Datasheet for Connolly et al. submitted "Using Neural Networks to Learn the Forced Response of the Jet Stream to Tropospheric Temperature Tendencies"

Released: November 17, 2022
Last updated: January 17, 2023

Charlotte Connolly
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
cconn@rams.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 contains two different kinds of data, a long control run from the CESM dry dynamical core and 18 thermally forced dry core runs. Data was created for the study Connolly et al. (in prep; DOI:), referred to throughout as C23. The long control run was created to train a convolutional neural network (CNN) to learn the correlation between a temperature tendency and a jet response. The thermally forced heating experiments were created to use a direct comparison between the CESM dry dynamical core and the trained CNN.

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.

Charlotte Connolly studying at Colorado State University and supported by NSF CAREER AGS-1749261 under the Climate and Large-scale program ran the CESM dry dynamical core to create the 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.

The writer of this datasheet was involved in the creation of the dataset. The author of the datasheet (Charlotte Connolly) ran the CESM dry dynamical core to create the data.

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 any 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.?)

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

This dataset contains data from two different styles of model runs from the CESM dry dynamical core, a long control run and 18 thermally forced dry core runs. The long control run contains time series of a zonally averaged smoothed temperature tendency (K/day), a smoothed jet stream location (deg. latitude), and a change in jet stream location (deg. latitude). The thermally forced dry core runs include a time series of the jet stream location from each of the 18 heating experiments. The composition of the saved data is outlined in the next question and detailed information about the data can be found in C23.

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

Motivation: Provide format and characteristics of the data.

Data is saved in netCDF format and contains four groups, Training_data, Validation_data, Testing_data, Heating_Experiements.

- Training_data group contains the data used to train the convolutional neural network (CNN) used in C23.
  - Variable Name: dTdt_training
    - Description: contains a zonally averaged smoothed temperature tendency. Each hemisphere is used as a separate sample. The zonally averaged smoothed temperature tendency is used as an input to train the CNN in C23.
    - Coverage: Stratosphere (200 hPa and above), removed to place focus on troposphere. Since each hemisphere is considered a separate sample and the hemispheres and symmetrical in a dry core, the latitude dimension goes from 0°-90°
    - Dimensions: (359280, 25, 32), where 359280 corresponds to the amount of samples used to train the CNN, 25 is the pressure levels (hPa), and 32 is the latitude bands (deg. latitude).
  - Variable Name: jet_lat_training
    - Description: contains the latitudinal location of the jet-stream (deg. latitude) calculated from zonally averaged zonal winds. The jet-stream location is used as an input to train the CNN in C23.
    - Dimensions: (359280), where 359280 corresponds to the amount of samples used to train the CNN in C23.
  - Variable Name: jet_shift_training
    - Description: contains the change in jet stream's location across the next 360 time steps and is used as the output to train the CNN in C23.
    - Dimensions: (359280), where 359280 corresponds to the amount of samples used to train the CNN in C23.
- Validations_data group contains the data used to validate the CNN.
  - Variable Name: dTdt_validation
    - Description: contains a zonally averaged smoothed temperature tendency. Each hemisphere is used as a separate sample. This data is used to validate the CNN in C23.
    - Coverage: Stratosphere (200 hPa and above), removed to place focus on troposphere. Since each hemisphere is considered a separate sample and the hemispheres and symmetrical in a dry core, the latitude dimension goes from 0°-90°
    - Dimensions: (199280, 25, 32), where 199280 corresponds to the amount of samples used to train the CNN, 25 is the pressure levels (hPa), and 32 is the latitude bands (deg. latitude).
  - Variable Name: jet_lat_validation
    - Description: contains the latitudinal location of the jet-stream (deg. latitude) calculated from zonally averaged zonal winds. This data is used to validate the CNN in C23.
    - Dimensions: (199280), where 199280 corresponds to the amount of samples used to validate the CNN in C23.
  - Variable Name: jet_shift_validation
    - Description: contains the change in jet stream's location across the next 360 time steps and is used as the output to validate the CNN in C23.
    - Dimensions: (199280), where 199280 corresponds to the amount of samples used to validate the CNN in C23.
- Testing_data group contains the data used to test the CNN.
  - Variable Name: dTdt_testing
    - Description: contains a zonally averaged smoothed temperature tendency. Each hemisphere is used as a separate sample. This data is used to test the CNN in C23.
    - Coverage: Stratosphere (200 hPa and above), removed to place focus on troposphere. Since each hemisphere is considered a separate sample and the hemispheres and symmetrical in a dry core, the latitude dimension goes from 0°-90°
    - Dimensions: (1399558, 25, 32), where 1399558 corresponds to the amount of samples used to test the CNN, 25 is the pressure levels (hPa), and 32 is the latitude bands (deg. latitude).
  - Variable Name: jet_lat_testing
    - Description: contains the latitudinal location of the jet-stream (deg. latitude) calculated from zonally averaged zonal winds. This data is used to test the CNN in C23.
    - Dimensions: (1399558), where 1399558 corresponds to the amount of samples used to test the CNN in C23.
  - Variable Name: jet_shift_testing
    - Description: contains the change in jet stream's location across the next 360 time steps and is used as the output to test the CNN in C23.
    - Dimensions: (1399558), where 1399558 corresponds to the amount of samples used to test the CNN in C23.
- The Heating_Experiments group contains the 18 thermal forcing patterns used to force the CESM dry dynamical core as well as the resulting jet location. The heating experiments are outlined in C23
  - Variable Name: thermal_forcing

    * Description: contains the zonally averaged thermal forcing used to force the additional heating experiments in the CESM dry dynamical core outlines in C23.
    * Dimensions: (18, 25, 32), where 18 corresponding to the 18 heating experiments used in C23. Table A1 #N-N in C23 contains the sizes and locations of the thermal forcings and Section NN describes how the heating experiements are created using Gaussians. 25 is the pressure levels and 32 is the latitude bands.
- Variable Name: heating_experiment_simulation_length
    * Description: contains a time series of the jet stream locations from each of the 18 forced heating experiments calculated from zonally averaged zonal winds.
    * Dimensions: (18, 16000), 18 corresponding to the 18 heating experiments used in C23 and are in the same order as in variable thermal_forcing. 16000 corresponds to the amount of time steps retained after the first 400 time steps are removed allow the system to reach its new equilibrium after a forcing is imposed.

