The Nature Conservancy × DATA SCIENCE SOCIETY AT BERKELEY

StreamSage 2.0

Final Deliverable

Meet the Team

Christiana Kang
Project Manager

Nathan Kuo
Project Manager

Marianne Choi
Consultant

Sooyeon Kim
Consultant

Mark Barranda
Consultant

Ashley Watanabe
Consultant

Background & Purpose

What?

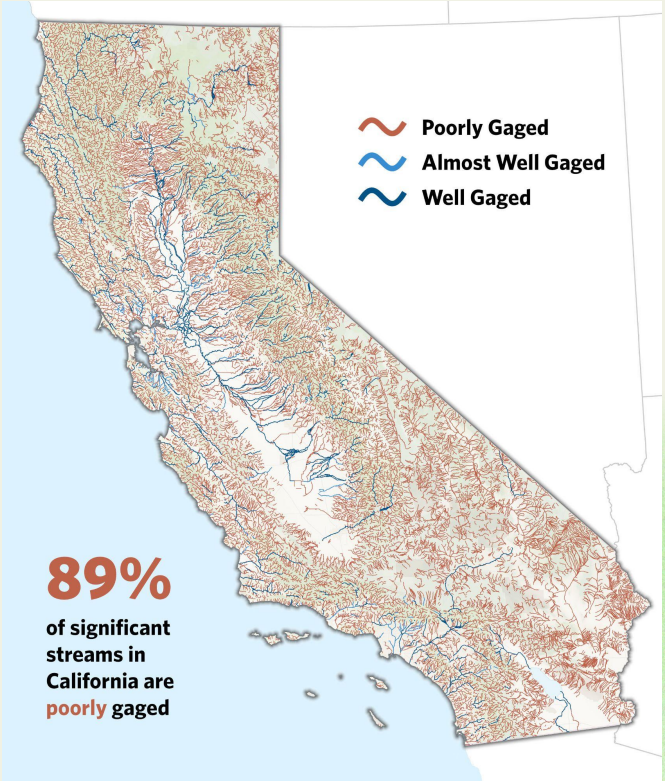
Predicting daily water flow in streams that may not have active gages

How?

Numerical modeling using similarity ranking and dynamic windowing

Why?

Inform water policy decisions and conservation efforts across California



Expert Rescale

Expert Recale Overview

$$Q_{ug} = Q_g \times \frac{A_{ug}}{A_g} \times \frac{I_{ug}}{I_g}$$

- **Q:** stream flow value
- **ug:** ungaged (target location), **g:** gaged (reference location)
- Ratio of **drainage areas (A)**
  - Larger drainage area → more flow relative to reference
- Ratio of **mean annual precipitation (I)**
  - Higher MAP → more flow relative to reference

Reference Gages Selection

Similarity Measure

Feature Embedding/Vectorization:

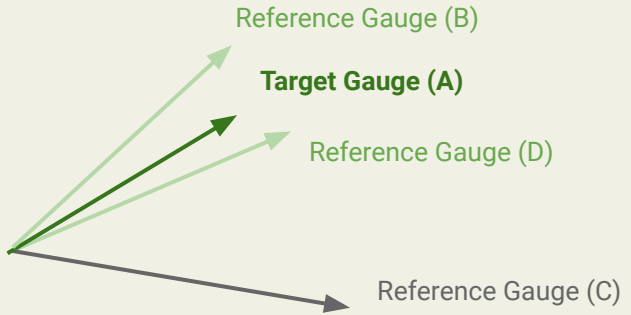
- Normalize feature vectors to account for different scales.

Similarity Metrics:

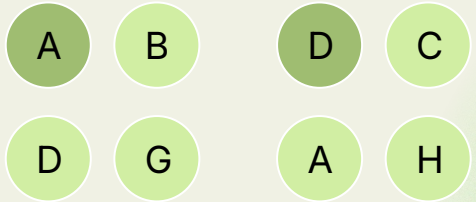
- Utilize *Euclidean distance* to quantify the similarity between gages.
- Closer (smaller) distances in vector space imply higher similarity.

