Datasheet for Connolly et al. 03/2023 "Using Neural Networks to Learn the Forced Response of the Jet Stream to Tropospheric Temperature Tendencies"

Released: November 17, 2022
Last updated: August 25, 2024

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).

Data contains two different kinds of data, a long control run from the Community Earth System Model (CESM) dry dynamical core and 18 thermally forced dry dynamical core runs. The long control run was created to train a neural network to learn the correlation between a thermal forcing and a jet response. A long control run with enough data to train a neural network did not exist before completing this project. The thermally forced heating experiments were created to use as a direct comparison between the dry dynamical core and the CNN from Connolly et al. 2023 (DOI: 10.1175/AIES-D-22-0094.1), referred to throughout datasheet as C23.

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 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 not, please describe their relation to the 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. 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 link.

Motivation: Document use cases of the dataset.

This dataset was generated to train the C23 convolutional neural network (CNN). To date, this data has not been used for any other tasks.

E. Any other comments?

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

2. Structure and Processing

This section concerns technical aspects of the dataset. If documented elsewhere, provide a brief description and stable link (permanent reference, e.g., a DOI) in the relevant question(s).

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

Motivation: Basic information about data classification.

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, metadata, spatiotemporal coverage). Is there important metadata in the data filenames? If so, document this here.

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 C23 convolutional neural network (CNN).
  - Variable Name: dTdt_training
    - Description: contains 359280 smoothed (via running average temperature tendency [HOW MUCH]) zonally averaged temperature tendency samples which are used as an input to train the C23 CNN.
    - 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) spanning from 0°-90°latitude. Since the hemispheres are symmetrical in a dry dynamical core, each hemispehere is used as an individual sample.
    - Coverage: Each sample includes only information from the troposphere. Stratosphere, defined as 200 hPa and above, is removed to focus on learning tropospheric dynamics.
  - 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 C23 CNN.
    - 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 199280 smoothed (via running average temperature tendency [HOW MUCH]) zonally averaged temperature tendency samples which are used as an input to validate the C23 CNN.
    - Coverage: Each sample includes only information from the troposphere. Stratosphere, defined as 200 hPa and above, is removed to focus on learning tropospheric dynamics.
    - 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. Is this dataset derived from a preexisting dataset? (e.g., variable[s] drawn from a modeling experiment). If so, please describe the process or link to the relevant paper.

Motivation: Describe whether a dataset is drawn or derived from a preexisting dataset.

This dataset is not derived from a preexisting dataset.

D. What processing, if any, has been applied to this data? Is any code used to process the data available? If so, please provide a stable link or other access point.

Motivation: Minimal description of the process to obtain the data described by this datasheet from its unprocessed form.

Methods described in detail in C23.

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 (60 day) zonally averaged zonal winds are used to calculate the jet location by finding the max wind speed from the 850 hPa level from the smoothed zonal winds.

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.

E. Is any unprocessed data available? If so, please provide a stable link.

Motivation: Clarify the location of unprocessed data to facilitate reproducibility or unforeseen future uses, if possible.

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

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

Motivation: Document data missing or lost from the dataset.

There is no information know to be missing from the dataset.

G. Are there known 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 known errors, sources of noise, or redundancies in the data.

H. Is the dataset self-contained? If external resources are involved, please describe them and any associated restrictions.

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.

I. Any other comments?

Motivation: Space for any other relevant information about the structure and processing of the dataset.

3. Distribution and Maintenance

A. Is the dataset available to others? If not, why? If so, how will it be distributed (e.g., FTP, Earth System Grid, personal communication)? Is there a stable link?

Motivation: Document availability and access to the dataset.

Data can be found on Zenodo titled "Data accompanying Using Neural Networks to Learn the Forced Response of the Jet-Stream to Tropospheric Temperature Tendencies". The link is https://doi.org/10.5281/zenodo.7796170

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

Motivation: Provide information about who is responsible for responding to inquiries about the dataset.

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

C. Will the dataset be updated in the future (e.g., to add new data)? If so, will older versions continue to be available?

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

The dataset will not be updated in the future.

D. What license or other terms of use apply to the dataset? 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.

Data is distributed under Creative Commons Attribution 4.0 International.

E. Is there a 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.

No erratum

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

Motivation: Document stable access to the datasheet.

The original datasheet is found in the GitHub repository (https://github.com/connollyc152/DDC jet sensitivity) along with the code used in C23. The data is not receiving any updates and therefore the datasheet is not expected to receive any updates.

G. Any other comments?

Motivation: Space for any other relevant information about data distribution and maintenance.

4. Data-dependent questions

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

A. How was the data generated or collected (e.g., model runs, reanalysis processes, observational measurements)? 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. Has the data been assessed against some baseline(s) (e.g., an observational product or physical laws)? If so, describe how, and provide any relevant citations.

Motivation: Document evaluation of the data 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. Has uncertainty quantification been carried out for this dataset? If so, describe how and provide citation(s).

Motivation: Provide information about known uncertainties.

No uncertainty quantifications have been carried out for this dataset.

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

Motivation: Describe important changes to instruments or methodology within the dataset.

Method of generation did not change.

E. Are there any unexplained but relevant numerical values ("magic numbers") that go into the data generation, collection, or processing? (e.g., calibration constants, hyperparameters)

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

There are no known unexplained numerical values that went into the data generation or processing of this data.

F. Was a model used to generate data in this dataset? If so, please describe the configuration and any modifications.

Motivation: Record the exact model setup used to create data.

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.

G. Is this dataset an ensemble? If so, how many members are there? Are there any differences between members? Describe the perturbation of the members, and any relevant sampling limitations (e.g., ocean states).

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

The data is not an ensemble.

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

Motivation: Document group definitions within the data.

No categories, groupings, or labels.

I. Can users contribute to this dataset (e.g., citizen science or human labeling)? If so, please describe the process including evaluation or verification.

Motivation: Describe if the data includes user contributions.

Users cannot contribute to the data.

J. 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.

No specific limitations, however future uses would be limited due to limited saved data.

K. What are the direct or downstream impacts on humans from this dataset? The non-comprehensive checklist below is intended to prompt the author 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.

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

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)

This data is meant to support reproducibility oc C23. The Computing power used to learn how to use the CESM dry dynamical core and generating this data contributed to CO2 emissions.

L. What biases were present in the construction or use of the dataset? The checklist below provides a non-comprehensive 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.

Motivation: Reflect on potential biases present in the dataset.

- ☐ Geographic bias (e.g., restricted or weighted to specific regions)
- ✓ Model bias (e.g., error relative to evaluation product)
- ☐ Sensor bias (e.g., calibration)
- ☐ Temporal bias (e.g., diurnal cycle, restrictions on detection)
- ☐ Seasonal bias (e.g., seasonal cycle)
- ☐ Bias towards extreme or standard conditions (e.g., catchment error, failure to represent extremes)
- ☐ Unbalanced sampling (e.g., unequal classes)
- ☐ Adversarial impacts on data (e.g., fraudulent samples in crowdsourced data)
- ☐ Label bias (e.g., subjective labeling)
- ☐ Threshold sensitivity (e.g., for an extreme index)
- ☐ Regime dependence (e.g., convective structure, mode of variability)
- ☐ Selection bias (e.g., case studies, survivorship effects, loss of historical data over time)

This data is from a dry dynamical core which is an idealized Earth system model. While the models resolves many dynamical features of the Earth system, there are also features not resolved in this model.

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.

Motivation: Space for any other relevant information.

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.
- 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.