CHRIS STELLATO
CAPSTONE PROJECT #3
APRIL 29, 2021

Flow Forecasting:

Forecasting river flows with machine learning

GALVANIZE
DATA SCIENCE IMMERSIVE

photo: TNC/Chip Carroon

PROJECT SUMMARY

Using time series data and machine learning to predict daily river flow rate for the Blue River near Breckenridge, Colorado.

Build models and compare prediction error rates:

- **Baseline:** Daily historical average
- **General Additive:** “Facebook Prophet” multivariate
- **Recurrent Neural Network:** multiple versions of Long Short Term Memory (LSTM) multivariate

TECHNICAL STACK

python™ TensorFlow Keras

PROPHET colab pandas

matplotlib NumPy scikit learn

photo: Vail Valley Partnership

PROJECT DATASET

Historical data from 1990-2020 (daily):

Hoosier Pass Snow Sensor
(SNOTEL #531, USDA-NRCS)

- Average temperature
- Accumulated Precipitation (rain and snow)
- Snowpack density (snow water equivalent)

Blue River Stream Flow Sensor
('BLUABLCO', USGS)*

- Stream flow rate in cubic feet per second

**Data includes effects of Goose Tarn Dam and peak season spillovers.*

Above: Example USGS stream gauge site
Left: Example SNOTEL sensor site

USING MULTIVARIATE FORECASTING

This project uses multivariate forecast models with four different input variables. Each variable affects predictions with varying amounts of time lag. As the models learn, they determine time windows for each variable that are most relevant to future predictions.

In the example below, snow and precipitation up in the mountains have a delayed effect on river flows, while average temperature may have a more immediate impact. For illustration purposes only, actual model logic is more complex!

	INPUT DATA						MODEL OUTPUT
	2020-04-25	2020-04-26	2020-04-27	...	2020-05-23	2020-05-24	2020-05-25
Accum Precip (in)	19.40	19.50	19.5	...	19.9	19.9	
SWE (in)	19.60	19.70	19.7	...	8.8	8.3	
Avg Temp (F)	32.00	36.00	40.0	...	42.0	31.0	
River Flow (cfs)	6.81	8.43	11.5	...	71.3	69.3	65.7