Reference Gages:

- Select similar gages (e.g., B and D) as reference points for each target gage to improve prediction accuracy.



With Similarity Measure:
→ Each gage has its own unique reference gage group of n size



Weight Assignment

Our Implementation:

- target variable is continuous → linear regression model
- coefficients assigned as weights to each feature

** cannot use ‘PRECIP’ or ‘totdasqmi’

Alternative Implementation:

- random decision trees and recursive feature elimination
- custom RFE function to select features and determine weights that maximize KGE and R2

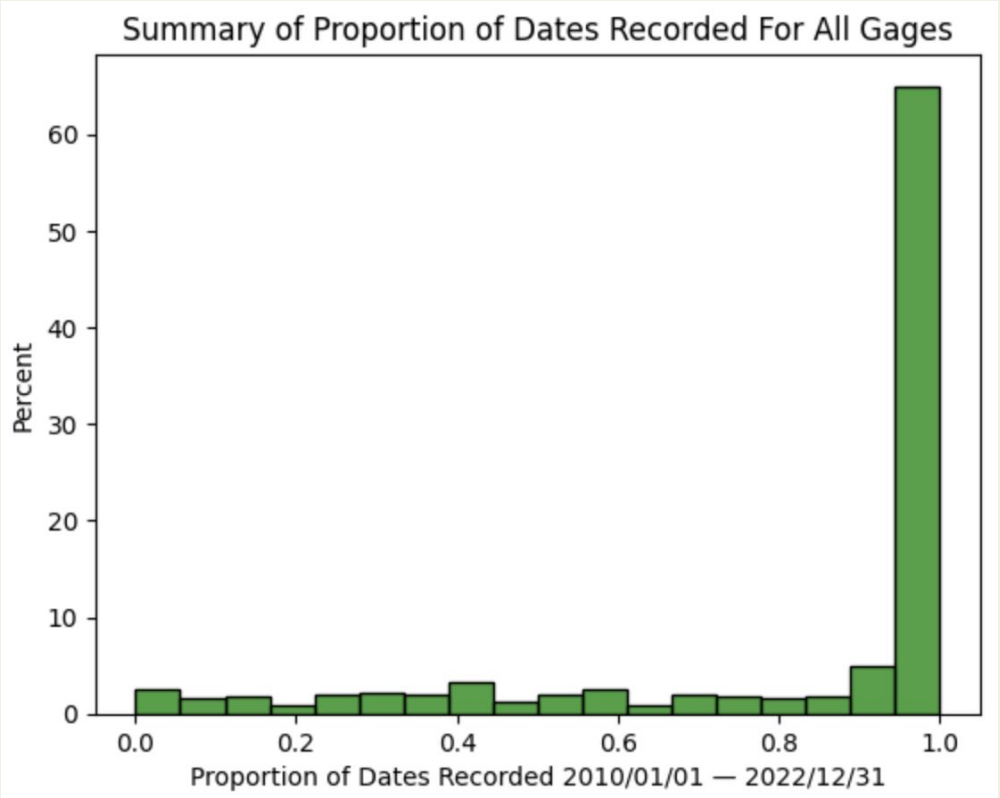
** computationally intensive

Features & Weights:

