Assignment 1 Check In

design the forecast system approach; get and process the data

B Steele

Progress:

- well-defined problem and functional preliminary NN architecture to work from

Challenges:

- Is there a way to constrain error due to water balance (or inbalance)? At least initially, I'll have to assume persistence for some inputs (inflow, outflow, etc)... or do we sacrifice precision and remove those from the operational forecast model? These don't show up as strong parameters in the preliminary model for 1m, but dropping them will definitely reduce fidelity of the model for the integrated depth value, since both the North Fork inflow and the Chipmunk interflow are in the top 20 feature values from the SHAP analysis.

- Forecasting requires acquisition and integration of NOAA GEFS forecasts, since many of the important features in the preliminary models are weather-related. This is mostly a coding lift more than anything.

Scientific Motivation and Problem Statement

Water temperature is often an indicator of water quality, as it governs much of the biological activity in freshwater systems. While temperature is an important parameter to monitor in freshwater lakes, manual monitoring of waterbodies (by physically visiting a site) and sensor network deployment to monitor water temperature are costly endeavors. Northern Water, the municipal subdistrict that delivers drinking water to approximately 1 million people in northern Colorado and irrigation water for ~600,000 acres of land, has had recurring issues with water clarity in Grand Lake, the deepest natural lake in Colorado. They believe that the clarity issues in Grand Lake are primarily due to algal and diatom growth in Shadow Mountain reservoir which are pushed into Grand when they initiate pumping operations. Clarity in Grand is regulated by Senate Document 80 which dates back to 1937 and the inception of the Colorado Big-Thompson project, however in 2016 stakeholders and operators adopted a

system of "goal qualifiers" for Grand. The goal qualifiers are defined through Secchi disc depth measurements (a measure of water clarity), requiring a 3.8-meter Secchi depth average and 2.5-meter Secchi depth daily minimum to be met throughout the July 1 to September 11 regulatory season.

Water in the Three Lakes System naturally flows from Grand into Shadow Mountain into Granby, but pumping operations reverse that by pumping cold water from Granby reservoir into Shadow Mountain and then into Grand and into the tunnel to serve the Front Range (Figure 1). Northern suspects there is a biological "sweet spot" for water temperature in Shadow Mountain Reservoir that may reduce algal and diatom growth and therefore mitigate clarity impacts during pumping operations.

The optimal temperature for reducing algal growth is to keep the upper 1m of water less than 15°C and to reduce diatom growth is to keep the average temperature of 0-5m ("integrated depth") greater than 14°C, which is a bit of a "Goldilocks" problem. A preliminary auto-regressive neural network model predicted tomorrow's water temperature reliably and better than or similar to a persistence model for the regulatory season during 2022 (Figure 2, Figure 3). Using SHAP analysis (a method of explainable AI for neural networks) we found that operational pumping has an impact on tomorrow's temperature (Figure 4, Figure 5), though the impact of operations on the integrated depth is stronger than the upper 1m. This leads me to believe that the operations could be used as a "knob" to control water temperature to some extent in the Three Lakes System.

The overarching goal is to create a decision support system that forecasts water temperature in Shadow Mountain Reservoir on a daily timestep to a horizon of seven days, since the operations of the pump over the previous seven days are the most influential operations variable from the preliminary model. Initially, this application will assume a constant operational pumping regime (where the previous day's pumping is continued throughout the forecast horizon), but the intention is to eventually add an operational "knob" that would alter pumping operations as a mechanism to mitigate water temperature within the forecast application and attempt to reach the "Goldilocks" range during the regulatory period. Adding that knob is likely out of scope for this class, so instead, I will focus on reliable 7-day forecasts using an auto-regressive neural network.

Description of the Data

- buoy data - there are two instrumented buoys that measure temperature in Shadow Mountain Reservoir. One is located near the interflow between Shadow Mountain and Grand Lake (deployed 2014) and the other is near the dam, at the opposite end of the reservoir (deployed 2012). The preliminary auto-regressive neural network only used data from the buoy closest to the interflow. Buoy data were aggregated to a daily average at <=1m depth and from 0-5m. We used previous day, 2 days previous, 7 days previous maximum and mean temperature per depth summary as inputs into the neural network.

- met data - temperature, solar radiation, and wind data originate from a meteorological station near Shadow Mountain. Precipitation data originate from a NCEI stations - one near Granby Reservoir, the other near Grand Lake. Solar radiation and wind data are gapfilled with NLDAS modeled data as needed. Data were prioritized by those available from the met station at Shadow Mountain, the NCEI data from Grand, the NCEI data from Granby, then the NLDAS data. Met data were summarized as previous 1, 3, 5, 10 days.

- pumping operations - pumping operations data were summarized to total volume per day, then were summarized as the maximum, mean, and total volume of the previous 1, 2, and 7 days.