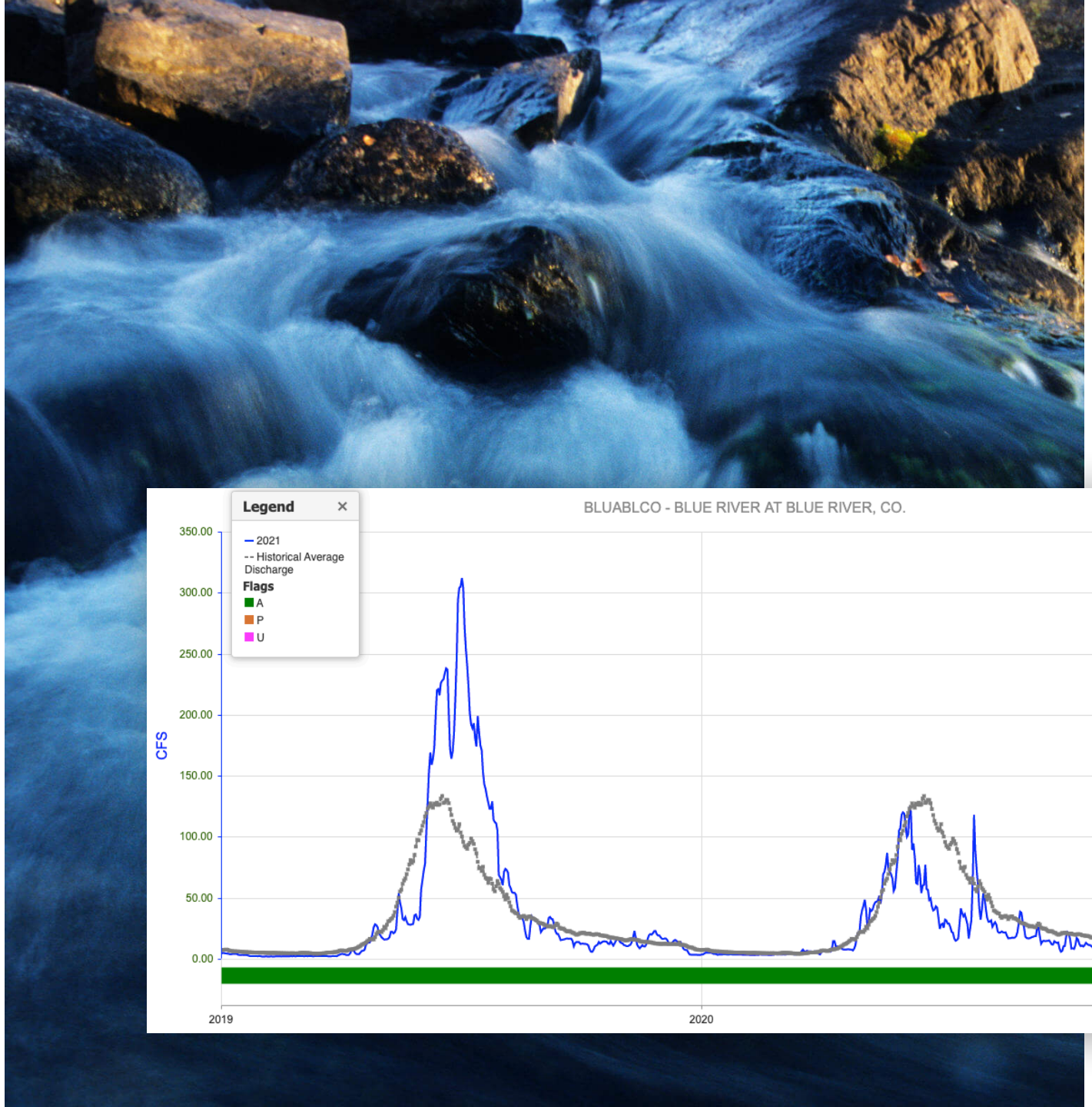
The value we are predicting

The model learns which data points are most relevant to the date it is trying to predict

Blue River Headwaters

Sensor locations and snowmelt drainage flow

Chris Stellato - Flow Forecasting



Historical Average Model

A daily historical average model is the USGS's publicly available tool to estimate future river flows. Comparing these averages to actual observations, it is clear this method could be improved upon.

The goal is to build a model that improves forecast accuracy over the historical average model.

https://dwr.state.co.us/Tools/Stations/BLUABLCO

A Quick Note:
Time Series Forecasting is Challenging!

Nobody has cracked the code to the stock market (yet), and the weather forecast is often incorrect. Using time series data to make predictions about the future is still a difficult endeavor, and only gets harder the further into the future you want to predict. Two key reasons:

1. In the real world, an infinite number of variables (features) are influencing future outcomes. This project used four different features, but this doesn't even begin to come close to the number of factors influencing river flows.

2. Time series data creates many problems a data scientist must solve. Challenges overcome in this project: accounting for leap years, handling missing data, matching historical averages to days of the year, and ensuring consistent time granularity at the daily level.

The author at 13,500, halfway through a long "sufferfest" hike in the Colorado mountains.

