THESIS

USE OF MACHINE LEARNING IN THE AQUATIC SCIENCES

&

A DATA-DRIVEN FORECAST SYSTEM AS AN OPERATIONAL DECISION TOOL IN A MANAGED RESERVOIR

Submitted by

Bethel Gene Steele

Department of Ecosystem Science and Sustainability

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2026

Master's Committee:

Advisor: Matthew R.V. Ross

Frances Davenport
Ed Hall

ABSTRACT

USE OF MACHINE LEARNING IN THE AQUATIC SCIENCES

&

A DATA-DRIVEN FORECAST SYSTEM AS AN OPERATIONAL DECISION TOOL IN A MANAGED RESERVOIR

The use of machine learning methods in aquatic sciences has increased greatly over the past few years, at least in part due to the rise of open-source packages that make it easy to apply these complex models in the complex environment we work in. In the current time of generative artificial intelligence and their integrations into coding interfaces, the process of creating machine learning models has become as easy as typing a request in plain language, making a coffee, and reviewing the results. This low barrier to entry has provided a great resource for researchers to better understand their systems and create global models with large datasets or niche models with limited data. This boom, however, has not yet been met with adequate resources for how to rigorously assess machine learning use in aquatic science, especially at the manuscript/journal interface.

The first chapter of this thesis presents an essay that draws attention to the democratization of coding tools and how it may impact aquatic sciences and the follow-on benchmarks created with these machine learning techniques. The call to action is that better infrastructure must be provided to the author, reviewer, and editors to be sure that the methods and results described in a manuscript match the code used to build and assess the models. This essay is aimed toward the system we work in as aquatic scientists: it is not aimed at solely authors (who are experts in aquatic

science) and reviewers (who volunteer their time to referee manuscripts) that need a framework, but also our journal infrastructure needs an update.

The second chapter presents a decision tool built on machine learning best practices as a case study: from conceptual model to stakeholder engagement and trust building. The tool was designed for the Three Lakes System (Grand Lake, Shadow Mountain Reservoir, Lake Granby) managed by Northern Water, who supplies drinking water to about one million people in northern Colorado via municipal utilities and irrigation for ~600,000 acres of agricultural land through the Colorado-Big Thompson Project. In this modeling effort, Northern Water and partner agencies work together to attain water quality and clarity goals in Grand Lake and Shadow Mountain Reservoir. Because the two waterbodies are hydrologically connected, water quality issues in Shadow Mountain Reservoir can directly impact water quality in Grand Lake. This application uses a machine learning tool to forecast water temperature, a primary biological control, in Shadow Mountain Reservoir at a 7-day time horizon that is responsive to operational changes in water delivery that consistently outperforms a persistence model, with a 7-day continuous rank probability skill score of 0.62 at the near-surface depth and 0.78 at the integrated depth relative to a persistence model, where a value of 1 is a perfect forecast at this horizon. Our end users can use this tool to test operational scenarios for real-time forecasting and it can be used to estimate how operations have impacted water temperature within the system. In a hindcast scenario the tool estimates that water temperature was reduced by an average of nearly 3 degrees at the near-surface and 4 degrees at an integrated depth over the course of the adaptive management season July 1 – September 11, 2025.

ACKNOWLEDGEMENTS

I am incredibly grateful for Lauren Mehl, who made sure I was fed, watered, and well-slept throughout the last big push of this effort. Lauren, you have been my rock through all of this and I can't express enough how lucky I am to have landed a catch like you. Also, to my local bike and queer community who probably don't even know how much they help keep me moving forward across other realms of existence, thank you.

Big thanks to Matt for removing the MS requirement for the Data Scientist position I currently have, for encouraging me to step into a leadership role, and for consistently pushing me to trust my expertise in this field. To Kak and Holly, and all the other women scientist that I've worked with over the years: thank you for setting an example of how to exist in a world that isn't "made for us" and helping me understand systems thinking and the power of interdisciplinary science. I wouldn't be the scientist I am today without your guidance over the past two decades. Gratitude to my committee for patience and great questions and for their time in service of applied science. To Water Boi Café and the ROSSyndicate, I am so happy I get to work with you all. Thanks for listening to me iterate over the same thing for the past three years and for being the best lab mates a guy could ask for.

Gratitude also to the Northern Water Water Quality Team (Becca, Juan, Matt G, Jen) and the Hydros Team (Jean Marie and Kevin) for coming along for this adventure and for your willingness to iterate. Your collective knowledge sharing and interest in this project has made it immensely rewarding. To Erin and Kelly, thank you for believing in me and treating me like a peer, I'm not sure I can adequately express how it's provided confidence in my day-to-day experience as a scientist.

Throughout the process of working on and writing my thesis, I used Claude AI and Claude Code to help keep me focused, provide code chunks to help in difficult applications, and provide clarity to existing text. I used Claude like a writing partner to bounce ideas off of and generally help scope sections, especially when I had too much text and needed to decide what could stay in the main text, what could go to the Appendix, and what could be removed altogether. When I got stuck with a sentence that felt awkward, I asked for help. When I needed help with literature examples to support claims, I asked Claude for help. When I had a difficult code implementation, I asked Claude for help. All of the code was reviewed for methodological alignment and the text content and motivation originated from my own words and ideas. The content provided to me through these tools was rigorously reviewed and edited prior to inclusion. I stand by the words and text in this document as a reflection of my own work and understanding.

During coding, this meant providing prompts like the following (copied from Claude logs):

> *Can we update section 5 to pull in the BAKC2 gauge data (data/NASA-NW_targets_output/e_bakc2_historical.csv) and estimate NI and EI from that like we will in the application using the coefficients at data/streamflow/bakc2_handoff.csv) instead of how it's calculated here?*
>
> *Can we create a new script where we evaluate how temperatures would be impacted if we did max or min pumping (where max is 440 through the tunnel, min is 0 through the tunnel, or the minimum for water balance, but kept the lake level even? Check out how we flag lake level impacts in the app.py file in the SMR_lagged_forecast folder.*

During writing, this meant providing prompts like the following:

> *What is missing from this introduction? (Providing text from the introduction of Chapter 2).*
>
> *I need help broadening the impacts of this research in the conclusion. I'm stuck in a very narrow space of application because I view this application as something that doesn't explicitly transfer to other systems. Can you help me zoom out?*

TABLE OF CONTENTS

CHAPTER 1: BUILT TO FLY, NOT TO LAND: THE DEMOCRATIZATION OF MACHINE LEARNING AND ITS RISKS FOR AQUATIC SCIENCE

The accessibility of machine learning (ML) methods and applications across scientific domains has dramatically increased over the past decade. Within the domain of aquatic science, ML presents a significant opportunity due to its ability to identify complex non-linear relationships in heterogeneous datasets across scales and drivers (J. D. Willard et al., 2025; M. Zhu et al., 2022).The growth of ML is, in part, due to increasing processing power in consumer-level laptops and desktops but also the evolution of open-source programming and the availability of code packages (R) or modules (Python) for easy ML implementation. Open-source programming has leveled the ML playing field with `caret` and `tidymodels` for R, `scikit-learn`, `tensorflow`, and `keras` for Python, and built-in machine learning functions in Google Earth Engine's code editor, all available for individuals to use with minimal effort and without deep understanding of the underlying methods. More recently, large language models (LLMs), generative artificial intelligence, and now agentic AI can assist in coding in minutes from a plain-language font (e.g. Claude Code, OpenAI Codex, Google Antigravity). Google Antigravity is a deliberate callback to the Python easter egg `import antigravity`, a reference to a comic in which a character flies simply by importing a single module, implying that Python makes the impossible effortless. This is a genuinely exciting time to be in science as the democratization of advanced modeling methods feels attainable for many researchers. The concern I raise here is specifically about the use of supervised, predictive ML and is not intended to insinuate aquatic scientists are misusing ML but rather note that these new tools make it possible to appear to apply ML correctly, while doing so incorrectly in substance, and our current domain infrastructure cannot tell the difference. In essence: we now have flight, but we must learn how to land.

In an effort to increase understanding of systems, researchers may attempt to increase predictive accuracy by aiming their results toward lower root mean square error (RMSE) or lower mean absolute error (MAE). Many domains of science explicitly reward low error metrics such as these, a fact that predates our current LLM and agentic AI reality by nearly 70 years (Sterling, 1959). This focus on low error metrics is a tempting carrot that can lead to developing a model away from physical reality in search of the best-performing model, a concept linked to equifinality that is well-documented in hydrology by Keith Beven of Lancaster University (Beven, 1993 and more recent publications). Beven notes that it is possible to train even physical models to the point where they are providing the correct solution but could arrive at that solution through the wrong mechanisms. Machine learning has both inherited this bias and amplified it, with agentic tools now automating the optimization of these metrics, removing the researcher's conscious participation in methodological decisions. Despite decades of awareness (Rosenthal, 1979) and genuine progress in some fields (e.g. pre-registration in clinical medicine, registered reports in psychology), publication is shaped by a preference for strong predictive metrics to which authors respond rationally, pushing their models towards the lowest possible, ideally defensible metrics.

This incentive has become particularly consequential as ML has proliferated across scientific domains. A growing body of literature documents overfitted models, weak baselines, inflated R^2, and leaky cross-validation: failures that produce artificially low error metrics without producing generalizable science (Kapoor & Narayanan, 2023). In the worst cases, these failures result in spurious correlations mistaken for ecological understanding (J.-J. Zhu et al., 2023). In aquatic science, these same pitfalls have been documented (Varadharajan et al., 2022; J. D. Willard et al., 2025) but to date, no coordinated response for our domain has emerged. Best

practices exist, including the comprehensive multi-disciplinary REFORMS checklist (Kapoor et al., 2024) and a companion framework for environmental scientists the Environmental Machine Learning, Baseline Reporting and Comprehensive Evaluation (EMBRACE; J.-J. Zhu et al., 2024); however, there is no dedicated framework for aquatic scientists, and little evidence of broad adoption of these frameworks across our field.

Aquatic systems present domain-specific challenges that generic frameworks were not designed to address, including the physical connectivity of aquatic systems that generates autocorrelation at multiple spatial and temporal scales, the reality that relationships between variables shift across seasons and climate-driven extremes including drought, flood, and wildfire, and the expectation that model outputs remain interpretable against known biogeochemical and hydrological mechanisms. The absence of such a framework may be a contributor to our collective landing problem. The same democratization that opened ML to aquatic scientists without formal training has introduced agentic tools that make consequential methodological decisions automatically, invisibly, and in service of optimizing low error metrics potentially at the cost of sound science or actual operational utility.

For the researcher who deploys ML through an agentic tool or LLM prompt without ML-specific methodological knowledge, the sequence of analytical decisions contains methodological choices they may never know were made and have no obvious way to detect. Critically, in many cases, reviewers of ML-based papers may not have the training or ability to dig into the codebase, making it harder to evaluate and catch errors arising from chasing lower error metrics. Unlike the most visible reproducibility crisis in psychology where researchers made questionable choices consciously through p-hacking (e.g. Stefan & Schönbrodt, 2023) or selective reporting (Franco et al., 2015), our current situation is fundamentally different: these

tools make decisions on the researcher's behalf optimizing for whatever metric the researcher specifies, or whatever the default setting is. An added complexity in this scenario is that these tools are designed to be helpful which is desirable until it manifests in sycophancy: the well-documented tendency of LLMs to give users what they want rather than what might be correct (Sharma et al., 2023). A researcher with an understanding of the underpinnings of ML can review the output provided to them and edit the code to assure that the sycophantic nature of AI has not crept into the code base and assuring that the methodological choices are sound.

Consider hyperparameter tuning, the method of selecting the settings for a given ML model architecture. A common approach is grid search, where a researcher – perhaps guided by an LLM generating the code, or an agentic tool acting independently – tests dozens or hundreds of hyperparameter combinations on the data to find the best-performing model settings. During this process, feature scaling or normalization process (necessary for many ML architectures) is applied to the full dataset before a training, validation, and test set split, and the researcher or agent selects the model with the lowest MAE. A LLM or agentic tool may now also write code to present a publication-ready figure, report error metrics, and maybe even a complete methods paragraph for a manuscript. The researcher has violated no current norms and even where AI disclosure is required, declaration of the tool says nothing about the soundness of what it produced. That some researchers do not declare AI use at all, despite journal policies requiring it, compounds the problem further but warrants its own discussion. This model, however, is almost certainly overfit and the error metrics reported would not hold up in an application to data because of data leakage in the pre-processing step (J.-J. Zhu et al., 2023). The output, when presented in a manuscript, is indistinguishable from one produced by a researcher who made every methodological choice deliberately and correctly. Had the researcher applied the

EMBRACE checklist (J.-J. Zhu et al., 2024), designed explicitly to catch errors of this kind, the data leakage would likely have been identified. But EMBRACE assumes the researcher understands what their tools did on their behalf, and no journal currently requires its completion, despite it being designed for this very purpose.

Code and data sharing is now standard practice in many journals, a genuine step forward designed to enable verification of the analytical process, but this requirement has been functionally neutralized. No journal currently requires that shared code runs, reproduces the reported results, or has been evaluated by anyone with ML expertise. This is not a critique of reviewers, who are unpaid and working under significant time constraints, it is a critique of a system that has not built the infrastructure to make verification possible. In ecology, only 26.9% of journals mandate code-sharing at all and crucially, this requirement typically applies at the time of publication rather than during peer review, meaning the analytical decisions embedded in the code may never be evaluated before acceptance of a manuscript (Ivimey-Cook et al., 2025). Even if the code is available and a reviewer does attempt to run it, the code itself could be the problem: AI-generated code is syntactically correct, but prone to logic errors that are domain-specific. For example, the code for the ML completes random subsampling instead of spatial/temporal blocking and that embeds data leakage into the workflow, in ways that may not be communicated within the methods section. Without awareness, this methodological choice could also result in suspiciously low error metrics and like the hyperparameter tuning example, would only be caught by using the EMBRACE checklist if the author knew that this the choice and that it was problematic. The result is reproducibility theater: the appearance of transparency without the substance of verifiability.

The field of aquatic science is not without precedent for self-correction: we have updated sampling standards, revised statistical norms, and built community infrastructure around data sharing. The call to action here is not individual expertise but collective infrastructure: standards, review mechanisms, and shared norms that make rigorous ML the default rather than the exception. Authors are well placed to lead by example because much of what rigorous ML requires is simply good scientific practice applied to a new context. For instance, pre-specifying an evaluation protocol before running models is no different in principle from declaring a sampling design before starting fieldwork. For those using LLMs or agentic tools, this is a matter of documenting what the tool did with the same rigor expected of a laboratory protocol. A growing literature provides practical guidance on method selection, evaluation design, and interpretability (Pichler & Hartig, 2023; J. D. Willard et al., 2025; J.-J. Zhu et al., 2023), and the EMBRACE checklist (J.-J. Zhu et al., 2024) translates these principles into actionable steps for environmental scientists.

Reviewers and editors serve as our primary mechanism for quality assurance and bring expertise in aquatic ecological interpretation, like whether a conclusion is ecologically meaningful or whether a proposed mechanism is plausible. Few reviewers are well-suited to assess whether the ML used to reach those conclusions was built correctly. At minimum, reviewers and editors evaluating ML-based aquatic science papers should be equipped with a short set of targeted questions to help identify when the reported results can be trusted (Box 1 provides some examples for supervised predictive ML methods but is not exhaustive). The burden of ML evaluation should not fall entirely on volunteer reviewers whose expertise lies in aquatic ecology rather than ML methodology, it should be built into editorial infrastructure,

through associate editors or data editors with quantitative expertise, as some journals are beginning to do (Ivimey-Cook et al., 2025).

Box 1. Assessing supervised, predictive ML methodology in aquatic science: reflective questions for reviewers and editors

For reviewers

General:

- What is the ML being used for (prediction, classification, exploration, or driver identification) and are the evaluation criteria appropriate for that goal?
- Is the choice of ML justified and is there a clear reason why a simpler statistical model would not suffice for this question?

Training, Validation, Testing:

- How was the training, validation, and test (TVT) split defined, and was temporal and/or spatial autocorrelation accounted for?
- Is the test set representative of the conditions the model is intended to generalize to?
- Are reported performance metrics based exclusively on the test set?
- Were preprocessing steps (scaling, imputation, feature selection) applied before or after the TVT split?
- Was any pre-training completed, and did it include samples from the test set?

For timeseries applications:

- Do reported metrics include visual inspection of predicted and observed values as a time series, not just aggregate statistics or predicted-observe plot? Systematic biases during ecologically critical periods (e.g. bloom onset, storm pulses, drought conditions) are often invisible in aggregate metrics but immediately apparent in time series plots.

For evaluation:

- Is the baseline comparison meaningful and does it outperform the best available existing approach for this system or question?
- Regardless of application, are explainability tools such as SHAP or feature importance used to evaluate whether model behavior is ecologically consistent?
- Some advanced architectures resist meaningful post-hoc interpretation; where this is the case, authors should justify why predictive performance alone is sufficient for the stated scientific goals, and whether a more interpretable model would have sufficed.

For editors:

- Was an LLM or agentic tool used in model development, and if so, is this disclosed and are the automated methodological decisions documented?
- Does the shared code match the methods as described? Are the preprocessing steps, split strategy, and model configuration consistent between the two?

Beyond individual authors and reviewers, our field needs infrastructure that makes rigorous ML the default rather than the exception. Two changes are particularly tractable. First, community-maintained benchmark datasets for common aquatic ML tasks (e.g., harmful algal bloom prediction, water quality index estimation, streamflow at ungauged sites) would allow model performance to be contextualized against known standards rather than isolated within-study claims. This is not a novel idea; benchmark datasets have accelerated progress and raised methodological standards in other ML-adjacent fields, and the data infrastructure to support them increasingly exists in aquatic science through monitoring networks like NEON, GLEON, and the USGS. For some applications this is not a realistic comparison, a ML application that is not trying to generalize to other systems should, at minimum, include comparison against simple baselines such as persistence forecasts or climatological means. Second, executable notebook requirements that demonstrably reproduce reported outputs are technically feasible and should become standard practice. Importantly, performance comparisons to previously published work should not be treated as a reliable benchmark at this stage, our field has not yet developed the review infrastructure needed to verify whether those prior results followed best practices. Unless a study explicitly adheres to the EMBRACE checklist or an equivalent standard, to assume that published performance metrics reflect rigorous methodology is premature.

The psychology reproducibility crisis was the result of accumulated non-reproducible results and public replication failures, and the field is still undergoing a slow and costly process of rebuilding. Aquatic science has a window to act before similar norms harden in our field. Whether ML is written line-by-line by an expert, aided by an LLM, or written and executed autonomously by an agentic tool, the practices we establish now will shape whether aquatic ML produces durable or disposable science. The incentive structure of academic publishing, what

(Smaldino & McElreath, 2016) describe as “no incentive to be right”, actively rewards the publication of positive results and high-performing models, whether or not the experimental design that produced them was sound. In aquatic ML, this creates a compounding risk that models built on poorly-conceived workflows become the published benchmark and that those benchmarks become the bar that subsequent work must clear to be published: “when a measure becomes a target, it ceases to be a good measure” (Smaldino & McElreath, 2016).