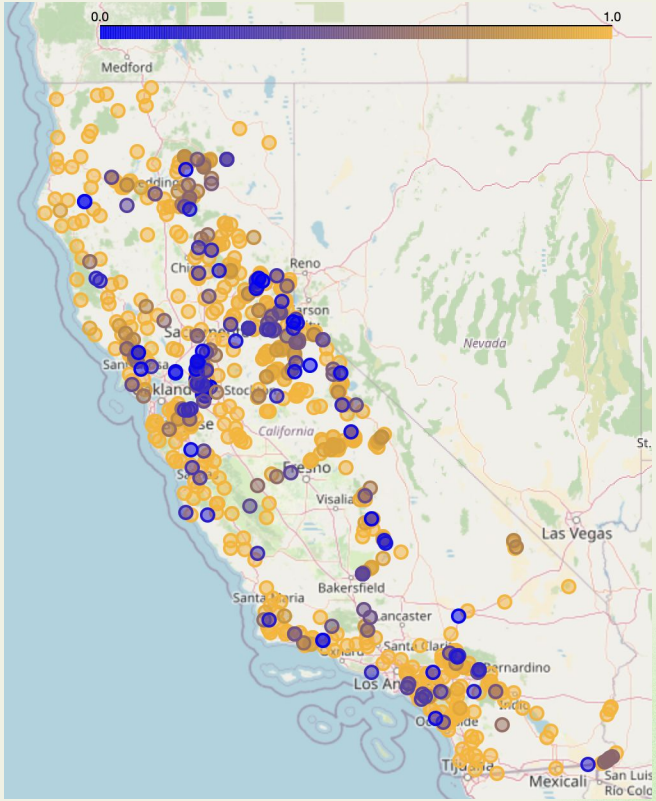
1 / DISTANCE	0.68	1 / distance between reference gage and target gage calculated using haversine formula
'MINBELEV'	0.16	minimum basin elevation
'JANMAXTMP'	0.16	maximum temperature in January of stream gauge
'RELIEF'	0.16	maximum - minimum elevation
'ELEVMAX'	- 0.16	maximum elevation near the stream gauge

Dynamic Windowing

The Problem



proportion of dates recorded



The Goal

Problem

Not all gages recorded flow data on every date

Solution

?

Goal

Constant *n* reference gages for each date to compute average

Dynamic Windowing

Solution

Analyze the top *n* most similar gages on a **day-by-day basis**

date	gage_1_id	gage_1_ pred_flow	gage_2_id	gage_2_ pred_flow	gage_3_id	gage_3_ pred_flow	avg_gage_ pred_flow
2010-01-01	1000	1	2000	2.5	3000	2.5	2
2010-01-02	1000	1.5	3000	4	4000	1.5	2.33
2010-01-03	1000	1.5	2000	2.5	3000	2	2
2010-01-04	2000	2	3000	2.5	4000	0.5	1.67

Performance

Measuring Expert Rescale Performance

KGE
(Kling-Gupta efficiency)

- Assesses the similarity between observed and simulated hydrological data
- Considers correlation, variability, and bias of data (less sensitive to outliers)

R^2
(coefficient of determination)

- Measures the proportion of the variance in the dependent variable that is explained by the independent variable
- Measures the strength of the relationship between observed and predicted values (sensitive to outliers)

Interpretation of KGE & R^2

KGE range

-∞ 0 1

- *KGE* = 1 indicates perfect agreement between observed and simulated data.
- *KGE* < 0 indicates that the model performs worse than a reference model that predicts the mean of the observed data.

R^2 range

0 1

- *R*2 = 1 indicates that the model perfectly predicts the dependent variable.
- *R*2 = 0 indicates that the model does not explain any of the variability in the dependent variable and is essentially equivalent to using the mean of the dependent variable as the predictor.

Models Comparison Pipeline

Model 1
(e.g. upstream)

Model 2
(e.g. expert recale with weight assignment)

Apply performance metric on a sample set of unique gages

gage id	Model 1 KGE	Model 1 R^2	Model 2 KGE	Model 2 R^2
11014000	-0.178879	0.168024	-13.30231	-3.729540
…	…	…	..	…
…	…	…	..	…
…	…	…	..	…
…	…	…	..	…
…	…	…	..	…
…	…	…	..	…
…	…	…	..	…