MACHINE LEARNING MODELS USED

01 Facebook Prophet

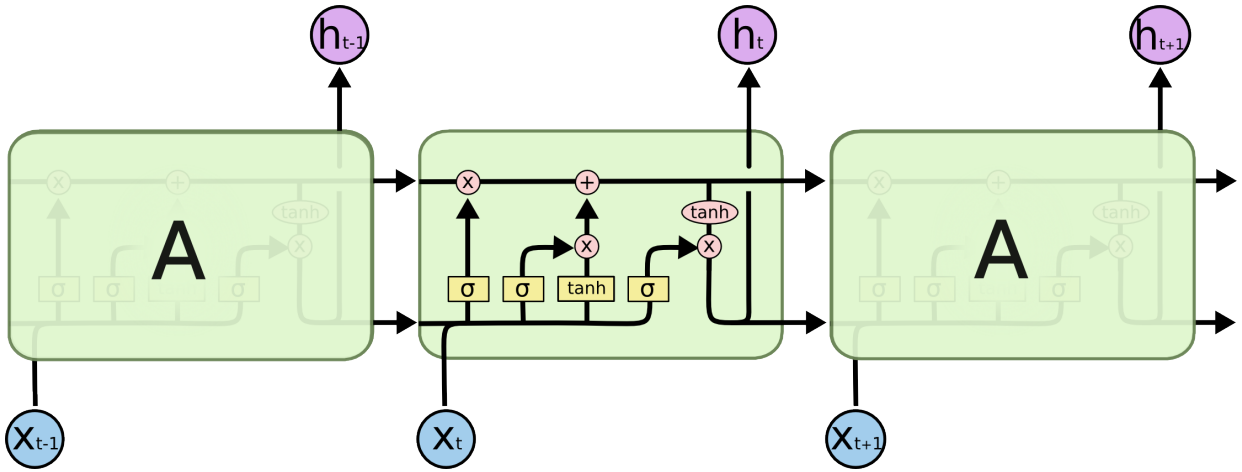
General Additive Model, works best with time series that have strong seasonal effects, picks up on signals from the data and considers both non-linear and seasonal trends when making predictions.

02 Long Short Term Memory (LSTM)

Recurrent Neural Network model, uses a network of neurons. Utilizes a feedback loop to pass information backwards and forwards, and a system of gates to control how information passes through the layer.

This project tested 30 different LSTM configurations and different hyper parameter settings to improve model performance:

- Number and configuration of layers
- Model optimizers “rmsprop” and “adam”
- Number of nodes in each layer (16, 32)
- Dropout Rates (0-0.4)
- Training Batch (Size 1-32)



Example flow of a LSTM model. Source: C. Olah, colah.github.io

EVALUATING MODEL PERFORMANCE

Model prediction errors compared to baseline

Root Mean Square Error was used to compare all models, including the historical average, to actual observations for years 2016-2020. The 10th LSTM variant tested (“LSTM10”) performed the best with the lowest average error.