The goal of this essay is not to slow the adoption of ML in aquatic science, nor require every user to be an expert in ML methods. We have powerful tools at our disposal to pursue scientific opportunity to ask interesting and difficult questions about aquatic systems. ML done well is a promising tool that we have to answer these questions (Reichstein et al., 2019), but powerful tools demand accountability, and accountability requires infrastructure. The norms we establish now – for authors, reviewers, editors, and journals – will shape the literature and the benchmarks that accumulate over the next decade. We have been given flight. The question is whether we build the instruments to navigate it wisely.

CHAPTER 2: A DATA-DRIVEN FORECAST SYSTEM AS AN OPERATIONAL DECISION TOOL IN A MANAGED RESERVOIR

2.1 Abstract

Water clarity in Grand Lake, Colorado is collaboratively managed during the Grand Lake Adaptive Management Season, in which Northern Colorado Water Conservancy District (Northern Water) and partner organizations work together to meet clarity goal qualifiers. Pumping operations that convey water from Lake Granby through Shadow Mountain Reservoir represent a primary operational lever for temperature-mediated water quality management in Grand Lake. Despite the importance of this system within the scope of water delivery to the Front Range of Colorado, no operational temperature forecast tool has previously existed to support near-term pump management decisions. We developed and evaluated a parsimonious, theory-guided neural network model for 7-day water temperature forecasting in Shadow Mountain Reservoir, driven by medium-range ensemble weather forecasts, inflow estimates, and pumping operations. The model simultaneously predicts temperature at near-surface (0–1 m) and integrated (0–5 m) horizons as a physically consistent structural constraint, trained on an 11-year observational record using five-fold temporal cross-validation with multiple compounding regularization strategies to mitigate overfitting. Model performance on the held-out 2025 test period was strong and consistent with cross-validation results, with a near-surface MAE 0.34°C, RMSE 0.45°C, NSE 0.956; integrated depth MAE 0.23°C, RMSE 0.28°C, NSE 0.982 and skill scores exceeding persistence across the full 7 day forecast horizon. Applied in a hindcast scenario framework using observed 2025 meteorological and inflow conditions, the model revealed that active pumping operations suppressed Shadow Mountain Reservoir temperatures

by nearly 3°C at the near-surface and 4°C at the integrated depth relative to a simulated no-pump baseline using the same observed meteorological and inflow conditions during the July 1 through September 11 adaptive management season. This demonstrates that pumping operations represent a meaningful and quantifiable thermal mitigation mechanism in this system. The forecast system provides pump operators and water managers with a computationally efficient tool for evaluating the thermal implications of operational scenarios at management-relevant timescales. More broadly, this work demonstrates the potential of theory-guided machine learning for operational water quality forecasting in managed reservoir systems.

2.2 Introduction

Water temperature governs biotic life and chemical cycles in freshwater ecosystems from photosynthesis and respiration to decomposition and nutrient cycling (Wetzel, 2024). Lake metabolism scales with water temperature, resulting in accelerated in-lake production (Brown et al., 2004; Gillooly et al., 2001), and lakes are currently warming faster than surrounding air temperatures (O'Reilly et al., 2015). Warming surface waters impact thermal stratification in lakes and reservoirs altering the vertical distribution of nutrients, which can further fuel biological production independent of external loading (Sondergaard et al., 2001; Wetzel, 2024). Specifically, phytoplankton growth rates generally increase with a rise in temperature across a range of nutrient conditions (Eppley, 1972). Increased primary production results in elevated chlorophyll *a* concentrations and particulate matter impacting clarity of a waterbody through the attenuation of light in the water column (Carlson, 1977). However, biotic sources are not the only factor impacting clarity. In snowmelt-dominated mountain watersheds, allochthonous inputs of inorganic suspended sediment during spring runoff can represent a dominant control on clarity that operates independently of temperature, potentially masking the biotic impact on clarity

(Bixby et al., 2015). Dissolved organic matter from terrestrial sources transported by precipitation and hydrological connectivity can similarly reduce water clarity through light absorption independent of in-lake biological production through pathways governed by watershed characteristics rather than water temperature (Davies-Colley & Smith, 2001; Williamson et al., 2015). In freshwater systems free from human flow alteration, water temperature is controlled by latent heat, inflow and outflow dynamics, and meteorological factors like solar radiation and air temperature (Edinger et al., 1968; Imberger & Patterson, 1989; Kettle et al., 2004; Wetzel, 2024). In engineered freshwater systems where thermal conditions can be actively controlled through flow regulation, pumps, and canals, water temperature represents a tractable lever for water quality management, a context directly relevant to Colorado's Three Lakes System.

The Three Lakes System, comprised of Grand Lake, Shadow Mountain Reservoir, and Lake Granby, is located in Grand County, Colorado on the western side of the Continental Divide. As a component of the Colorado-Big Thompson Project (C-BT), the system stores and conveys water for delivery to the Front Range of Colorado, managed by Northern Water in partnership with the Bureau of Reclamation to serve more than one million individuals and over 600,000 acres of irrigated farmland (GEI Consultants, 2013). Water clarity in Grand Lake is subject to established goal qualifiers under the Grand Lake Clarity Stakeholders Memorandum of Understanding (MOU No. 16-LM-60-2568, 2016), where all parties aim to meet at least a minimum average Secchi depth of 3.8 meters and a daily minimum of 2.5 meters during an adaptive management period spanning July 1 through September 11 (referred to as the Grand Lake Adaptive Management period, GLAM). Temperature can be manipulated in this system by pumping water from deep in Lake Granby (the pump draws from 85 feet below the surface at full

pool) into Shadow Mountain Reservoir, where the introduction of cooler water reduces temperature, may suppress rates of biological production (Eppley, 1972). Clarity in Grand Lake is sensitive to the quality of water conveyed from Shadow Mountain Reservoir through pumping operations, making temperature-mediated biological production in Shadow Mountain Reservoir a meaningful operational lever for meeting these goal qualifiers during the GLAM season (Grand Lake Adaptive Management Committee, 2023). While both biotic and abiotic processes govern clarity in mountain reservoir systems like this one, this study focuses on water temperature as the operational entry point into that chain. The development of an operational temperature forecast model sensitive to pumping operations represents a direct response to this management need, providing pump operators at the Bureau of Reclamation and water managers at Northern Water with data-driven decision support at timescales relevant to operations, a capability that does not currently exist for the Three Lakes System.

We worked with the GLAM Committee, which is responsible for week-to-week operational decisions in the Three Lakes System, to create a tool they could use at their weekly meetings to test operational scenarios and their impacts on Shadow Mountain Reservoir water temperature to assist in their operational decision making. Our hypothesis, guided by our end-users, is that managing near-surface temperature for algal growth and integrated temperature for diatom growth would assist in managing clarity in Grand Lake. This tool operates under the assumption that those biotic communities would be pushed into Grand Lake when the Farr Pump plant was active. Evaluating the thermal implications of operational decisions in Shadow Mountain Reservoir requires a predictive modeling framework capable of resolving temperature response quick enough to test scenarios during the meeting and providing a week-long forecast. Although process-based hydrodynamic models offer physically explicit simulation of lake

thermal dynamics, their computational requirements often make them poorly suited for operational forecasting contexts where rapid, repeatable predictions are needed to inform near-term management decisions (Read et al., 2019; J. Willard et al., 2022). The tool we co-developed with our end users is a 7-day water temperature forecast model built on a neural network architecture and driven by weather forecasts, inflow forecasts, and pumping operations, providing a computationally efficient tool through which managed water transfers can be assessed in an operational decision support context. We adopt the theory-guided data science paradigm (Karpatne et al., 2017), integrating system-level physical knowledge into the model architecture as structural constraints rather than relying solely on data-driven pattern recognition. Specifically, the model simultaneously predicts water temperature at two depths (0-1 meter "near-surface" and 0-5 meter "integrated") enforcing a physically consistent vertical temperature structure as an internal constraint to the model. This multi-output architecture reflects the known physical relationship between surface heating and vertical thermal stratification, intended to ensure predictions remain physically plausible across varying operational, meteorological, and streamflow conditions.

This work emerged from close collaboration with operational end users, where building trust in model outputs was as essential as predictive accuracy. The result is a 7-day water temperature forecasting system for Shadow Mountain Reservoir that responds directly to pumping operations, a capability that supports real-time decision-making and enables scenario analysis that cannot be tested through observation alone. We offer two contributions to operational water quality management in the Three Lakes System through the development of this tool. The first is the development and evaluation of a parsimonious theory-guided neural network model driven by weather forecasts, inflow estimates, and Adams Tunnel delivery

operations. We assess forecast performance independently at near-surface (0–1 m) and integrated (0–5 m) depths across both the 1-day-ahead and full 7-day horizons, and build stakeholder trust in model behavior through SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) feature attribution analysis. The second contribution applies the validated model in a hindcast framework, using observed 2025 meteorological and inflow conditions to evaluate the thermal response of Shadow Mountain Reservoir to a range of static Adams Tunnel delivery scenarios, from no-pump conditions to near-maximum delivery. This analysis provides a quantitative foundation for assessing the potential of managed water transfers to optimize temperatures in Shadow Mountain Reservoir during the GLAM Season, a scenario space that cannot be evaluated through direct observation.

Water temperature forecasting tools for operational reservoir management have developed primarily within large-scale decision support frameworks. RiverWare (Zagona et al., 2001) was developed at the University of Colorado as a generalized tool for complex reservoir system modeling, with early applications supporting Bureau of Reclamation and Tennessee Valley Authority operations. It has since become the backbone of the Bureau of Reclamation's Colorado River Simulation System, representing the water supply and river network across dozens of reservoirs, inflow nodes, and water users, and is now used extensively for operational scheduling, scenario analysis, and stakeholder engagement across multi-reservoir systems in the western United States. For systems where thermal stratification and selective withdrawal are central management levers, physics-based two-dimensional hydrodynamic models such as CE-QUAL-W2 have been widely applied to simulate temperature control operations and evaluate downstream thermal outcomes; however, these tools are designed primarily for retrospective simulation and scenario analysis rather than operational near-term forecasting (Barthlow et al.,

2001; Hanna et al., 1999). Physics-based iterative forecasting systems represent a parallel development: (Thomas et al., 2020) developed FLARE (Forecasting Lake and Reservoir Ecosystems), a real-time system integrating sensor data, data assimilation, and a one-dimensional hydrodynamic model to generate ensemble water temperature forecasts with quantified uncertainty at horizons up to 35 days. While FLARE demonstrates the feasibility of near-term mechanistic forecasting in lakes and reservoirs, its primary orientation is ecological water quality prediction rather than forecasting conditioned on infrastructure operations. More recently, machine learning approaches have demonstrated capacity to deliver operational temperature forecasts at management-relevant timescales. Zwart, et al. (2023) developed a process-guided deep learning and data assimilation system producing 7-day probabilistic forecasts of daily maximum water temperature in the Delaware River Basin, driven by weather forecasts and observed reservoir releases, in direct support of decisions about cold-water reservoir releases. That forecast system demonstrated that data-driven forecasting can directly inform cold-water reservoir release decisions, the same operational logic underlying our tool for adaptive management in the Three Lakes System. Despite the established goal qualifier framework used to manage clarity in Grand Lake and the demonstrated potential for thermal mitigation through pumping operations, this study represents the first operational temperature forecasting tool developed specifically for the Three Lakes System.

2.3 Study Area and Decision-Making Context

2.3.1 Physical Setting and Operational Infrastructure

The Three Lakes System is a hydraulically connected network of three human-made reservoirs and one natural lake situated in Grand County, Colorado on the western side of the Continental Divide. The system was developed as a component of the C-BT to provide water to the growing populations of the Front Range (Autobee, 1996) and it serves dual purposes by providing recreational and ecological value as a natural resource while functioning as a critical component of the C-BT infrastructure for Front Range water delivery. The Three Lakes System is comprised of Grand Lake, Shadow Mountain Reservoir, Lake Granby (a reservoir), and Willow Creek Reservoir. We exclude Willow Creek Reservoir from this analysis, as it is not affected by the pumping operations described here. Grand Lake is the deepest natural lake in Colorado at approximately 81 meters. Shadow Mountain Reservoir (constructed 1946) is a shallow conveyance waterbody of approximately 7 meters whose limited depth renders it particularly sensitive to temperature fluctuations and biotic growth. Lake Granby (constructed 1950) is the largest storage reservoir in the C-BT system with a total capacity of 539,800 acre-feet (United States Bureau of Reclamation, 2012). Grand Lake receives natural inflows from the North and East Inlets, both originating in

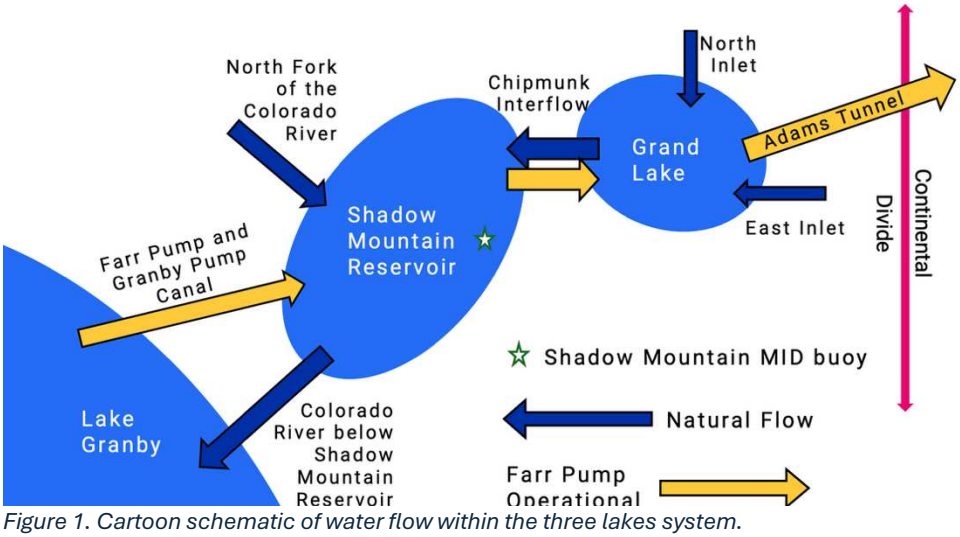


Figure 1. Cartoon schematic of water flow within the three lakes system.

Rocky Mountain National Park. Shadow Mountain Reservoir is fed primarily by the North Fork of the Colorado River, originating in the Kawuneeche Valley of the upper Colorado River headwaters. During spring runoff, the North Fork delivers substantial sediment loads to Shadow Mountain Reservoir, a seasonal dynamic with direct implications for water quality during subsequent pumping operations (Figure 1).

Water is delivered to the Front Range via the Alva B. Adams Tunnel, which travels 13.1 miles under the Continental Divide. Under natural flow conditions, water moves from Grand Lake into Shadow Mountain Reservoir via the Chipmunk Lane interflow and continues to Lake Granby via the Colorado River. Front Range demand first met by this natural inflow via the Adams Tunnel, with pumping operations engaged only when natural flow is insufficient for the Front Range requirements, a condition that occurs in all operational years; in instances where natural inflow meets full demand, pumping is not initiated and Grand Lake clarity is largely unaffected by operations. The Farr Pumping Plant draws water from approximately 26 meters depth in Lake Granby (approximately 85 feet at full pool), typically below the thermocline, providing hypolimnetic water to Shadow Mountain Reservoir that is generally cooler than ambient reservoir temperatures. Water then moves from Shadow Mountain Reservoir into Grand Lake via the Chipmunk Lane connecting channel and onward through the Adams Tunnel for Front Range delivery (Figure 1).

In practice, pumping operations are not initiated immediately following spring runoff. Operators often implement an informal settling period to allow inorganic sediment delivered by the North Fork of the Colorado River to settle out of the water column in Shadow Mountain Reservoir before water is conveyed into Grand Lake, reducing the abiotic sediment load conveyed to Grand Lake during pumping operations. As a result, temperature-mediated

biological production in Shadow Mountain Reservoir represents the dominant water quality concern during active pumping operations, providing the mechanistic basis for the temperature forecast model developed in this study.

2.3.2 Regulatory Context

When the C-BT was approved, Senate Document 80 (1937) established the principle “to preserve the fishing and recreational facilities and the scenic attractions of Grand Lake, the Colorado River, and the Rocky Mountain National Park,” a goal that has governed water storage and delivery operations since the project’s inception. Efforts to formalize this principle into quantifiable standards began in earnest in 2008, when the Colorado Water Quality Control Commission (WQCC) established a narrative standard alongside a strict 4-meter numeric clarity standard; however, the numeric standard was given a deferred effective date and never went into effect. Recognizing the persistent challenge of meeting clarity objectives under existing operations, the Bureau of Reclamation formally evaluated alternatives to the current system configuration to reduce operational impacts on Grand Lake clarity (United States Bureau of Reclamation, 2012). In 2014, the WQCC amended the narrative standard to its current form: "the highest level of clarity attainable, consistent with the exercise of established water rights, the protection of aquatic life, and protection of water quality throughout the Three Lakes System." Because the 4-meter numeric standard was deemed operationally infeasible, this updated narrative standard became the foundation for stakeholder negotiations, directly prompting the creation of the Grand Lake Clarity Stakeholders Memorandum of Understanding (MOU No. 16-LM-60-2568) in 2016. The MOU, signed by the United States Bureau of Reclamation, Northern Colorado Water Conservancy District (Northern Water), Grand County Board of Commissioners, Northwest Colorado Council of Governments, and the Colorado River Water

Conservation District, operationalizes the narrative standard through Secchi disk depth measurements, with goal qualifiers of a minimum average Secchi depth of 3.8 meters and a daily minimum of 2.5 meters throughout the July 1 through September 11 Grand Lake Adaptive Management (GLAM) period.

2.3.3 Decision-Making Context

Water quality management decisions in the Three Lakes System are made collaboratively through weekly GLAM meetings during the GLAM Season, attended by representatives from Northern Water, the Bureau of Reclamation, Grand County, and other stakeholders. Pump operators and water managers currently rely on a combination of real-time sensor observations, manual measurements and observations (particularly Secchi depth), and seasonal simulations produced by CE-QUAL-W2 to inform these decisions. While the monitoring network provides timely observational data, it supports a reactive rather than predictive management approach. The seasonal CE-QUAL-W2 simulations, though physically explicit and holistically calibrated, require days to run and do not update with real-time conditions, creating a temporal mismatch between the timescale of available predictions and the weekly decision cadence of GLAM operations. Near-term iterative forecasting frameworks have emerged as a promising approach for bridging this gap between static seasonal simulations and dynamic in-season decision making (Dietze et al., 2018; Hydros Consulting, 2025). Despite growing interest in near-term freshwater quality forecasting (Lofton et al., 2023), operational temperature forecast tools for managed reservoir systems remain rare, motivating the development of a rapid, operationally responsive temperature forecast model capable of informing pump operation decisions at management-relevant timescales.

