Assignment 3

final testing forecasts and insights into sources of predictability/skill

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. 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 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 in the Summer and Fall and to reduce diatom growth is to keep the average temperature of 0-5m ("integrated depth") greater than 14°C in the Spring and early Summer, which is a bit of a "Goldilocks" problem.

The overarching goal of this forecast system 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 at two depths (near-surface 0-1 meter, integrated 0-5 meter). When operational, we would add an operational "knob" that would alter pumping operations (while maintaining water balance) 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 out of scope for this class, so instead, I will focus on reliable 7-day forecasts using an auto-regressive neural network.

Data-Driven Forecast System

Model development

Northern Water has collected extensive data at the Three Lakes System for many years. In 2014, they deployed an instrumented buoy in Shadow Mountain near Chipmunk Lane, the connection between Grand and Shadow Mountain. Each season the buoy is deployed in late may and removed from the lake in early October, creating relatively independent years of data. The data-driven forecast relies on aggregated daily data from the buoy, volume data from inflows (North Fork into Grand Lake, North Inlet into Shadow Mountain), and volume of the interflow between Shadow and Grand (Chipmunk Lane). In addition to these hydrologic data, I use lagged meteorology data from a met station at the southern end of Shadow Mountain.

To determine the optimal use of the meteorological data, I experimented with two feature sets to predict the 2022 season (where 2022 was a hold-out test set). One feature set used each individual 3h data point (mimicking the GEFS 0.25 output) and another used aggregated daily values from those 3h data points. The observed met data were summarized to mimic the 3 hour GEFS data for instantaneous air temperature, wind speed, and relative humidity, minimum and maximum of the previous 3 hours for temperature, and mean of the previous 3 hours for solar radiation. Using each of the 3h data as features created a very large feature set (>100 features). I compared the performance of two feature sets, one where the 3h data were each used as features and one where the 3h data were used to create daily summaries (Table 1). From this, I determined that I should use the daily aggregations instead of the individual 3h data as it performed most similar to our baseline of a persistence forecast ('yesterday-is-today').

Table 1: Summary of model performance for aggregated and unaggregated met variables for 2022 test data.

Met Features	MSE 0-1m	MAE 0-1m	percent bias 0-1m	MSE 0-5m	MAE 0-5m	percent bias 0-5m
persistence (baseline)	0.23	0.37	2.44	0.12	0.28	2.15
3h data	0.34	0.44	3.00	0.20	0.37	2.87
3h aggregated to daily	0.29	0.43	2.90	0.16	0.33	2.53

The architecture of the initial model to determine input features and the operational forecast system is a fully-connected neural network comprised of a 10% drop out layer followed by 2 layers of 10 nodes each with leaky relu activations. The model was optimized through an Adam optimization function and the loss function was mean square error. Batch size was 64 and learning rate 0.001. To assure that the model was not overfit, I aimed to use hyperparameters that balanced losses across all cross validation data sets, assuring that the validation error was similar to the training error. For the 2022 testing for feature sets and the 2023 operational model, I used an ensemble method for prediction resulting from the leave-one-out validation (resulting in 8 members for 2022 and 9 for 2023).

Operational Model Testing Results

On a one-day time horizon, the operational model performs similarly to the persistence model, indicating that it should have reasonable skill upon roll out (Table 2). The near-surface (0-1 meter) prediction was more accurate during the entire buoy deployment period (mid May until early October) when compared to the persistence model than the integrated depth (0-5 meter depth), which only was comparable to the persistence when you only consider the regulatory period (Jul 1-Sept 11).