LOWER ERROR IS BETTER

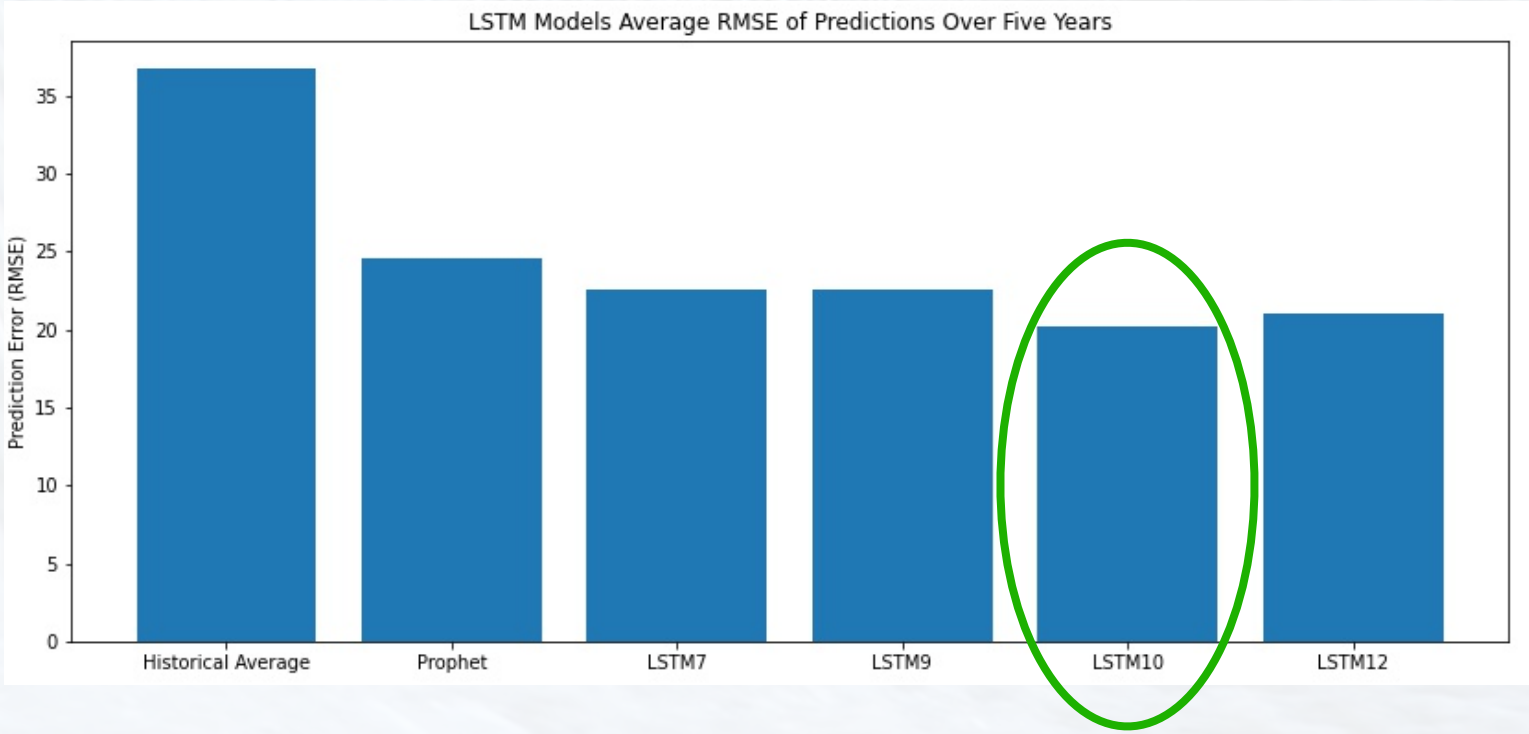


photo: Bill Frank

LSTM10 MODEL ARCHITECTURE

Establishes the Keras Sequential framework, allowing data scientists to add a variety of layers in desired order.

Conv1D layers: convolutional kernel “slides” along a single dimension (in this case, time).

LSTM layers: controls how information flows through the layer using gates.

Dense layer: each neuron receives input from all neurons of previous layer.

```
model = keras.Sequential()
model.add(keras.layers.Conv1D(filters=30, kernel_size=3, input_shape=(None, 5)))
model.add(keras.layers.LSTM(32, return_sequences=True))
model.add(keras.layers.Conv1D(filters=30, kernel_size=3, input_shape=(None, 5)))
model.add(keras.layers.LSTM(32, return_sequences=False))
model.add(keras.layers.Dense(predict_steps, activation='linear'))
model.compile(optimizer='adam',
              loss='mse')
```