2.4 Methods

Our analytical approach consists of two primary components: the development and performance assessment of a theory-guided neural network model for 7-day water temperature forecasting in Shadow Mountain Reservoir, and the application of that model in a hindcast scenario framework to evaluate the thermal effects of Adams Tunnel delivery operations. I first describe the physical drivers of water temperature in Shadow Mountain Reservoir to motivate model architecture and feature selection decisions, followed by a description of data sources, neural network design, and cross-validation framework. Model performance is then assessed at both 1-day ahead and 7-day forecast horizons at near-surface and depth-integrated scales before applying the validated model to evaluate thermal response across a range of static delivery scenarios during the July 1 through September 11 adaptive management window.

Understanding the limnological processes governing water temperature in Shadow Mountain Reservoir is the foundation of our modeling approach. I use domain expertise to define the physical drivers of water temperature and constrain model architecture, an approach consistent with the theory-guided data science paradigm introduced above (Karpatne et al., 2017). Water temperature in Shadow Mountain Reservoir is governed by three categories of forcing: meteorological variables (air temperature, wind speed, and solar radiation), flow variables (inflows and outflows), and operational mechanisms (Farr Pump operations and Adams Tunnel deliveries). At near-surface depths, temperature is driven primarily by air temperature, solar radiation, wind, and precipitation (Edinger et al., 1968; Wetzel & Likens, 2000). At integrated depths, temperature reflects the combined influence of wind-driven mixing, inflow magnitude, and inflow temperature, with previous-day temperature providing an additional constraint on diurnal change due to the latent heat properties of water (Imberger & Patterson,

1989; Kettle et al., 2004; Wetzel & Likens, 2000). While inflow temperature and precipitation represent known physical drivers of water temperature, consistent direct measurements of these variables were unavailable for this system; their influence is assumed to be captured indirectly through inflow volume and lagged temperature features included in the model. Finally, pumping operations from Lake Granby represent the primary operational lever for temperature mitigation in the system and are treated as an inflow, introducing water into Shadow Mountain Reservoir. While interflow between Shadow Mountain Reservoir and Grand Lake via the Chipmunk Lane connecting channel is observed historically, it must be derived for forecast periods using a daily water balance incorporating Adams Tunnel delivery scenarios, Colorado River minimum flow requirements, and pumping operations as inputs. The 1-foot year-round elevation constraint on the Shadow Mountain Reservoir and Grand Lake complex stipulated by Senate Document 80 allows the assumption of negligible storage change, reducing the water balance to a straightforward mass balance calculation. These physical relationships inform both the selection of model inputs and motivate the multi-output architecture that simultaneously predicts temperature at near-surface and integrated depths, ensuring predictions reflect the physically connected nature of thermal dynamics across the water column.

2.4.1 Data Sources

Meteorological Data. Meteorological forcing for both model training and operational forecast application was derived from the NOAA Global Ensemble Forecast System version 12 (GEFSv12) rather than from local observed meteorological stations, providing instantaneous forecasts of air temperature, wind speed, and downward shortwave radiation. Because GEFSv12 serves as the operational meteorological forecast driver for this system, training on GEFSv12 output ensures consistency between the data used during model development and the data

available during operational forecast application, minimizing distributional mismatch between training and forecast inputs (Quiñonero-Candela et al., 2009). Debiasing GEFSv12 output against local observed meteorological data was not attempted, as the coarse resolution of GEFS topography relative to the complex mountain terrain of the study area introduces systematic offsets that are difficult to characterize reliably (Cui et al., 2025; Dabernig et al., 2017; Scheuerer & Hamill, 2019), and because maintaining a single consistent data source across training and application is preferable for an operational forecasting context.

Two GEFSv12 data products were used to span the full study period: the GEFSv12 reforecast archive (Guan et al., 2022), which provides retrospective ensemble forecasts from 2000–2019 for model training, and the GEFSv12 operational archive, which provides data from late September 2020 onward for model training continuation and forecast application (National Centers for Environmental Prediction, 2020). GEFS meteorological data were retrieved programmatically from Amazon Web Services (AWS) using the Python package Herbie (Blaylock, 2022) and extracted to the study location using nearest-neighbor interpolation on the native 0.25° grid. Wind speed was derived from zonal and meridional wind components and temperature was converted from Kelvin to degrees Celsius. To reduce processing complexity and maintain consistency across training and application data, only instantaneous near-noon forecasts were used through training, testing, and forecasting steps. Precipitation was excluded from the meteorological feature set as its acquisition would require processing all forecast timesteps; its influence on reservoir temperature is assumed to be captured indirectly through inflow volume changes, which are included as model inputs. Detailed data retrieval procedures, AWS access dates, and processing methodology are provided in Appendix A.

Streamflow Data. Observed streamflow data used for model training were obtained from two sources. Daily mean flow for the North Fork of the Colorado River above Shadow Mountain Reservoir, the Farr Pump, and the Adams Tunnel were acquired from Northern Water's Kisters Water Information System API. Sub-daily flow data for the Chipmunk Lane connecting channel were obtained using the R package *dataRetrieval* (DeCicco et al., 2025) and aggregated to daily means for model inputs. All site identifiers are listed in Appendix B.

For forecast periods, streamflow must be estimated. Deterministic 7-day streamflow forecasts for Baker's Gulch were obtained for 2025 through the Colorado Basin River Forecast Center (CBRFC) via personal communication (B. Alcorn, CBRFC, personal communication, 2026). Baker's Gulch is upstream of the North Fork flow measurement location and is the only flowing site in the system with an operational streamflow forecast available. Linear relationships between observed flow at Baker's Gulch and the North Fork of the Colorado River, North Inlet, and East Inlet were developed to extend the Baker's Gulch forecast signal across the system, under the assumption that streamflow patterns at Baker's Gulch are representative of broader inflow variability across the Three Lakes System. If the linear model provided a negative streamflow value, it was replaced with zero.

Chipmunk Lane flow for forecast periods was estimated using a simple daily mass balance of the Grand Lake and Shadow Mountain Reservoir complex. Treating the Grand Lake and Shadow Mountain Reservoir complex as having negligible storage change, which is consistent with the 1-foot year-round elevation constraint stipulated by Senate Document 80, daily Chipmunk Lane flow was estimated as:

$$Q_{Chipmunk} = Q_{NorthInlet} + Q_{EastInlet} - Q_{AdamsTunnel}$$

where Q represents daily mean discharge in cubic feet per second for each respective term. Negative values of $Q_{Chipmunk}$ indicate net flow from Shadow Mountain Reservoir into Grand Lake, consistent with active pumping operations, while positive values reflect natural flow conditions. North Inlet and East Inlet flows were estimated from the Baker's Gulch linear relationships described above, and Adams Tunnel deliveries were specified by the hindcast scenario being evaluated. The performance of all streamflow estimates used during forecast periods was evaluated against observed data and is reported in Appendix B. The contribution of streamflow estimation uncertainty to overall forecast performance was assessed during hindcast evaluation by comparing model performance using observed versus estimated streamflow inputs.

Water Temperature. Water temperature observations used as both model target variables and lagged input features were collected from an instrumented buoy deployed in Shadow Mountain Reservoir near the interflow with Grand Lake. The buoy operates as a profiler, ascending from maximum depth at its location (approximately 5.5 meters) over the course of 15 minutes and recording temperature measurements at approximately 0.5 meter intervals six times per day. Daily mean water temperature was calculated for near-surface (≤1 meter) and depth-integrated (0–5 meter) depth horizons, corresponding to the two target depths of the forecast model. Lagged values of these daily means were included as model inputs as a theory-guided structural constraint, reflecting the role of thermal inertia in constraining day-to-day temperature change (Kettle et al., 2004; Karpatne et al., 2017).

Forecast Data Preprocessing. Meteorological inputs are drawn directly from all 31 members of the near-noon GEFSv12 operational forecast described in Appendix A, with each forecast day corresponding to a successive lead time of the meteorological forecast. Streamflow inputs for forecast periods are derived from the Baker's Gulch linear relationships described in

Appendix B, extending the CBRFC deterministic streamflow forecast for that site across the North Fork of the Colorado River, North Inlet, and East Inlet. Chipmunk Lane flow for each forecast day is derived from the daily mass balance described in the "Streamflow" section above, using the estimated inflows and specified Adams Tunnel deliveries as inputs. Lagged temperature inputs are initialized from the most recent observed buoy data at the time of forecasting, a method that anchors the forecast to current observed conditions before propagating forward using model predictions for subsequent forecast days.

For both forecast testing and hindcast scenario analysis, all 31 available GEFSv12 operational archive ensemble members are used as meteorological forcing, generating an ensemble of temperature predictions that characterizes forecast uncertainty attributable to meteorological forcing. To maintain computational tractability for an operational decision support context, forecast rollout uses the ensemble mean of the one-day ahead temperature prediction as the lagged input for subsequent forecast days rather than propagating the full ensemble distribution forward. This approach preserves the meteorological uncertainty signal across forecast days without compounding uncertainty through the rollout in ways that would be difficult to interpret in an operational setting. As a result, forecast uncertainty reported here reflects meteorological forcing uncertainty and interannual variability characterized through cross-validation (described in Section 2.4.2), and should be considered a partial characterization of total forecast uncertainty.

2.4.2 Neural Network Design, Feature Selection, and Cross-Validation Framework

Water temperature at near-surface and integrated depths are governed by overlapping physical processes and are functionally connected across the water column, a relationship that motivated the use of a multi-output architecture that simultaneously predicts both target depths

rather than treating them as independent modeling problems. A fully-connected feedforward neural network was selected as the model architecture, as the multi-output requirement and the desire to enforce physically consistent co-estimation of the two depth horizons precluded simpler regression approaches (Karpatne et al., 2017; Read et al., 2019). Recurrent architectures such as long short-term memory networks, which implicitly require continuous temporal sequences (J. Willard et al., 2022), were not suitable due to annual winter gaps in the Shadow Mountain Reservoir buoy record. Neural network applications with modest sample sizes are susceptible to overfitting, particularly when the observation-to-feature ratio is low (Goodfellow et al., 2016; Srivastava et al., 2014). With approximately 887 observations across an 11-year record, the model was intentionally designed to be parsimonious, using SHAP (Lundberg & Lee, 2017; Marcílio & Eler, 2020) applied prior to hyperparameter tuning and multiple compounding regularization strategies employed to reduce the likelihood of overfitting: dropout and L2 regularization as architectural constraints, a compact architecture limited to two hidden layers of modest size, and a model selection criterion that explicitly penalizes training-validation divergence. I used the EMBRACE checklist (J.-J. Zhu et al., 2024) to assess our methodology as conforming to best practices, the assessment is provided in Appendix C.

Training, validation, and test sets were delineated by year with a five-fold temporal cross-validation applied to the training and validation period spanning 2014–2024, with data from 2025 held out entirely as the test set. Folds were balanced by the total number of observations while retaining whole years within each group. Temporal data leakage between folds was precluded by two factors: the feature set contains no cross-year lagged variables, and the annual winter gap in the buoy record created a natural physical discontinuity between successive seasons, ensuring that no information from one year's observations propagates into the following

year's training data. Prior to model training within each fold, all features were preprocessed using an inline Yeo-Johnson power transformer with standardization (Pedregosa et al., 2011; Yeo & Johnson, 2000), fit exclusively on training data to prevent leakage of validation or test data characteristics into the transformation step. This preprocessing approach also reduces feature preprocessing requirements during operational deployment by embedding the transformation parameters directly into the model pipeline.

Feature engineering followed the theory-guided data science paradigm (Karpatne et al., 2017), with input features organized into two categories: global variables (meteorological forcing, streamflow, and operational flow data) and target variables (near-surface and integrated depth water temperature). Model memory was provided through independent lagged variables rather than statistical summaries of prior days, reducing computational load and simplifying the data ingestion process during forecast rollout. A feature set with a global memory of 5 days and a target memory of 2 days was developed, producing an initial set of 39 features. Independent lagged variables were used in preference to rolling summaries (e.g. mean wind speed over the previous 5 days) because they allow individual forecast days to be assembled from a rolling buffer of observed and forecasted values without requiring recomputation of statistical summaries at each forecast step. Only the control member (c00) of the GEFS forecast was used to train the model.

Feature reduction was conducted using SHAP applied to a preliminary fixed-architecture neural network with a 0.1 dropout layer followed by two layers of 8 nodes each, trained with a learning rate of 0.01 and early stopping with a patience of 100 epochs (details in Appendix D). Feature importance was aggregated across all five cross-validation folds as the mean absolute SHAP value per feature. Features in the lower 30% of importance were eliminated, with two

exceptions retained on the basis of domain knowledge: the forecasted 1-day-ahead Chipmunk Lane flow and the 1-day lag of North Fork flow, both of which represent operationally meaningful inputs to the temperature modeling system despite falling below the importance threshold. Following feature reduction, the final feature set consisted of 21 features. Final feature names and SHAP importance values are reported in Appendix D.

Hyperparameter tuning was performed using Bayesian optimization with the Keras Tuner Python package (O'Malley et al., 2019) over 30 trials, using the third cross-validation fold under the assumption that optimal hyperparameters should be relatively stable across folds. Each trial was allowed up to 10,000 epochs with early stopping patience of 100 and mean squared error (MSE) loss. The hyperparameter search space included a dropout rate between 0.1 and 0.2 (in 0.05 increments), up to two hidden layers of 8–24 nodes (in increments of 8), L2 regularization strength sampled from a logarithmic distribution between 0.0001 and 0.01, and learning rate sampled from a logarithmic distribution between 0.0001 and 0.01. The dropout randomly deactivates between 10 and 20% of features per training epoch to reduce dependence on any single feature, while L2 regularization adds a penalty term to the loss function that discourages large node weights and reduces multicollinearity. Trial results were ranked by the difference between training and validation loss where a lower difference indicates better generalization. Training and validation loss curves were also visually inspected as a diagnostic, with overtraining indicated by rising validation loss concurrent with continued decline in training loss (Goodfellow et al., 2016). The final model was selected as the most parsimonious and conservative configuration among the top 10 trials ranked by the difference in loss between the training and validation sets (top 10 trail results are listed in Appendix D). The selected model

consisted of 2 hidden layers of 8 nodes each, with a dropout rate of 0.1 and L2 strength of 0.0005.

2.4.3 Performance Assessment

Model performance was assessed in two stages reflecting the sequential structure of the forecast system: first at the 1-day-ahead horizon using deterministic metrics, and then across the full 7-day forecast horizon using probabilistic skill metrics. In both stages, forecast skill was evaluated relative to a persistence baseline using a "yesterday-is-today" model that assumes water temperature remains static at the most recently observed value for the entirety of the forecast period. The persistence baseline provides a standard benchmark for evaluating the added predictive value of the operational forecast system (Pappenberger et al., 2015; Zwart, Diaz, et al., 2023), with the 1-day-ahead model required to perform at least as well as persistence before evaluation at extended forecast horizons.

One-day-ahead performance was first assessed across the five cross-validation folds to evaluate generalization prior to application on held-out data, with final performance reported against observed buoy temperatures for the 2025 test period. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Nash-Sutcliffe Efficiency (NSE; Nash & Sutcliffe, 1970) were calculated at both near-surface and integrated depths for each stage of evaluation. Time series of predicted versus observed temperatures were also visually inspected to identify potential over-reliance on lagged target variables, which can manifest as lagged prediction of day-to-day temperature departures, a known artifact in autoregressive forecast systems. SHAP values were additionally examined to confirm that feature importance is consistent with the conceptual model of the system: specifically, that pumping operations from Lake Granby are associated with decreased temperatures in Shadow Mountain Reservoir to

provide a qualitative check that the model has learned physically meaningful relationships rather than spurious correlations.

Seven-day forecast performance was assessed using the Continuous Ranked Probability Score (CRPS; Gneiting and Raferty 2007; Hersbach 2020) and Continuous Ranked Probability Skill Score (CRPSS), which evaluate probabilistic forecast skill across the full GEFSv12 ensemble, consistent with emerging standards for probabilistic freshwater quality forecast evaluation (Lofton et al., 2022; Olsson et al., 2025). CRPS is analogous to MAE for probabilistic forecasts, measuring the integrated difference between the forecast probability distribution and the observed value, while CRPSS expresses forecast skill relative to the persistence baseline with positive values indicating improvement over persistence. Both metrics were calculated at each forecast lead time from day 1 through day 7 at both depth horizons, characterizing how forecast skill degrades with increasing lead time.

The contribution of forecast input uncertainty to overall temperature prediction skill was assessed by independently substituting all GEFSv12 forecast data with observed meteorological inputs and all Baker's Gulch-derived estimates with observed streamflow inputs, evaluating the change in forecast skill at each lead time from day 1 through day 7 for each substitution independently. Because the model was trained on GEFSv12 output rather than local observed meteorology, the GEFSv12 day zero forecast serves as the observed meteorological baseline in this comparison rather than station observations. This sensitivity analysis isolates the relative contribution of meteorological and streamflow forecast uncertainty to overall forecast uncertainty across the 7-day horizon.

2.4.4 Hindcast Scenario Analysis

Predicted water temperatures in Shadow Mountain Reservoir were evaluated across June–September 2025 under four operational scenarios to assess the thermal response of the system to varying Adams Tunnel delivery rates. The pre-scenario period of June 1 through June 30 was run using observed operations across all scenarios to confirm model stability and consistent initialization prior to the scenario sub-window. Within the July 1 through September 11 adaptive management window, three static delivery scenarios were imposed: maximum Adams Tunnel delivery held constant at 550 cfs, middle delivery held constant at 220 cfs, and a no-pump scenario in which Farr Pump operations were eliminated except to maintain consistent elevation. The 220 cfs scenario represents a plausible near-future delivery rate informed by anticipated changes to system-wide operations following planned 2027 upgrades to Pole Power Plant (Bureau of Reclamation, personal communication, 2026). Observed gauge data for East Inlet, North Inlet, and North Fork were used as inputs across all scenarios and for the full hindcast period. Within the scenario sub-window, only Adams Tunnel deliveries and Farr Pump operations were varied between scenarios, with Chipmunk Lane connecting channel flow recomputed from the mass balance for each scenario to maintain internal consistency. A fourth scenario using observed 2025 Adams Tunnel deliveries and Farr Pump operations served as the operational baseline against which scenario effects were quantified, representing modeled temperature under observed conditions rather than directly observed temperature.

