Data-Driven Forecasting: Assignment 2

prototype of the forecast system; initial evaluation of the data with baselines for comparison

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. Despite the importance of water quality to determine water-system health, consistent and frequent monitoring of waterbodies (by physically visiting a site) or sensor network deployment to monitor water temperature are both costly endeavors (CITE). 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 hypolimnetic water (cold water) from Granby reservoir into Shadow Mountain and then into Grand and into the tunnel to serve the Front Range (Figure 4). 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.

Northern Water has collected extensive data at the Three Lakes System for many years. In 2015, they deployed an instrumented buoy in Shadow Mountain near Chipmunk Lane, the connection between Grand and Shadow Mountain. The data-driven forecast relies on aggregated daily data from the buoy, as well as volume data from inflows (North Fork into Grand Lake, North Inlet into Shadow Mountain) as well as the interflow between Shadow and Grand (Chipmunk Lane). In addition to the hydrologic data, I use lagged meteorology data from a met station at the southern end of Shadow Mountain, which is gap-filled with NCEI data from a local weather station on the north side of Grand Lake, which I assert are similar enough to be interchangable. To create an operational model that we can assimilate forecasted meteorology into, we added tomorrow's meteorology at noon MST to the model architecture described in Assignment 1.

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.

Example Forecast

My goal was to create three types of forecasts (operational, null, and persistence) so that I could enumerate forecast skill of the operational model and determine if the forecasted model also exhibited any downstream effects of pump operations like in Figure 9 and Figure 6. The operational forecast will use actual pumping/inflow/interflow data for future days/forecast outlook, but assimilated predicted water temperature for the autoregressive features and assimilated NOAA GEFS forecasts for the lagged met data.

While I was able to create the architecture for making an operational forecast, it took quite some time, and I was unable to complete the other forecast versions to be able to calculate some of the desired evaluation metrics, so I'll have to work more on that for the final presentation. Examples of operational forecasts can be seen in Figure 1, Figure 2, and Figure 2

I would like to add two additional forecasts once I fix out the bugs in the operational forecast:

