README

This README describes the dataset titled 'TTI-HydroMet: A Decade of High-Resolution Rainfall and Streamflow for the Tamanduateí River Watershed, Brazil,' which can be accessed at https://doi.org/10.5281/zenodo.17654660. The dataset is divided into two primary folders: Code and Data. Further details are given below.

1 Folder: Code

This folder includes four different codes for data processing.

- *Create_Radar_Input*: this Python script is responsible for constructing the final modeling datasets by integrating weather radar observations with river stage measurements.
  - ➢ It reads a preprocessed radar dataset containing radar data (gridded variables) at 10-minute resolution and a time series of observed stage data ("nível").
  - ➢ The code aligns both datasets temporally using the timestamp column ("datahora"), verifies that radar data exist for all stage observation times, and converts stage values from millimeters to kilometers.
  - ➢ For each specified forecast lead time (e.g., 10, 60, 120, and 240 minutes), the script generates a target output column that represents the future stage shifted forward by the corresponding number of time steps.
  - ➢ The resulting datasets contain the timestamp, all radar grid variables, the current stage, and one forecast target column (e.g., "out00h10m", "out01h00m").
  - ➢ Each lead time is saved as an independent CSV file following a standardized naming convention ("iFast_Radar_Obs_2015_2025_<lead_time>.csv"), which serves as direct input to the machine-learning modeling framework.
- *Model_v6_SP*: this Python script coordinates the full machine learning experimentation process for training, validating, and evaluating river stage forecasting models.
  - ➢ It loads previously generated CSV datasets, dynamically selects input variables based on the experiment configuration, and prepares the data for modeling by defining input–output structures and temporal splits.
  - ➢ The script executes the ML4FF framework (Soares et al., 2025) using nested cross-validation (inner and outer loops) and a holdout dataset to ensure robust performance assessment.
  - ➢ Multiple machine-learning or deep-learning algorithms can be tested, with automatic hyperparameter optimization performed via Bayesian search.
  - ➢ For each experiment, the script saves trained models, prediction results, performance metrics (NSE, RMSE, and KGE), execution diagnostics, and full configuration metadata.
  - ➢ Outputs are organized into timestamped result directories and exported as CSV and Excel summary files, along with configuration logs.
- *RadarDataMissingCode*: this Python script analyzes missing data in the 10-year time series from 01/04/2015 to 29/03/2025. It generates a distribution of missing data based on duration or period — specifically, the number of missing records multiplied by the length of missing data — considering only precipitation exceeding 0.2 mm.

- *Stage_and_Gauge_data_Pre_Processing*: this script offers general Python code for reading, analyzing, and preprocessing rain-gauge and river stage data. It loads all data from specified files (organized by station), aligns them on a common timestamp grid (at 10-minute intervals), fills missing data with NaNs, and performs analysis to extract key information about each station's time series. A collection of data frames for active and high-quality stations is created. The stage data undergoes outlier removal using the Hampel identifier and smoothing with a low-pass zero-phase filter to produce a refined time series. The processed stage and rain-gauge data (with rainfall treated separately) are saved into organized CSV files, and training datasets for the ML4FF framework (Soares et al., 2025) are prepared.

2 Folder: Data

2.1 Folder: Radar

It includes all the information regarding the radar data.

1) For files: dataList_2015, dataList_2016, dataList_2017, dataList_2018, dataList_2019, dataList_2020, dataList_2021, dataList_2022, dataList_2023, dataList_2024, and dataList_2025. They have two columns: 'datahora' and 'dado'.

- 'datahora' --> YYYYMMDDHHMM (YYYY=year, MM=month, DD=day, HH=hour, MM=minute). Example: 201504010000 (2015/04/01 00:00)
- 'dado' --> 0 or 1, where 0 means no data available and 1 means there is data available

2) For file 'NaN_times_radar_2015-2025': it is presented the day and time of Not a Number (NaN) values, i.e., no value available for the quantitative precipitation estimation (QPE) in the dates and times.

3) For file 'Radar': it contains the raw radar data. The header includes 'datahora' and the radar grid IDs over the study area. 'datahora' represents YYYY-MM-DD HH:MM:SS (for example, 2015-04-01 00:00:00), and the radar grid ID contains the radar QPE for the grid (1 km x 1 km). It is important to note that radar data are collected every 5 minutes during significant precipitation events within the coverage area and every 15 minutes during surveillance mode (when no rain is observed). As a result, many entries in the time series are 'empty' because the system operates in surveillance mode, indicating that no rainfall is detected in the sky.

4) For File 'radar_20150401_to_20250329_0000': it contains the preprocessed radar data, with empty entries filled with 0 (zero) to support further statistical analysis. The header includes 'datahora' and the radar grid IDs over the study area. 'datahora' represents YYYY-MM-DD HH:MM:SS (for example, 2015-04-01 00:00:00), and the radar grid ID contains the QPE (mm) for the radar grid (1 km x 1 km). It is important to note that 29 timestamps are missing from the raw radar data because they were identified as failures; however, this represents only 0.006% of the dataset.

5) For File ‘radar_failure_series_2015-04-01_to_2025-03-29’: it presents the failure in the radar data series. It has two columns: ‘datahora’ and ‘present’.

- ‘datahora’ --> DD/MM/YYYY HH:MM (Example: 01/04/2015 00:00)
- ‘present’ --> 0 or 1, where 0 means no failure and 1 means it presents failure

----

2.2 Folder: SHP

This folder includes vector files (.SHP): rainfall gauge stations (points), station 413 - stage (point), radar location (point) and coverage (polygon), Tamanduateí River (line), supplementary hydrography - tributaries (line), and Tamanduateí River Watershed (polygon). The data is referenced in EPSG 31983 - SIRGAS 2000 / UTM zone 23S coordinate system.