For the maximum and middle delivery scenarios, the Farr Pump was adjusted daily to satisfy the Colorado River minimum outflow obligation below Shadow Mountain Reservoir:

$$Pump = max(0,\ Q_{Adams\ Tunnel} + Q_{CR\ minimum} - Q_{East\ Inlet} - Q_{North\ Inlet} - Q_{North\ Fork})$$

For the no-pump scenario, Adams Tunnel delivery was adjusted daily to meet the Colorado River obligation using available natural inflow only:

$$Tunnel\ Delivery = max(0,\ Q_{East\ Inlet} + Q_{North\ Inlet} + Q_{North\ Fork} - Q_{CR\ minimum})$$

Monthly Colorado River minimum outflow requirements are defined in Appendix B and applied consistently across all scenarios. In all scenarios, Shadow Mountain Reservoir storage was held effectively flat by enforcing outflow equal to the Colorado River minimum, consistent with the negligible storage change assumption described in "Streamflow Data" under Section 2.4.1.

To ensure internal consistency between imposed operational constraints and model inputs, Chipmunk Lane connecting channel flow was recomputed for each scenario as:

$$Q_{Chipmunk} = Q_{East\ Inlet} + Q_{North\ Inlet} - Q_{Adams\ Tunnel}$$

rather than drawn from the observed dataset, ensuring that scenario-imposed conditions propagated consistently through the lagged predictor structure of the model. Within the scenario sub-window, a rolling lag buffer stored each scenario's daily pump flow, North Fork flow, Chipmunk Lane flow, and ensemble-mean predicted temperatures, with lag features drawn from this buffer at offsets of 1–5 days for hydrological variables and 1–2 days for temperature. Prior to the scenario sub-window, all scenarios shared the observed lag buffer.

Scenario effects were quantified as the difference between each scenario's ensemble-mean daily prediction and the corresponding baseline prediction at both near-surface and depth-integrated horizons, where negative values indicate scenario-driven cooling relative to observed operations. To assess model extrapolation risk, the distribution of scenario-derived Chipmunk Lane flow values was compared against the training data distribution; results are reported in Appendix E and confirm that all scenarios operated within the range of conditions represented in the training record.

2.5 Results

2.5.1 1-day ahead performance

The theory-guided neural network model achieved strong and consistent 1-day-ahead predictive skill at both depth horizons across cross-validation folds and the held-out 2025 test period. At near-surface depth (0–1m) NSE where 1 is a model that predicts observed values perfectly, 0 indicates the model does not perform better than the mean value, and a negative value means the model is worse than guessing the mean, ranged narrowly from 0.941–0.956 across the five cross-validation folds, with MAE ranging from 0.345–0.397°C and RMSE from 0.451–0.521°C. At the integrated depth (0–5m), the model achieved NSE of 0.974–0.981, MAE of 0.229–0.276°C, and RMSE of 0.291–0.337°C. The ensemble of five cross-validation models achieved the strongest overall performance at both depths (near-surface MAE 0.345°C, NSE 0.956; integrated MAE 0.228°C, NSE 0.982, Appendix D). On the held-out 2025 test period, the model maintained nearly identical performance, near-surface MAE 0.34°C, RMSE 0.45°C, NSE 0.956, bias 0.03°C; integrated MAE 0.23°C, RMSE 0.28°C, NSE 0.982, bias 0.05°C, demonstrating no meaningful degradation on held-out data (Figure 2).

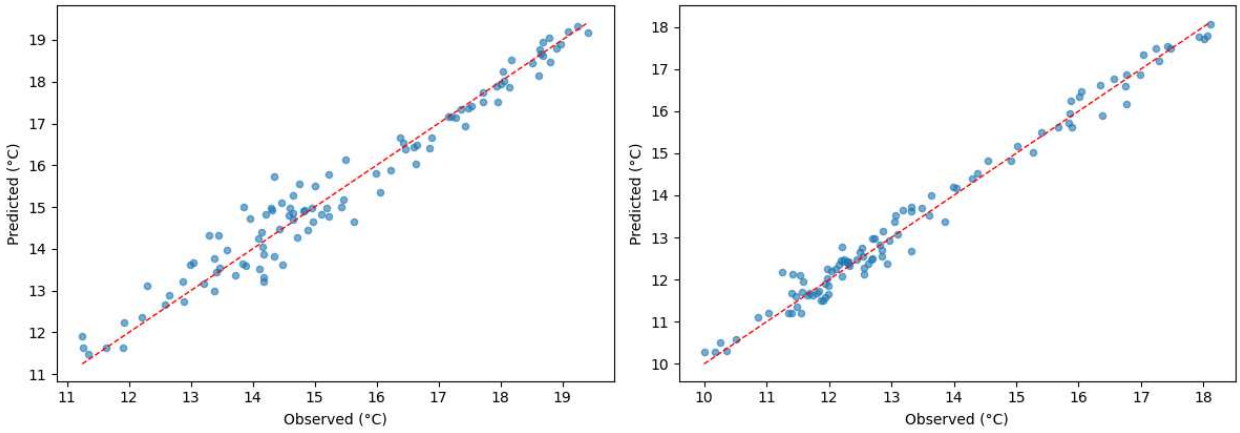
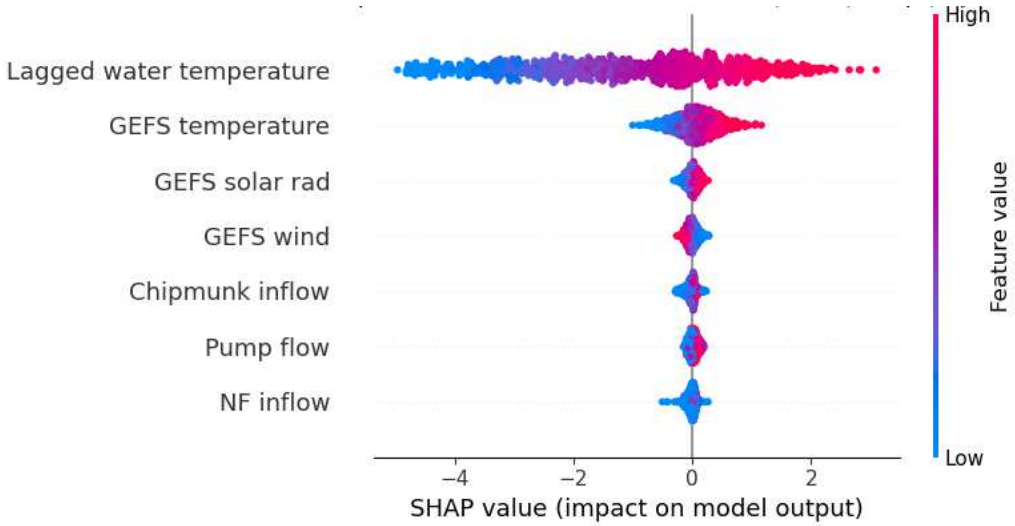


Figure 2. Performance of one-day-ahead ensemble (mean only) at near-surface (left) and integrated depth (right) as predicted temperature (y-axis) versus observed temperature (x-axis). Red dashed line is the 1:1 line.2

SHAP values (Figure 3) confirmed that feature importance aligned with the conceptual model of the system. Lagged water temperature was the dominant predictor for both output variables, reflecting the strong thermal inertia of a waterbody. For near-surface temperature, GEFS air temperature ranked second in importance, consistent with direct surface heat exchange as the primary meteorological driver of near-surface conditions. The integrated depth target showed a different ordering: pump flow ranked second, with higher Farr Pump deliveries associated with decreased integrated temperatures. Solar radiation ranked prominently for both targets, while wind exerted relatively greater influence on near-surface temperature. Inflow terms exhibited more coherent feature-response relationships for the integrated target, consistent with cold inflows having a more direct influence on depth-integrated conditions than on near-surface temperatures, which are more strongly mediated by atmospheric forcing. The alignment between feature importance rankings and process understanding indicates that the model learned physically meaningful relationships rather than spurious correlations.



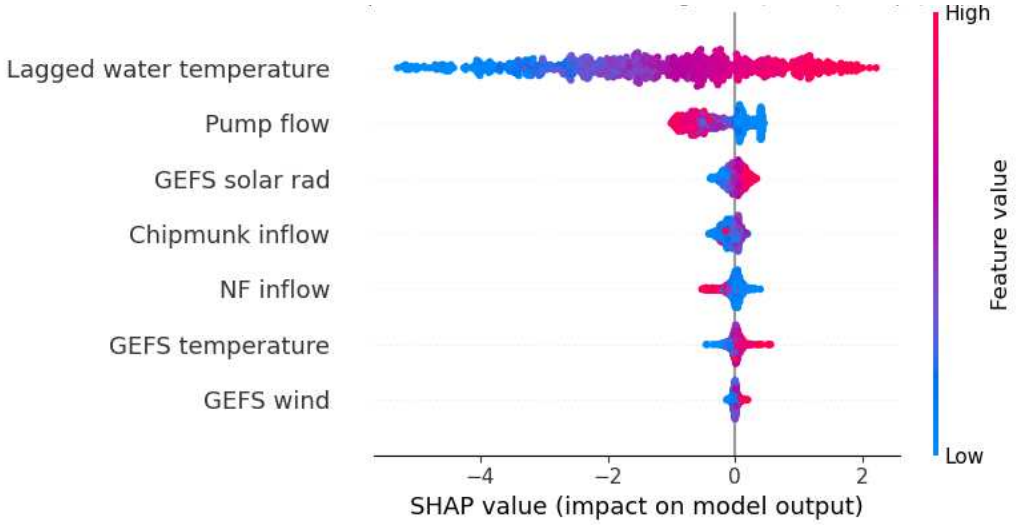


Figure 3. Grouped SHAP beeswarm plots for near-surface (0–1 m; top) and depth-integrated (0–5 m; bottom) water temperature predictions. Each point represents a single forecast instance where horizontal position indicates the magnitude and direction of each feature group's contribution to the model output, and color indicates relative feature value (red = high, blue = low). Feature groups aggregate individual input SHAP values: lagged water temperature includes prior observed temperatures at both depth intervals; pump flow, North Fork inflow, and Chipmunk inflow aggregate same-day and lagged flow features; and GEFS temperature, solar radiation, and wind aggregate same-day and lagged GEFSv12 forecast features.

The model outperformed the persistence baseline at both depths in the test set with near-surface MAE 0.35°C versus persistence 0.38°C, and integrated MAE 0.23°C versus persistence 0.25°C, a consistent 9–10% reduction in MAE that, combined with near-zero bias and NSE exceeding 0.95 at both depths, confirms the model adds meaningful predictive skill beyond a simple yesterday-is-today assumption. The model achieved lower error at the integrated depth than at the near-surface, with integrated MAE approximately 35% lower across both evaluation periods, consistent with the greater thermal inertia and reduced day-to-day variability of the depth-integrated layer. Predicted and observed temperature time series for the 2025 test period showed that the model tracked the seasonal temperature trajectory (Appendix D). Some day-to-day temperature departures exhibited a slight lag relative to observations, a known artifact of autoregressive forecast systems with lagged temperature inputs, but this behavior did not reflect a one-to-one persistence response, as the model outperformed the yesterday-is-today baseline at both depths. The model achieved MAE skill scores of 0.098 and 0.092 at near-surface and

integrated depths respectively, reducing MAE by 9.8% and 9.2% over the persistence baseline. This further confirms that the model adds meaningful predictive value beyond a simple yesterday-is-today assumption while acknowledging that the margin of improvement is modest at the 1-day horizon. While the model occasionally underestimated the magnitude of sharp day-to-day temperature changes (especially later in the season), it accurately reproduced the seasonal temperature trajectory and multi-day trends that are most relevant to operational decision making. Monthly statistics for the 2025 test period revealed greater model performance variability at the near-surface depth than at the integrated depth (Table 1). Near-surface performance was strongest in July (MAE=0.226°C, NSE=0.926) and weakest in August (MAE=0.434°C, NSE=0.052), where the model captured mean temperatures reasonably well but failed to reproduce day-to-day variability, as reflected in the near-zero NSE despite moderate MAE. Integrated depth performance remained consistently strong across all months (NSE=0.629–0.977).

Table 1. Monthly one-day-ahead model performance statistics for near-surface (0–1m) and integrated depth (0–5m) temperature predictions during the 2025 test period.

	Near-Surface Performance			Integrated Depth Performance		
Month	MAE	RMSE	NSE	MAE	RMSE	NSE
June	0.387	0.476	0.893	0.289	0.359	0.964
July	0.226	0.337	0.926	0.234	0.278	0.977
August	0.434	0.534	0.052	0.214	0.267	0.629
September	0.362	0.437	0.739	0.184	0.215	0.888

2.5.2 Seven-day time horizon performance

The model demonstrated consistently improving probabilistic forecast skill relative to the persistence baseline across the full 7-day forecast horizon at both depth horizons (Table 2). At the near-surface depth, CRPSS increased from 0.133 at day 1 to 0.617 at day 7, while at the integrated depth CRPSS increased from 0.130 at day 1 to 0.773 at day 7 indicating that the model's advantage over persistence grows with forecast lead time. This pattern reflects the rapid degradation of persistence skill at extended horizons as temperatures evolve away from the most recently observed value, while the model's ensemble-based probabilistic forecasts maintained relatively stable CRPS values across the forecast window. Near-surface CRPS increased modestly from 0.324 at day 1 to 0.387 at day 7, and integrated depth CRPS increased from 0.219 at day 1 to 0.243 at day 7, a small degradation in absolute forecast error across a 7-day horizon that reflects the model's ability to capture the dominant thermal forcing drivers at both depths (Figure 4).

Table 2. Performance statistics for the 2025 hold out (test set) at near-surface and integrated depths for June-Sept.

		Near-surface depth (0-1m)			Integrated depth (0-5m)		
Horizon	N dates	CRPS predicted	CRPS persistence	CRPSS	CRPS predicted	CRPS persistence	CRPSS
1	100	0.3236	0.3732	0.1328	0.2185	0.2511	0.1299
2	98	0.3320	0.4945	0.3287	0.2157	0.3951	0.4540
3	96	0.3452	0.6066	0.4310	0.2317	0.5508	0.5794
4	94	0.3490	0.7064	0.5059	0.2305	0.6898	0.6658
5	93	0.3631	0.7914	0.5412	0.2261	0.8289	0.7272
6	92	0.3789	0.8905	0.5745	0.2439	0.9566	0.7450
7	91	0.3873	1.0102	0.6166	0.2428	1.0716	0.7734

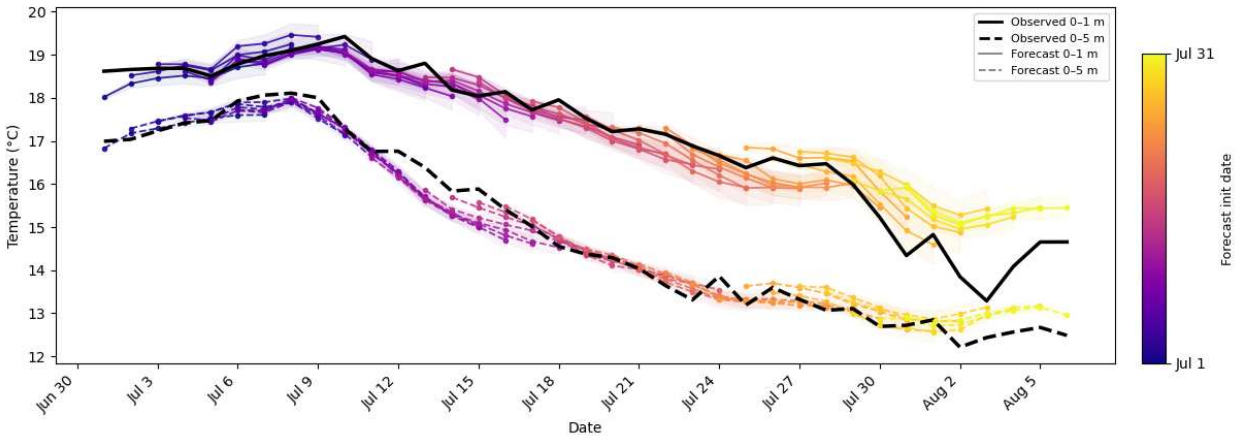


Figure 4. 7-day forecasts for the month of June of near-surface (0–1 m, solid) and depth-integrated (0–5 m, dashed) water temperature in Shadow Mountain Reservoir under observed 2025 operations, initialized daily from July 1–31. Forecast lines are colored by initialization date (purple = July 1, yellow = July 31). Shading represents ±1 SD of the GEFSv12 ensemble spread. Observed temperatures are shown in black for reference.

The integrated depth performance was consistently better than the near-surface across all forecast lead times, with integrated CRPS approximately 35% lower than near-surface CRPS at each horizon. This is consistent with the 1-day-ahead performance pattern and the greater thermal inertia of the depth-integrated layer. By day 7, the model reduced probabilistic forecast error by 62% over persistence at the near-surface and 77% at the integrated depth, demonstrating that the ensemble-based forecast system provides substantial added value over a naive baseline at operationally relevant timescales. Monthly CRPS was more consistent across the forecast horizon at integrated depths than at near-surface, where June and August forecasts showed higher absolute error than other months, likely reflecting greater thermal variability during the early summer transition and peak summer heating periods respectively. Despite this variability, all months outperformed the persistence baseline by the 2-day forecast horizon at both depths (Figure 5), confirming that the model adds consistent predictive value across the full June-

September evaluation window regardless of monthly thermal conditions.

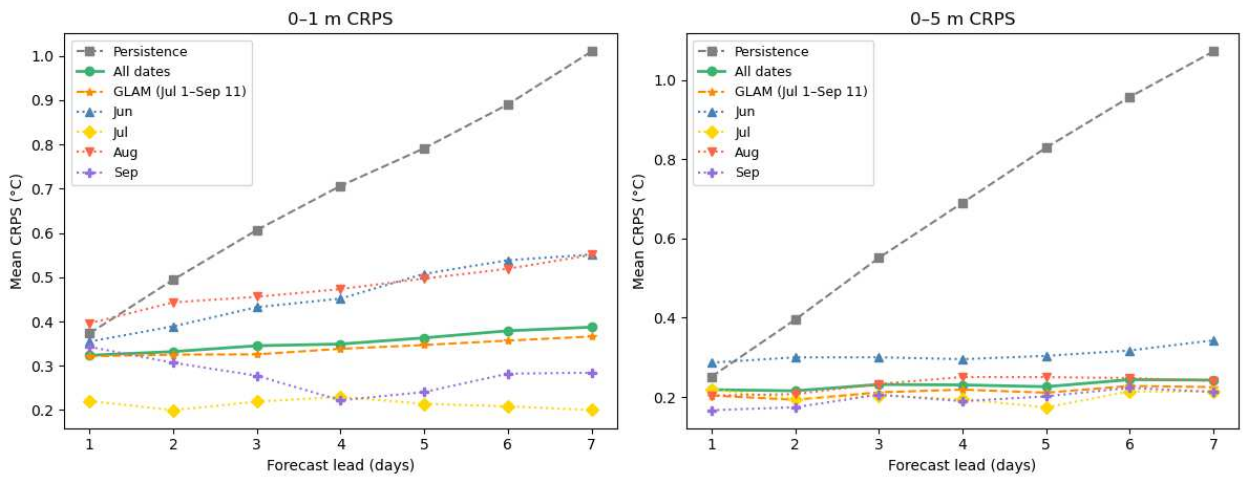


Figure 5. Continuously ranked probability score for the persistence forecast that assumes that the forecasted temperature is static for the 7-day horizon (grey dashed line and boxes) and the forecast (green circles and solid green line) for hindcasts made for June through September of 2025. The GLAM season (Jul 1 – Sept 11) performance is indicated with orange stars and dashed orange lines. Dotted lines represent data subsets by month.