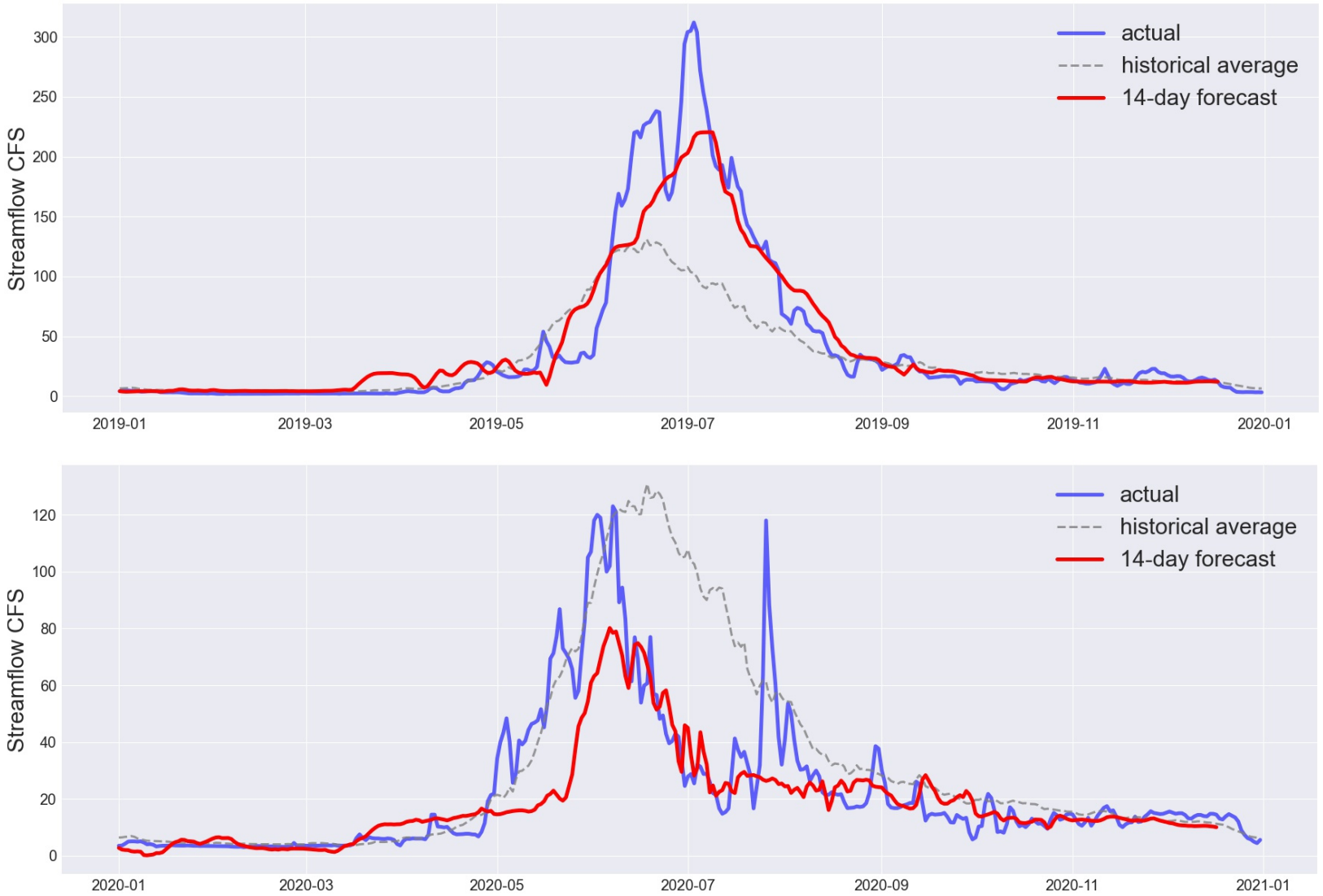
Optimizer determines how a model learns, including parameters like learning rate.

The loss metric determines what the model will try to minimize, in this case mean square error.

LSTM10 MODEL IN ACTION

The LSTM10 model is used here to predict river flows for holdout years 2019 and 2020. Models were tested on a range of +1 day to +30 days forecasts.

Forecasting only one day in advance may not have useful real-world applications. The +14 day plot is an example of a prediction time window where users could get more value from the forecast.



Plots showing model predictions for 2020 and 2019 vs. actuals and historical average, an example of how model performance will vary from one forecast period to another.