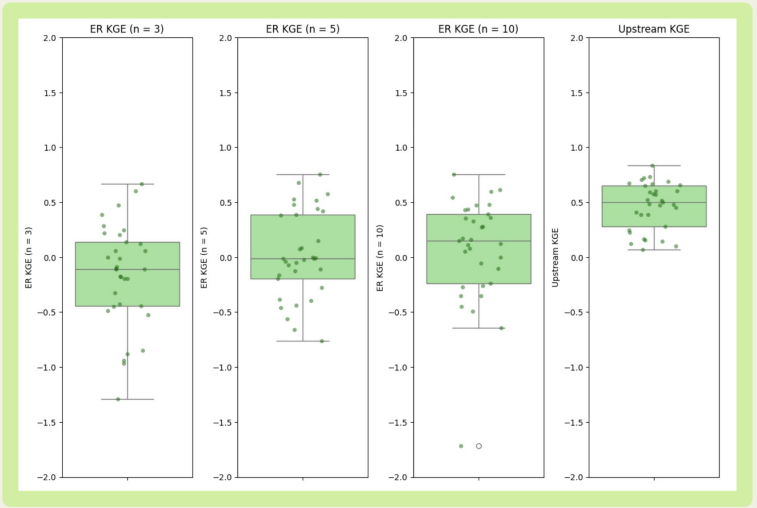
```
# Calculate R2 score
r_squared = r2_score(original_gage_flows, expert_rescaled_flows)
print("R2 Score:", r_squared)

# Calculate KGE
kge_value = kge(original_gage_flows, expert_rescaled_flows)
print("KGE Value:", kge_value)
```

```
R2 Score: -2.2304485865226162
KGE Value: -2.9745042382467815
```

start date = 01/01/2022
end date = 12/30/2022

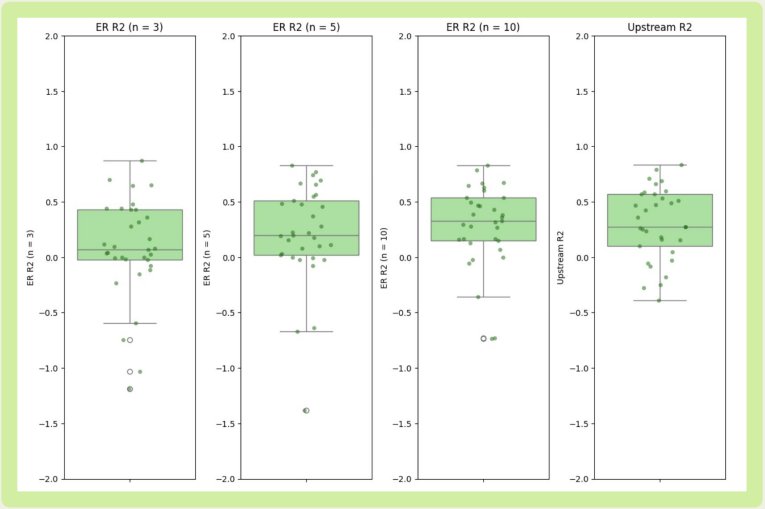
Upstream vs Expert Rescale

Average KGE Score:



Expert Rescale (n = 3)	-0.160
Expert Rescale (n = 5)	0.022
Expert Rescale (n = 10)	**0.067**
Upstream	**0.466**

Average R2 Score:



Expert Rescale (n = 3)	0.076
Expert Rescale (n = 5)	0.204
Expert Rescale (n = 10)	**0.283**
Upstream	**0.302**