2.4.3 Performance degradation due to estimated features

To evaluate the robustness of the forecast system to estimated inputs, model performance was assessed under progressive substitution of observed data with forecasted or derived counterparts. Replacing observed streamflow with water balance-derived Chipmunk Lane

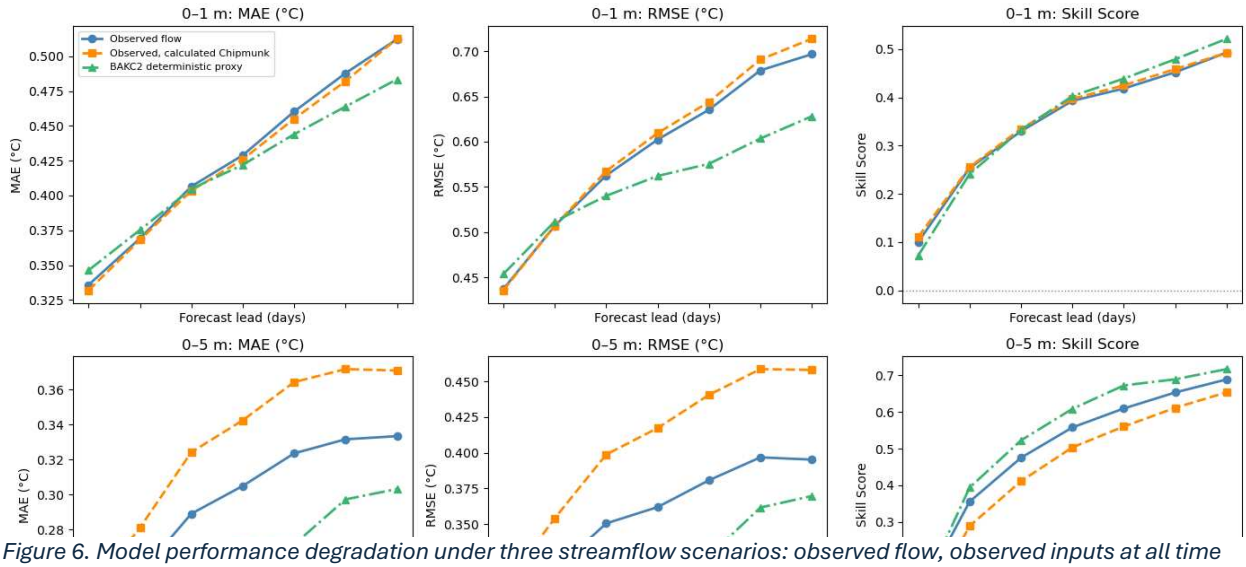


Figure 6. Model performance degradation under three streamflow scenarios: observed flow, observed inputs at all time horizons (blue circles), observed inputs at all time horizons, but using water balance equation to estimate Chipmunk (orange squares), using the BAKC2 flow as a proxy to calculate streamflow variables and estimate chipmunk flow, all other inputs observed (green triangles).

estimates and Baker's Gulch-based inflow forecasts introduced modest degradation in forecast skill across the 7-day horizon at both depths, but skill scores remained consistent across the substitution confirming that the streamflow estimation approach does not meaningfully compromise forecast skill (Figure 6). Similarly, substituting day-0 GEFSv12 meteorological inputs with forecasted GEFSv12 ensemble data produced negligible differences in forecast skill at both depths (Figure 7), suggesting that meteorological forecast uncertainty is not the primary driver of performance degradation relative to day-0 inputs at extended horizons.

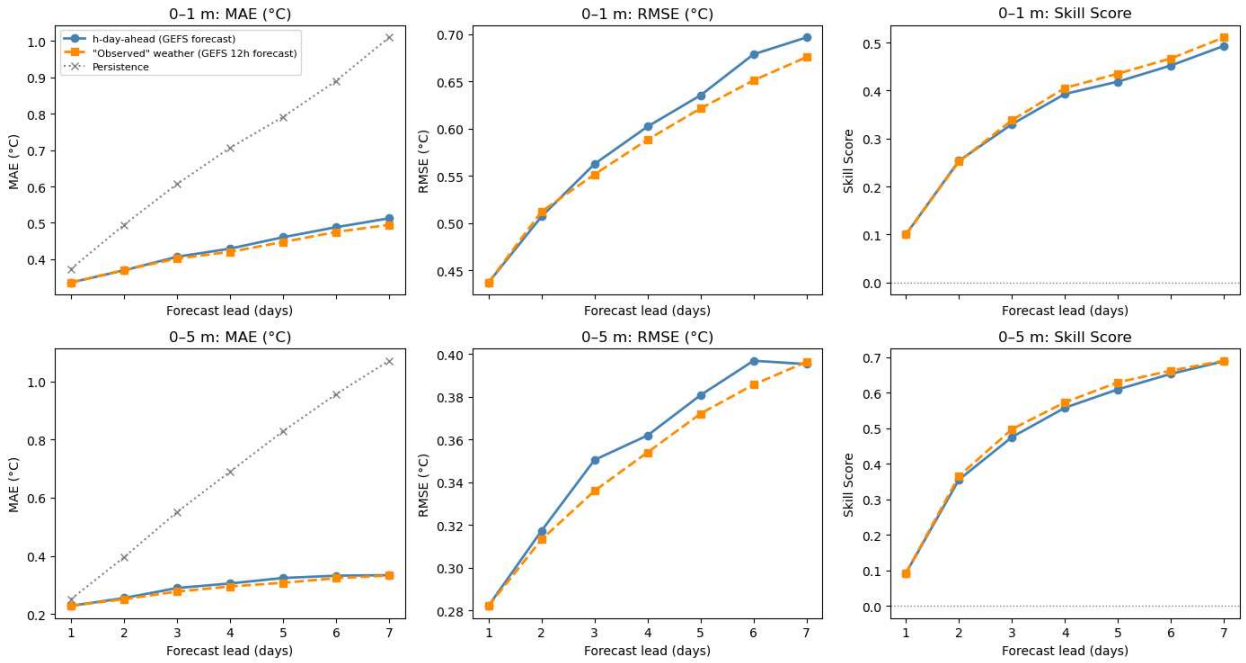


Figure 7. Model performance degradation relative to "observed" and forecasted meteorological inputs. The orange squares and orange dashed line represent day zero GEFSv12 meteorology, which was used as "observed" in the training process and the GEFSv12 noon-time forecasts in the blue line with blue circles.

2.5.4 Hindcast Scenario Analysis

Hindcast scenario analysis revealed a clear response relationship between Three Lakes System operations and delivery volume and predicted water temperature in Shadow Mountain Reservoir during the July 1–September 11 adaptive management window (Table 3, Figure 6). Under the maximum delivery scenario (550 cfs), the model predicted mean near-surface cooling of 0.35°C and mean integrated depth cooling of 0.46°C relative to the observed baseline, with

peak cooling reaching 1.85°C at the near-surface and 3.14°C at the integrated depth in early July. Under the middle delivery (220 cfs) and minimum pump scenarios, the model predicted warming relative to the observed baseline, with mean near-surface deltas of +1.26°C and +2.96°C and mean integrated depth deltas of +1.57°C and +4.04°C respectively (Table 3). These warming results reflect the fact that observed 2025 operations averaged approximately 464 cfs, substantially exceeding both alternative scenarios, a consequence of the limited operational flexibility in a low snowpack year (a timeseries of the deliveries and Farr Pump operations over the GLAM period are provided in Appendix E). The minimum pump scenario produced the largest departure from baseline, with near-surface and integrated depth temperatures reaching maximum deltas of +4.71°C and +5.74°C respectively during peak summer conditions in late August and early September (Figure 6), somewhat later in the season than the middle delivery scenario peak, consistent with the greater thermal inertia of a system receiving no cold hypolimnetic inflow.

Table 3. Summarized daily temperature deviation from observed baseline for the near surface and integrated depth layers for three hindcast scenarios during the scenario sub-window (1 Jul – 11 Sep 2025). Negative deltas indicate cooler predicted water temperatures than observed; positive deltas indicate warmer. The minimum pump scenario constrains Adams Tunnel delivery to natural flow from the North Fork, East Inlet, and North Inlet and pumps only as needed to maintain surface elevation, while the max and mid Adams Tunnel delivery scenarios adjust delivery and pump rates to hold lake level constant.

Scenario	Mean pump (cfs)	Mean pump adj. vs baseline (cfs)	Mean delta 1m (ºC)	Min delta 1m (ºC)	Max delta 1m (ºC)	Mean delta 0-5m (ºC)	Min delta 0-5m (ºC)	Max delta 0-5m (ºC)
Max Adams Tunnel Delivery (550 cfs)	490.6	26.7	-0.352	-1.849	0.037	-0.464	-3.138	0.062
Middle Adams Tunnel Delivery (220 cfs)	161.0	-302.9	1.255	-0.283	2.083	1.565	-0.600	2.216

Minimum pump operation	0	-463.9	2.959	-0.081	4.708	4.039	-0.049	5.742

Figure 8. Predicted departure from baseline (predicted values using observed inputs) based on three operational scenarios. Scenario 1 is maximizing the Adams Tunnel Deliveries (550cfs, orange line), Scenario 2 is consistent, but low deliveries (220 cfs, green line), Scenario 3 with minimal pump operation, and delivery is determined by natural inflow (blue line) with pumping only to maintain water level.

Temperature deltas grew through the summer across all scenarios, peaking in late August before tapering as temperatures declined toward the end of the adaptive management window (Figure 8). The maximum delivery scenario maintained negative temperature deltas throughout the scenario sub-window at both depths, while the mid and minimum pump scenarios maintained positive deltas for the majority of the window. Extrapolation risk assessment confirmed that all scenario-derived Chipmunk Lane flow values fell within the training data distribution (Appendix E).

2.6 Discussion

This study presents the first operational water temperature forecast tool developed specifically for the Three Lakes System, addressing a gap between the temporal resolution of

existing decision support tools and the weekly cadence of GLAM decisions. The development of near-term iterative forecast systems for freshwater quality management represents a growing area of applied limnological research (Dietze et al., 2018; Lofton et al., 2023), and this work operates within that paradigm by demonstrating that a parsimonious, theory-guided neural network can deliver consistent, operationally relevant temperature predictions in a data-limited mountain reservoir system. The model achieved predictive skill across a range of hydrological conditions, and hindcast scenario analysis revealed a physically coherent and meaningful thermal response to Adams Tunnel delivery operations, together providing a foundation for temperature-informed pump management in the Three Lakes System. Critically, the model did not merely perform well statistically, SHAP value analysis confirmed that it learned physically meaningful relationships consistent with the conceptual model of the system, with near-surface water temperature importance dominated by meteorological forcing and integrated depth importance shifting toward operational and hydrological variables, including pumping operations from Lake Granby as an influential operational input on integrated depth temperature.

This shift in feature importance across depth targets confirms that the dual-target architecture, where I predict both near-surface and integrated depths together, captures a genuine physical distinction in the dominant forcing at each depth horizon, one that would not have emerged from a model trained on either depth alone. The indirect influence of pumping on near-surface temperature, mediated through its effect on integrated temperature and propagated through the lagged autoregressive feature structure, was evident in the hindcast scenario departures, where near-surface and integrated depth deltas tracked each other across all scenarios. This alignment between learned feature importance and physical understanding provides a basis for operational trust in the tool that extends beyond statistical performance

metrics alone (Karpatne et al., 2017). End users expressed initial skepticism toward a machine learning approach, consistent with broader challenges in translating data-driven forecast tools into operational practice (Carey et al., 2022; Dietze et al., 2018; Lofton et al., 2023). SHAP-based feature importance analysis was central to addressing this concern, providing a transparent accounting of model behavior that grounded stakeholder confidence in the physical meaningfulness of model predictions.

The forecast system demonstrated consistent skill beyond the persistence baseline across the full 7 day forecast horizon at both depth horizons, the timescale most relevant to weekly GLAM operational decisions. This consistent improvement over persistence at decision-relevant timescales represents a core operational utility of the tool, providing pump operators and water managers with daily temperature forecasts that are more informative than assuming conditions will remain unchanged. Critically, while operators understand that initiating or increasing Farr Pump operation or Adams Tunnel deliveries will suppress Shadow Mountain Reservoir temperatures, the magnitude and rate of that thermal response under specific operational configurations and meteorological conditions has not previously been quantifiable in near-real time. The forecast system addresses this gap directly, providing daily ensemble-based temperature predictions driven by GEFSv12 meteorological forcing that give end users a range of predicted thermal outcomes rather than a single deterministic value, even under the imperfect near-surface performance characteristic of convectively active summer months.

The magnitude of temperature suppression attributable to pumping operations is particularly noteworthy in the context of 2025 hydrology: a low snowpack year with limited operational flexibility. By suppressing water temperatures in Shadow Mountain Reservoir by nearly 3°C at the near-surface and 4°C at the integrated depth relative to natural inflow

conditions, observed pumping operations likely reduced rates of biological production during the period when temperature-mediated growth would otherwise be highest, directly targeting the biotic pathway through which water temperature influences clarity in Grand Lake. While this study does not directly measure clarity outcomes, the thermal suppression documented here represents a meaningful operational intervention along the mechanistic chain linking water temperature, biological production, and water clarity that motivated the development of this tool. I can estimate the whole-system metabolic suppression due to the temperature decrease using the equation for Boltzmann-Arrhenius temperature scaling of metabolic rate ($B \propto e^{-E_a/kT}$) as described in (Gillooly et al., 2001):

$$\frac{B_2}{B_1} = e^{E_a/k(\frac{1}{T_1}-\frac{1}{T_2})}$$

Where E_a is the activation energy, k is Boltzmann's constant, and T is absolute temperature (Kelvin). Using an activation energy value of 0.65 and k as 8.617×10^{-5} eV K^{-1}, the 2025 operations resulted in an estimated 31% suppression of whole-system metabolism at the near-surface and an estimated 46% suppression in the integrated depth over the GLAM period.

Importantly, this counterfactual, what temperatures would have been without pumping, is inherently unobservable from temperature records alone, and the hindcast framework provides a unique means of quantifying the thermal contribution of current operations to reservoir management outcomes. The warming predicted under the 220 cfs scenario relative to the baseline reflects the operational context of 2025: actual operations substantially exceeded 220 cfs throughout the season because of limited snowpack-driven flexibility, making the observed baseline a high-delivery regime against which reduced delivery scenarios necessarily appear warmer. This result has direct relevance, as 220 cfs represents a plausible near-future delivery

rate following anticipated 2027 upgrades to Pole Power Plant (Bureau of Reclamation, personal communication, 2025).

Beyond its core function as a 7-day temperature forecast, the primary operational value of the forecast system lies in its capacity for on-demand scenario evaluation. The seasonal question of when to initiate Farr Pump operations is a key decision context, as snowmelt-driven natural inflow declines and rising Shadow Mountain Reservoir temperatures increase the risk that biologically productive water will be conveyed into Grand Lake upon pump initiation. Initiating pumping earlier in the season, for example, could produce additional temperature suppression during the critical early GLAM period, though early initiation involves a water quality tradeoff: pumping before adequate settling of North Fork sediment inputs could introduce turbid water into Grand Lake, potentially offsetting the clarity benefits of thermal mitigation. This tradeoff between thermal mitigation and sediment-driven clarity degradation illustrates precisely the kind of decision context for which an operational forecast tool adds value, providing managers with quantitative thermal context to inform the informal settling period judgment described in Section 2.3.3. Water temperature represents a tractable first modeling target within a broader decision support architecture aimed at operational prediction of water clarity in Grand Lake, the ultimate management outcome motivating adaptive operations in the Three Lakes System.

2.6.1 Model Constraints and Sources of Uncertainty

As with all data-driven models, the operational envelope of this forecast system is defined by the range of conditions represented in the training record. While the theory-guided architecture and rigorous cross-validation framework provide some protection against extrapolation risk, the model may perform less reliably under conditions substantially outside historical experience, such as extreme drought years with anomalous thermal regimes or

operational configurations not represented in the training data. The 2025 test period, produced performance nearly identical to cross-validation results, providing some confidence in the model's generalization capacity under typical hydrological conditions. Further, the operational model predicts temperature at a single monitoring location in Shadow Mountain Reservoir, and point predictions at the SM-MID buoy location may not fully represent thermal conditions throughout the waterbody, particularly given the shallow and well-mixed nature of the reservoir. Additionally, daily average water temperature predictions do not capture the diurnal temperature extremes, afternoon highs and overnight lows, that are relevant to biotic growth processes. These spatial and temporal averaging limitations should be considered when interpreting model outputs in the context of water quality management.

Near-surface forecast performance showed greater monthly variability than integrated depth performance, with August producing the largest performance divergence between depth targets (Table 1). Near-surface NSE dropped to 0.052 in August despite reasonable MAE, suggesting the model captured mean August temperatures but missed high-frequency variability, consistent with the influence of convective afternoon storms not directly represented as model inputs. Precipitation is not included as a model input and enters the model only indirectly through streamflow response, limiting the model's ability to capture rapid near-surface thermal responses to episodic summer convection. The integrated depth target is less sensitive to this forcing, consistent with the damping of high-frequency atmospheric signals at depth (Wetzel, 2024). The poor August near-surface NSE should be interpreted in context: ensemble CRPS across the 7-day forecast horizon remained nearly flat relative to persistence in August at both depth targets (Figure 5), indicating that the model retained skill in tracking multi-day temperature trends even when day-to-day variability was poorly captured. For weekly GLAM

pump management decisions, this multi-day trend tracking represents the operationally relevant performance characteristic, and the August near-surface limitation is therefore less consequential than the one-day-ahead NSE alone would suggest. The noon-time instantaneous GEFS inputs used in this study may compound this limitation by excluding the diurnal forcing cycle most relevant to near-surface temperature dynamics. Incorporating daily minimum and maximum meteorological forecasts rather than noon-time instantaneous values only could improve near-surface temperature predictions by capturing the full diurnal forcing cycle. However, adding features to an already parsimonious model trained on a modest observational record risks reintroducing the overfitting concerns that motivated the conservative architecture design. Additional features would require careful evaluation against the low observation-to-feature ratio that underpins the current model's generalization performance, potentially necessitating compensating reductions elsewhere in the feature set. Higher temporal resolution meteorological forecasts, such as the High Resolution Rapid Refresh (HRRR) model, could provide improved meteorological specificity in complex mountain terrain, but are limited to a 72-hour forecast horizon, which is insufficient for an operational tool requiring 7-day predictions. This presents a fundamental tradeoff between meteorological specificity and forecast horizon length that is not easily resolved within the current operational constraints of the system.

