Data-Driven Forecasting: Assignment 1
design the forecast system approach; get and process the data

B Steele

[GitHub link](#)

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 that incorporates the previous day's water temperature as an input feature.

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. This will remain the same for the new NN that will predict *today's* temperature.

- 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. Since the primary driver (other than the previous-day's water temperature) is meteorological variables and we are going to use forecasted data to estimate water temperature on a given day, met data will now be summarized as the current day, yesterday, and 3, 5, and 10 days previous. For the application of the model, forecasted met data will be treated as observed met data and incorporated in the summaries that are input into the model.

- 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

2

1, 2, and 7 days. For implementation, we will assume that pumping operations is static from today until the end of the forecast horizon.

- 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. For implementation, we will assume that inflow and outflow data are static from today until the end of the forecast horizon. I will likely try developing a model without these data, though I have found the inclusion of these data to impact my ability to predict tomorrow's temperature.

- NOAA GEFS - While NOAA GEFS is not the best option for forecasts in a topography like the Front Range, it is the only forecast that offers a horizon of at least 7 days (as compared with the HRRR which is 2 days and the NAM of < 4 days). While forecast accuracy drops precipitously after a few days, initial models seem to indicate that a longer forecast horizon is necessary for operational decision support (the previous 7 days of pumping operations is one of the most important operational variables). I will be using the control and all 30 perturbed forecasts from the GEFS model. From the 0.25 degree product, I will use temperature, wind, categorical precipitation (and possibly forecasted precipitation amount, depending on it's impact on accuracy) and from the 0.5 degree product I will use short and long wave radiation since it is not present in the archived 0.25 degree product and it is an important variable in the model. Given the heterogeneity of terrain in this area, it is very likely that there are biases in the forecast that I will need to deal with. For the time being, I'm going to pretend that these forecasts can be used as-is, and will try some de-biasing methods if time allows.

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.

For predicting *today's* water temperature, I may need to sacrifice some accuracy in the initial model to reduce the input parameters (the meteorological parameters, specifically) so they are more inline with the GEFS forecast data. I'm hopeful that I will have better accuracy for *today's* temperature than for the preliminary NN that predicted *tomorrow's* temperature

which will provide some room to make more conservative decisions about the input parameters for the operational model.

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 2014-2021/2014-2022, creating an ensemble prediction of 8/9 models trained on all but one year of data, validated on that hold out data.

- testing: 2022, 2023

I am going to do initial testing on 2022 data, just to check to see if the forecasting approach works, and then move to integrating the 2022 data into the model to forecast for 2023.

Description of the forecast approach

The first step to this forecast approach is to create a NN similar to preliminary model described here. In this case, I will change the targets to *today's* water temperature at 0-1m and 0-5m (instead of *tomorrow's*) so that when I plug in the forecast information into my NN I am predicting that day's temperature. Additionally, I will either need to predict min/mean/max for each depth summary (instead of the current mean value per depth only) or change the NN inputs to only use the mean 0-1m and 0-5m temp (for forecast implementation) instead of any additional summary statistics. I will also be testing implementation of categorical rainfall (instead of standardized absolute values). Upon reasonable performance when compared with the baseline (yesterday-is-today), I will roll out the model using the GEFS forecast for noon local time (18UTC) for a horizon of 10 days from the midnight local time cycle (06UTC), keeping all other input variables static (inflow/outflow/pumping), and recalculating the met summaries at each time horizon. This should hypothetically conserve water balance within the system (ignoring rainfall), since the reservoir height does not change significantly throughout the season (no more than 1 foot).

When analyzing performance of the forecast, I will need to consider that pumping operations are considered static for the forecast horizon in the proposed regime, however, in reality, pumping operations often change over time in the historical record on a daily basis. I know that the preliminary model is sensitive to pumping (see Figure 6, Figure 7), but it is unclear how sensitive the forecast accuracy will be to the static pumping regime relative to the actual operations. To truly test the operational capacity of this decision support system and build