2.3 Folder: Telemetric_Time_Series

This folder includes all data related to rain gauge and hydrological gauge stations. It has 37 *.csv* files; however, only 23 rain gauges and one hydrological station (river stage) were utilized. The .csv files of the stations are P_143_INST, P_275_INST, P_279_INST, P_280_INST, P_283_INST, P_413_INST, P_511_INST, P_563_INST, P_629_INST, P_1000350_INST, P_1000370_INST, P_1000390_INST, P_1000400_INST, P_1000410_INST, P_1000420_INST, P_1000430_INST, P_1000490_INST, P_1000500_INST, P_1000510_INST, P_1000550_INST, P_1000837_INST, P_1000839_INST, and P_1000867_INST. It is important to mention that *‘P_413_INST.csv’* contains information for both rain gauge and hydrological station (i.e., this file contains rainfall and river stage data).

- ‘P_143_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, and ‘BateriaV’. ‘Posto’ contains the station ID (143), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_275_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, ‘Qm3s’, and ‘BateriaV’. ‘Posto’ contains the station ID (275), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, ‘Qm3s’ presents the station flow (m3/s), and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_279_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, ‘Qm3s’, and ‘BateriaV’. ‘Posto’ contains the station ID (279), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, ‘Qm3s’ presents the station flow (m3/s), and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_280_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, ‘Qm3s’, and ‘BateriaV’. ‘Posto’ contains the station ID (280), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00),

‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, ‘Qm3s’ presents the station flow (m3/s), and ‘BateriaV’ presents the battery voltage (V). NA values mean *‘Not Available’*.

- ‘P_283_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, ‘Qm3s’, and ‘BateriaV’. ‘Posto’ contains the station ID (283), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, ‘Qm3s’ presents the station flow (m3/s), and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_413_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, and ‘BateriaV’. ‘Posto’ contains the station ID (413), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_511_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘BateriaV’. ‘Posto’ contains the station ID (511), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, and ‘BateriaV’ contains the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_563_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, and ‘BateriaV’. ‘Posto’ contains the station ID (563), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_629_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, and ‘BateriaV’. ‘Posto’ contains the station ID (629), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), ‘PLUmm’ contains the rainfall (mm) for the date and time, ‘FLUm’ shows the river stage in meters, and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_1000350_INST.csv’: the column headers are ‘Posto’, ‘DATA’, and ‘PLUmm’. ‘Posto’ contains the station ID (1000350), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), and ‘PLUmm’ contains the rainfall (mm) for that date and time.
- ‘P_1000370_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000370); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA indicates *‘Not Available’*.
- ‘P_1000390_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000390); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA indicates *‘Not Available’*.

- ‘P_1000400_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000400); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000410_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000410); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000420_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000420); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000430_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000430); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000490_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000490); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000500_INST.csv’: the column headers are ‘Posto’, ‘DATA’, and ‘PLUmm’. ‘Posto’ contains the station ID (1000500); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); and ‘PLUmm’ contains the rainfall (mm) for that date and time. NA indicates *‘Not Available’*.
- ‘P_1000510_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000510); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000550_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, and ‘FLUm’. ‘Posto’ contains the station ID (1000550); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; and ‘FLUm’ shows the river stage in meters. NA values indicate *‘Not Available’*.
- ‘P_1000837_INST.csv’: the column headers are ‘Posto’, ‘DATA’, and ‘PLUmm’. ‘Posto’ contains the station ID (1000837); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); and ‘PLUmm’ contains the rainfall (mm) for that date and time. NA indicates *‘Not Available’*.

- ‘P_1000839_INST.csv’: the column headers are ‘Posto’, ‘DATA’, ‘PLUmm’, ‘FLUm’, and ‘BateriaV’. ‘Posto’ contains the station ID (1000839); ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00); ‘PLUmm’ contains the rainfall (mm) for the date and time; ‘FLUm’ shows the river stage in meters; and ‘BateriaV’ presents the battery voltage (V). NA values indicate *‘Not Available’*.
- ‘P_1000867_INST.csv’: the column headers are ‘Posto’, ‘DATA’, and ‘PLUmm’. ‘Posto’ contains the station ID (1000867), ‘DATA’ shows the date and time of the recorded data in the format YYYY-MM-DD HH:MM:SS (for example, 2011-10-13 15:00:00), and ‘PLUmm’ contains the rainfall (mm) for that date and time. NA indicates *‘Not Available’*.

2.4 File: RainGauges.csv

The file includes raw data from all rain gauge stations within the study area. The columns are labeled ‘Posto’, ‘DATA’, and ‘PLUmm’. ‘Posto’ indicates the station ID (e.g., 1000500), ‘DATA’ presents the date and time of the measurement in the format YYYY-MM-DD HH:MM:SS (such as 2015-04-01 00:00:00), and ‘PLUmm’ storages the rainfall amount (mm) for that specific timestamp.

REFERENCE

SOARES, J. A. J. P.; OZELIM, L. C. S. M.; BACELAR, L.; RIBEIRO, D. B.; STEPHANY, S.; SANTOS, L. B. L. (2025). ML4FF: A machine-learning framework for flash flood forecasting applied to a Brazilian watershed, Journal of Hydrology, 652, 132 674. DOI 10.1016/j.jhydrol.2025.132674