The performance degradation analysis showed remarkably little skill loss across the 7-day forecast horizon, a result best understood in the context of the training data decision described in “Meteorological Data” heading of Section 2.4.1. Because the model was trained on GEFSv12 day-0 output rather than local observed meteorology, the terrain offset error between gridded GEFS output and local conditions was present in the training data from the outset. Switching from day-0 to forecasted GEFS inputs therefore introduces relatively little additional

uncertainty beyond what the model was already trained to handle, not because the model is robust to meteorological error in a general sense, but because the specific character of that error is already embedded in its learned relationships. This consistency comes at a cost: training on gridded rather than locally observed meteorology likely depresses the overall performance baseline relative to what a locally calibrated model could achieve. Data assimilation approaches such as Kalman filtering represent one promising avenue for addressing systematic meteorological bias at the point of forecast issuance, allowing model state to be updated with incoming observations as the forecast period progresses (Carey et al., 2022; Dietze et al., 2018; Zwart, Oliver, et al., 2023). Incorporating such approaches into future iterations of this forecast system could reduce the performance ceiling imposed by training on gridded rather than locally observed meteorology.

Another source of uncertainty within this system is from the use of the Bakers Gulch flow forecast and the estimation of Chipmunk Lane flow as inputs into the system. Streamflow estimation for forecast periods relied on linear relationships between Baker's Gulch and the North Fork, East Inlet, and North Inlet, and a mass balance equation to derive Chipmunk Lane connecting channel flow. These approaches simplify the hydrological complexity of the system. One source of variability not captured by the mass balance assumption of negligible storage change is the intentional operational manipulation of reservoir storage within the 1-foot Senate Document 80 constraint. This same constraint defines the outer bounds of the scenario space evaluated here, as any operational alternative must maintain the required water level relationship between Shadow Mountain Reservoir and Grand Lake, a management reality that the scenario design honors by adjusting Farr Pump operations to maintain consistent elevation across all delivery scenarios. While this operational flexibility represents a real source of model

uncertainty, SHAP values indicated that Chipmunk Lane flow was among the lower-importance features at both depths (Figure 3), and the performance degradation analysis demonstrated modest skill loss when observed streamflow was replaced with estimated values (Figure 6), suggesting that this simplification does not substantially compromise forecast skill under typical operational conditions.

The 2025 low snowpack year also raises a related physical consideration that warrants acknowledgment. The Farr Plant intake is fixed at approximately 26 meters below the surface of Granby at full pool elevation, a depth that reliably draws from the hypolimnion under normal storage conditions. In low water years where Lake Granby storage is substantially reduced, the water surface drops toward the fixed intake depth, increasing the risk that the thermocline descends below the intake and that pumped water is drawn from above the hypolimnion rather than below it, reducing the thermal contrast between pumped and ambient Shadow Mountain Reservoir temperatures. While operational records and internal modeling suggest the intake remains below the hypolimnion even at reduced storage levels and that the thermal impact would be minimal (Northern Water, unpublished data), this represents a source of uncertainty not currently incorporated in the model and warrants monitoring.

2.7 Conclusion

Water temperature management in Shadow Mountain Reservoir through pumping operations represents a meaningful lever for water quality in the Three Lakes System, where Grand Lake clarity is subject to established goal qualifiers which is influenced by the water quality of a connected waterbody. This study developed and evaluated the first operational water temperature forecast tool for the Three Lakes System, demonstrating that a parsimonious, theory-guided neural network driven by weather forecasts, inflow estimates, and pumping operations can deliver consistent and physically interpretable temperature predictions in a data-limited

mountain reservoir system. This tool provides pump operators and water managers with a quantitative, data-driven foundation for temperature-informed decision making during the adaptive management period.

The most direct extension of this work is direct forecasting of Secchi disk depth, the target variable of the goal qualifiers, which would connect the thermal management framework developed here to the clarity outcomes that motivated it. Improved uncertainty quantification also represents a priority: a more complete characterization of forecast uncertainty, evaluated against standard probabilistic benchmarks such as the 90% prediction interval coverage rate (Olsson et al., 2025), would strengthen operational credibility and better support uncertainty-informed decision making among end users.

The results demonstrate that parsimonious, theory-guided models grounded in the operational and regulatory context of a managed system are tractable even under data-limited conditions, a finding that has implications beyond this system. Most managed reservoirs lack the long observational data required by physical models like a general lake model or CE-QUAL-W2, but they may have operational logs and routine measurements that could allow for them to build a forecast system under a similar framework presented here. The central architectural decision of embedding management levers directly in model structure, rather than treating them as external or post-hoc considerations, is generalizable to any system where an operational decision has a measurable mechanistic result such as inter-basin transfers, or dam releases, or hypolimnetic oxygenation. The focus on water quality and clarity in Grand Lake that motivated this tool in the Three Lakes System have analogs across managed freshwater systems worldwide, as do the stakeholder trust barriers that constrain ML adoption in operational settings. The stakeholder trust building and SHAP analysis is a tool within the context of this architecture. SHAP-based

interpretability and iterative co-development with end users are not Three Lakes-specific solutions, either; they are generalizable strategies for building the institutional confidence that translates forecast skill into operational action.

REFERENCES

Autobee, R. (1996). *Colorado-Big Thompson Project* (Reclamation Project Histories). Bureau of Reclamation. https://www.usbr.gov/history/ProjectHistories/Colorado-Big-Thompson-Project.pdf

Barthlow, J., Hanna, R. B., Saito, L., Lieberman, D., & HORN, M. (2001). Simulated Limnological Effects of the Shasta Lake Temperature Control Device. *Environmental Management*, *27*(4), 609–626. https://doi.org/10.1007/s0026702324

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Bixby, R. J., Cooper, S. D., Gresswell, R. E., Brown, L. E., Dahm, C. N., & Dwire, K. A. (2015). Fire effects on aquatic ecosystems: An assessment of the current state of the science. *Freshwater Science*, *34*(4), 1340–1350. https://doi.org/10.1086/684073

Blaylock, B. K. (2022). *Herbie: Retrieve Numerical Weather Prediction Model Data* (Version 2022.9.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4567540

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a Metabolic Theory of Ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/10.1890/03-9000

Carey, C. C., Woelmer, W. M., Lofton, M. E., Figueiredo, R. J., Bookout, B. J., Corrigan, R. S., Daneshmand, V., Hounshell, A. G., Howard, D. W., Lewis, A. S. L., McClure, R. P., Wander, H. L., Ward, N. K., & Thomas, R. Q. (2022). Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. *Inland Waters*, *12*(1), 107–120. https://doi.org/10.1080/20442041.2020.1816421

Carlson, R. E. (1977). A trophic state index for lakes. *Limnology and Oceanography*, *22*(2), 361–369. https://doi.org/10.4319/lo.1977.22.2.0361

Cui, L., Wang, J., SADEGHI TABAS, S., & Carley, J. (2025). *A Machine Learning-Based Bias Correction Method for Global Forecast System Products* (p. 23) [Office Note 520]. NOAA National Centers for Environmental Prediction. https://doi.org/10.25923/6266-E822

Dabernig, M., Mayr, G. J., Messner, J. W., & Zeileis, A. (2017). Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 909–916. https://doi.org/10.1002/qj.2975

Davies-Colley, R. J., & Smith, D. G. (2001). Turbidity Suspeni)ed Sediment, and Water Clarity: A Review. *JAWRA Journal of the American Water Resources Association*, *37*(5), 1085–1101. https://doi.org/10.1111/j.1752-1688.2001.tb03624.x

DeCicco, L. A., Hirsch, R. M., Lorenz, D., Read, J. S., Walker, J., Platt, L., Watkins, D., Blodgett, D., Johnson, M., Krall, A., & Stanish, L. (2025). *dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services* (Version 2.7.18) [R]. U.S. Geological Survey. https://doi.org/10.5066/P9X4L3GE

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Edinger, J. E., Duttweiler, D. W., & Geyer, J. C. (1968). The Response of Water Temperatures to Meteorological Conditions. *Water Resources Research*, *4*(5), 1137–1143. https://doi.org/10.1029/WR004i005p01137

Eppley, R. (1972). Temperature and phytoplankton growth in the sea. *Fisheries Bulletin*, *70*(4), 1063–1085.

Franco, A., Malhortra, N., & Simonovits, G. (2015). Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results. *Political Analysis*, *23*(2). https://doi.org/10.1093/pan/mpv006

GEI Consultants. (2013). *Grand Lake Water Clarity Technical Review and Work Plan* (Order No. R12PX60331). Bureau of Reclamation. https://www.usbr.gov/gp/ecao/final_grand_lake_clarity.pdf

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M., & Charnov, E. L. (2001). Effects of Size and Temperature on Metabolic Rate. *Science*, *293*(5538), 2248–2251. https://doi.org/10.1126/science.1061967

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://mitpress.mit.edu/9780262035613/deep-learning/

Grand Lake Adaptive Management Committee. (2023). *Grand Lake Clarity Adaptive Management 2023*. Northern Water, Bureau of Reclamation, Grand County, NWCCOG, and Colorado River Water Conservation District. https://www.northernwater.org/getattachment/b3b31d04-c115-4cd7-ad28-ee705011e76e/2023_Grand_Lake_Adaptive_Management_Annual_Report

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., Xue, X., Hou, D., Peng, J., Nageswararao, M. M., Tallapragada, V., Hamill, T. M., Whitaker, J. S., Bates, G., Pegion,

P., Frederick, S., Rosencrans, M., & Kumar, A. (2022). *GEFSv12 Reforecast Dataset for Supporting Subseasonal and Hydrometeorological Applications*. https://doi.org/10.1175/MWR-D-21-0245.1

Gupta, H. V., & Kling, H. (2011). On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research*, *47*(10). https://doi.org/10.1029/2011WR010962

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., Sinsky, E., Zhu, Y., Tallapragada, V., Guan, H., Zhou, X., & Woollen, J. (2022). *The Reanalysis for the Global Ensemble Forecast System, Version 12*. https://doi.org/10.1175/MWR-D-21-0023.1

Hanna, R. B., Saito, L., Bartholow, J. M., & Sandelin, J. (1999). Results of Simulated Temperature Control Device Operations on In-Reservoir and Discharge Water Temperatures Using CE-QUAL-W2. *Lake and Reservoir Management*, *15*(2), 87–102. https://doi.org/10.1080/07438149909353954

Hydros Consulting. (2025). *2024 Operational and Water-Quality Summary Report for the Three Lakes*. https://www.northernwater.org/getattachment/a1148476-0e02-4777-b720-be62a2168282/2024_Operational_and_Water_Quality_Summary_Report_for_the_Three_Lakes

Imberger, J., & Patterson, J. C. (1989). Physical Limnology. In *Advances in Applied Mechanics* (Vol. 27, pp. 303–475). Elsevier. https://doi.org/10.1016/S0065-2156(08)70199-6

Ivimey-Cook, E. R., Sánchez-Tójar, A., Berberi, I., Culina, A., Roche, D. G., A. Almeida, R., Amin, B., Bairos-Novak, K. R., Balti, H., Bertram, M. G., Bliard, L., Byrne, I., Chan, Y.-C., Cioffi, W. R., Corbel, Q., Elsy, A. D., Florko, K. R. N., Gould, E., Grainger, M. J., …

Moran, N. P. (2025). From policy to practice: Progress towards data- and code-sharing in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2055), 20251394. https://doi.org/10.1098/rspb.2025.1394

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, *10*(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kettle, H., Thompson, R., Anderson, N. J., & Livingstone, D. M. (2004). Empirical modeling of summer lake surface temperatures in southwest Greenland. *Limnology and Oceanography*, *49*(1), 271–282. https://doi.org/10.4319/lo.2004.49.1.0271

Lofton, M. E., Brentrup, J. A., Beck, W. S., Zwart, J. A., Bhattacharya, R., Brighenti, L. S., Burnet, S. H., McCullough, I. M., Steele, B. G., Carey, C. C., Cottingham, K. L., Dietze, M. C., Ewing, H. A., Weathers, K. C., & LaDeau, S. L. (2022). Using near-term forecasts and uncertainty partitioning to inform prediction of oligotrophic lake cyanobacterial density. *Ecological Applications*, *32*(5), e2590. https://doi.org/10.1002/eap.2590

Lofton, M. E., Howard, D. W., Thomas, R. Q., & Carey, C. C. (2023). Progress and opportunities in advancing near-term forecasting of freshwater quality. *Global Change Biology*, *29*(7), 1691–1714. https://doi.org/10.1111/gcb.16590

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: Assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. https://doi.org/10.1109/SIBGRAPI51738.2020.00053

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

National Centers for Environmental Prediction. (2020). *Global Ensemble Forecast System version 12 (GEFSv12) operational archive* (Version 12) [Dataset]. NOAA Open Data Dissemination Program. https://registry.opendata.aws/noaa-gefs

Olsson, F., Carey, C. C., Boettiger, C., Harrison, G., Ladwig, R., Lapeyrolerie, M. F., Lewis, A. S. L., Lofton, M. E., Montealegre-Mora, F., Rabaey, J. S., Robbins, C. J., Yang, X., & Thomas, R. Q. (2025). What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge. *Ecological Applications*, *35*(1), e70004. https://doi.org/10.1002/eap.70004

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., & Invernizzi, L. (2019). *KerasTuner* [Computer software]. https://github.com/keras-team/keras-tuner

O'Reilly, C. M., Sharma, S., Gray, D. K., Hampton, S. E., Read, J. S., Rowley, R. J., Schneider, P., Lenters, J. D., McIntyre, P. B., Kraemer, B. M., Weyhenmeyer, G. A., Straile, D., Dong, B., Adrian, R., Allan, M. G., Anneville, O., Arvola, L., Austin, J., Bailey, J. L., … Zhang, G. (2015). Rapid and highly variable warming of lake surface waters around the globe. *Geophysical Research Letters*, *42*(24). https://doi.org/10.1002/2015GL066235

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., & Salamon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*(null), 2825–2830.

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset shift in machine learning*. MIT Press.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J. A., Hanson, P. C., Watkins, W., Steinbach, M., & Kumar, V. (2019). Process-Guided

Deep Learning Predictions of Lake Water Temperature. *Water Resources Research*, *55*(11), 9173–9190. https://doi.org/10.1029/2019WR024922

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Scheuerer, M., & Hamill, T. M. (2019). Probabilistic Forecasting of Snowfall Amounts Using a Hybrid between a Parametric and an Analog Approach. *Monthly Weather Review*, *147*(3), 1047–1064. https://doi.org/10.1175/MWR-D-18-0273.1

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023, October 13). *Towards Understanding Sycophancy in Language Models*. The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=tvhaxkMKAn

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Sondergaard, M., Jensen, P. J., & Jeppesen, E. (2001). Retention and Internal Loading of Phosphorus in Shallow, Eutrophic Lakes. *The Scientific World Journal*, *1*(2), 676350. https://doi.org/10.1100/tsw.2001.72

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346. https://doi.org/10.1098/rsos.220346

Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Thomas, R. Q., Figueiredo, R. J., Daneshmand, V., Bookout, B. J., Puckett, L. K., & Carey, C. C. (2020). A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time. *Water Resources Research*, *56*(11), e2019WR026138. https://doi.org/10.1029/2019WR026138

United States Bureau of Reclamation. (2012). *Colorado-Big Thompson Project West Slope Collection System, Grand County, Colorado: Preliminary Alternatives Development Report*. United States Department of the Interior. https://www.usbr.gov/gp/ecao/1208_final_prelim_alt_dev.pdf

Varadharajan, C., Appling, A. P., Arora, B., Christianson, D. S., Hendrix, V. C., Kumar, V., Lima, A. R., Müller, J., Oliver, S., Ombadi, M., Perciano, T., Sadler, J. M., Weierbach, H., Willard, J. D., Xu, Z., & Zwart, J. (2022). Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrological Processes*, *36*(4), e14565. https://doi.org/10.1002/hyp.14565

Watters, A. (2023). *cdssr: R Client for CDSS REST API services* (Version 1.1.4.310) [R]. https://anguswg-ucsb.github.io/cdssr/

Wetzel, R. G. (2024). *Wetzel's limnology: Lake and river ecosystems* (I. D. Jones & J. P. Smol, Eds.; Fourth edition). Academic Press, an imprint of Elsevier.

Wetzel, R. G., & Likens, G. E. (2000). The Heat Budget of Lakes. In R. G. Wetzel & G. E. Likens (Eds.), *Limnological Analyses* (pp. 45–56). Springer. https://doi.org/10.1007/978-1-4757-3250-4_4

Willard, J. D., Varadharajan, C., Jia, X., & Kumar, V. (2025). Time series predictions in unmonitored sites: A survey of machine learning techniques in water resources. *Environmental Data Science*, *4*, e7. https://doi.org/10.1017/eds.2024.14

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Computing Surveys*, *55*(4), 66:1-66:37. https://doi.org/10.1145/3514228

Williamson, C. E., Overholt, E. P., Pilla, R. M., Leach, T. H., Brentrup, J. A., Knoll, L. B., Mette, E. M., & Moeller, R. E. (2015). Ecological consequences of long-term browning in lakes. *Scientific Reports*, *5*(1), 18666. https://doi.org/10.1038/srep18666

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zagona, E. A., Fulp, T. J., Shane, R., Magee, T., & Goranflo, H. M. (2001). Riverware: A Generalized Tool for Complex Reservoir System Modeling. *JAWRA Journal of the American Water Resources Association*, *37*(4), 913–929. https://doi.org/10.1111/j.1752-1688.2001.tb05522.x

Zhu, J.-J., Boehm, A. B., & Ren, Z. J. (2024). Environmental Machine Learning, Baseline Reporting, and Comprehensive Evaluation: The EMBRACE Checklist. *Environmental Science & Technology*, *58*(45), 19909–19912. https://doi.org/10.1021/acs.est.4c09611

Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environmental Science & Technology*, *57*(46), 17671–17689. https://doi.org/10.1021/acs.est.3c00026

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, *1*(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001

Zwart, J. A., Diaz, J., Hamshaw, S., Oliver, S., Ross, J. C., Sleckman, M., Appling, A. P., Corson-Dosch, H., Jia, X., Read, J., Sadler, J., Thompson, T., Watkins, D., & White, E. (2023). Evaluating deep learning architecture and data assimilation for improving water temperature forecasts at unmonitored locations. *Frontiers in Water*, *5*. https://doi.org/10.3389/frwa.2023.1184992

Zwart, J. A., Oliver, S. K., Watkins, W. D., Sadler, J. M., Appling, A. P., Corson-Dosch, H. R., Jia, X., Kumar, V., & Read, J. S. (2023). Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. *JAWRA Journal of the American Water Resources Association*, *59*(2), 317–337. https://doi.org/10.1111/1752-1688.13093

APPENDIX A: GEFS Data Overview

The GEFSv12 reanalysis (Hamill et al., 2022) product provides retrospective ensemble forecasts generated with the operational GEFSv12 model configuration initialized at 00:00 UTC at 6 hour forecast intervals for one control (c00) and four perturbations (p01-p04) to a 16-day time horizon for the two decades spanning 2000-2019 for training and calibration purposes. The GEFSv12 operational archive provides data beginning in late September 2020. The AWS operational archive features initializations every 6 hours and 31 forecast members (one control, thirty perturbations) at a 3-hour forecast interval to a 35-day time horizon.