Streamlit

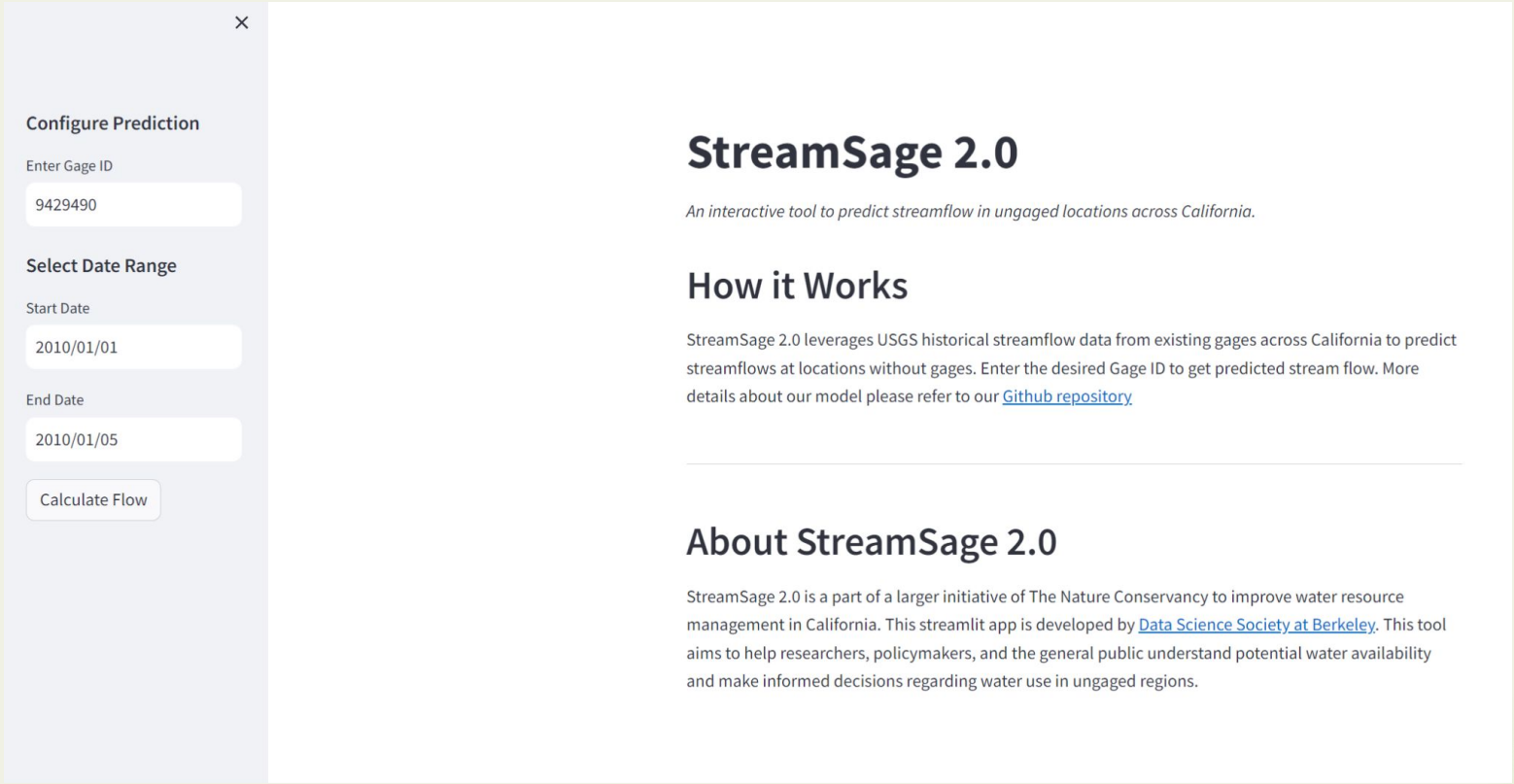
What is Streamlit?

- Open-source Python framework to create interactive app
- Can be deployed on Streamlit Community Cloud, Docker, Kubernetes
- Have integration with Snowflake (Support SQL & Big database)

Flexible
Easy to code

```
import model
if calculate_button:
    prediction = model.predictedflow(gage_id, window_start, window_end)
    st.write(prediction)
```

Demo


Configure Prediction
Enter Gage ID
9429490
Select Date Range
Start Date
2010/01/01
End Date
2010/01/05
Calculate Flow
StreamSage 2.0
An interactive tool to predict streamflow in ungaged locations across California.
How it Works
StreamSage 2.0 leverages USGS historical streamflow data from existing gages across California to predict streamflows at locations without gages. Enter the desired Gage ID to get predicted stream flow. More details about our model please refer to our Github repository
About StreamSage 2.0
StreamSage 2.0 is a part of a larger initiative of The Nature Conservancy to improve water resource management in California. This streamlit app is developed by Data Science Society at Berkeley. This tool aims to help researchers, policymakers, and the general public understand potential water availability and make informed decisions regarding water use in ungaged regions.