trust in it, it will likely be necessary to run a static pumping operations model and one that incorporates the known pumping regime to truly test performance of the forecast. If the two forecasts are different when the static and actual pumping regimes are different, it will indicate that pumping operations could indeed be used as a "knob" at some point in this application. I will also need to assess whether/how to incorporate the actual inflow/outflow data, since those will remain static in the operational model, but do provide important information according to the SHAP analysis (both North Fork inflow and Chipmunk interflow are present as important variables in the 0-5m output).
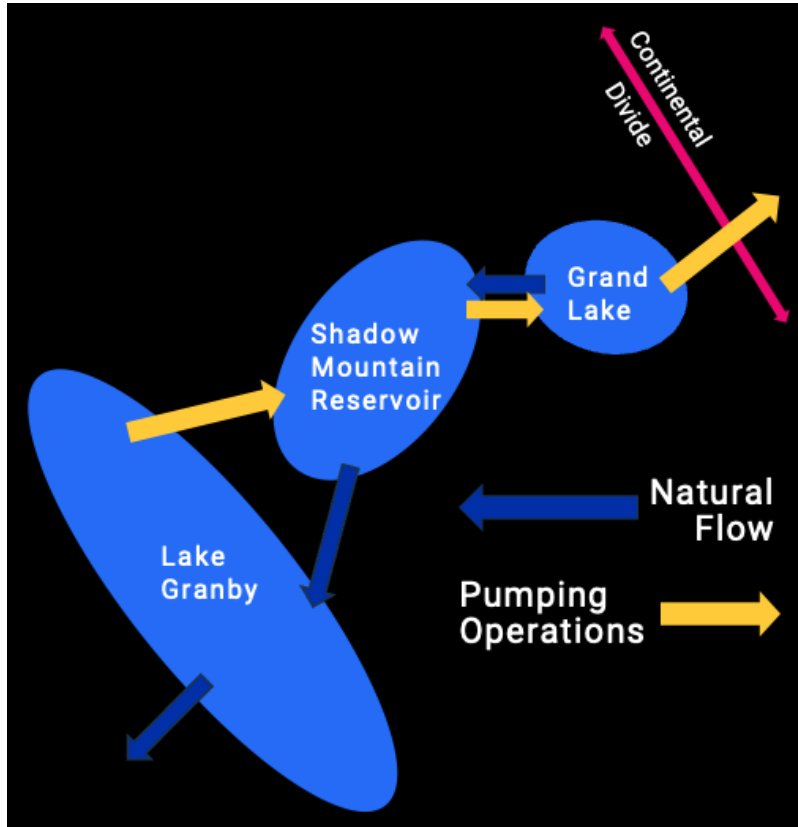
Figure 1: Cartoon schematic of water flow in the Three Lakes System
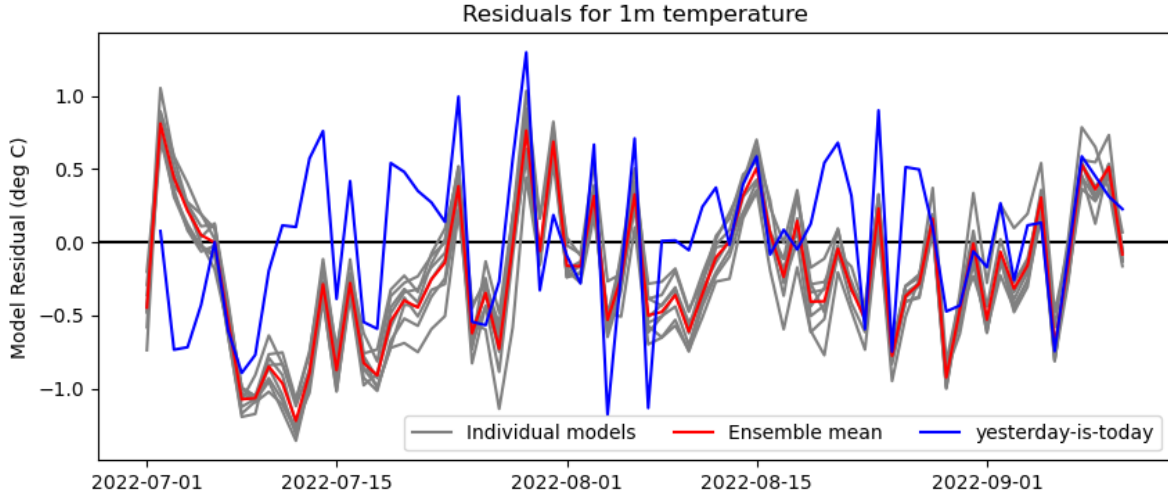
Figure 2: Residuals (predicted minus observed) at upper 1m of auto-regressive neural network ensemble (grey/red) and persistence model (yesterday is today). Initial model performance for the test set was slightly worse than persistence with a MSE of 0.27°C (persistence 0.24°C), MAE of 0.43°C (persistence 0.40°C), MAPE: 2.49% (not calculated for persistence).
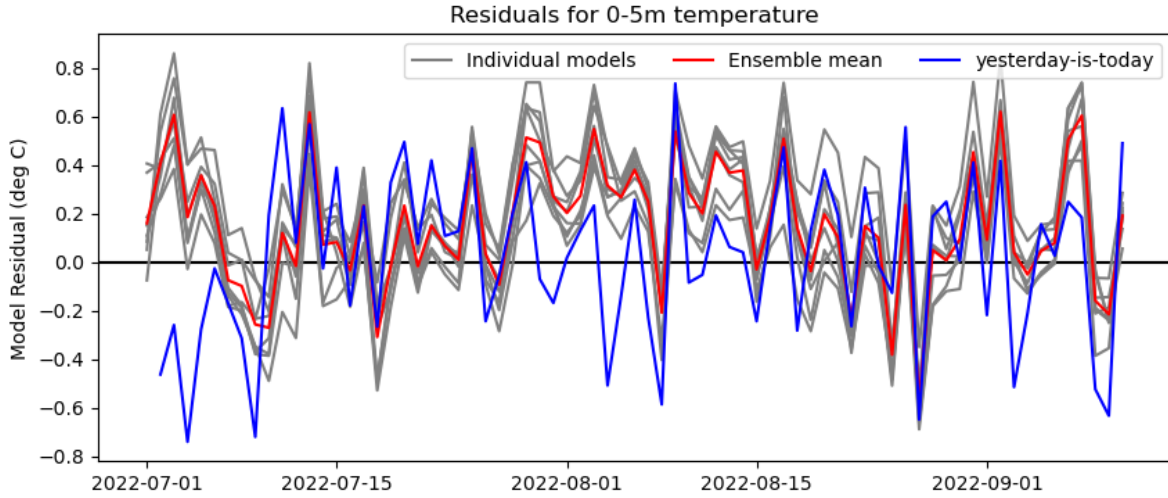