Table 2: Testing metrics for operational forecast model at 1 day time horizon across the full season of data and the regulatory period (Jul 1- Sept 11) for the 2023 test set.

model	MSE 0-1m	MAE 0-1m	percent bias 0-1m	MSE 0-5m	MAE 0-5m	percent bias 0-5m
persistence full season	0.26	0.42	2.45	0.12	0.28	2.00
operational full season	0.25	0.40	2.89	0.22	0.37	3.18

model	MSE 0-1m	MAE 0-1m	percent bias 0-1m	MSE 0-5m	MAE 0-5m	percent bias 0-5m
persistence regulatory period only	0.24	0.38	2.25	0.11	0.26	1.79
operational regulatory period only	0.20	0.37	2.20	0.14	0.30	2.20

An aspect of this forecast system is to determine whether it is sensitive to pumping operations. Using SHAP analysis (a method of explainable AI for neural networks as described in Lundberg and Lee (2017)) I found that operational pumping has an impact on tomorrow's temperature (Figure 2, Figure 3). The impact of operations on the integrated depth is stronger than the upper 1m and the impact of 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. Chipmunk Lane ('chip_q', today's volume passing the interflow between Grand and Shadow Mountain) is also impacted by pumping operations (when volume is negative, it indicates reverse flow between the two waterbodies), so it is possible that a water balance type knob (that incorporates operations, and water flow, might elicit a stronger response as an operational knob.

Implementing the Forecast System

In Assignment 2, I acquired noon-time met data from GEFS 0.25°, attempted to debias it, and then implemented the forecast system at a few select dates. After applying the forecast system across all 2022 dates, the forecasts were medicore, with minimal skill. In order to assess whether it was possible to implement a 7 day forecast with skill, I decided to first try a 'control' forecast where I used observed data for all input features *except* the previous days' water temperature. This would allow me to see what the 'best' possible outcome of the implementation could be before adding additional sources of uncertainty as I did in Assignment 2. Unfortunately, attempting a simple debias for all 3h GEFS time horizons was beyond my capacity for this assignment, which meant I could not implement the forecast outside of the control version. I decided it was more useful to spend time assessing this control forecast and sources of predictability and unpredictability than attempting to accurately debias the GEFS and implement a rollout with those data without knowing what the 'ceiling' of performance is. Because my assessment of preliminary skill of the operational model is based on only one day of forecasts averaged across the year, the performance when rolled out could run away even with observed data. My initial debiasing of the GEFS data for Assignment 2 was based

on the control forecast at midnight only, and I assumed the debias was the same at all time horizons (which I suspect is not true). I also believe that by only using noon-time forecasted variables, I was missing a lot of the drivers for near-surface temperature which are driven by late-afternoon air temperatures and total solar radiation throughout the entire day.

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. Because I'm using an ensemble method across the leave-one-out cross validation, I can still calculate CRPS, since there are 9 members and a probability distribution function can be meaningful as long as the values encompass a full range of anticipated values. While I can't guarantee that this subset represents all possible initial conditions, there is some variability in predictions across the ensemble members, especially in the early and late season (Figure 4, Figure 5). It is possible that 9 years is insufficient for incorporating more decadal patterns (ENSO) that could add to the range of predicted values and provide a more robust PDF for CRPS assessment.

Results and Skill of Forecast System

The control forecast for both near-surface and integrated depths shows some visible skill, even during periods of observable day-to-day variability (late August at 0-1m, Figure 6). The forecast system generally underestimates temperature during times of waterbody thermal instability (mid June through mid July) and especially at near-surface during the warmest observed temperatures (Figure 6, Figure 8).

To better assess forecast skill, we can examine the CRPS across time. Figure 7 and Figure 9 display CRPS per forecast date across the 7-day time horizon. While the ECMWF suggests that any CRPS value at/above 1 is wholly inaccurate which indicates that a fair deal of the forecasts made are quite inaccurate.

The rollout of this forecast (Figure 6, Figure 8) was consistent with the original test performance (Figure 4, Figure 5). Because the most influential features for each of these forecasts are the previous days' temperature (Figure 2, Figure 3), and we allow those values to propagate in the forecast, I feel confident that the forecast is stable at a 7 day time horizon since the rollout did not induce greater variability.

