UPSTREAM TECH

Powering smart conservation on a changing planet

Utilizing HydroForecast to Model Historic Ungauged Flows and Gap-Fill Time-Series

Contact: team@hydroforecast.com

Created December 4, 2020

Overview

Upstream Tech is a technology company that builds decision-making tools for environmental conservation to improve natural resource management. We do so by harnessing technological advancements in remote sensing, computer science, and machine learning to create customizable planning and monitoring platforms for conservation organizations focused on water management, agriculture, and beyond.

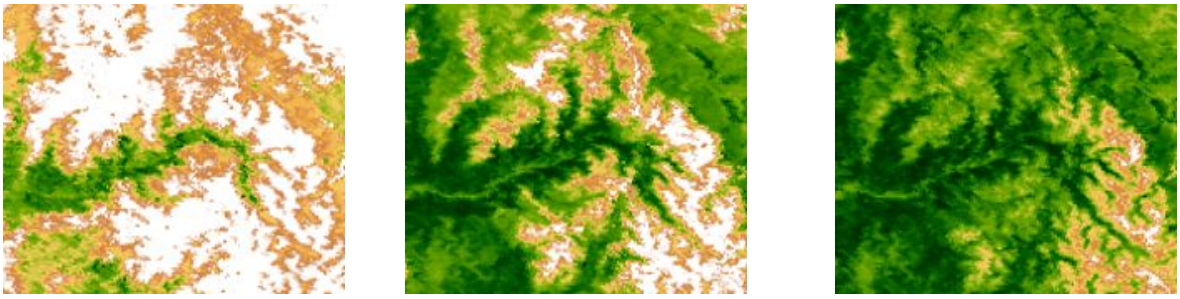
Upstream Tech has developed a process using machine learning hydrology models to predict stream and river flows at locations where the flow is modified by humans and where gauge data is not available, conditions we call Ungauged Actual Flows. A key motivating factor for our work predicting Ungauged Actual Flows is to gap-fill holes in the gauged data record at locations with short or entirely unavailable flow gauge data records.

This document presents our modeling approach, our gap-filling workflow, and finally shares results for prediction and validation results at several example sites. This report focuses on the extension and gap-filling of **historic** time series and validation. We follow a similar set of methods for generating forecasts in this context.

Ungauged Actual Flows Modeling Approach

Input Data

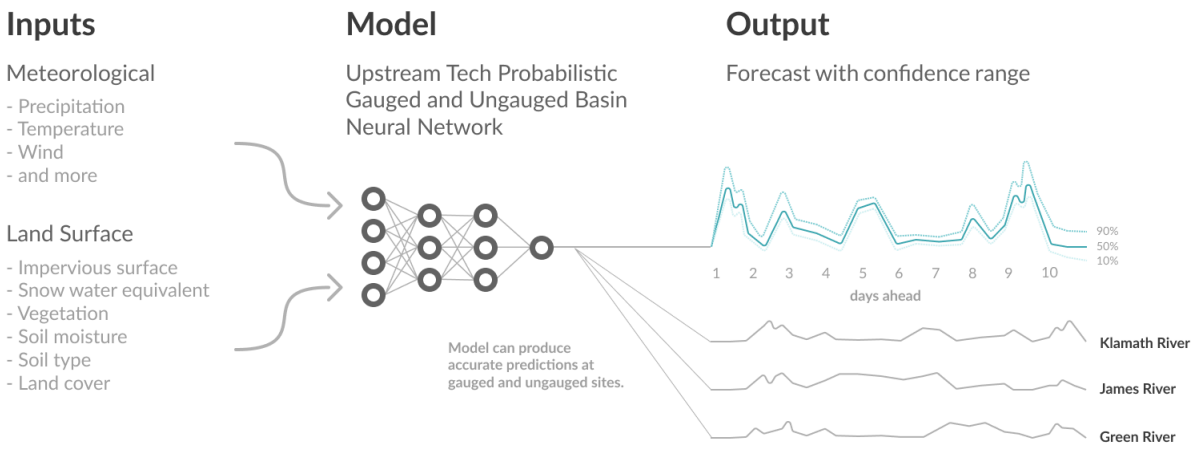
Upstream Tech's Ungauged Actual Flows hydrologic models use satellite data inputs and weather data, which provide the model with an up-to-date description of hydrologically relevant conditions in the basin. The satellite inputs used in this study describe vegetation vigor, snow extent, land surface temperature (helpful for determining soil moisture and snow conditions), and the fraction of the basin covered in a range of land cover classifications, primarily derived from NASA's MODIS satellite. The vegetation vigor values are calculated from raw satellite reflectance data using the Normalized Difference Vegetation Index, a unitless metric where higher values indicate greater leaf cover and vegetation photosynthetic activity. The snow cover values are calculated using the Normalized Difference Snow Index which reports the fraction of each pixel that is covered in snow.