- inflow/outflow data - flow data from the North Fork of the Colorado that flows into Shadow Mountain ("nf"), the interflow between Grand and Shadow Mountain ("chip), and outflow from Shadow Mountain into Granby were summarized over the previous 1, 2, and 7 days.

Description of pre-processing

Highly left-skewed data were transformed by squaring the value, right-skewed data were transformed by taking the log of the value. We first added 0.1 to any value that contained zeros and required log transformation to avoid NaNs. Data were then standardized around zero using the mean and standard deviation of each parameter for best ingestion into a neural network architecture.

The data summaries described above resulted in a massive amount of data as input into the model. After transforming and standardizing the data, I then eliminated redundant variables. To do this, I removed variables which were correlated with another by more than 90%, retaining the one that seemed more 'sensical' as a practitioner. This reduced the number of input features from 68 to 52.

Data split

The data here are timeseries data that begin around May or June and end in October of each year. There are discrete gaps in data over the winter, so I can assert some independence between each of the years of data (and we won't talk about 'memory' in lakes, which is probably more important for biogeochemical/biological variables like phosphorus and chlorophyll than temperature).

- training/validation will be leave one year out for years 2014-2022, creating an ensemble prediction of 9 models trained on all but one year of data, validated on that hold out data

- testing: 2023

Description of the forecast approach

I think I'm going to use a NN approach similar to preliminary model described here. I'll likely repeat that NN for forecasted weather and persistence inflow data using the forecasted next day temperature for 7 days forward. While it may constrain error less to use a model that forecasts all 7 days for each depth, I don't believe I have sufficient data to train that kind of model (but probably worth a try?!) I'll likely give direct forecasting a go with the data for 2022 (since this wouldn't require any steady-state assumptions of inflow/interflow/outflow) and see how that goes before leaning into an iterative/rollout approach.
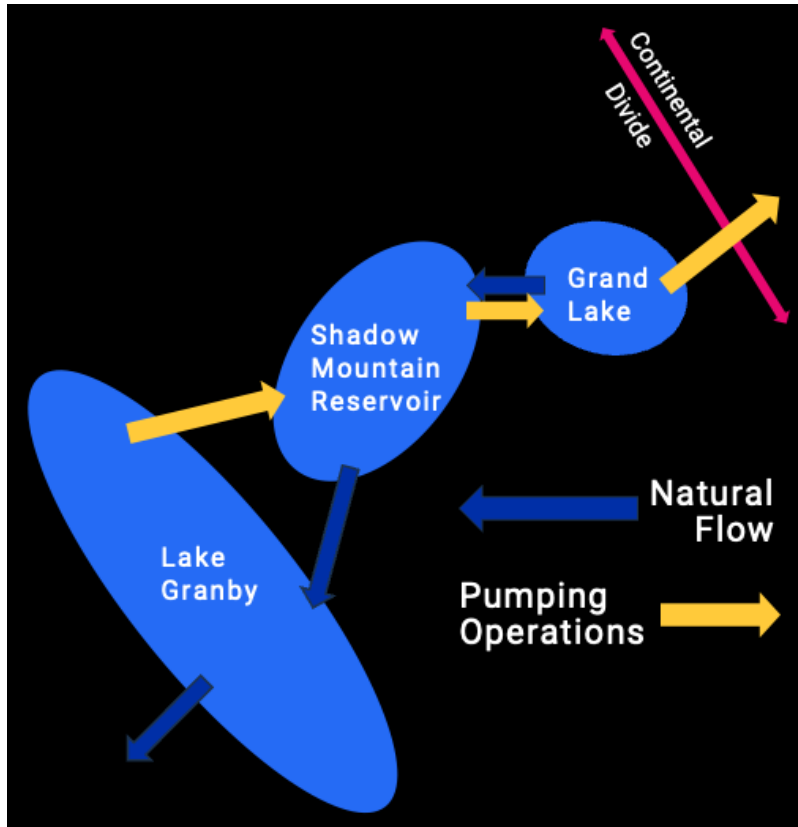
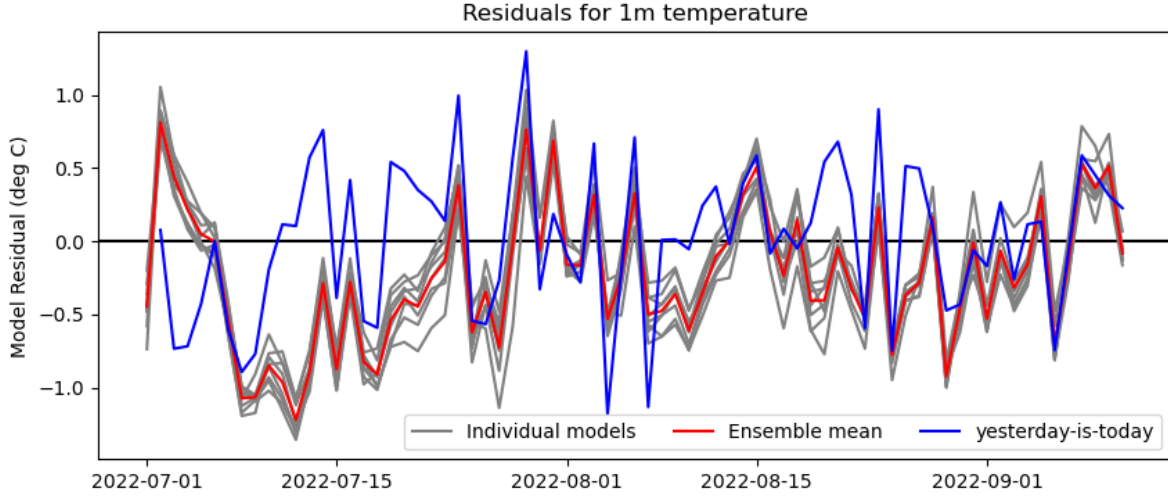Figure 1: Cartoon schematic of water flow in the Three Lakes System