To train the model, I obtained near-noon control forecast (c00) meteorological data (instantaneous air temperature, wind speed, downward shortwave radiation) from both forecast sources from the AWS data archive, but at two different initialization times due to data availability. For both products, only the control and a single forecast valid time per initialization was used to train the model. For the reforecast data, I acquired the f018 data from the 00:00 UTC initialization. NOAA Global Ensemble Forecast System (GEFS) Re-forecast was accessed from March 9, 2026 to March 10, 2026 from https://registry.opendata.aws/noaa-gefs-reforecast. For the operational archive data, I acquired the f012 forecast from the 06:00 UTC initialization. NOAA Global Ensemble Forecast System (GEFS) was accessed from March 10, 2026 to March 28, 2026 from https://registry.opendata.aws/noaa-gefs. Point extraction was performed using the Python module Herbie function `pick_points()` with nearest-neighbor interpolation on the native 0.25° grid. Processing was parallelized across initialization dates using Python's `Pool` function in the `multiprocessing` module, allocating all but two CPU cores to maintain system responsiveness. Per-date output files were written for each date to avoid reprocessing on reruns.

Wind speed was derived from the raw *u* (zonal) and *v* (meridional) components using the equation:

$$wind\ speed\ (mps) = \sqrt{\{u^2 + v^2\}}$$

Temperature values were converted from Kelvin to degrees Celsius by subtracting 273.15 from the GEFS temperature value.

For application and forecast assessment, I acquired the near-noon forecast for each of the GEFS members to seven days ahead using the same Python methodology outlined above.

APPENDIX B: Streamflow Data Overview

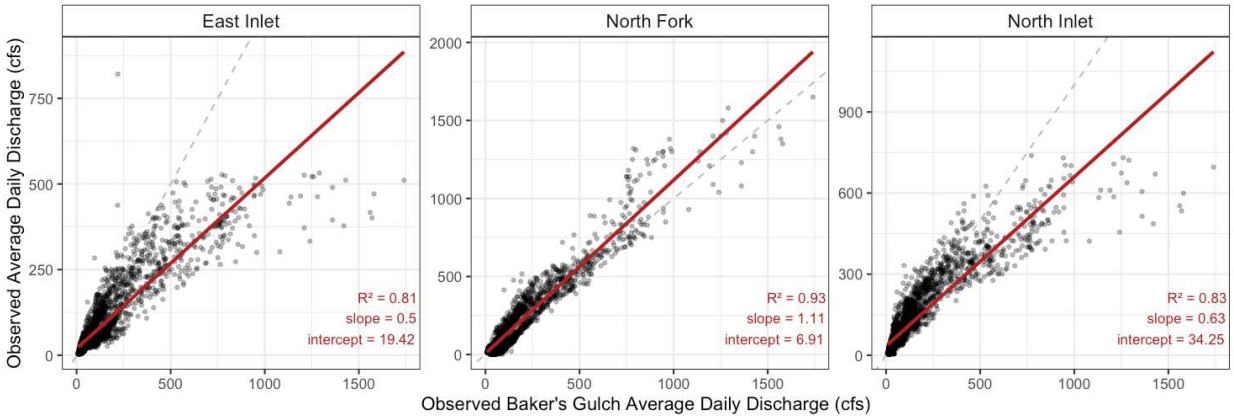
Streamflow data that were used to train and assess the model were pre-aggregated to a daily mean and acquired from Northern Water's Kisters Service API (https://data.northernwater.org/KiWIS/) for the North Fork of the Colorado River above Shadow Mountain Reservoir (NW site M-0009), the Farr Pump (US Bureau of Reclamation site EX-0054), the Adams Tunnel (US Bureau of Reclamation site EX-0047). The US Bureau of Reclamation sites' data is pulled into Northern Water's system for ease of use by downstream users. I used the R package `dataRetrieval` to gather sub-hourly flow data for the Chipmunk Lane connecting channel (USGS site 09014050) and aggregated these data to daily averages for use in the model.

I obtained the NOAA's deterministic forecast streamflow data for the Baker's Gulch (site bakc2) for 2025 through personal communication with Brenda Alcorn of NOAA's Colorado Basin River Forecast Center (cbrfc.webmasters@noaa.gov). The Baker's Gulch site is upstream of the North Fork flow data location but is the only flow site with an off-the-shelf flow forecast available. I used a linear relationship between the observed flow at Baker's Gulch and the North Fork, North Inlet and East Inlet sites to apply the Baker's Gulch forecast data across the system for forecast assessment. Observed flow for this analysis was obtained using `get_telemetry_ts()` function from the `cdssr` package (Watters, 2023) for the four sites (COLBAKCO, COLGRAND, EASTINLET, NORINLET). I assume that any pattern of stream flow predicted at Baker's Gulch is likely to be similar across the other inflows. While I do not use the North Inlet and East Inlet flow directly in the model, these estimates alongside the expected Adams Tunnel deliveries are used to estimate flow through Chipmunk Lane. The Colorado River below Shadow

Mountain must meet minimum flow requirements throughout the year. These requirements are 20 cfs January through May, 50 cfs in June and July, 40 cfs in August, 35 cfs in September and October, and 45 cfs November through December.

Baker's Gulch Linear Relationships

Using the observed streamflow from 2010-2025 for Baker's Gulch, the North Fork of the Colorado River that the North Inlet and East Inlet into Grand Lake, I calculated a simple linear model to establish a general relationship between Baker's Gulch Flow and the inputs for the model. All linear relationships have high R^2 values ranging from 0.81 (East Inlet) to 0.93 (North Fork) (Appendix Figure 1), providing a reasonable method for estimating streamflow at these three locations from the forecasted flow at Baker's Gulch, despite being an overly simplistic proxy.



Appendix Figure 1. Linear models for relating Baker's Gulch discharge to the discharge observed at three other locations in the Three Lakes Region: (left) East Inlet into Grand Lake, (center) North Fork into Shadow Mountain Reservoir, (right) North Inlet into Grand Lake. The red line represents the linear model for each location and the dashed grey line indicates the 1:1 line, provided for visual reference.

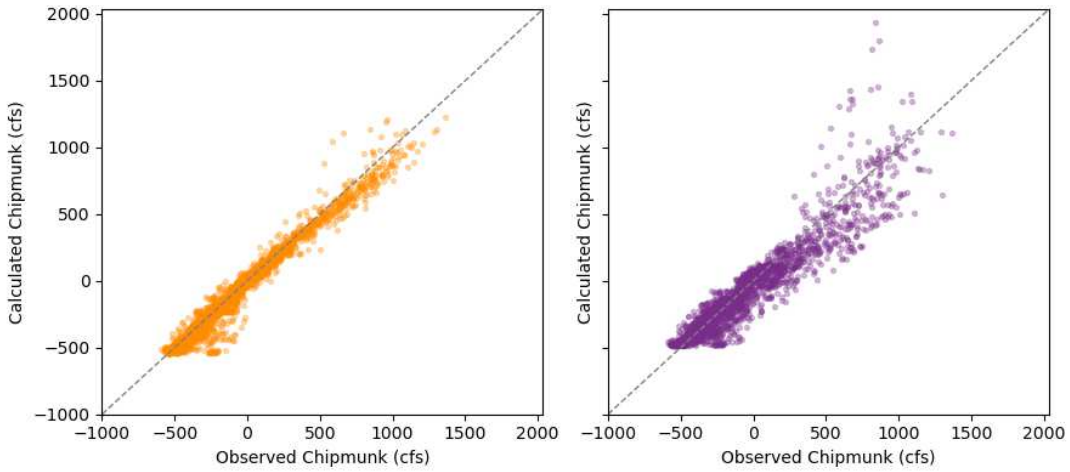
Chipmunk Lane Water Balance Validation

I used a simple water balance equation to estimate Chipmunk flow for assessment of the forecast system. Assuming that the volume of the Grand Lake/Shadow Mountain Reservoir

complex is constant (due to the 1 foot principle of Senate Document 80), I can estimate average daily Chipmunk flow as the sum of inflows to Grand Lake (East Inlet and North Inlet) minus the delivery to the Front Range via the Adams Tunnel. To assess this assumption, I compared this simple water balance estimate of Chipmunk Lane Flow with observed Chipmunk Lane Flow to validate this method.

Estimations of North Fork, North Inlet, and East Inlet flow were evaluated using Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970) and Kling-Gupta Efficiency (KGE) (Gupta & Kling, 2011) in addition to standard performance metrics like RMSE, MAE, and bias. Both NSE and KGE account for correlation, variation, and bias and are typically used in hydrological model assessment. Chipmunk Lane flow estimates were assessed using only NSE because the channel's bidirectional flow produces a near-zero observed mean, which destabilizes the KGE bias ratio term.

I validated the simplistic water balance equation to calculate flow at Chimpunk Lane by using historical discharge and delivery data, estimating flow and comparing the estimated flow with the observed (Appendix B, Figure 2). The relationship had a NSE of 0.963, RMSE of 72.2 cfs, and a MAE of 46.5 cfs. The relationship remains strong, even when using the Baker's Gulch proxy for flow at North Inlet and East Inlet with a NSE of 0.873, RMSE of 134.2 cfs, and a MAE of 88.9. Both the version that uses observed data and the version using a proxy to estimate flow had a slightly negative bias, indicating systematic underprediction of flow in Chipmunk. While using this proxy method to estimate flow introduces some level of uncertainty and has a structured bias, it retains a strong positive relationship, and provided this variable is not of high importance in the model it should not negatively impact the model substantially.



Appendix Figure 2. Comparison of observed and proxy-derived estimates of Chipmunk Lane Flow. The left panel calculates Chipmunk flow as observed East Inlet + observed North Inlet – observed Adams Tunnel delivery, the right panel calculates the flow using estimated East Inlet + estimated North Inlet – observed Adams Tunnel delivery using the linear relationship developed from the Baker's Gulch site.

APPENDIX C: EMBRACE Checklist

Environmental Machine-learning, Baseline Reporting, And Comprehensive Evaluation

The EMBRACE Checklist

Version 1.0. Last Updated Date: September, 2024.

The main purpose of EMBRACE Checklist is to facilitate the reporting of basic data and methods used for supervised machine learning (ML) research in environmental science and engineering (ESE). It also aims to provide a comprehensive evaluation of important preparation and analysis steps. We warmly encourage researchers to include the checklist as a reference, and if interested list it as supplementary materials in their publications.

Detailed instructions and additional resources are available open access for download from the author's GitHub: https://github.com/starfriend10/EMBRACE

- Future updates and new resources will be available in the above GitHub repository, which is also served as a community-owned platform to share data and codes, refine forms/checklists, discuss ML-ESE topics, promote research outcomes, and continuously identify/avoid pitfalls and follow good practices.
- We encourage you to direct share your checklists, but you can also save it as a read-only document. See instructions for detailed steps.
- Please cite the Viewpoint when using the checklist. If you have any questions or suggestions, please contact Dr. Jun-Jie Zhu at Princeton University (junjiez@princeton.edu or ranmuweijie@gmail.com).

Project Overview					
Project Title	DATA-DRIVEN FORECAST SYSTEM AS AN OPERATIONAL DECISION TOOL IN A MANAGED RESERVOIR				
Contributing Authors	Bethel Steele, Matt Ross				
Date	June 30, 2026	Completed by	Bethel Steele		
Contact Email	b.steele@colostate.edu	DOI (if published)			
Domain Category	Water Quality and Treatment	Or Specify:			
Learning Type	Regression ◉	Classification ○	Regression+ Classification ○		
Prediction Type	Deterministic ○	Probabilistic ◉	Deterministic+ Probabilistic ○		
Other Information					

I. Problem Formulation							
Project Objective(s)	Provide a 7-day forecast of water temperature at two depth intervals at a managed reservoir.						
Feasibility Assessment	Data Availability	Model Accessibility	Computation Resources	Knowledge Preparation	Time Availability	Financial Availability	Risk Tolerance
Levels	Medium	Low	Low	Low	Low	Low	Medium

II. Data Collection								
Data Sources	☑	Time-series Monitoring	☑	Field Campaigns	☑	Government Database	☐	Scientific Literature
	☐	Laboratory Experiments	☐	Simulation Outputs	☐	Others. Specify:		
Source Types	☐	Ordinary	☑	Time-series	☐	Literature-based	☐	Others. Specify:
Data Types	☑	Continuous Numerical	☐	Categorical	☐	Textual	☐	Visual
	☐	Graphical	☐	Auditory	☐	Others. Specify:		
Data Ethics	☑	No special ethical considerations are required, or followed the necessary requirements when needed						
	☑	No special permissions are required, or obtained the necessary permissions when needed.						
	☑	Provided appropriate credit(s) to the data source(s)						
Data Availability	☐	Data are publicly available through a PURL or DOI	☑	Data are available as requested				
Raw Data Quantity	Raw Sample Size	Raw Variable Size	Raw Total Data volume	Raw Sample-Variable Ratio				
	990	31		30:1				
Raw Data Quality	The raw data had an internal QA/QC before data retrieval?		Yes ◉	No ○				
	The raw data went through a peer-review or similar process?		Yes ○	No ◉				

III. Data Preprocessing										
Data Cleaning	☐	Abnormal Data Management	☑	Statistical Outlier Management	☑	Technical Outlier Management				
	☐	Data Correction due to Technical Limitations	☐	Others. Specify:						
Data Enrichment	☐	Missing Data Management	☐	Data Augmentation	☑	Data Aggregation	☐	Data Balancing		
	Upper sample ratio among groups:		☐	Others. Specify:						
Feature Engineering	☑	Initial Variable(s) Exclusion	☑	Feature Importance Analysis	☐	Feature Extraction				
	☐	Categorical Feature Transformation	☐	Dummy Feature Removal						
	☐	Ordinal Encoding	Encoded variable(s):							
	☑	Feature Selection	☐	Feature Scaling (for X)	☑	Feature Scaling (for Y)	☐	Circular Feature Scaling		
	☐	New Feature Development	☑	Ordinary Mathematical Transformation(s)	☑	Time-Dependent Feature Transformation(s)				
			☐	Spatial Feature Transformation(s)	☐	Others. Specify:				
	Sparse Dataset?	Yes ○	No ◉	Preprocessed to increase information density?	Yes ○	No ◉				
	☐	Normality Test (before/after transformation)	☐	Others. Specify:						
Data Splitting	Framework	○	Training-Testing	○	(Pre-)Training-Validation-Testing	◉	Cross-Validation-Testing			
	Splitting Method	☐	Simple Random	☐	Grouped Random	☐	Simple Block	☑	Sub-Group R/B	
	☑	Ensured training dataset covered most possible scenarios (e.g., different seasons)								
Final Data Quantity	Testing Ratio	Preprocessed Sample Size	Preprocessed Feature Size	Sample-Feature Ratio	Training SFR					
	0.125	887	21	40						

IV. Methods Selection and Model Initialization									
Methods Selection	☑	Selected supervised learning methods to be tested based on case needs (e.g., data, computation resources, etc.)?							
	Methods Studied: N ML and M Statistical			N = 1	M = 0				
	ML Family	☐	Tree-based	☑	Neural Network	☐	Others		
	ML Type	☐	Shallow ML	☑	Deep ML	Final Optimal Method(s):			
Model Uncertainty	☐	Examined Random Seeds	☐	Replicated Modeling Process					
	☐	Assessed probabilistic prediction results	☐	Others. Specify:					

V. Model Evaluation and Model Optimization									
Model Evaluation	☑	Set a goal model performance before evaluation							
	Specify Evaluation Metrics: MAE, CRPS, CRPSS against baseline of persistence								
	☑	Studied under-fitting issue	☑	Studied over-fitting issue					
	☑	Final evaluation was based on the testing data	☐	Weighted metrics were used for imbalanced data					
	☐	ML significantly outperformed statistical models	☐	Statistical perform similar or even better than ML models					
Hyper-parameter Optimization	HPO Method	☐	No HPO	☐	Trial-and-Error/Experience	☐	Grid/Random Search (similar)	☑	Metaheuristics
		☐	Other HPO. Specify:						
	Number of methods tuned by HPO: 30				Number of hyperparameters searched: 4				
	Number of values in each hyperparameter:	◉	Continuous Search	○	Discrete Search. Specify the number:				
	CV Method	○	No CV	○	k-fold CV	○	Leave-one-out CV	○	Stratified CV
		◉	Pre-defined or other CV. Specify:	used fold 3 as validation					

VI. Model Explanation				
Model Interpret-ability	☑	Reported model interpretability as intrinsic property	☑	Described tradeoff between interpretability and complexity
	☐	Explained limitations of low model interpretability	☐	Compared the methods with different interpretabilities
Model Explain-ability	☑	Reported the explanation method(s) used for FIA	Explanation method(s):	SHAP
	☐	Reminded high accuracy might not be trustable explanation	☑	Described application of explanation on the optimal model
	Reflected on understanding that explanations:	☑	couldn't cover all mechanisms	☑ might be incorrect
Causality	☑	Reflected on higher accuracy not implying logical causal relationships		
	☑	Reported understanding that explanation results do not indicate causality		
	☑	Described alignment of explanations with environmental domain knowledge for causality		
	☐	Identified illogical or counterintuitive parts based on environmental domain knowledge		
	☐	Described study of illogical or counterintuitive parts based on environmental domain knowledge		

VII. Data Leakage Management						
General/ Data Source	Verified no:	☑ response variable Y leakage	☑ future-to-current data leakage	☑ current-to-current data leakage		
	Confirmed that no overlapped data between training and testing	Yes ◉	No ○			
	Data splitting method for literature-source data:					
Data Enrichment	Verified that missing data replacement utilized only seen dataset	N/A ◉	Yes ○	No ○		
	Verified that missing data interpolation utilized neighbor data in same subset	N/A ◉	Yes ○	No ○		
	Verified that model-based imputation utilized only seen data	N/A ◉	Yes ○	No ○		
Feature Engineering	Confirmed that feature engineering utilized only seen dataset	N/A ○	Yes ◉	No ○		
	Confirmed that feature scaling utilized only seen dataset	N/A ○	Yes ◉	No ○		
	Confirmed that data splitting method for time-dependent data:	annual split				
CV Loop	Verified that missing data replacement utilized only seen data in each split	N/A ◉	Yes ○	No ○		
	Verified that missing data interpolation with neighbor data in each split	N/A ◉	Yes ○	No ○		
	Verified that feature engineering utilized only seen data in each split	N/A ○	Yes ◉	No ○		
	Verified that feature scaling utilized only seen data in each split	N/A ○	Yes ◉	No ○		
	Data splitting method for literature-source data in CV loop:					
	Data splitting method for time-dependent data in CV loop:	anual split				
Others	☐	Self-specify:				

VIII. Additional Items		
☐	PURL or DOI are publicly available. Data:	Code:
☐	Email or URL are available to request. Data:	Code:
☐	Self-specified Item:	

APPENDIX D: Machine Learning And Interpretation

	Data Source	Processing	Model training	1-day-ahead assessment	7-day assessment
Weather	GEFS control forecast + 30-member ensemble		☑ Control forecast only	☑	☑
Thermal Inertia	Observed Temperature		☑	☑	☑
2-day "memory" via autoregressive temperature	Model Output	Combined with observed during rollout of forecast			☑
Inflow	Observed inflow		☑		
	Baker's Gulch deterministic forecast	Per-inflow linear relationship		☑	☑
Operations	Observed operational pump operation and Adams Tunnel Deliveries		☑	☑	☑

Appendix Figure 3. Data source overview for the training and evaluation of the model.

