Flow-DB: A new large-scale dataset of stream and river flows

Isaac Godfried
Kriti Mahajan
Kevin Li
Maggie Wang
Pranjalya Tiwari

CORONAWHY

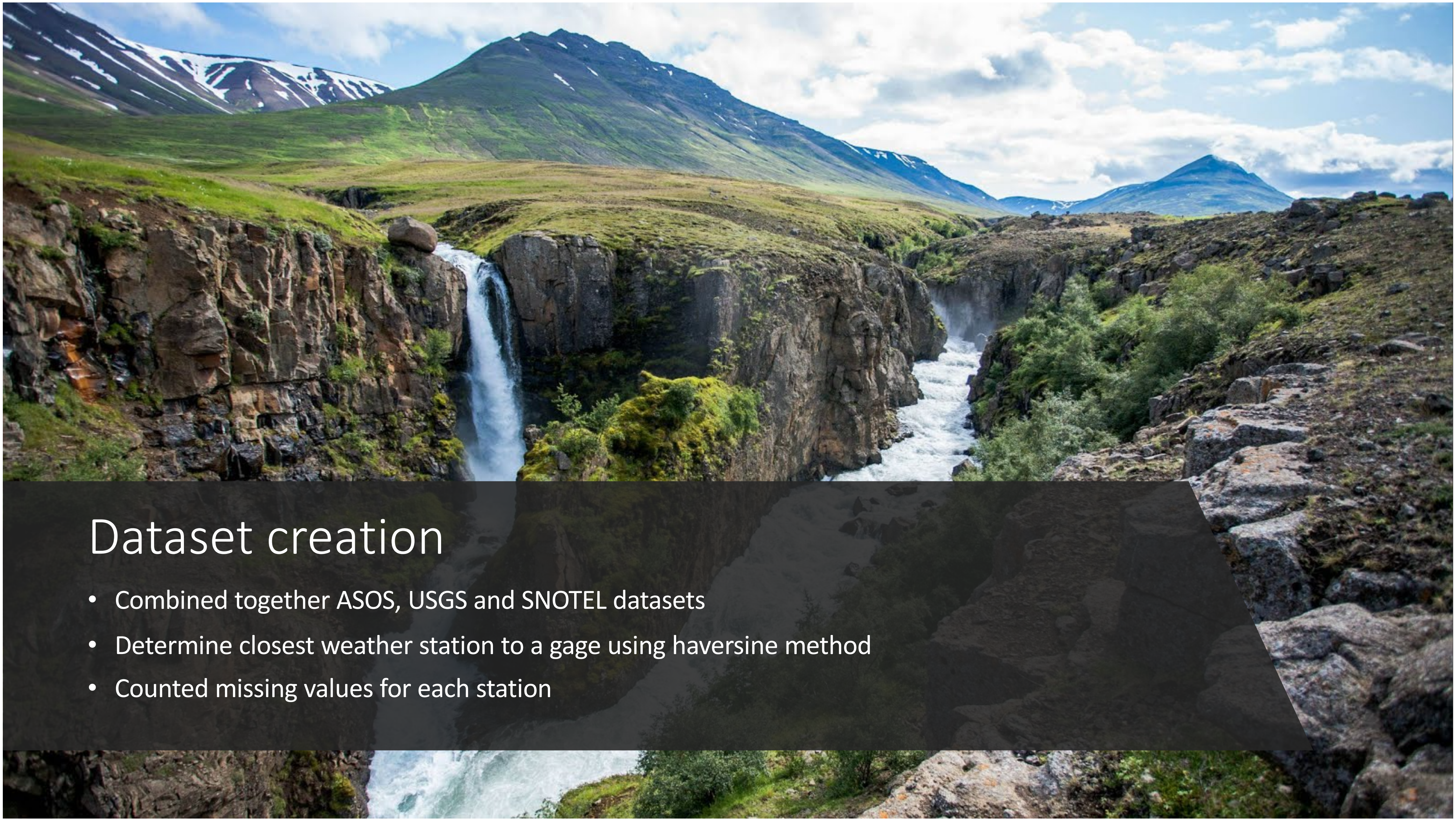
Societal Problems

- Floods result in the most lives lost of any natural disaster in the US.
- In 2011 the government declared 58 flood disasters, totaling $8 billion dollars.
- The most common cause of floods are large-scale precipitation events
- Accurately forecasting river flows, precipitation, and adverse weather events can help government officials plan responses, warn residents, and mitigate the damage.
- In the opposite direction forecasting low flows can help plan for droughts.

https://www.americanrivers.org/rivers/discover-your-river/10-facts-about-flooding/

Prior research

- CAMELs dataset
  - Contained 671 catchments
  - Data reported daily
- Despite limited size of CAMELS several papers found ML useful at predicting river flows.
  - F Kratzert et al 2019 (LSTMs)
  - Gauch et al Aug 2020
- Other research has studied flash flood and natural diaster damage estimates



Dataset creation

- Combined together ASOS, USGS and SNOTEL datasets
- Determine closest weather station to a gage using haversine method
- Counted missing values for each station

Core dataset

- Contains hourly flow, temperature and precipitation data.
- Collected for 2014-2019 with goal to automate ingestion of new data.
- Data for more than 9,000+ streams and rivers around U.S.
- Gage meta-data (i.e. lat/lon, mean snow fall, slope, soil depth, etc)