Figure 2: Residuals at upper 1m of auto-regressive neural network ensemble (grey/red) and persistence model (yesterday is today). Initial model performance for the test set was slightly worse than persistence with a MSE of 0.27°C (persistence 0.24°C), MAE of 0.43°C (persistence 0.40°C), MAPE: 2.49% (not calculated for persistence).
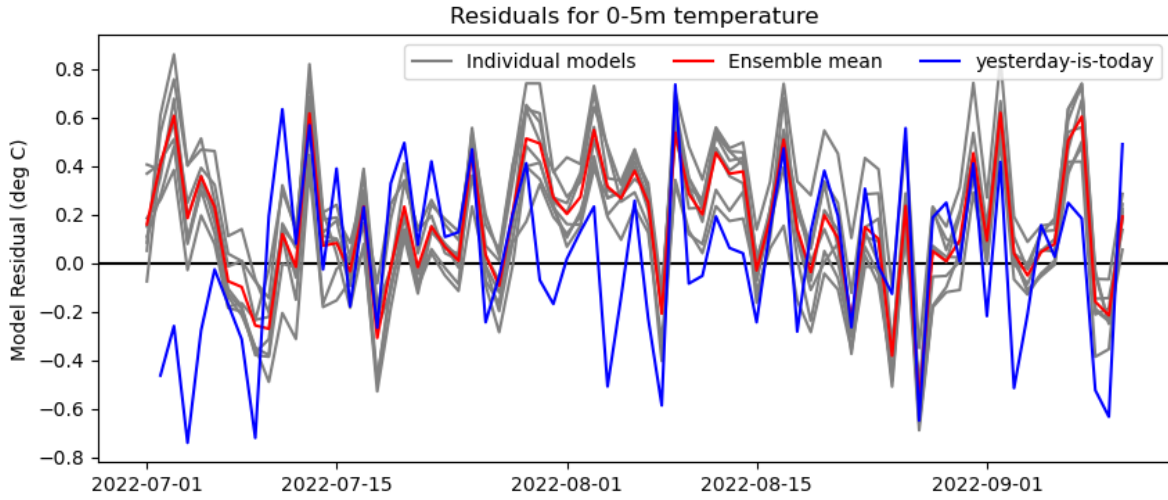
Figure 3: Residuals at the integrated depth of auto-regressive neural network ensemble (grey/red) and persistence model (blue). Initial model performance for the test set was better than persistence with a MSE of 0.09°C (persistence 0.11°C), MAE of 0.24°C (persistence 0.24°C), MAPE: 1.73% (not calculated for persistence).