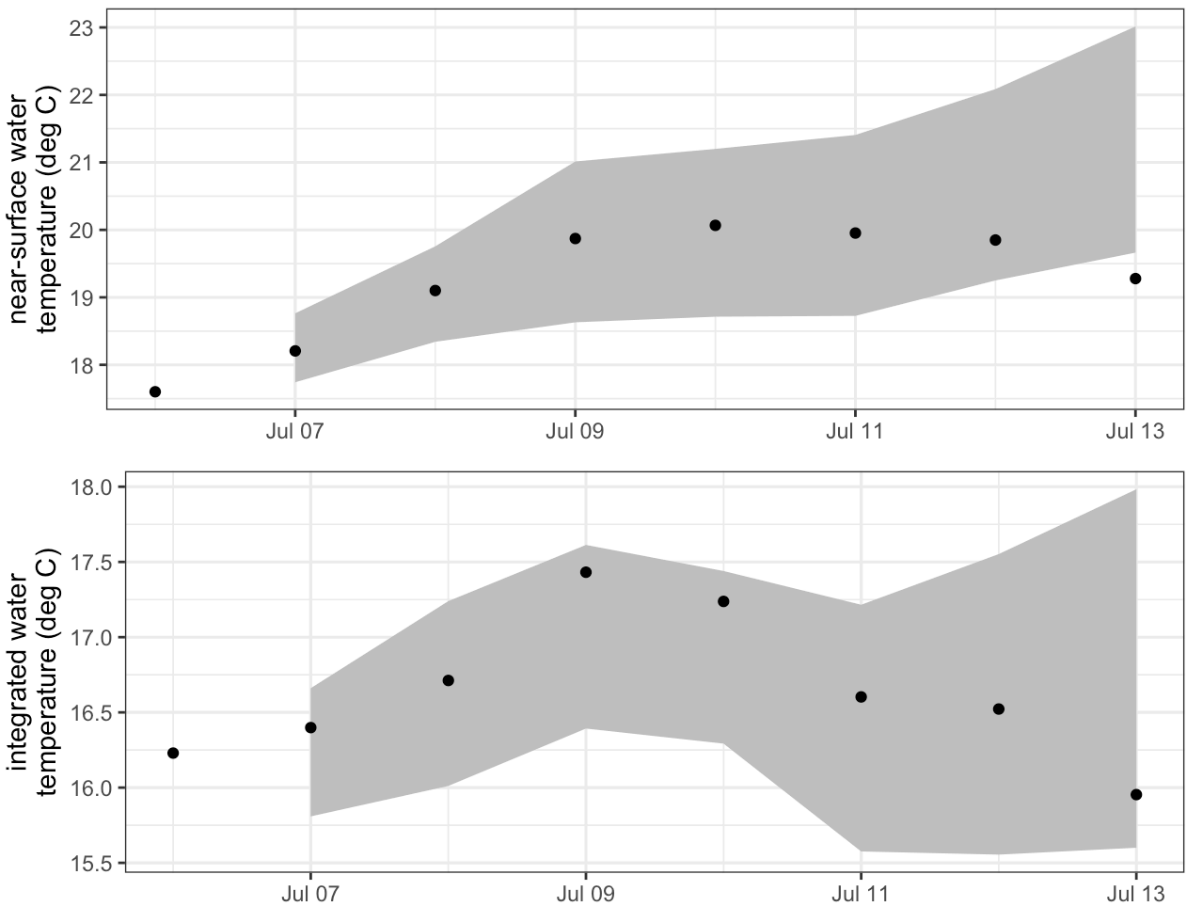

Figure 1: Seven-day forecast of water temperature at Shadow Mountain Reservoir Jul 07-Jul 13. Forecasted range is depicted in grey, observed values as black dots.

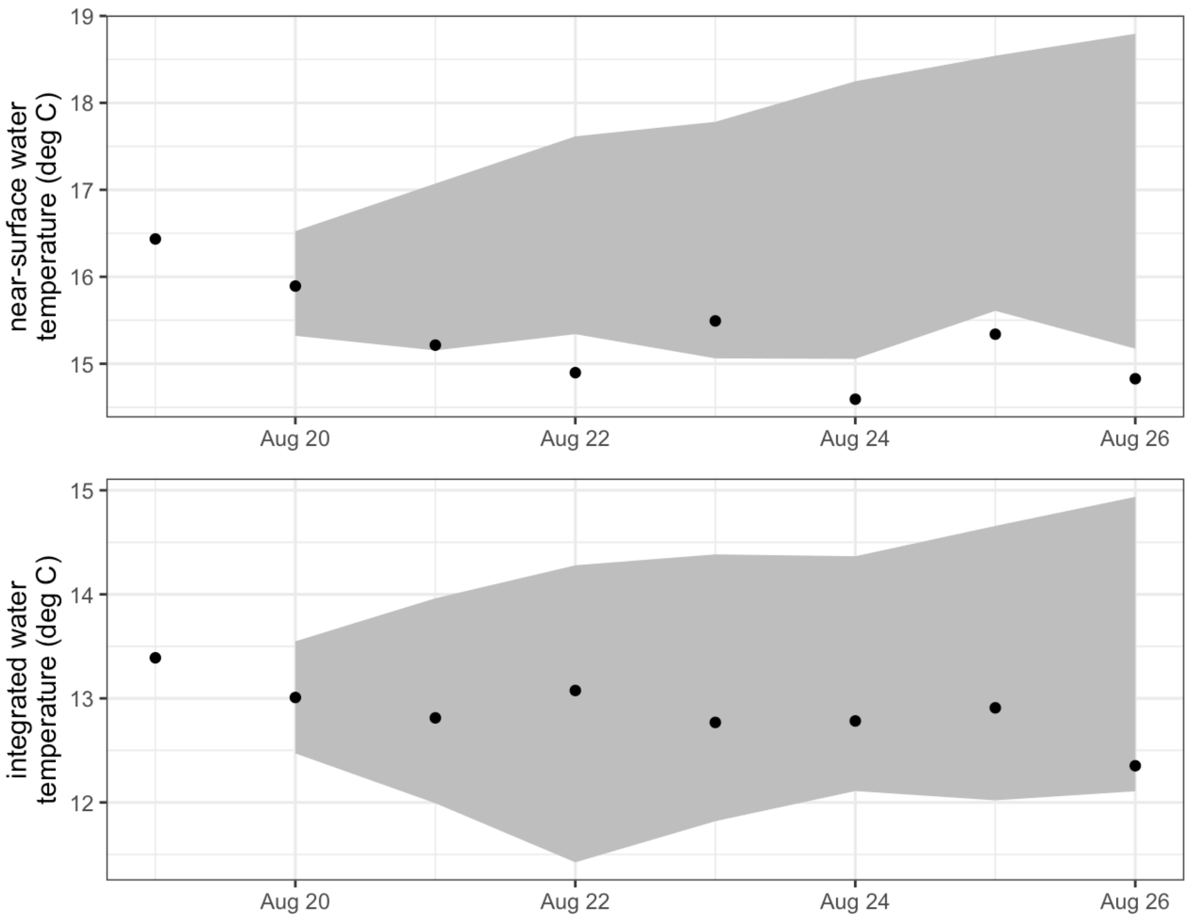

Figure 2: Seven-day forecast of water temperature at Shadow Mountain Reservoir Aug 20-Aug26. Forecasted range is depicted in grey, observed values as black dots.