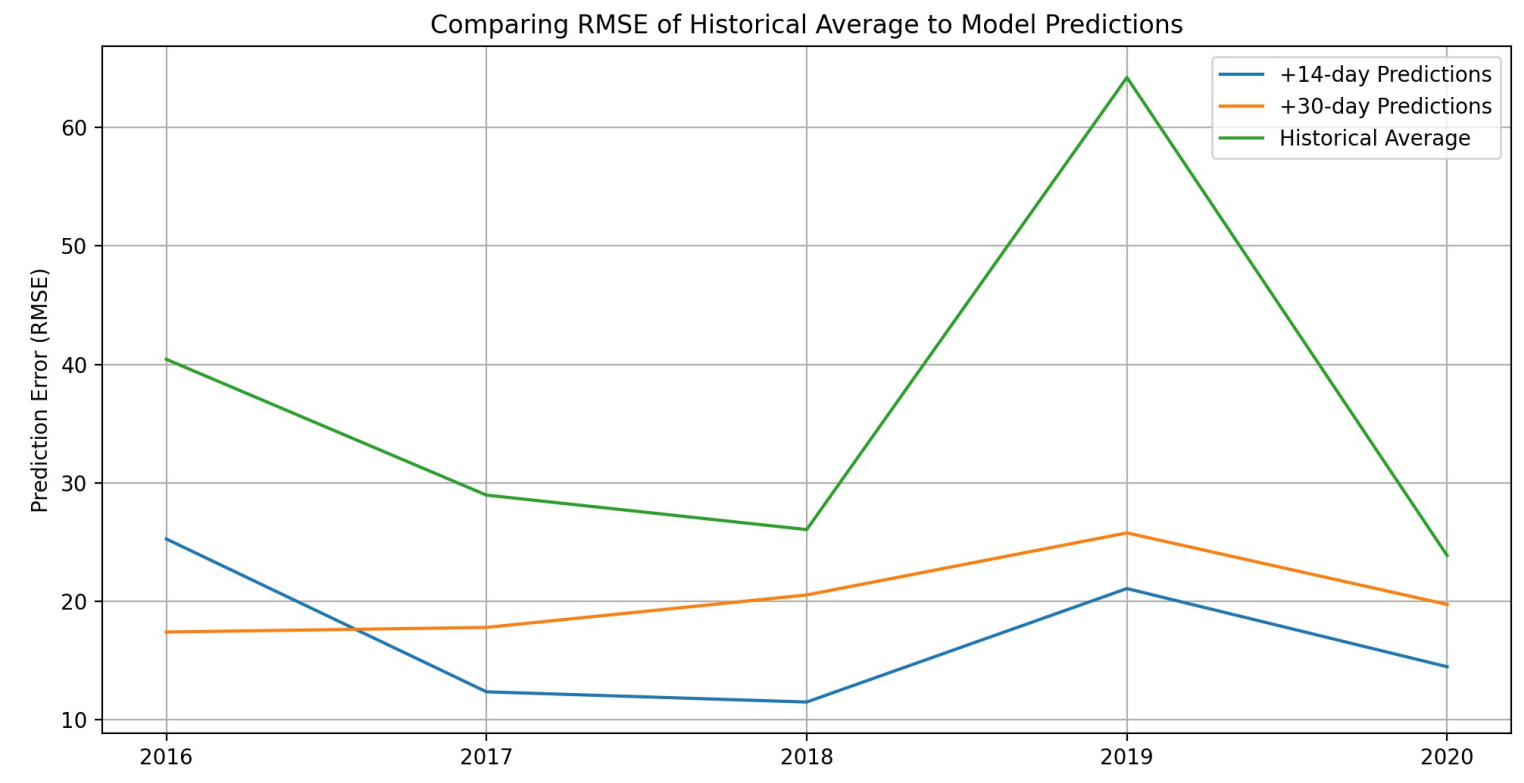
LSTM10 vs. HISTORICAL AVERAGE

Prediction error comparison, yearly for 2016-2020

The LSTM10 model performed an average 39% better than the historical average model when predicting 30-days out.
The LSTM10 model performed an average 52% better than the historical average model when predicting 14-days out.

Results are based on sample data. Changing conditions may cause model performance to vary in the future.
Models used in production environments should be reevaluated and updated at regular intervals.

LOWER ERROR IS BETTER



DATA FOR BETTER DECISION MAKING

A river flow prediction model could be one part of a system that forecasts flows for users. A more accurate model could help individuals and businesses who interact with the river make better decisions when planing projects or trips.

For Colorado's Blue River, there are many stakeholders:

- **Residents of the greater Denver area** receive a portion of their drinking water from the Blue River.
- **Water managers** at the Dillon Reservoir and throughout the Colorado water system.
- **Construction managers** with projects near the river.
- **Recreational users** rafting, kayaking, fishing, hiking, and camping.

Questionable decision making on a river.
photo: Red Bull

FUTURE EXTENSIONS

How could this project be extended for other use cases?

The framework established in this project can be adapted to any other pair of SNOTEL and river flow sensors. Adjustments could be made to use additional data features if they were available, for example data from the stream gauge of a tributary river above its confluence with the forecast river.

Forecasting is not limited to natural phenomenon. **This framework could be adapted to forecast any value, given that historical data is available to train the model.**

Melted snow feeds the river: Looking down at the Blue River headwaters from Quandry Peak.

Download the models:

Trained model .h5 files and full repo available at github.com/chris-stellato/capstone3

CHRIS STELLATO
BOULDER, CO
stellatocjs@gmail.com

github.com/chris-stellato

linkedin.com/in/cstellato

Thank You

CHRIS STELLATO
APRIL 29, 2021

photo: CO Photo Project