Cross-Validation Structure

Data for training, validation, and test sets were delineated by calendar year. Because there is a data gap over the winter (where the buoy is not deployed) and we can consider the winter ice cover a "reset" of the system regarding temperature, we consider each year of data to be independent. No previous-year lagged data were included as features. 2025 was selected as the test year (n = 103), assuming that 2025 would be a reasonable performance indicator for any future year. Data from 2014-2024 (n = 887) were then randomly assigned to one of five cross validation folds while attempting to balance the number of samples in each fold (Appendix D, Table 1). Very little data was available for the 2020 season due to a late deployment of the buoy, but any available data was included in the training of the model.

Appendix Table 1. Five-fold cross validation sets with years and number of rows assigned to each fold.

Fold	Years	Number of rows of data
1	2018, 2024	171
2	2017, 2023	187
3	2016, 2021	184
4	2014, 2022	180
5	2015, 2019, 2020	165

Feature Generation

Given the physical mechanisms for the control of temperature within this system described in 2.4, we generated a lagged feature set of temperature at the two depth horizons (autoregressive) at a 1- and 2-day lag and up to a 5-day lag from initialization date for all other features (meteorology, streamflow, operational variables). One day ahead forecasted meteorology, streamflow, and operations were also provided to the model as forcing. This resulted in a preliminary feature set of 39 features.

Water Temperature 0-5m: lag 2, lag 1 (n =2)
Water Temperature 0-1m: lag 2, lag 1 (n =2)
North Fork: lag 5-lag 1, d1 (n=6)
GEFS solar radiation: lag 5-lag 1, d1 (n=6)
GEFS Temp: lag 5-lag 1, d1 (n=6)
GEFS Wind: lag 5-lag 1, d1 (n =6)
Pump: lag 5-lag 1 (n=5)
Chipmunk: lag 5-lag 1, d1 (n=6)

For the purposes of feature generation, we consider day 1 to be the firts date of the forecast and lag 1 is the equivalent of the initialization date (e.g, Lag 5 – lag 4 – lag 3 – lag 2 – lag 1 – day 1). For a forecast initialized on June 29th, lag 5 would be data from June 25th and the forecast would be valid for June 30th (day 1).

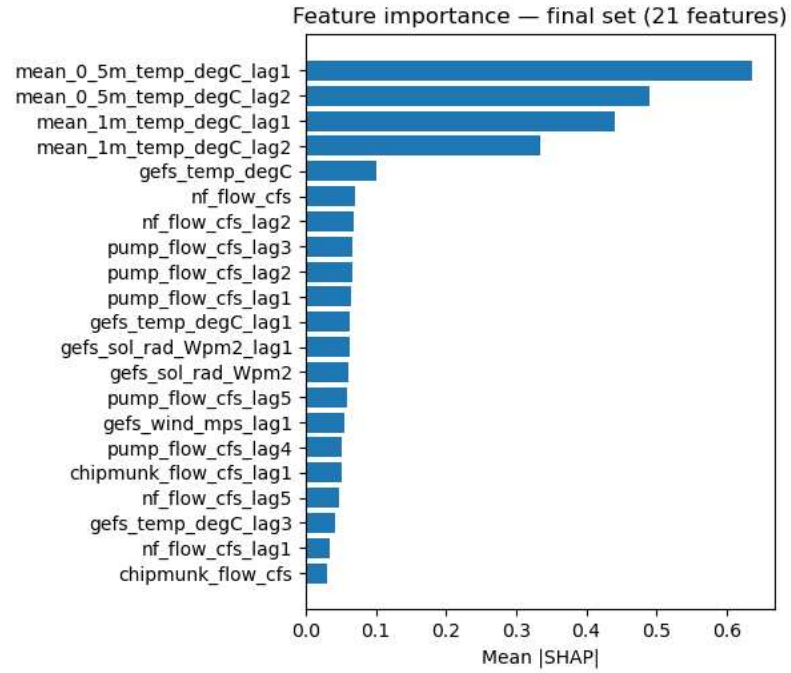
Feature Reduction

Following the feature generation step, I completed a feature reduction exercise where I eliminated features based on SHAP values from a neural network without tuning the

hyperparameters (using a 0.1 dropout layer followed by 2 layers of 8 nodes each, with a 100 epoch persistence, and a 0.01 learning rate). Feature importance was aggregated across all five of the training and validation folds as the average absolute SHAP value per feature. Any features that were in the lower 30% of importance were eliminated, unless we considered them important to include through domain knowledge. For instance, I kept the forecasted Chipmunk Lane flow and the 1-day lag of North Fork flow in the feature set even though they were below the 30% threshold. Following the feature reduction step, there were 21 features in the dataset.

Water Temperature 0-5m: lag 2, lag 1
Water Temperature 0-1m: lag 2, lag 1
North Fork: lag 5, lag 2, lag 1**, d1
GEFS solar radiation: lag 1, d1
GEFS Temp: lag 3, lag 1, d1
GEFS Wind: lag 1
Pump: lag 5, lag 4, lag 3, lag 2, lag 1
Chipmunk: lag 1, d1**

** indicates features that were retained in the model despite falling below the threshold for inclusion.



Appendix Figure 4. Mean SHAP feature importance across the 5-fold validation set for the retained features.

Hyperparameter Tuning Results

Hyperparameter tuning was completed on fold 3 of the CV scheme, assuming it was representative of all other folds. Given our desire to constrain overtraining, we ranked the models by composite loss defined as validation loss minus training loss. The top 10 hyperparameter runs are listed in Appendix Table 2. The hyperparameters from the model in rank 3 was chosen due to its simplistic architecture (2 layers of 8 nodes each) and comparable composite loss to higher-ranked hyperparameters.

*Appendix Table 2. Top 10 hyperparameter runs ordered by composite loss (validation loss – training loss). Selected model indicated by "**".*

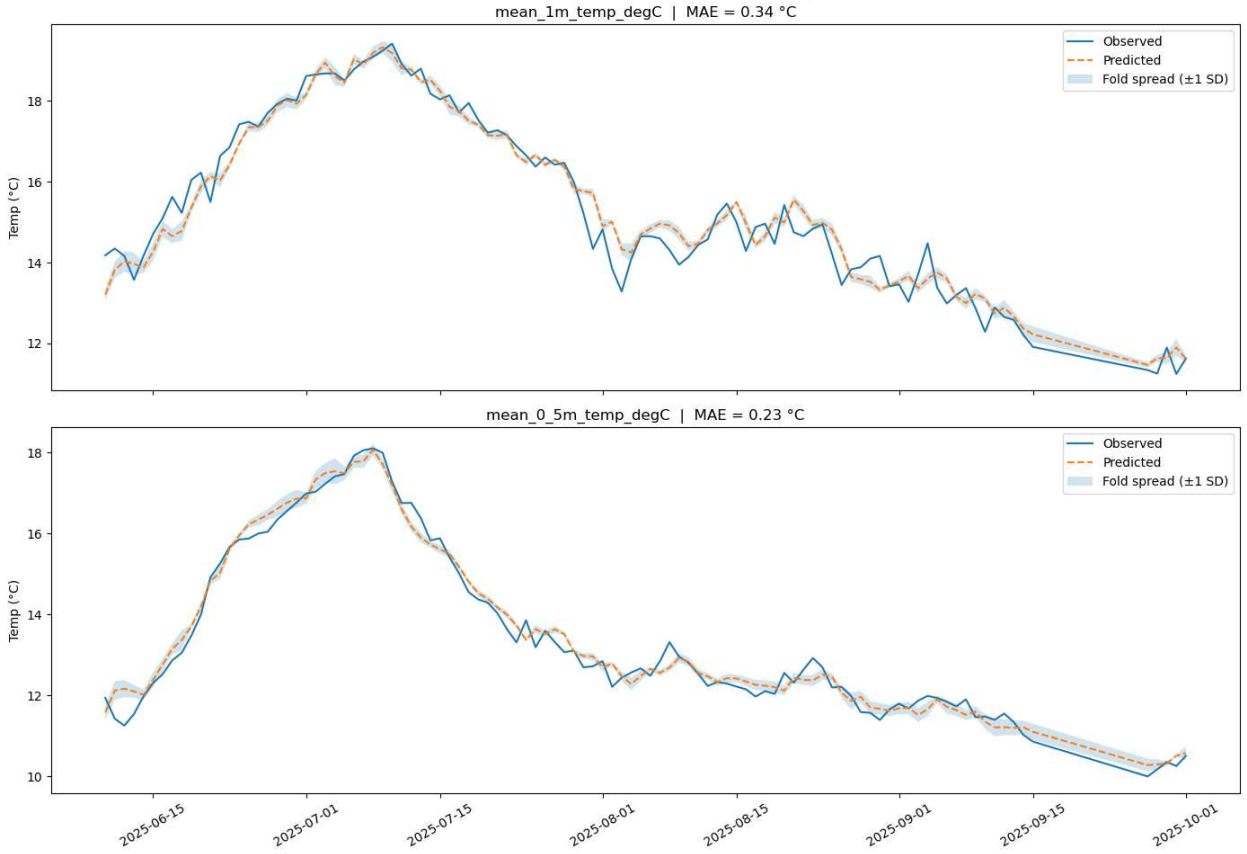
Rank	Composite Loss	Units	Dropout	L2 regularization	Learning rate
1	0.1605	8, 16	0.1	4.13e-04	9.75e-04
2	0.1610	24,16	0.1	1.13e-04	4.60e-03
**3 **	**0.1621**	**8, 8**	**0.1**	**4.80e-04**	**9.68e-04**
4	0.1634	8, 16	0.1	3.29e-04	8.40e-04
5	0.1652	8, 8	0.1	2.04e-04	6.59e-04
6	0.1679	24, 24	0.1	1.00e-04	6.66e-03
7	0.1711	24, 8	0.1	1.00e-04	1.04e-03
8	0.1732	24, 8	0.1	1.00e-04	1.00e-02
9	0.1736	24, 24	0.1	5.30e-04	1.00e-02
10	0.1744	8, 8	0.1	2.48e-04	2.45e-03

Cross-Validation Performance

We assessed cross validation performance on the test set by reviewing the five contributing models. All five models performed similarly and the ensemble outperformed persistence for both depths (Appendix Table 3). The 2025 timeseries did not display consistent one-day lagged prediction (Appendix Figure 5) where the predicted line would be consistently lagged one day behind the observed, further indicating that the model was not simply predicting "yesterday is today."

Appendix Table 3. Per-fold and ensemble performance of the 5-fold cross validation neural network for near-surface and integrated depths at a one-day-ahead horizon for test set (2025). Ensemble performance meets or exceeds performance of individual folds.

Model	Near-surface (0-1m) temperature			Integrated depth (0-5m) temperature		
	MAE (°C)	RMSE (°C)	NSE	MAE (°C)	RMSE (°C)	NSE
CV Fold 1	0.353	0.452	0.956	0.260	0.321	0.977
CV Fold 2	0.356	0.453	0.955	0.276	0.337	0.974
CV Fold 3	0.397	0.521	0.941	0.233	0.291	0.981
CV Fold 4	0.348	0.463	0.953	0.229	0.295	0.980
CV Fold 5	0.345	0.451	0.956	0.254	0.329	0.976
Ensemble	0.345	0.449	0.956	0.228	0.282	0.982
Persistence	0.382		0.951	0.251		0.979



Appendix Figure 5. Timeseries of the one-day-ahead ensemble displayed as the mean temperature (orange dashed line), and the fold spread (+/- 1 standard deviation of the mean) with the observed temperature (solid blue line).

7-day Performance Metric Definitions

The seven-day forecasts are assessed using the Continuous Ranked Probability Score (CRPS) which is a similar statistic in practice to MAE at the one-day-ahead horizon and the

Continuous Ranked Probability Skill Score (CRPSS) which compares the skill of the model relative to a baseline. CRPS calculate error across a probabilistic forecast integrating the range of predictions against the observed value. For our purposes, we use persistence (assuming today's value is tomorrow's) as the baseline for comparison in the CRPSS calculation.

$$CRPS = \frac{1}{M}\sum_{i=1}^{M}|\widehat{y_i} - y_{obs}| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|\widehat{y_i} - \widehat{y_J}|$$

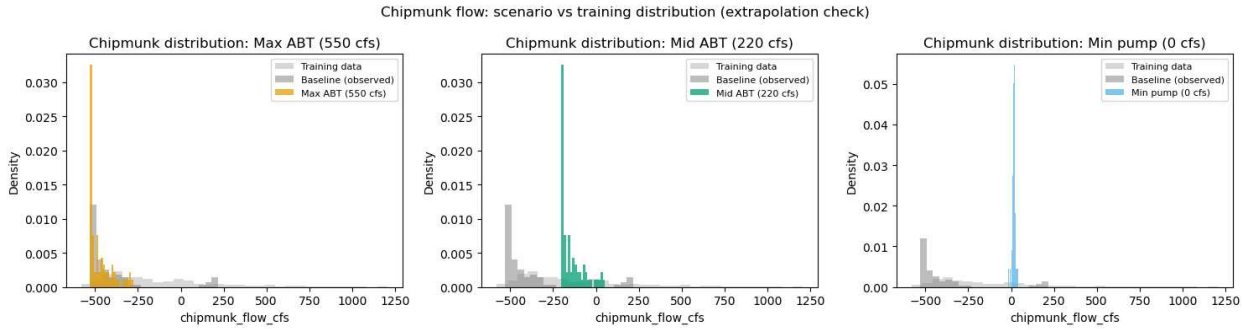
Where *M* is the total number of ensemble members, *i* is an individual forecast ensemble member, and *j* is every other ensemble member, the first half represents the average MAE across all forecasts and the second half represents the average spread of the forecasts, where a narrow spread is penalized more than a spread that encompasses the observed values.

$$CRPSS = 1 - (\frac{CRPS_{forecast}}{CRPS_{persistence}})$$

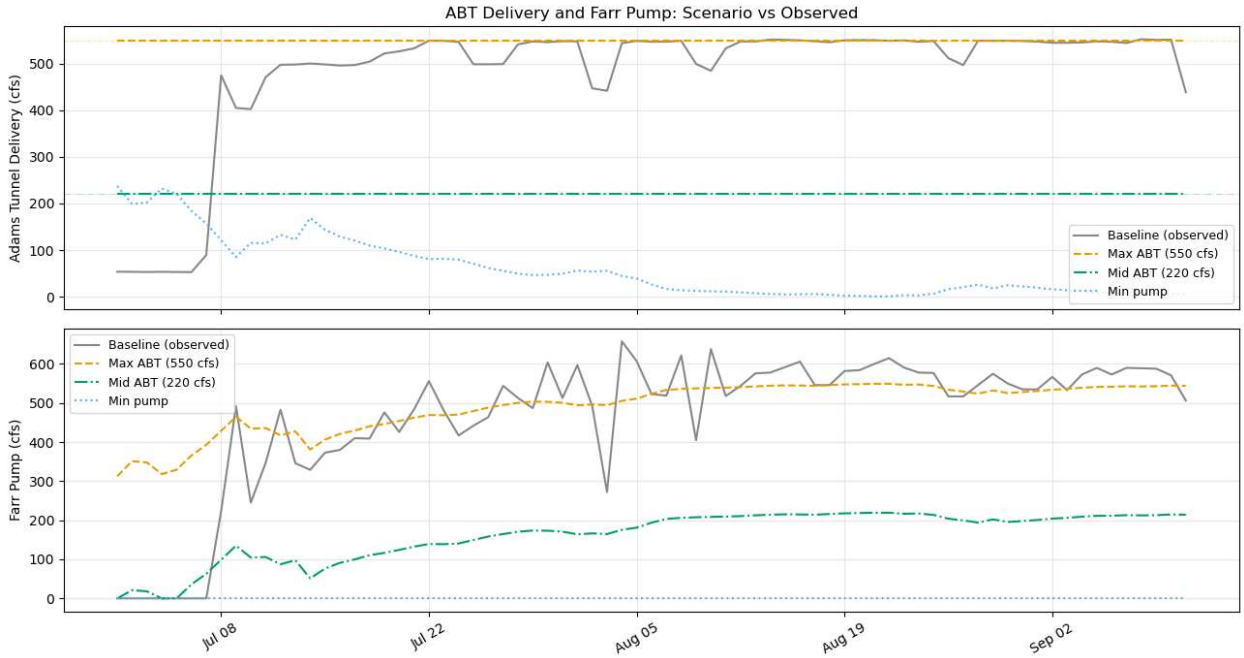
In this context, a lower CRPS indicates better performance and a higher CRPSS indicates a more skillful forecast relative to the baseline. A CRPS of zero and a CRPSS of 1 is a perfect forecast. A CRPSS of 0 indicates that the model has no skill relative to the persistence model, a negative value indicates that the forecast is less skillful than the baseline.

APPENDIX E: Hindcast Scenario Analysis

I assessed extrapolation risk in the hindcast scenario by examining the histogram of Chipmunk flow for the computed scenarios against the data used to train the model (Appendix Figure 6). This inspection confirms that the distribution falls within the historical range and could predict plausible results. Appendix Figure 7 shows the Adams Tunnel deliveries as well as the actual pump operations required to maintain the deliveries for each scenario.



Appendix Figure 6. Proportional distribution of Chimpunk Flow in the training dataset (light grey bars), the baseline (dark grey bars), and the three operational scenarios (colored bars).



Appendix Figure 7. Adams Tunnel deliveries (top) and Farr Pump operations (bottom) as observed and (grey line) and in the three operational scenarios.