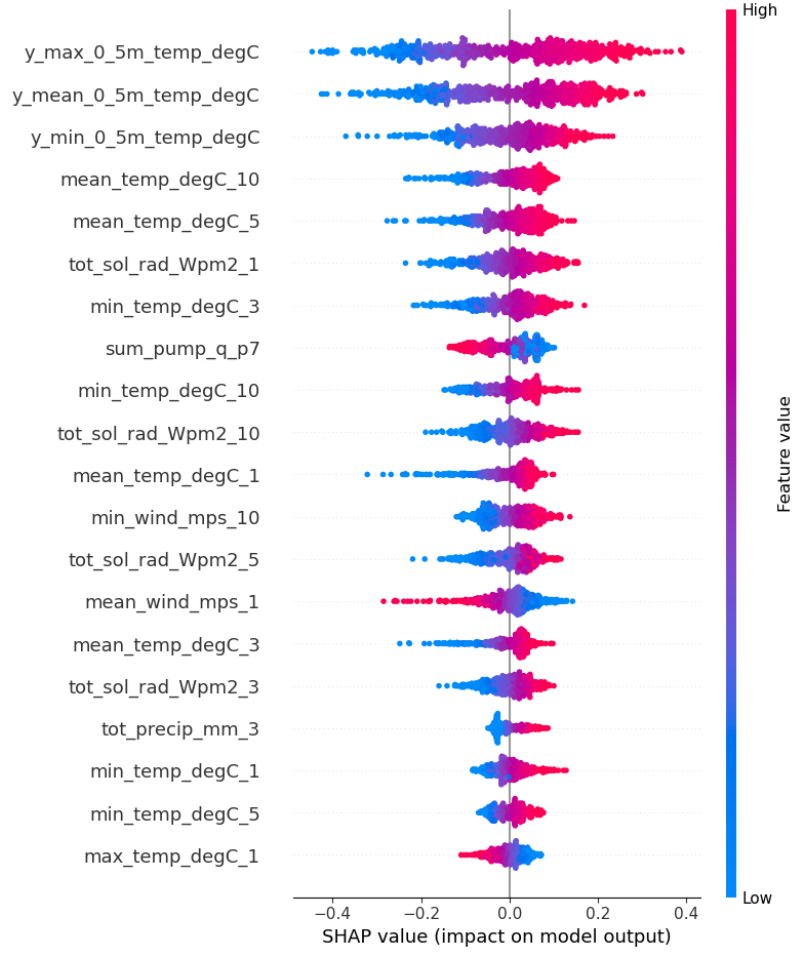
Figure 4: SHAP analysis for predicting the top 1m water temperature at Shadow Mountain from a fully-connected, auto-regressive neural network. Note "sum_pump_q_p7" (sum of pumping volume over the previous seven days), which indicates some sensitivity of near-surface temperature to pumping operations.
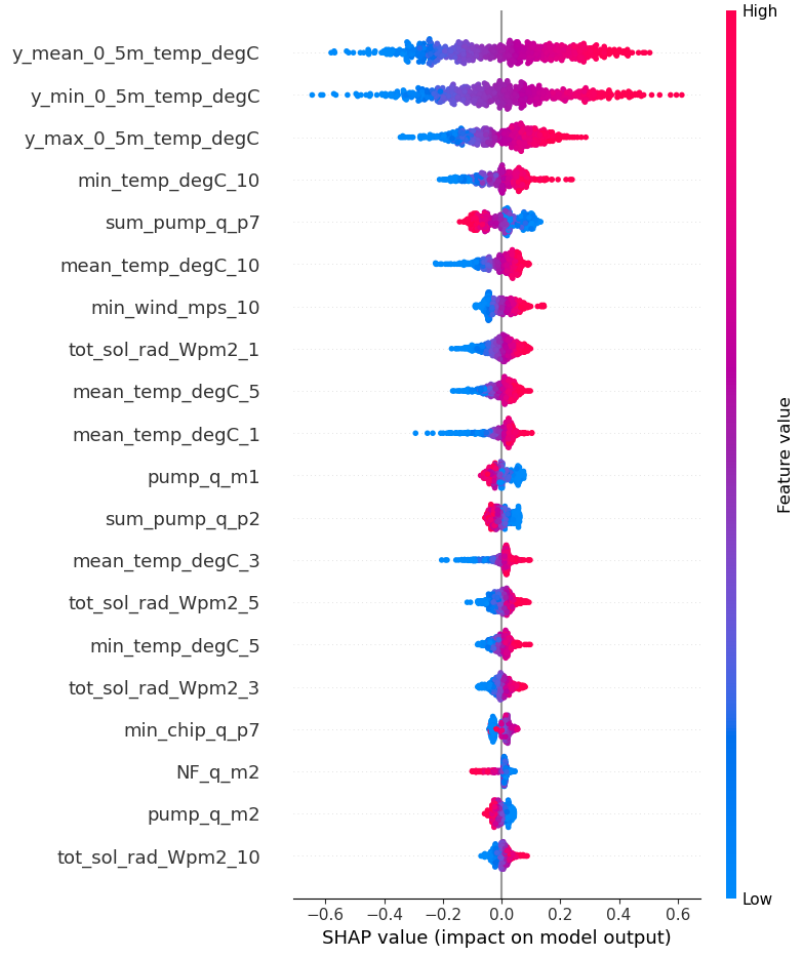
Figure 5: SHAP analysis for predicting the average water temperature (0-5m) at Shadow Mountain from a fully-connected, auto-regressive neural network. Note "sum_pump_q_p7" (sum of pumping volume over the previous seven days), which indicates a relatively strong response in predicted integrated depth water temperature to pumping operations.