Figure 3: Residuals (predicted minus observed) at the integrated depth of auto-regressive neural network ensemble (grey/red) and persistence model (blue). Initial model performance for the test set was better than persistence with a MSE of 0.09°C (persistence 0.11°C), MAE of 0.24°C (persistence 0.24°C), MAPE: 1.73% (not calculated for persistence).
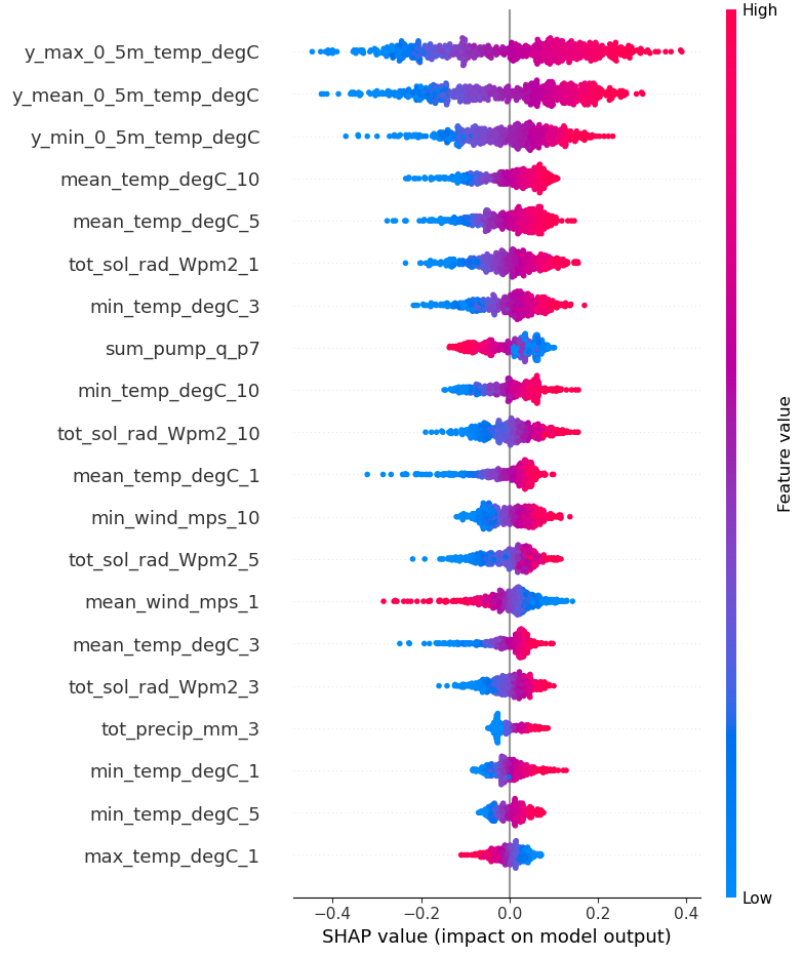
Figure 4: SHAP analysis for predicting the top 1m water temperature at Shadow Mountain from a fully-connected, auto-regressive neural network. Note "sum_pump_q_p7" (sum of pumping volume over the previous seven days), which indicates some sensitivity of near-surface temperature to pumping operations.
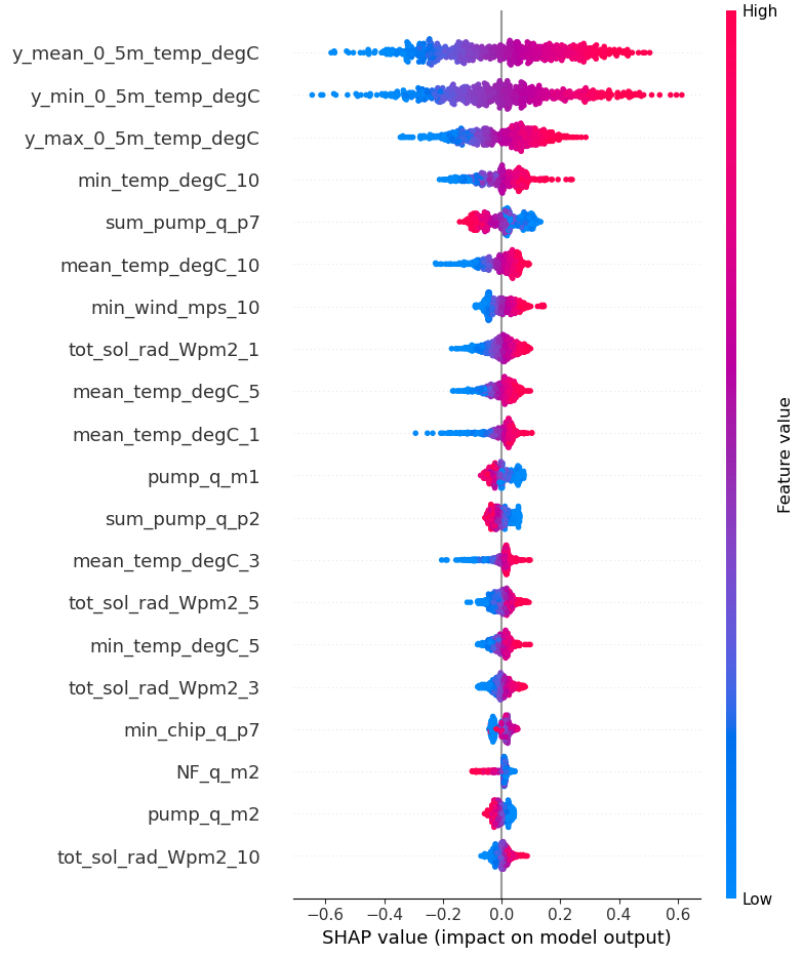
Figure 5: SHAP analysis for predicting the average water temperature (0-5m) at Shadow Mountain from a fully-connected, auto-regressive neural network. Note "sum_pump_q_p7" (sum of pumping volume over the previous seven days), which indicates a relatively strong response in predicted integrated depth water temperature to pumping operations.