While the CRPS assessment indicates that the forecast system is poor performing (Figure 7, Figure 9) and mostly shows increasingly worse skill over the 7 day time horizon, I suspect that some of the lack of skill is attributable to a lack of variability in the ensemble. This lack of variability results in a narrow PDF which decreases apparent skill via CRPS. At times where the ensemble has greater variability (specifically at the beginning of the season), CRPS indicates greater skill, which would support this assertion. The general convergence of all ensemble members during the warm-up and cool-down periods of the water body could also indicate that a) the training and validation data are very similar and do not have sufficient variation to generalize well, b) the models are overfit and overly confident in those predictions,

c) the way the waterbody behaves during warmup/cool down is consistent across years, but the features used inadequately capture the physical behavior of the water.

Future Work

I suspect that setting my baseline as persistence and trying to match performance on a single day time horizon may have lead me astray a bit in this endeavor. By focusing on a single-day metric, it may have reduced variability through over parameterization of the features included in my model. While the models aren't overfit by the most traditional definitions (e.g., I employed a dropout layer, monitored testing and validation loss curves, set hyperparameters to have balanced models across timeseries cross validations), by including a large feature set for model development, the individual ensemble members may have picked up on nuances in the data that were specific to a single year/subset of data reducing generalizability of the forecast system.

For future model development, I coud reduce the feature set which may result in slightly poorer performance at a 1-day time horizon, but may allow for better generalizability during implementation of the ensemble. This could also backfire and create a model that does not perform well at all just to create a wider PDF to serve the CRPS calculation. In Assignment 2, the PDF area increased upon rollout with all 31 ensemble members of the GEFS forecast, perhaps that would be a better use of time to see how useful the GEFS can be in future model development, or whether a different ensemble forecast system with higher temporal and spatial resolution (like NAM or HRRR) and should be used despite the constraint of forecast horizon.

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.

