinnison, P. J. Kushner, V. E. Larson, M. C. Long, S. Mickelson, J. K. Moore, E. Nienhouse, L. Polvani, P. J. Rasch, and W. G. Strand. The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001916.