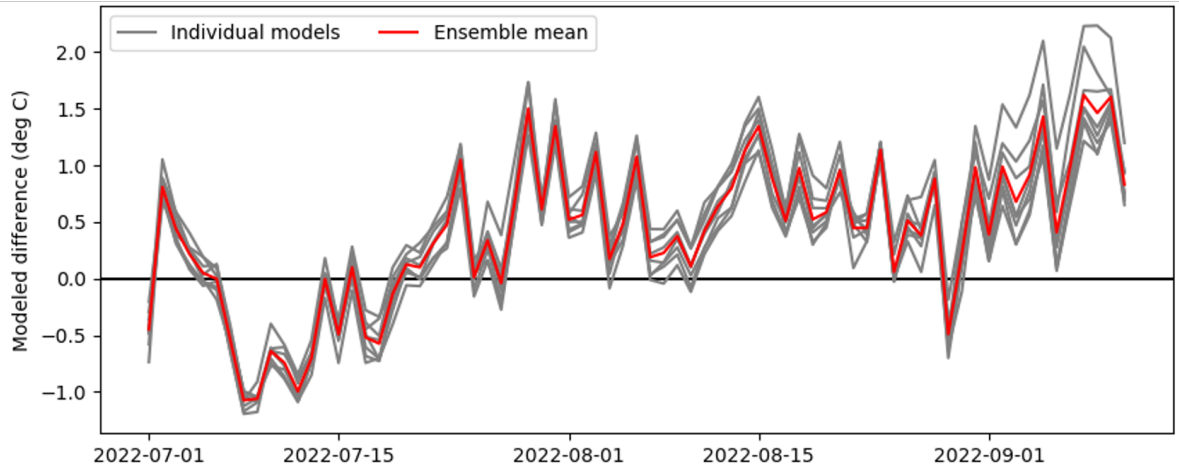
Figure 6: Residuals (predicted minus observed) of test set when pumping operations set to zero for water temperature at 0-1m. This indicates that the model is sensitive to pumping and that the result is sensical (if there is no introduction of cold water, the tempearture is predicted to be higher).
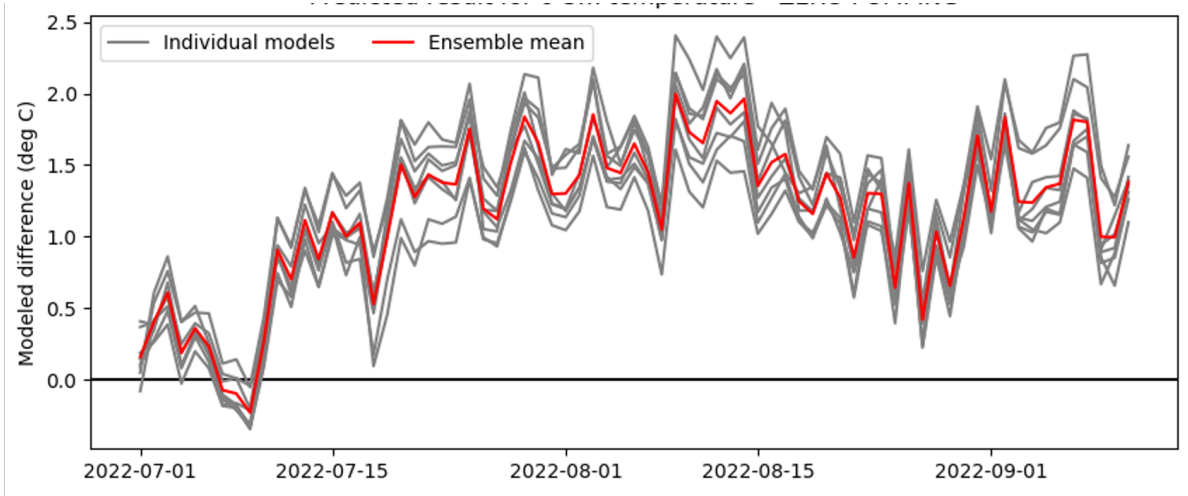

Figure 7: Residuals (predicted minus observed) of test set when pumping operations set to zero for water temperature at 0-5m. This indicates that the model is sensitive to pumping and that the result is sensical (if there is no introduction of cold water, the tempearture is predicted to be higher), additionally the integrated temperature is more dramatically impacted by no pumping operations than the upper 1m temperature in the previous figure.