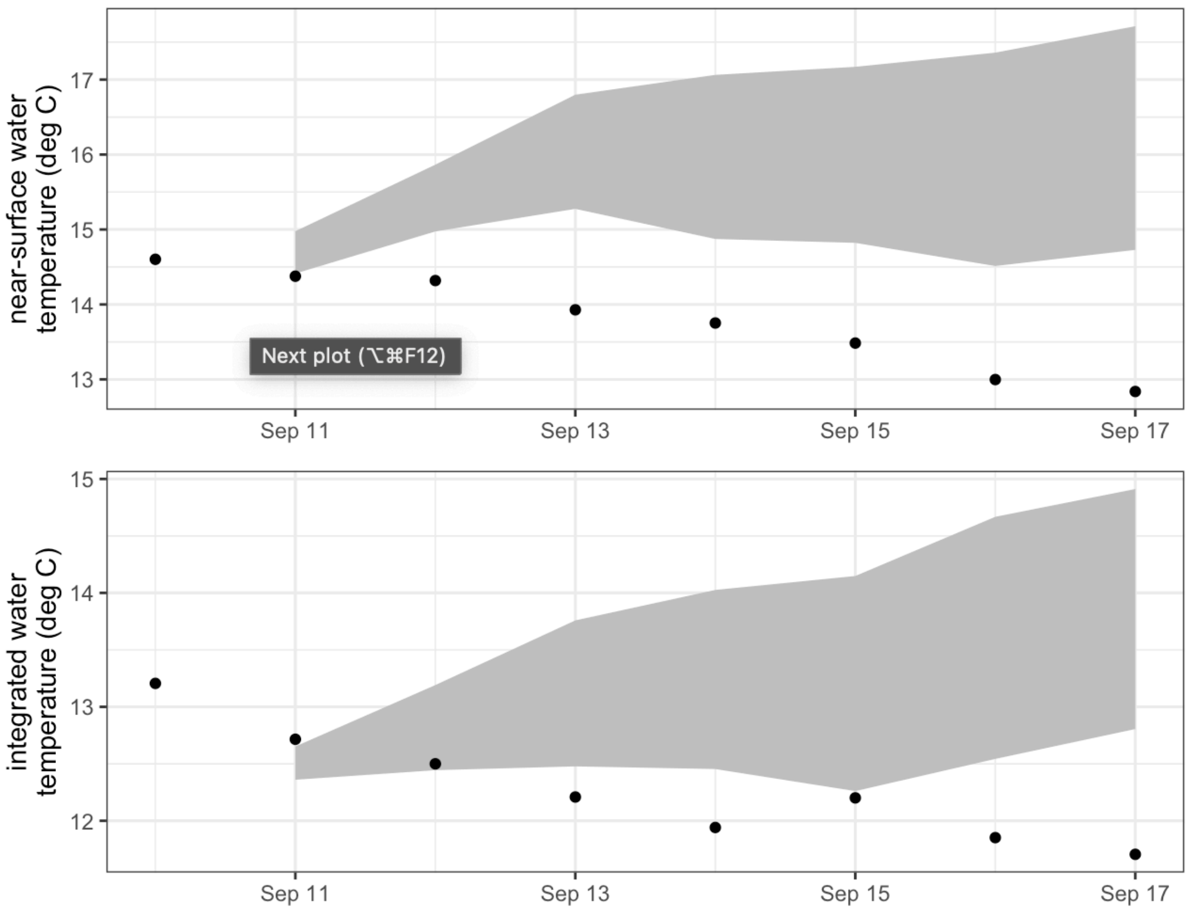

Figure 3: Seven-day forecast of water temperature at Shadow Mountain Reservoir Sept 11-Sept 17. Forecasted range is depicted in grey, observed values as black dots.

1) persistence forecast: use static/zero pump/inflow data but otherwise rollout the forecast like operational - this will start to get at whether or not the 'pump' lever has any impact on our forecasts

2) null forecast: yesterday is today and tomorrow and the day after, but with noise. This helps qualify the value of the operational model, as it serves as the baseline.

Metrics and baselines

To determine the accuracy of the initial data-driven model, I will compare the performance for a single day forecast against the baseline of 'yesterday is today'. As long as the model performs as well, or nearly as well, the forecast rollout is likely to be relatively skillful.