Example visualizations of the Normalized Difference Vegetation Index (NDVI) computed from NASA MODIS satellite data on three selected days during a spring snowmelt period in a mountainous region of Idaho. NDVI values indicating healthy vegetation (dark green), bare ground (brown), and snow (white) demonstrate the spatially-detailed and dynamic information these observations provide.

In addition to the satellite data inputs, our model also incorporates weather information from multiple weather prediction models. Beyond time-varying inputs, our model also receives basin-specific characteristics. This allows the model to learn how conditions like elevation and slope affect flows, and to use that information when moving to a new ungauged location to provide more accurate predictions.

To create training and validation datasets, *in situ* streamflow observations are paired with relevant meteorological and land surface inputs to create a historic input dataset ready for the model.



HydroForecast modeling workflow

Machine Learning Model Structure and Training

The model used in this analysis is built with a type of neural network building block called Long Short-Term Memory (LSTM). The model works with time-based sequences of data. At each time step in the sequence, the LSTM takes in new inputs, updates a set of internal states it maintains which represent the hydrologic basin conditions, and then finally outputs a prediction.

Training a model is the process of providing data to the model and allowing it to learn the relationships between model inputs and streamflow output. The model is trained providing historic training samples (each sample is a time series of inputs paired with gauge measurements) and iteratively updating the strength of the internal connections in the neural network to more accurately predict the desired outputs. Once the model has been trained, it can be used to make predictions at new locations and points in time.

The Ungauged Actual Flows model training process begins by training the model on many basins within a region to learn the general relationships between our inputs and streamflow. Critically, the satellite-derived inputs allow the model to distinguish differences between basins — such as fraction of cropland, average slope, etc. — allowing the model to learn how land characteristics interact with the weather inputs to produce streamflow. In an optional second step, the model can be further trained (or "tuned") with gauge data from hydrologically- and/or geographically-similar locations to the Ungauged Actual Flows prediction sites. This tuning process improves model accuracy at ungauged locations and is utilized in the results presented in this report.

After training is complete, we evaluate the Ungauged Actual Flows models on locations and time ranges for which they have never seen data before. This allows us to understand their accuracy in truly ungauged settings. For example, we built a model that predicted and was evaluated at Deer Creek, CA for water years 2015 – 2019 and had only been trained with data from different basins during water years 2001 – 2014. See the Model Validation section for the results from this test.

Data Gap-filling Workflow

To create streamflow predictions at entirely new, ungauged locations or to fill gaps in the gauge data record we employ the following process. In this report, we use this process to provide historical predictions. In a follow-up report, we describe how a similar process can be used to produce streamflow forecasts.

1. **Train Ungauged Actual Flows model.** The first phase of our model training process uses a training dataset that includes a broad diversity of drainage basins and hydrologic conditions. In this phase, the model learns the relationships between weather, land conditions, and streamflow across geographies. This process is done once and repeated only when the underlying model structure or the types of inputs are changed. The model created in this step is called the General Ungauged Actual Flows Model.

