Datasheet for *ClimART: A Benchmark Dataset for Emulating Atmospheric Radiative Transfer in Weather and Climate Models*

SALVA RÜHLING CACHAY*, TU Darmstadt & Mila
VENKATESH RAMESH*, Université de Montréal & Mila
JASON N. S. COLE, Environment and Climate Change Canada
HOWARD BARKER, Environment and Climate Change Canada
DAVID ROLNICK, McGill University & Mila

1 Dataset Motivation

- **For what purpose was the dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* Numerical weather prediction and global climate models require very large allocations of computer resources. Estimation of atmospheric radiative flux profiles can amount to a sizeable portion of an entire model's computer time. As such, optimization of radiative transfer (RT) models has been an ongoing project for several decades now, and Machine Learning (ML) renditions of RT models offer an attractive way of realizing this. In order to produce a useful ML-based RT algorithm, in whole or in part, a wide range of sample atmosphere-surface systems is required to train such an ML model. The synthetic dataset used to train an ML-based RT algorithm was produced by the Canadian Earth System Model.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  The raw data was generated by the Canadian Centre for Climate Modelling and Analysis (CCCma) as part of Environment and Climate Change Canada.

- **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number.* Funding for this activity was provided by Environment and Climate Change Canada via a grant to Mila – Quebec AI Institute.

2 Dataset Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
  The instances of the dataset represent snapshots (state of the atmosphere at any particular time) of atmospheric variables for the purposes of radiative transfer calculation. Each training or testing example is a 1-D column (vertical profile of the atmosphere for a given latitude and longitude) which consists of 49 layers and 50 levels (the layer interfaces) of the atmosphere, as well as non-spatial information that we coin *globals* data in the paper.

---

* Equal contribution.

Authors' addresses: Salva Rühling Cachay*, TU Darmstadt & Mila ; Venkatesh Ramesh*, Université de Montréal & Mila; Jason N. S. Cole, Environment and Climate Change Canada; Howard Barker, Environment and Climate Change Canada; David Rolnick, McGill University & Mila .

- **How many instances are there in total (of each type, if appropriate)?** There are approximately 352256 individual columns at different time intervals over the course of a year. In total there are over 10 million columns/examples in ClimART.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).* The dataset is sampled from climate model simulations done at a frequency of 205 hours over 1979-2014 (and 1850-52, 2097-99 for the out-of-distribution sets). We believe that the dataset has enough temporal diversity to cover all possible ranges of atmospheric conditions.
- **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.* Each instance consists of the extracted state of the Canadian Earth System Model over the levels and layers of the atmosphere.
- **Is there a label or target associated with each instance?** If so, please provide a description. Yes, the target in this case is either the flux (shortwave and/or longwave) or heating rates (shortwave and/or longwave). We provide both of those in our dataset.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.* We have proposed and used a dataset split that we describe in the main paper (1979-90 + 1993-2004 as potential training years, 2005-06 as potential validation years, 2007-14 for testing). However, the user should feel free to experiment with different splits depending on their needs and downstream application.
- **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.* None to our best knowledge.

3 Collection Process

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?* The raw data was sampled from simulations using version 5 of the Canadian Earth System Model (CanESM5). The experimental design for the simulations are the AMIP (present), historical (past) and SSP5-85 (future) scenario for the Sixth Coupled Model Intercomparison Project (CMIP6).
- **Over what timeframe was the data collected?** *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.* This data was extracted from CanESM5 simulations performed between January and August 2021.

4 Dataset Preprocessing, Cleaning, Labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.* The pre-processing procedure (reshaping, and concatenating of arrays) is described in our main paper in detail.
- **Was the "raw" data saved in addition to the pre-processed/cleaned/labeled data (e.g., to support unanticipated future uses)?***If so, please provide a link or other access point to the "raw" data.* The raw data was saved in netCDF format. We process it and save it to Hdf5 format to

make it easier for machine learning experiments. Unfortunately, we have not hosted the netCDF format due to storage constraints. If required, we're willing to make subsets of the netCDF format available on request.

4.1 Dataset Use Cases

- **Has the dataset been used for any tasks already? If so, please provide a description.** The dataset has been used for the purposes of radiative transfer emulation using machine learning. We use this dataset to compute shortwave flux for clear-sky and pristine-sky cases for various ML models.
- **What (other) tasks could the dataset be used for?** The dataset could be used for evaluation of improvements to the CanESM5 radiative transfer model, which is publically available, as well as input for other radiative transfer models.
- **Are there tasks for which the dataset should not be used?** *If so, please provide a description.* Given the limited number of variables and how it was sampled, it should not be used to evaluate the climate simulated by CanESM5 nor climate change. If there is an interest in this type of research, then we suggest using CanESM5 outputs published by the Canadian Centre for Climate Modelling and Analysis (CCCma). A large number of experiments and output produced by CanESM5 for CMIP6 can be downloaded from the Earth System Grid Federation (ESGF): https://esgf.llnl.gov/.

5 Dataset Distribution

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?* The dataset will be made available on Microsoft Azure blob storage with a DOI, and a script to download the dataset will be provided. There will be a corresponding GitHub repository associated with the dataset to provide starter code to use the dataset and recreate the baseline methods.
- **When will the dataset be distributed?** The dataset will be made available to the reviewers along with the submission and later released to the public with a companion GitHub repository.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.* No, the dataset won't be under a copyright or IP. We provide the dataset under Creative Commons Attribution 4.0 International (CC BY 4.0) where users are allowed to share and adapt the dataset so long as they give credit.

6 Dataset Maintenance

- **Who is supporting/hosting/maintaining the dataset?** We'll be hosting the dataset on Azure blob storage/
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The authors can be contacted via their emails mentioned in the paper.
- **Is there an erratum?** *If so, please provide a link or other access point.* Not to our best knowledge.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?* Yes, we plan to update the dataset by adding more atmospheric conditions like all-sky which take clouds into consideration, as well as more data from future climate simulations. The corresponding GitHub page will be updated regularly.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.* There will only be addition of new data (with data versioning as supported by Azure). The current version will thus remain intact and available to the public.