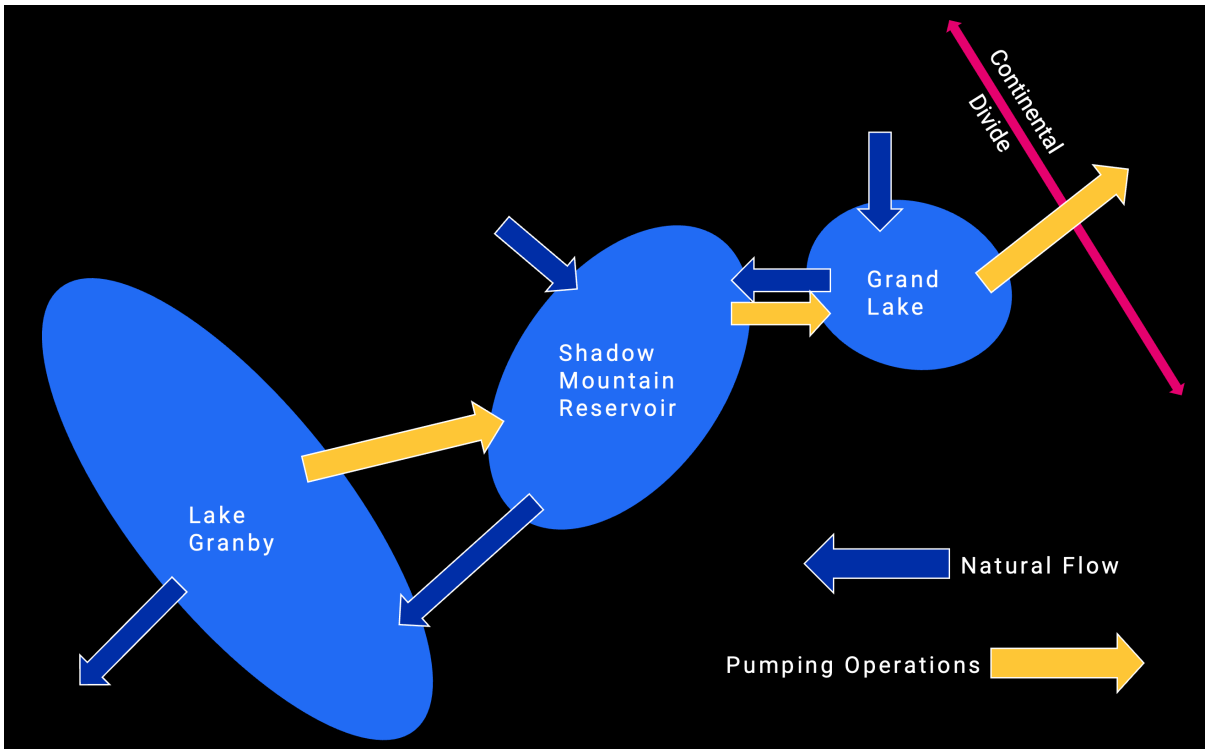

Figure 1: Cartoon schematic of water flow in the Three Lakes System

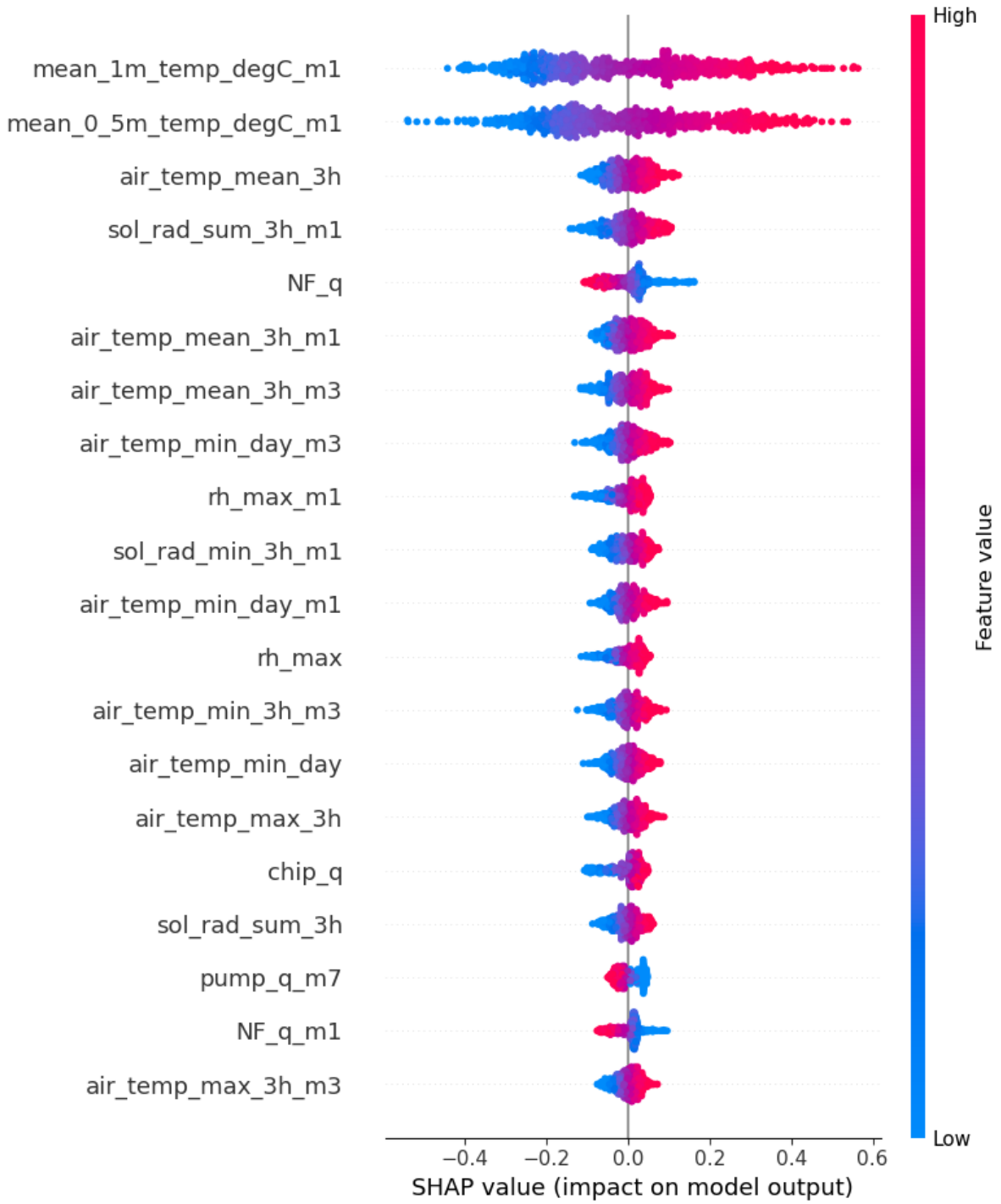

Figure 2: SHAP analysis for predicting the top 1m water temperature at Shadow Mountain from a single member of the ensemble of a fully-connected, auto-regressive neural network. Note "pump_q_m7" (the pump volume seven days prior), which indicates some sensitivity of near-surface temperature to pumping operations.