2. **Delineate the drainage basin and prepare input data.** To make an Ungauged Actual Flows prediction at a new site, we begin by delineating the drainage basin using a digital elevation model. Next, we use our in-house automated data preprocessing systems to: 1) collect the satellite and weather data, 2) perform cloud removal or additional data cleaning if necessary, 3) summarize the data into time series to record the input value for each model input at each point in time, and 4) finally save these prepared values to be used by the

model. This system is fully automated at Upstream Tech and can process decades of data across hundreds of locations in parallel.

3. **Tune using regional data (optional).** If gauge data from the region in question is available, it can optionally be used to further train ("tune") the General Ungauged Actual Flows model. This allows the model to pick up nuances of the region that may not have been captured in the first step of model training. This new "tuned" model is called the Regional Ungauged Actual Flows model.

4. **Predict.** Finally, the saved model (either the General or appropriate Regional Ungauged Actual Flows model) and the prepared dataset for the desired site are loaded. The model takes the dataset as input, passes it through the neural network, and returns a streamflow prediction. The output is saved into a database, CSV file, and/or visual plot. This final prediction step takes less than a second, allowing us to make predictions across multiple sites at many different points in time in parallel.

Selected Sites

In this phase we created hindcasts for three tributaries of the Sacramento River in California: Antelope Creek, Deer Creek (DVD), and Mill Creek Sacramento (MCH). On Antelope Creek, the first site (A1 in the map below) is just upstream from a dam (Edwards Dam) and is largely unimpaired. The second site (A2) is further downstream, just above the confluence with the Sacramento River, and is impaired by the Edwards Dam as well as agricultural withdrawals. The DVD and MCH sites are both impacted by human modifications such as agricultural withdrawals.

Site map: The locations of the two Antelope Creek sites (A1 and A2), the Mill Creek Sacramento (MCH), and Deer Creek (DVD) gauges used in this study.