To evaluate the forecasts, I will be calculating CRPS (continuous ranked probability score) where a lower CRPS indicates a more accurate forecast and a higher CRPS indicates a less accurate forecast. I will also be using Continuously Ranked Probability Skill Score (CRPSS) to assess forecast skill:

$$CRPSS = 1 - CRPS_o/CRPS_p$$

$CRPS_o$ = *CRPS of operational forecast*

$CRPS_p$ = *CRPS of persistence forecast*

Like with CRPS, a lower value for CRPSS is a more skillful forecast relative to the persistence and a higher value for CRPSS is a less skillful forecast relative to the persistence.

Within the context of this forecast system, a persistence model is likely the 'best' baseline due to interannual variablility in water temperature on any given day of the year. In order to actually assess skill at temperature mitigation via the pump operations, though, we need to address the value/amount of impact pumping has on the forecast skill. As noted in 'Example Forecast', running the forecast model with pumping operations set to zero will help us learn about this type of skill.

Forecast Evaluation

First evaluation: assess the data-driven model's ability to predict tomorrow's temperature based on observed data compared to 'yesterday is today' baseline.

Model	MSE 1m	MAE 1m	RMSE 1m	MAPE 1m	MSE 0-5m	MAE 0-5m	RMSE 0-5m	MAPE 0-5m
Operational Model	0.37	0.49	0.61	2%	0.08	0.22	0.28	1.56%
Persistence Baseline	0.23	0.38	0.48	2.46%	0.11	0.28	0.33	2.09%

This "operational" auto-regressive neural network model predicted tomorrow's water temperature better than or similar to a persistence model for the regulatory season during 2022 (Figure 5, Figure 6). While the performance at the surface is not better than persistence, it is still quite good. The additional error is a bit expected, as 2022 had some of the highest water temperatures on record.

Using SHAP analysis (a method of explainable AI for neural networks as described in Lundberg and Lee (2017)) we found that operational pumping has an impact on tomorrow's temperature (Figure 7, Figure 8) as does much of the the noon-time met features. The impact of operations on the integrated depth is stronger than the upper 1m and the impact of noon meteorology is stronger in the upper 1m. These are both sensical results that leads me to believe that the forecast implementation should be successful and that the pumping operations could be used as a "knob" to control water temperature to some extent in the Three Lakes System. Further, setting the pumping operations to zero (but not accounting for water balance within the system) shows sensitivity to pumping operations (Figure 9, Figure 10) through increased temperature for both outputs, which is expected.

Next evaluation: calculate CRPS for the forecasts.

The CRPS for the July forecast is 0.33, August is 1.17, and September is 1.85. There is a noticeable difference in model skill across these randomly-chosen dates, with early season (before pumping operations begin) being more skillful than later-season dates.

Without the null model baseline, it is difficult to empirically asses the skill of the forecast relative to it, but just from a visual analysis, the operational forecast is more skillful than a null (yesterday is today plus noise) forecast in all cases except September, where forecasted temperature increased, but the actual temperature decreased for both targets.

While the output of these models is completely reasonable, it is clear that there are varying degrees of skill throughout the year. I would like to plot the CRPS as a timeseries to see if there is any structure in the accuracy over time. Such seasonal performance differences may indicate that the basic debiasing I performed on the GEFS forecast data may be flawed. When debiasing, I assumed that bias was consistent across season and time horizon, which may not be an accurate assumption. It could also indicate that… I completely flubbed something in my code.