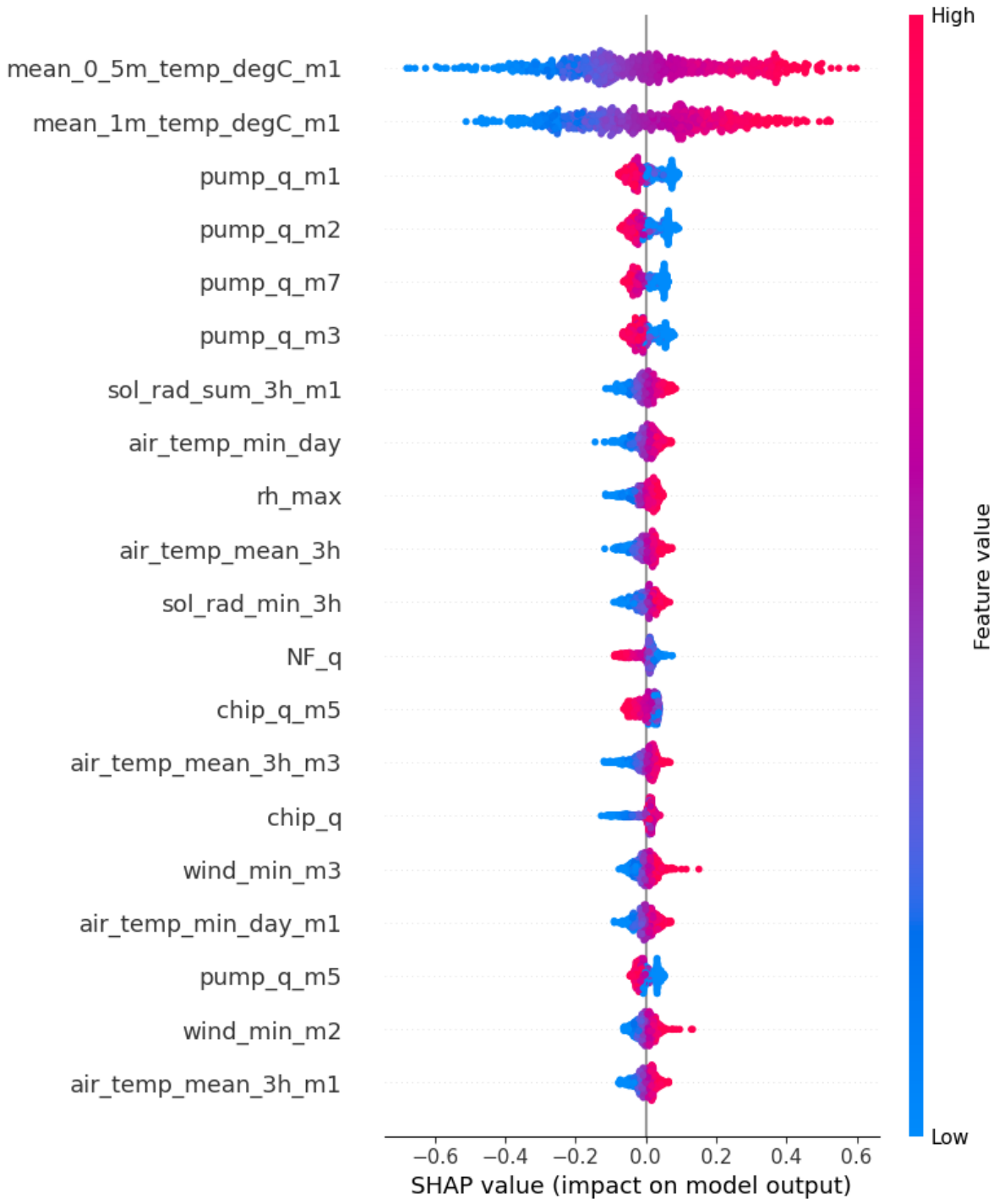

Figure 3: SHAP analysis for predicting the average water temperature (0-5m) at Shadow Mountain from a single member of the ensemble of fully-connected, auto-regressive neural network. Note "pump_q_m*" (lagged pumping volume), which indicates strong response in predicted integrated depth water temperature to pumping operations.

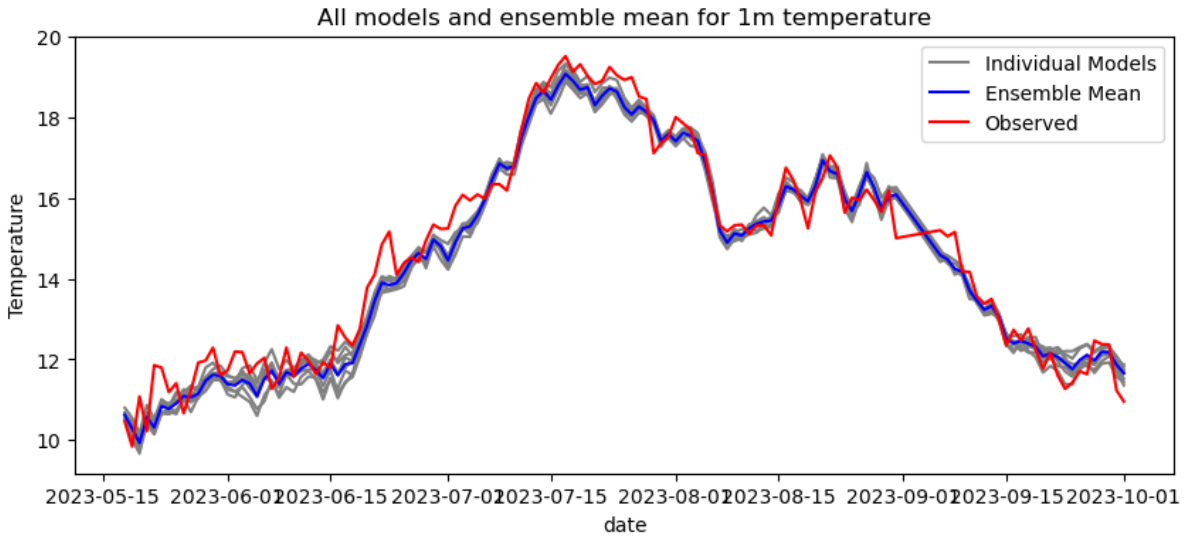

Figure 4: Operational model results for the near-surface temperature at a one-day time horizon, where the observed is in red, the ensemble members in grey, and the ensemble mean in blue.