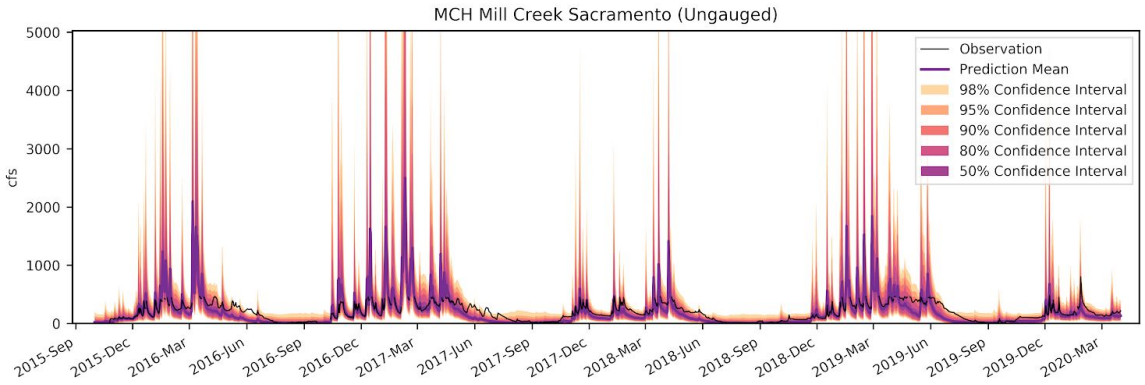
Model Validation

Gauge Data Considerations

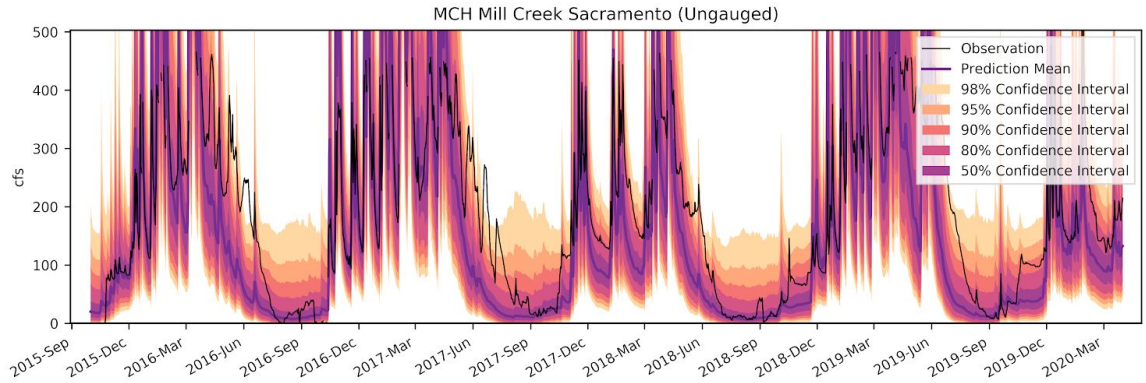
The data from both DVD and MCH gauges have several gaps. These gaps are most common during peak flows above 800 cfs at Deer Creek and above 500 cfs at Mill Creek. Despite these gaps in the data, we can still be reasonably confident in our model evaluation because data is available for the majority of dates and a long record of data is available. The following hydrographs show the validation (2015-2020) of the model at four "ungauged" sites highlighted in the map above (i.e. at sites where the observations were entirely *hidden* from the model). Given the focus on environmental flows, a detailed view of the low flow period is provided.

MCH: Mill Creek Sacramento (human impacted flows)

Validation Period (2015-Oct to 2020-Mar)

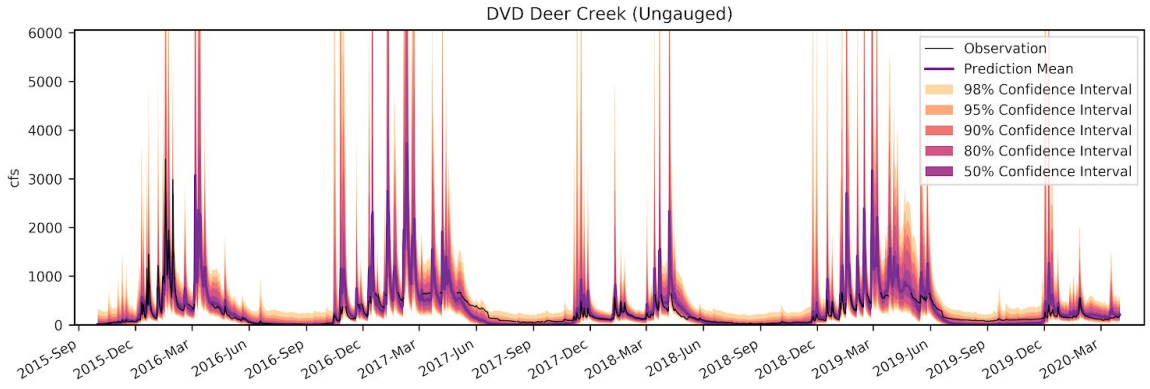


Validation Period: Low Flow View (2015-Oct to 2020-Mar)

