Predicted Daily Flow Data (Unimpaired and Existing Models)

Summary

This dataset contains predicted daily streamflow time series for river and stream reaches across California, developed jointly by The Nature Conservancy (TNC) and Upstream Tech. Predictions are provided for nearly all stream segments in California as represented in the NHD+ Medium Resolution Flowline dataset and are indexed using the unique COMID identifier for each reach.

Two complementary modeled flow products are included:

- **Unimpaired (Natural) Flows** – estimates of streamflow that represent conditions in the absence of human alteration.
- **Existing (Actual) Flows** – estimates of streamflow that reflect real-world conditions, including the effects of dams, diversions, land use, and other anthropogenic influences.

Flows are provided at the daily timestep for a 21-year period covering water years 2002–2022 (October 1, 2001 – September 30, 2022). The dataset is intended to support hydrologic analysis, ecological flow assessments, conservation planning, and water management applications, particularly in data limited or ungauged basins.

Limitations and Disclaimer: Predictions are not provided for approximately 1% of NHD+ flowlines where drainage area attributes were missing or inconsistent between NHD+ metadata and watershed delineation. These data are not observed streamflow measurements, rather they are model generated estimates from Long-Short Term Memory (LTSM) Models. Model performance varied by hydrologic regime, basin size, and degree of alteration. Internal testing indicated that model skill at validation gages was often low, with Nash–Sutcliffe Efficiency (NSE) values ranging from negative values to moderate skill (Unimpaired flows mean NSE -0.4; Existing flows mean NSE -1.35). Negative mean NSE values indicate poor overall performance relative to observed mean flow. Users should apply these data with caution, particularly in highly intermittent, flashy, small, or in heavily altered basins. We advise that users should consult the accompanying accuracy and methods documentation prior to use. For more information on accuracy, see https://rivers.codefornature.org/reports/modeling_actual_flows_evaluation_oct2025_report3.pdf (note: The “Existing” flows model is called “Actual” in this report).

Description

The dataset was created to provide a spatially and temporally consistent record of daily streamflow across California, including both natural reference conditions and existing, altered conditions. The work was conducted to advance statewide analyses of streamflow patterns, support ecological flow science, and enable comparison across basins where observed gage data are sparse or unavailable.

The modeling approach employed LSTM neural networks, a class of machine learning models that are well suited for time-series hydrologic data. The two models shared the same architecture, but one was trained for unimpaired systems and one trained for systems with human alteration. Model development proceeded in three stages: (1) base model development, (2) flow routing through the river network, and (3) downscaling of predictions to NHD+ reach-scale flowlines. For more information on the modeling approach, see https://rivers.codefornature.org/reports/reach_based_daily_actual_and_natural_flow_predictions_feb2025_report1.pdf (note: The “Existing” flows model is called "Actual" in this report).

Date

Temporal coverage: October 1, 2001 – September 30, 2022 (Water Years 2002–2022)

Publication date: 2026

Point of Contact

Organization: The Nature Conservancy, California Program

Contact: Kirk Klausmeyer

Email: kklausmeyer@tnc.org

Field Definitions

The data are distributed as comma separated values (CSV) files, bundled by the first four digits of the COMID. Each file contains daily predicted flows for one model scenario. Fields include:

- **comid** – Unique identifier for the NHD+ Medium Resolution flowline (reach).
- **date** – Date of the daily prediction (MM/DD/YYYY).
- **discharge_cfs_mean** – Mean predicted daily discharge in cubic feet per second (cfs).
- **discharge_cfs_q0.01** – 1st percentile predicted daily discharge (cfs).

- **discharge_cfs_q0.025** – 2.5th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.05** – 5th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.1** – 10th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.25** – 25th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.5** – 50th percentile (median) predicted daily discharge (cfs).
- **discharge_cfs_q0.75** – 75th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.9** – 90th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.95** – 95th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.975** – 97.5th percentile predicted daily discharge (cfs).
- **discharge_cfs_q0.99** – 99th percentile predicted daily discharge (cfs).

Percentile values represent modeled uncertainty bounds around the daily flow estimates.

Abbreviation Definitions

- **comid** – Unique identifier for NHD+ flowline features.
- **cfs** – Cubic feet per second.
- **q -** Streamflow.

Access Constraints

None.

Use Constraints

None. Users are requested to use the following citation when using these:

The Nature Conservancy and Upstream Tech. Predicted Daily Flow Data, Version 1.0.0. 2026. San Francisco, CA. https://rivers.codefornature.org/#/data. (Date Accessed).

Data Distribution

The dataset is distributed by The Nature Conservancy as downloadable CSV files at https://rivers.codefornature.org/#/data.

Progress

Complete.

Update Frequency

As needed.

Keywords

Daily flow; streamflow; hydrology; ecological flows; machine learning; LSTM; California

Projection

Not applicable. (Tabular data linked to NHD+ flowline identifiers.)

Datum

Not applicable.