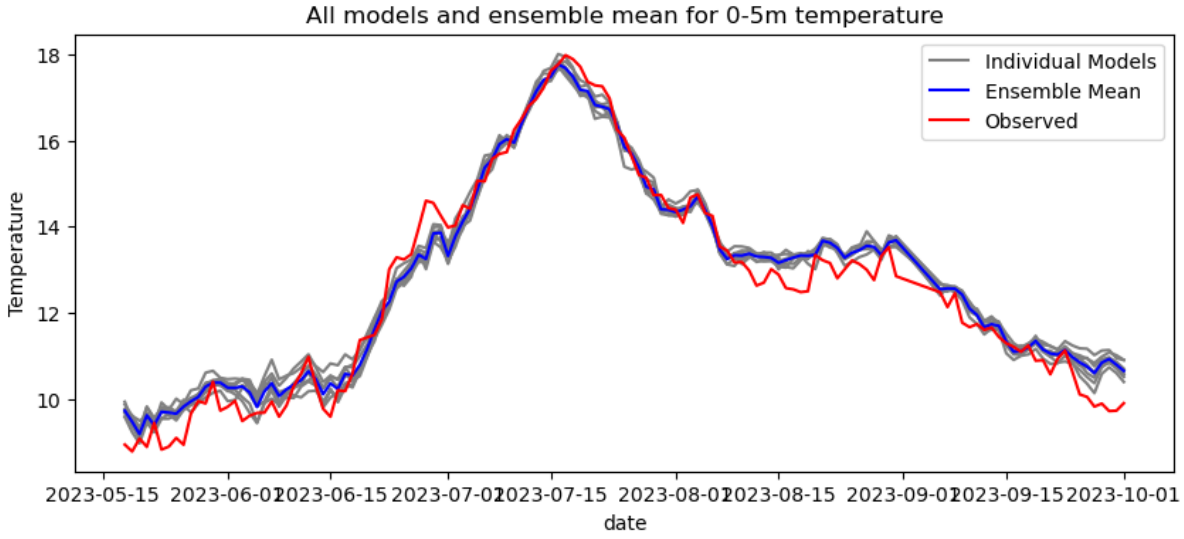

Figure 5: Operational model results for the integrated depth temperature at a one-day time horizon, where the observed is in red, the ensemble members in grey, and the ensemble mean in blue.

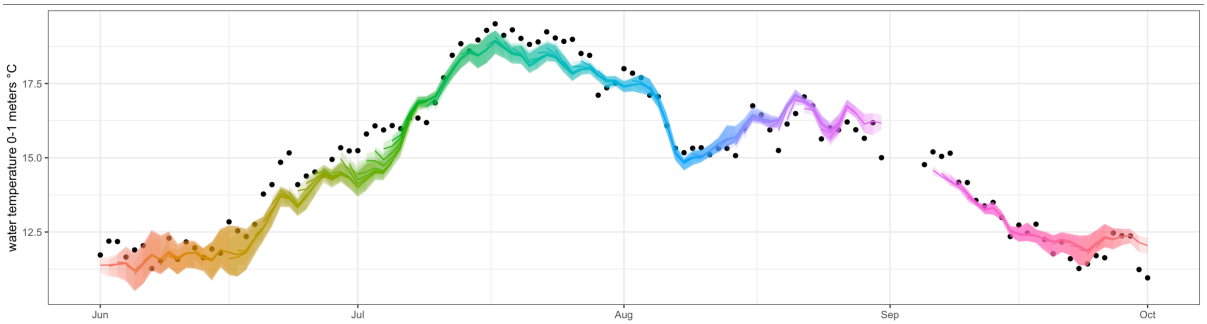

Figure 6: Control forecast rollout results for the test set (2023) at near surface. Observed temperature in black, solid lines are the ensemble mean of the forecast, and shaded area is the range of predicted values.