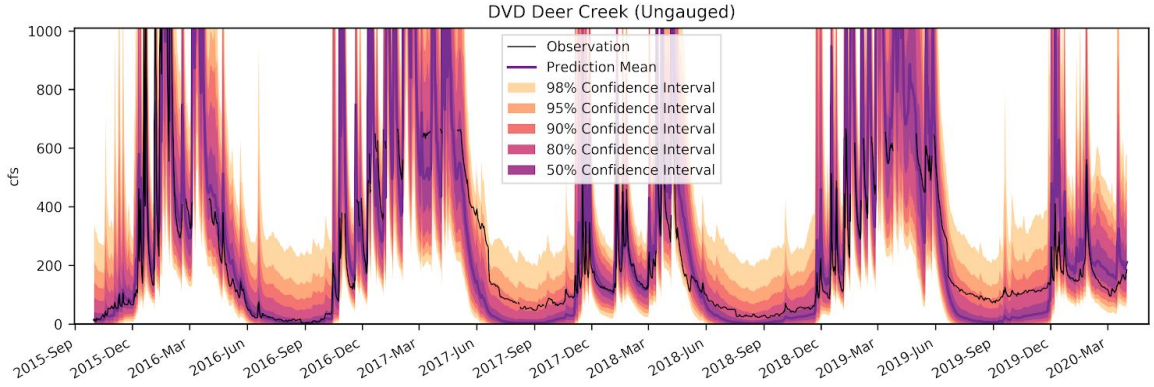


DVD: Deer Creek (human impacted flows)

Validation Period (2015-Oct to 2020-Mar)

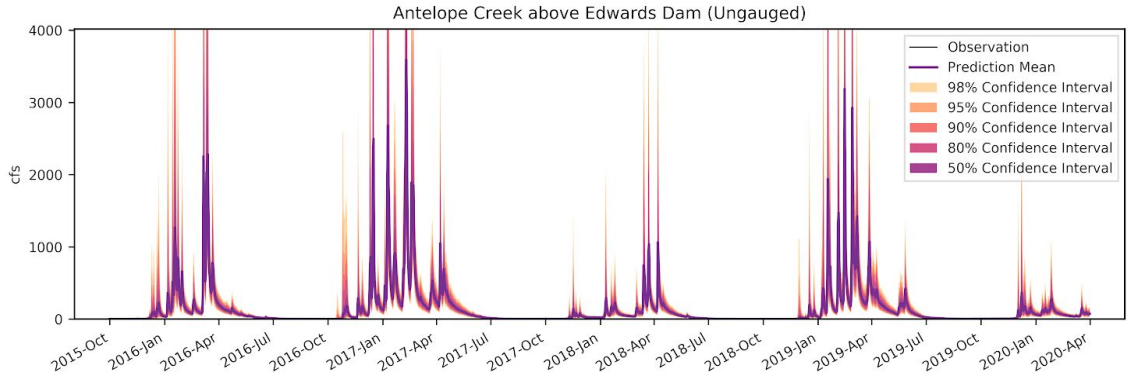


Validation Period: Low Flow View (2015-Oct to 2020-Mar)

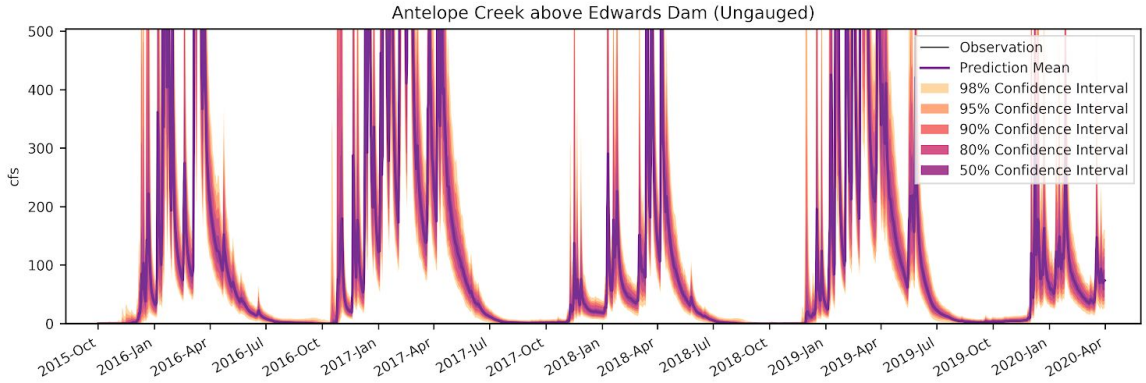


Antelope Creek above Edwards Dam (largely unimpaired)