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.

Methods described in detail in C23 (DOI:)

Briefly, from the long control run, zonally averaged temperature and zonally averaged zonal winds are used. Using a backward running mean, a temperature tendency is calculated from the zonally averaged temperature. The zonally averaged temperature tendency and zonally averaged zonal winds are smoothed with a 60 day running average.

The smoothed zonally averaged zonal winds are used to calculate the jet location by finding the max wind speed at the 850 hPa level.

The jet shift is calculated by subtracting the jet location 90 days ago from the current jet location. The 90 days ensure no overlap between the input and output caused by the 60 day running mean. This results in a 30 day gap between input and output.

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.

Due to the large size of the data, the raw data was not saved.

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 is available in GitHub. (not actually available yet. Link and how to find the code will be added once available.)

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).

Not derived from a larger dataset.

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 missing or mislabeled data.

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.

No errors, sources of noise, or redundancies in the data.

I. Is the dataset self-contained, or does it rely on external resources? Please provide descriptions of 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.

Dataset is self contained, no older versions of the data exist (i.e. no archive), and there are no restrictions.

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.

Data has not been used for any other tasks.

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.

Only the processed temperature tendency, the jet stream location, and the jet stream shift were saved. This limits future uses of data to uses that focus on the relationship between a temperature tendency and a jet location.

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.

The saved variables limits the use of this data to solely be used to investigate the jet response to a temperature tendency. In addition, the simulations contain no seasons, diurnal cycle, or surface topography. This further limits future applications of the data as this data can not be used to investigate seasonality, and influences of surface topography on the jet response.

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.

The dry dynamical core simulates only dry dynamics and while that is adequate for the work in C23 and many other dynamical studies, this data misses important feedbacks in the Earth's atmosphere that could have potentially important impacts. Therefore, conclusions that pertain to climate change using this data should be considered in the context of the dry core and elsewhere before any climate policies or climate change mitigation strategies are created and/or implemented.

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.

The work from C23 is in prep and data will be made available upon submission.

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.

Charlotte Connolly, created that dataset and the datasheet and is the point of contact (email: cconn@rams.colostate.edu).

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

Motivation: Clarify whether this version of the data is final.

The dataset is complete and will not be updated 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.

Motivation: Provide information about whether the dataset is receiving ongoing support.

The dataset is not receiving ongoing maintenance.

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).

Motivation: Provide information about what future uses of the data are permitted.

The work from C23 is in prep, this information will be made available upon submission.

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

Motivation: Document any corrections to the dataset.

No erratum.

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

Motivation: Describe whether any specific version of the dataset will always be accessible.

There are no older versions of this dataset.

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

Motivation: Clarify stable access to the datasheet and whether it will be updated.

The work from C23 is in prep and data will be made available upon submission.

I. 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 Community Earth System Model CAM Eulerian spectral-transform dynamical core with the Held Suarez setup was used to generate the data.

- https://doi.org/10.1002/2014MS000329 : a paper that describes the dry dynamical core.
- https://www.cesm.ucar.edu/models/simpler-models/dry-dynamical-core.html : CESM page that details the dry dynamical core used to simulate this data as well as the default parameteres and the dry core set up. This is not a stable link and can break.

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.

The dry dynamical core has been previouslt shown to reproduce the majority of the Northern Hemisphere's jet response to heating perturbations despite simulating only dry dynamics (Mbengue and Schneider, 2013).

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.

There are no relevant additional biases beyond the discussion in question 3D.

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.

The CAM Eulerian spectral-transform dynamical core was ran with the held-suarez configuration (FHS94) with the default configurations. The model was ran with T42 resolution and 30 pressure levels (T42z30 T42 mg17) with 6 hour time steps. The long control run was ran for 1,000,000 time steps.

The 18 thermally forced experiments are initiated from the end of the long control run by imposing a temperature tendency. The forced experiments are ran under the same conditions as the long control run for an additional 20,000 time steps.

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.

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.

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

N/A

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.

The method of generating did not change.

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 data is smoothed with a 60 day running mean. This value was specific to the goal of C23 to learn a forced response from internal variability.

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.

N/A

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.

N/A

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).

Users cannot contribute to the data.

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

- Connolly, C., Barnes, E. , Hassanzadah, P., Prichard M., submitted: Using Neural Networks to Learn the Forced Response of the Jet-Stream to Tropospheric Temperature Tendencies. Artif. Intel. Earth System.
- Evans, K. J., and Coauthors, 2014: A spectral transform dynamical core option within the commu- nity atmosphere model (CAM4). J. Adv. Model. Earth Syst., 6 (3), 902–922.
- Mbengue, C., and T. Schneider, 2013: Storm track shifts under climate change: What can be learned from Large-Scale dry dynamics. J. Clim., 26 (24), 9923–9930.