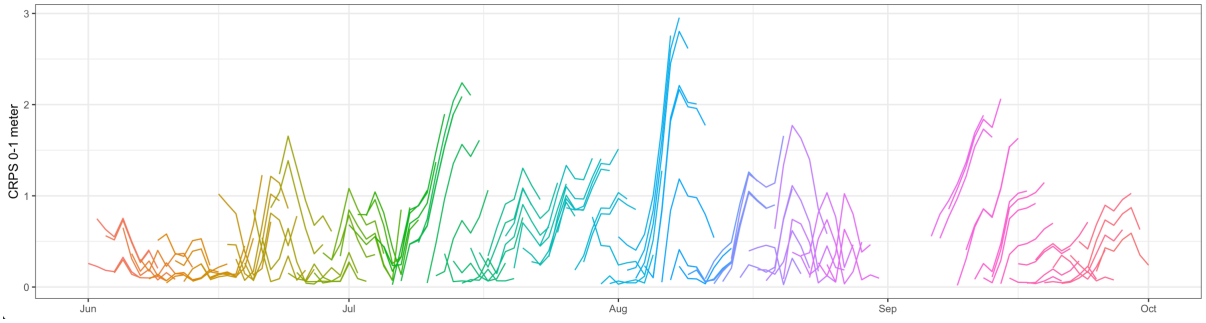

Figure 7: Timeseries of CRPS (continuous ranked probability score) for near surface temperature. Each colored line represents an individual forecast date.

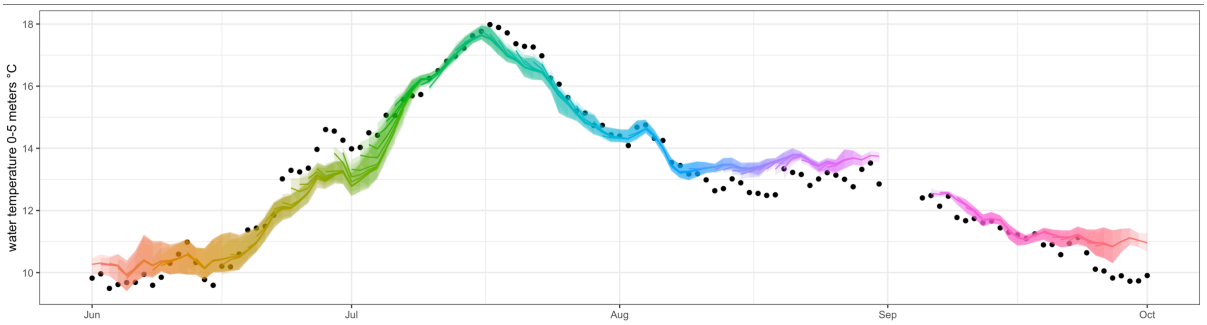

Figure 8: Control forecast rollout results for the test set (2023) at integrated depth. Observed temperature in black, solid lines are the ensemble mean of the forecast, and shaded area is the range of predicted values.

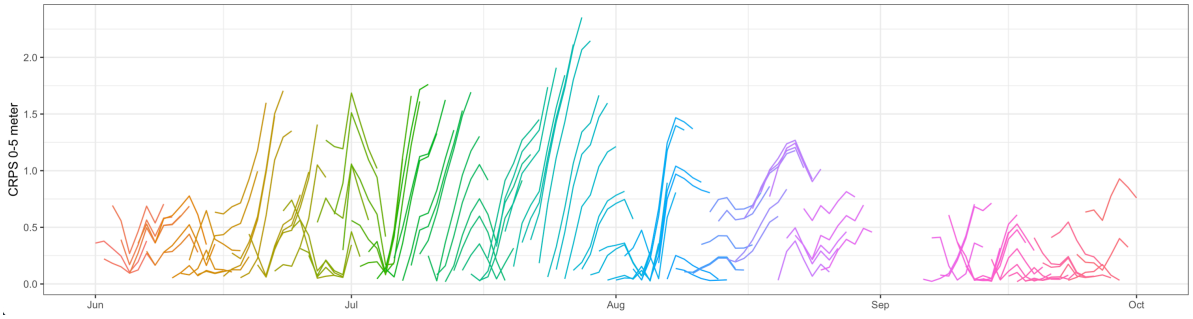

Figure 9: Timeseries of CRPS (continuous ranked probability score) for the integrated depth. Each colored line represents an individual forecast date.