Figures

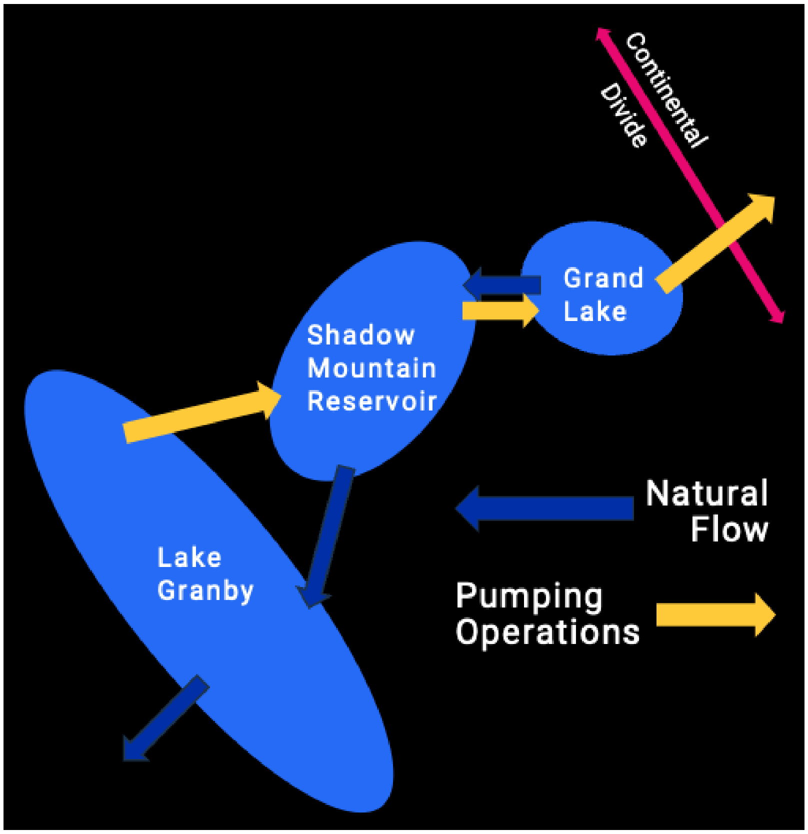

Figure 4: Cartoon schematic of water flow in the Three Lakes System

References

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

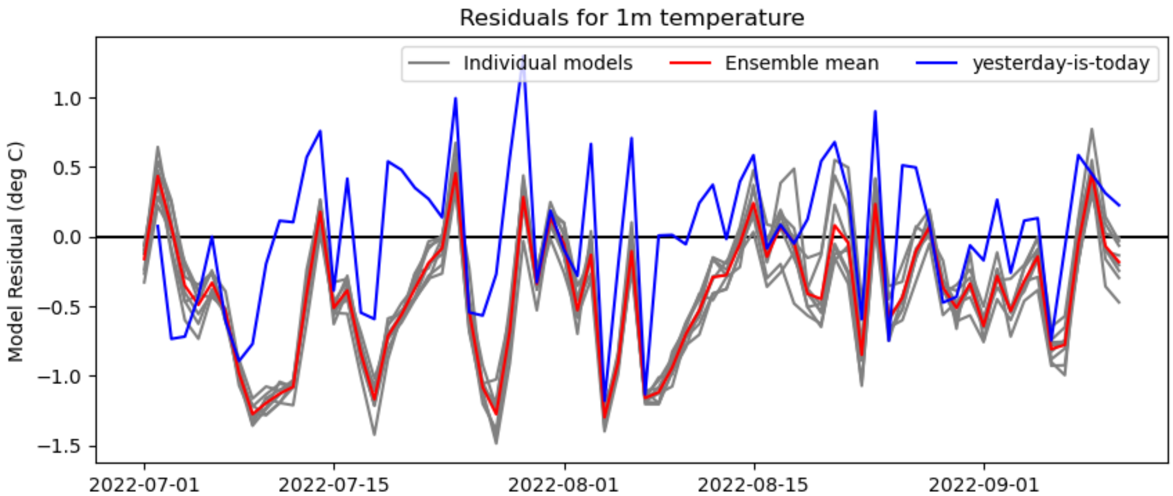

Figure 5: Residuals (predicted minus observed) at upper 1m of auto-regressive neural network ensemble (grey/red) and persistence model (yesterday is today, blue). Operational model performance for the test set was slightly worse than persistence with a MSE of 0.37°C (persistence 0.23°C), MAE of 0.49°C (persistence 0.38°C), MAPE: 2.80% (persistence 2.46%).