Demo

Configure Prediction

Enter Gage ID

9429490

Select Date Range

Start Date

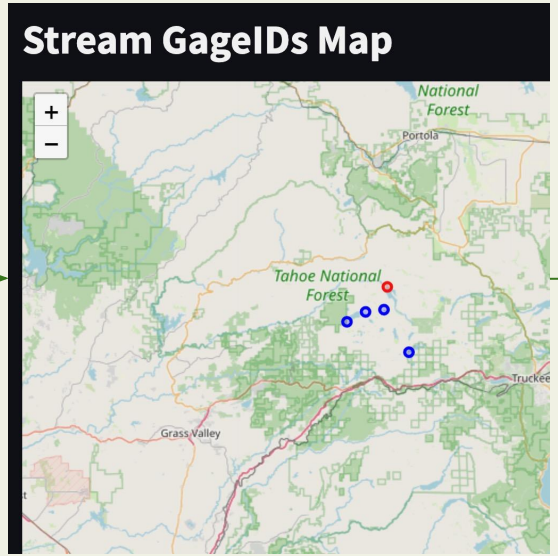
2010/01/01

End Date

2010/01/05

Calculate Flow

Calculating streamflow for the selected gage...



StreamSage 2.0

An interactive tool to predict streamflow in ungaged locations across California.

How it Works

StreamSage 2.0 leverages USGS historical streamflow data from existing gages across streamflows at locations without gages. Enter the desired Gage ID to get predicted str details about our model please refer to our Github repository

	date	predicted flow value
0	2010-01-01	1,458.9371
1	2010-01-02	1,420.0086
2	2010-01-03	1,396.5982
3	2010-01-04	1,349.8441
4	2010-01-05	1,342.0851

Conclusion

Final Deliverables

Deliverable	Description
Expert Rescale Notebook	Cleaned implementation of expert rescale
Our Final Model Notebook	Similarity Measure+Dynamic Windowing+Expert Rescale
Models Performance	KGE and R^2 scores of our model and Upstream
Streamlit Github Repo	Streamlit codes to set up dashboard UI+Model+Data

Main Takeaways

- Our model Streamsage 2.0 achieved improved performance with better systematic reference gages selection (similarity measure) and dynamic windowing.
  - **Comparable with Upstream**
  - **Low computation cost**

- *Streamlit* is accessible and can be coded with only python
  - **Good framework to build & deploy quick tools**

Thank you!
Questions?

Next Steps

Weight Optimization: GMM-XGBoost

- Gaussian Mixture Model (GMM): Cluster streamflow data into homogeneous groups
  - Makes probability
- Extreme Gradient Boosting (XGBoost): Models trained on these clusters to predict actual stream flow with each XGBoost model specialized to predict within specific conditions of that cluster
  - Uses probability as weights
- Output: Predicted Streamflow based on underlying probabilistic structure provided by GMM

https://www.sciencedirect.com/science/article/abs/pii/S0022169420303619

Models Comparison Pipeline

	A	B	C	D	E
1	gage_id	kge_score_upstr	r2_score_upstre	kge_score_expe	r2_score_expert_
2	9423350	-0.6175327133	-0.01224186415	-45.01698546	-111.5148121
3	10258000	-0.126492274	0.158842388	-18.21104554	-50.33432821
4	10258500	-0.2490337081	0.08193953643	-0.2674489726	0.7336160979
5	10259200	-0.1054396197	0.08801253127	-0.1985742293	0.3458830833
6	10261500	-0.2102334255	0.06398295113	-2.522626989	-3.283192469
7	10308783	0.7415193649	0.7502190034	-0.865161511	-1.010137687
8	10308794	0.2547049789	0.4103702921	0	0
9	10310000	0.7101325149	0.8356935558	0.6407751753	0.6279507289
10	10336660	0.5315629527	0.7204954552	0.4249320224	0.47808826
11	10340500	0.3457426419	0.3491838596	0.5564769812	0.4033459318
12	10344400	0.5617646102	0.4569850525	0.4801289941	0.0626309718