- Working on incorporating snow-pack, soil moisture data, and aerial imagery


Flash Flood Subset
Small subset of ~10,000 floods across USA

Evaluation Methods

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points, f_i the value returned by the model and y_i the actual value for data point i.

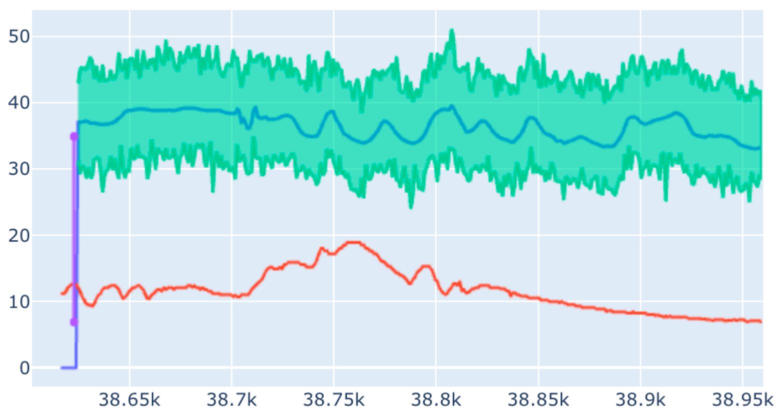
$$MASE = \frac{MAE}{MAE_{in-sample,\, naive}}$$

Key ML Challenges

- What deep learning architectures can effectively incorporate static river basin meta-data with dynamic (hourly) time series data?
- Is transfer learning effective and to what extent?
- What ways can we most effectively impute missing data?
- How can we effectively incorporate seasonality into model forecasts?
- How can we ensure models will preform well in the face of out of distribution events (e.g. 1000-year flood)?

Models and methods

- Information is saved to Weights and Biases
- We have tried many models: LSTM, DA-RNN, and GRUs.
- Difficult for models to fully learn seasonal patterns on some gages.
- Particularly hard for models to generalize to out of distribution events



Using dataset for pre-training

- We found success in using river flow data to pre-train large transformer models for time series follow by fine-tuning to a target task:
  - COVID-19 forecasting
  - Solar forecasting
- We believe there could be even more positive transfer for other climate and/or agriculture tasks

Can your model do better?

Visit to find out how to test your model on our dataset.

- pytorchforecasting.com/flow
- http://github.com/AIStream-Peelout/flow-forecast