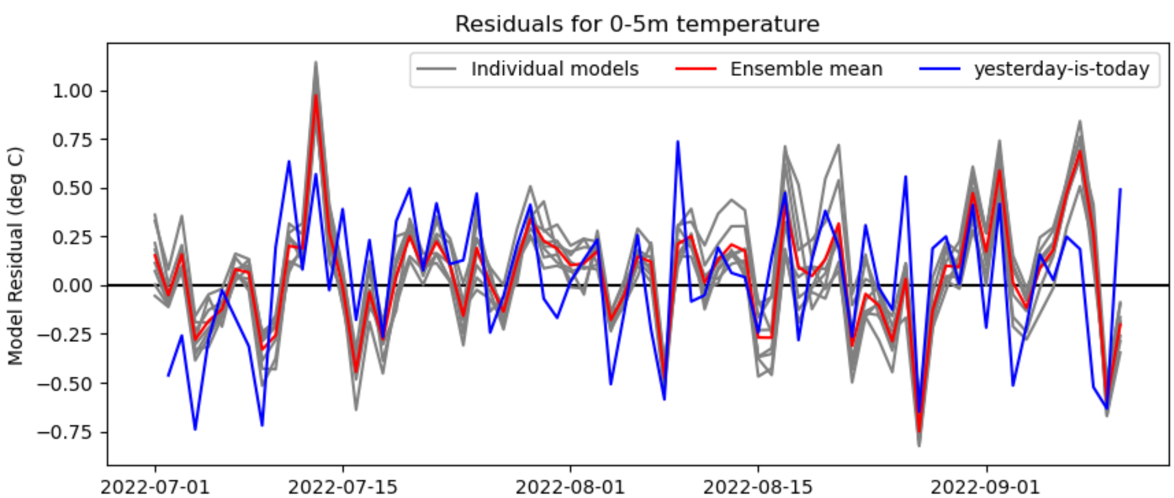

Figure 6: Residuals (predicted minus observed) at the integrated depth of auto-regressive neural network ensemble (grey/red) and persistence model (blue). Operational model performance for the test set was better than persistence with a MSE of 0.08°C (persistence 0.11°C), MAE of 0.22°C (persistence 0.28°C), MAPE: 1.56% (persistence 2.09%).

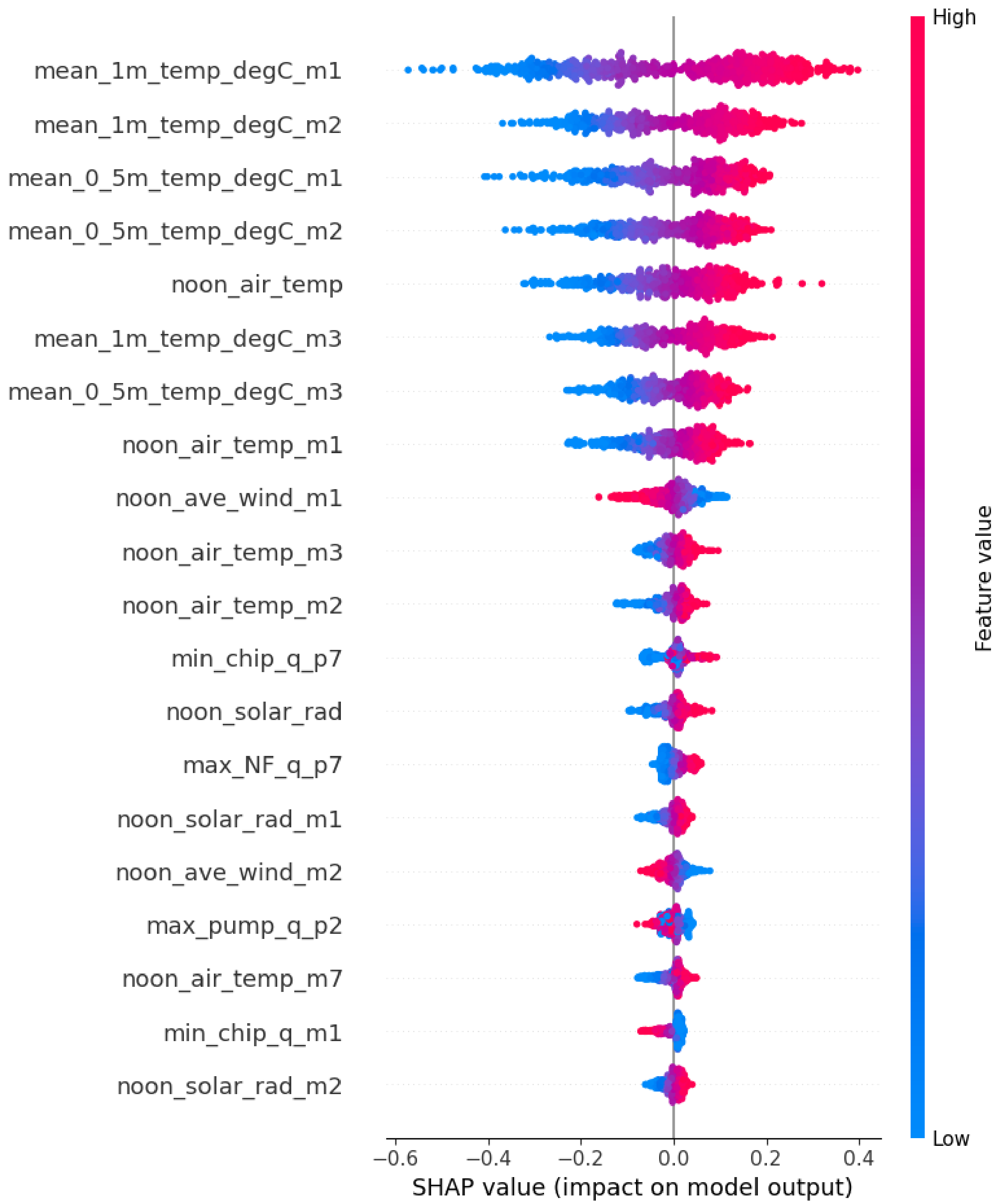

Figure 7: SHAP analysis for predicting the top 1m water temperature at Shadow Mountain from a fully-connected, auto-regressive neural network. Note “sum_pump_q_p7” (sum of pumping volume over the previous seven days), which indicates some sensitivity of near-surface temperature to pumping operations.