(2015-Oct to 2020-Mar)

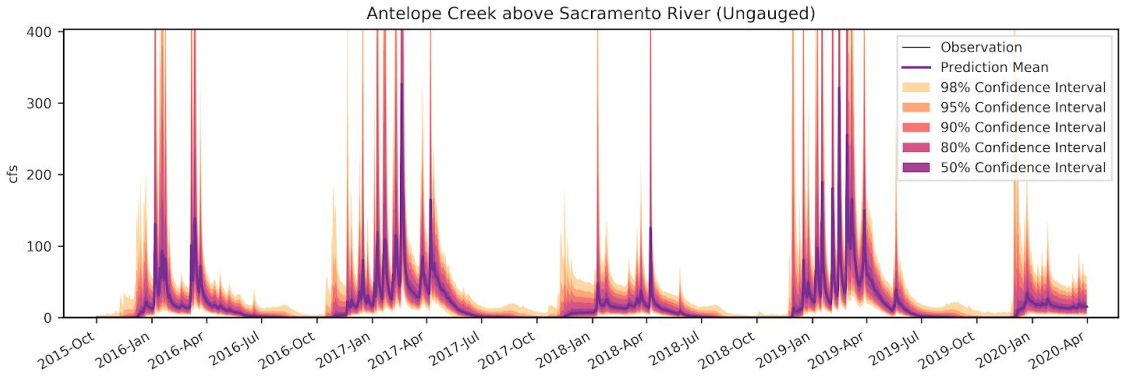


Low Flow View (2015-Oct to 2020-Mar)

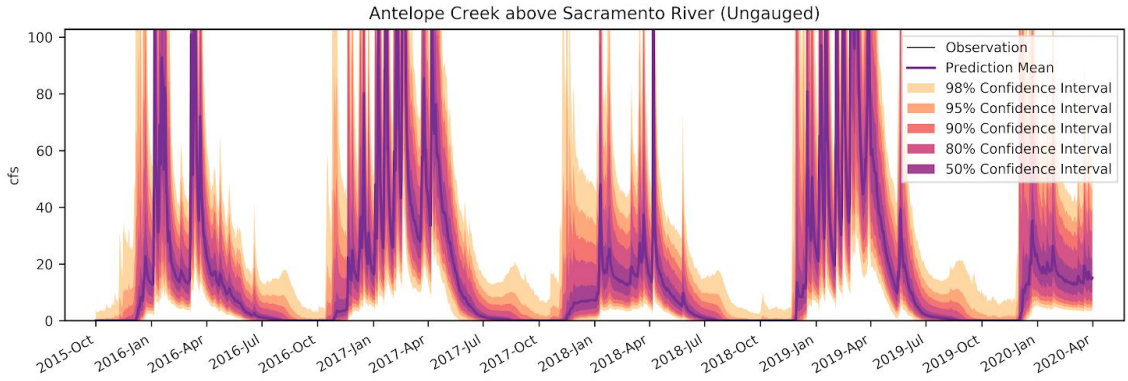


Antelope Creek above Sacramento River (human influenced flows)

(2015-Oct to 2020-Mar)



Low Flow View (2015-Oct to 2020-Mar)



Conclusions

We have built and tested a workflow to predict scalable Ungauged Actual Flows. We have developed this capability on a scalable computing architecture to support Ungauged Actual Flows predictions broadly. The combination of machine learning, hydrologic science, and cloud computing allows us to understand actual flows across ungauged locations with improved accuracy and scalability.

We also improved our understanding of how this method performs in different situations. The accuracy of our Ungauged Actual Flows models is highest when flows are not directly regulated by a dam, and decreases with the strength of dam regulation.

Finally, through this project we have learned that our approach supports the deployment of a well-informed base model in making accurate predictions in ungauged locations, even when data are missing from the historic record. Ungauged Actual Flows prediction is an exciting step towards the ability to understand the true streamflow in any stream reach.