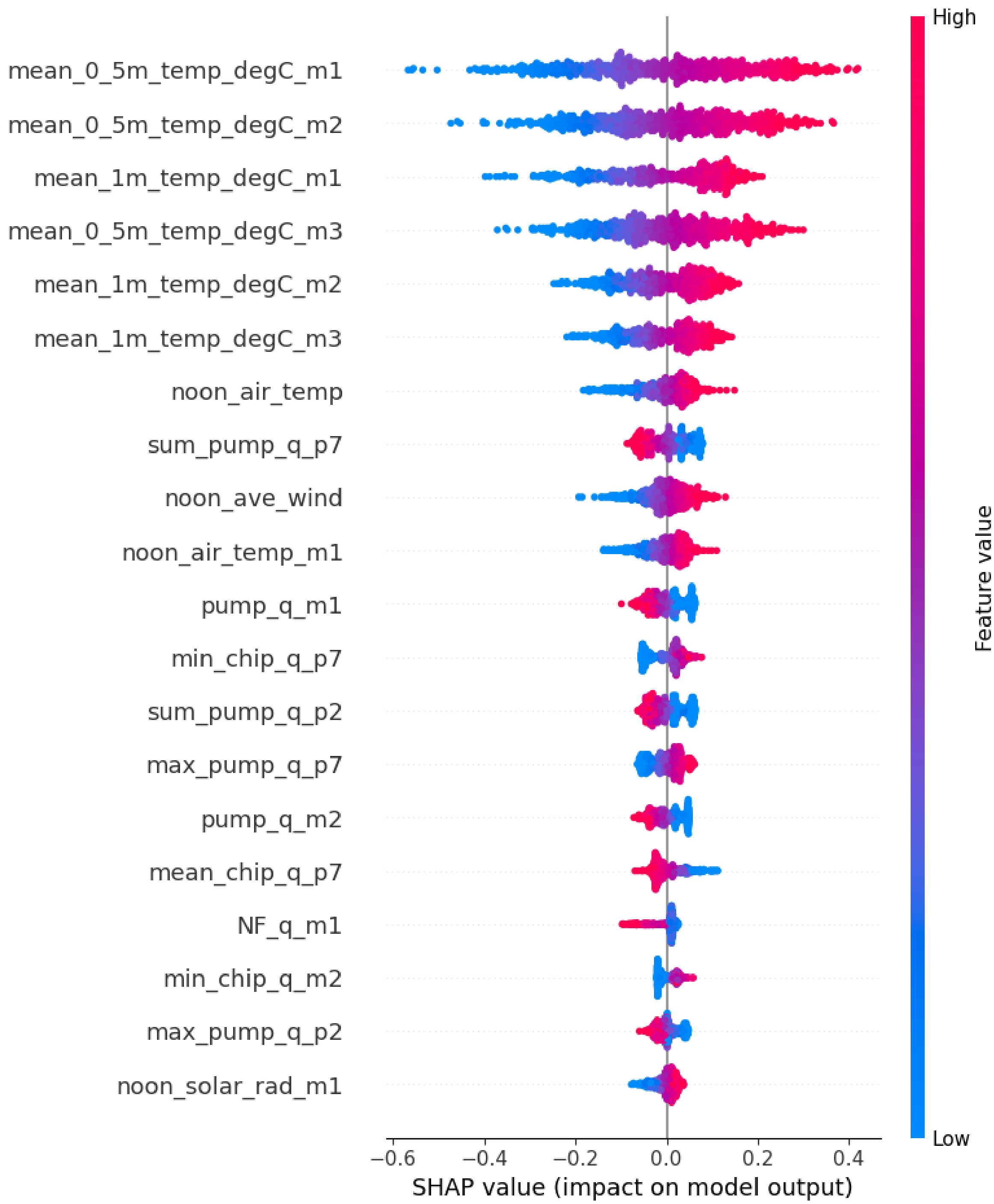

Figure 8: SHAP analysis for predicting the average water temperature (0-5m) at Shadow Mountain from a fully-connected, auto-regressive neural network. Note "sum_pump_q_p7" (sum of pumping volume over the previous seven days), which indicates a relatively strong response in predicted integrated depth water temperature to pumping operations.

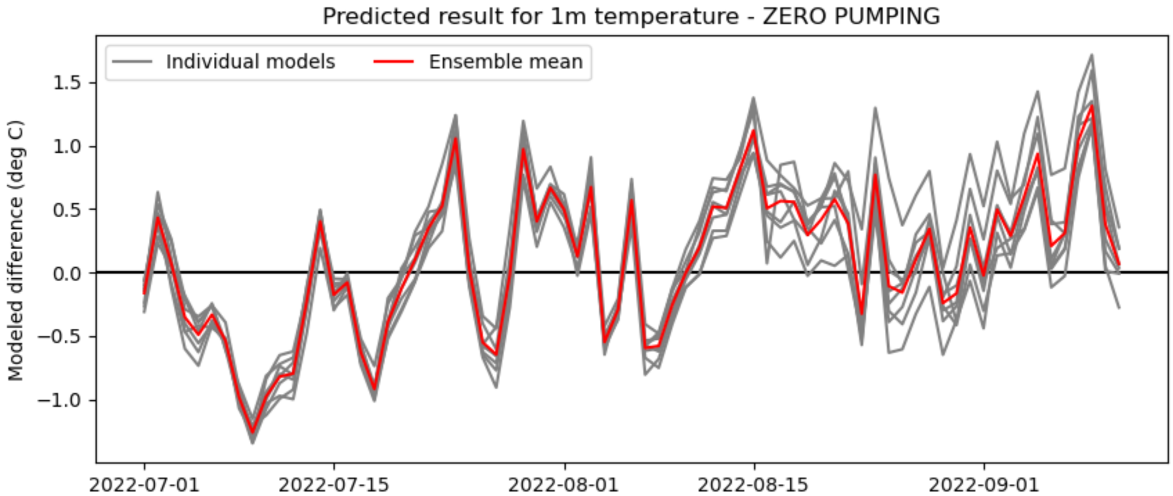

Figure 9: Residuals (predicted minus observed) of test set when pumping operations set to zero for water temperature at 0-1m. This indicates that the model is sensitive to pumping and that the result is sensical (if there is no introduction of cold water, the tempearture is predicted to be higher).

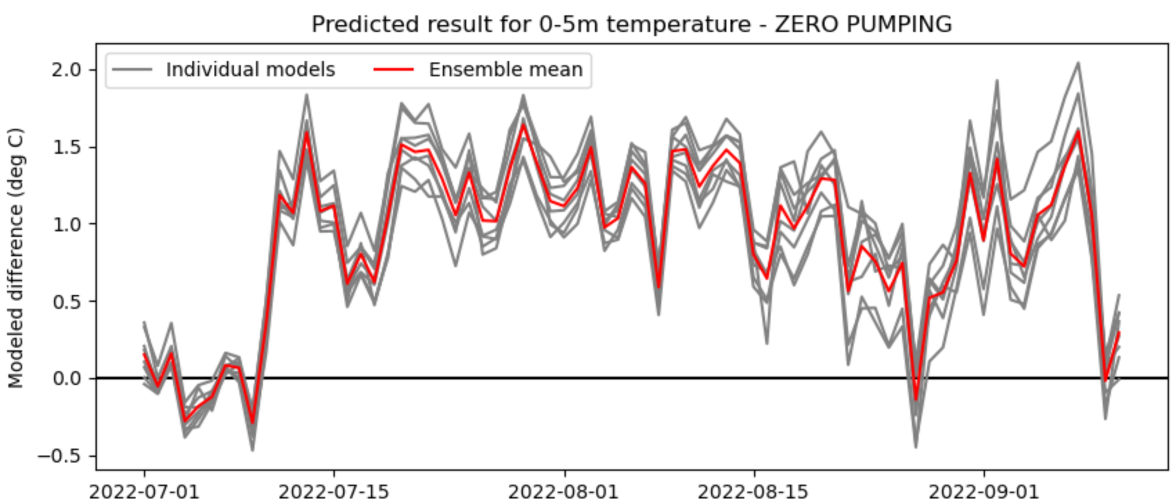

Figure 10: Residuals (predicted minus observed) of test set when pumping operations set to zero for water temperature at 0-5m. This indicates that the model is sensitive to pumping and that the result is sensical (if there is no introduction of cold water, the temperature is predicted to be higher), additionally the integrated temperature is more dramatically impacted by no pumping operations than the upper 1m temperature in the previous figure.