A Comparison of Neural Network Models for Water Quality Forecasting

A Preprint

Marcus Lapeyrolerie
Department of Environmental Science, Policy, and Management
University of California, Berkeley
Berkeley, California
mlapeyro@berkeley.edu

Carl Boettiger
Department of Environmental Science, Policy, and Management
University of California, Berkeley
Berkeley, California
cboettig@berkeley.edu (corresponding author)

June 13, 2024

Abstract

1.
2.
3.
4.

***K*eywords** Artificial Intelligence · Forecasting · Machine Learning · Time Series · Limnology

1 Introduction

The ability to accurately forecast water quality variables has become increasingly important in the era of global change. Freshwater ecosystems have been disproportionately impacted by anthropogenic activities, a trend that is expected to continue in the future [**?**]. Since water quality variables, like water temperature, dissolved oxygen and chlorophyll concentration, influence many biological and physical processes in freshwater ecosystems [**?**, **?**, **?**], preemptive water quality forecasts could allow managers to evade situations with

far-reaching negative consequences [?]. Limnologists have historically used statistical, process-based and machine learning models to forecast water quality metrics [?]; but, in recent years, researchers have shown that machine learning models are particularly well suited to take advantage of recent advancements in computer science and the rising availability of water quality data [?, ?]. A shortcoming of past limnological forecasting studies that support the use of machine learning is that they tend to focus on a selection of sites that are not representative of freshwater systems across a broad geographic scale. This paper has the primary aim of making a comprehensive comparison of state-of-the-art machine learning methods by evaluating forecasts at 34 different sites across North America at different times of the year.

The time series that are used in this study are taken from the National Ecological Observatory Network's (NEON) Ecological Forecasting Challenge, an open challenge where teams can submit forecasts to data that is collected and made publicly accessible by NEON [?]. A common finding across the challenge is that a day of year historical mean model (also referred to as the climatology model) commonly produces top scoring forecasts [?, ?]. For instance, in a model comparison that examined the forecasts for phenological time series, [?] found that the climatology model outperformed all except one of the submitted models; and, the best performing model only marginally outperformed the climatology model. Thus, a primary consideration in this paper is comparing the performance of the machine learning models against the historical null model.

There are many promising neural network architectures that have not yet been evaluated for water quality forecasting. Most applications of neural networks to limnological time series have focused on the long-short-term-memory (LSTM) network, a model that was published in 1997 [?]. Over the nearly 30 years since LSTM's were introduced, researchers in computer science have built upon these earlier neural network architectures, introducing new models that achieve state of the art performance [?, ?, ?, ?]. Many recent studies that use neural networks for water quality forecasting have not explored more contemporary neural network architectures. We will address this gap in the literature by comparing 8 neural network models including LSTM's as well as more recently published neural network models.

2 Materials and Methods

We evaluated the forecasting performance of 12 different forecasting models on water temperature, dissolved oxygen and chlorophyll-A time series recorded by in-situ sensors at 34 freshwater sites across North America. These sites consist of Lakes, Non-wadeable Rivers and Wadeable Streams, subtypes classified by NEON. There is variability across sites in which target variables were observed. Additionally, maintenance issues led to gaps in the data which also varied across locations. Provided the lack of time series and large gaps at certain sites, we evaluated the forecast performance for water temperature, dissolved oxygen and chlorophyll-A at 33, 32 and 9 sites, respectively.

The imputation of missing data proved to be critical to the performance of the ML models. After experimenting with data filling methods that resulted in poor model performance, we developed an imputation method inspired by the climatology model. If the gap size was less than 5 days, then the gap was filled using a Gaussian Process Filter. For gaps over 5 days, missing data was estimated using the daily historical mean when this statistic is available; and, when there are no data collected for a day of the year, either the monthly median, quarterly median, previous observation or global median was used in this order of preference. The intuition behind this method for missing data impuation is that the neural network models will be biased towards the model that you use to fill in missing data. Since it has been established that a climatology model is a top performing model throughout the NEON forecasting challenge, we reckoned that biasing the neural network models towards the climatology model will likely bias the models towards making better predictions.

We compared the performance of the neural network models to the climatology and naive persistence model. The climatology model generates forecasts by finding the daily mean and standard deviation and draws samples from a Gaussian Distribution with these parameters. The naive persistence model finds the last observed value of the target variable and predicts this value for each day in the forecast horizon. For each model, we computed the Continuous Ranked Probability Score (CRPS) and the Root Mean Square Error (RMSE). To gauge how models performed in relation to the climatology model, we computed the Continuous Ranked Probability Skill Score (CRPSS) which we defined as

$$CRPSS_{model} = 1 - \frac{CRPS_{model}}{CRPS_{clim}}. \tag{1}$$

Similarly, using the naive persistence model as a reference, we computed the RMSE Skill Score,

$$RMSE - SS_{model} = 1 - \frac{RMSE_{model}}{RMSE_{naive}}. \tag{2}$$

If a target model outperforms the reference model, then the target model will have a positive skill score. If the target model performs worse than the reference model, the skill score will be negative.

2.1 Method Group 1: Statistical Models (Theta)

To compare the performance of neural network models to statistical models, we evaluated forecasts generated by the Theta model. Provided a seasonally adjusted univariate time series, the Theta model creates forecasts by modifying the second differences of the data [**?**]. The magnitude of local curvature modifications is given by the Theta coefficient,

$$\nabla^2 Z_t(\theta) = \theta \nabla^2 X_t, \tag{3}$$

where ∇ is the difference operator, Z_t is defined as a theta line and X_t is the original data. When $\theta < 1$, the second differences are reduced from the data, yielding a Theta line that amplifies the long term trends

of the data. For instance, if $\theta = 0$, the theta line will be a linear regression of the data. And if $\theta > 1$, the short term behavior of the data will be magnified in the Theta line. The Theta model generates a forecast by extrapolating the linear combination of two or more Theta lines. Although the Theta model is relatively simple, it has performed remarkably well in prominent forecasting competitions [**?**]. In the time since the Theta model was originally presented in 2000, new variations of the Theta model have been presented which have been shown to outperform the original Theta model [**?**]. Throughout this study, we use the StatsForecast AutoTheta model which selects the best performing model from a selection of variations of the Theta model.

2.2 Method Group 2: Neural Network Models

All the machine learning algorithms that are used in this study are based on neural network architectures. Neural networks have the property of being universal function approximators, so in theory, it is possible that all of these models could exactly approximate the data generating process [**?**]. Yet, it is often true that neural networks greatly underperform their function approximation capabilities in practice [**?**, **?**]. This performance gap can be due to a variety of reasons including overfitting and insufficient hyperparameter tuning [**?**]. The discrepancy between theory and practice in function approximation with neural networks motivates research on how machine learning models perform in specific domains.

The 8 machine learning models that we compare take a variety of approaches with the design of their neural networks. We will not go into detail on how these different models work, but for those interested to learn more, we list the models and their references in table **??**. For readers who do not specialize in ML, an important concept to understand is that the neural network models learn directly from the data and are not instilled with domain knowledge. This non-mechanisitic basis is at once very powerful as it does not restrict the models with misleading assumptions, but neural networks are also limiting in that they are not readily interpretable and require more data than knowledge-guided methods [**?**]. For the management of critical resources like water, the lack of interpretability could be a deterrent to the adoption of NN methods, and there may not be enough data for some water systems to accommodate ML approaches [**?**].

The neural network models are configured to generate probabilistic forecasts. The neural network architectures that are used in this study map real-valued vectors to real-valued vectors, so they are not able to directly produce probabilistic forecasts. We work around this limitation by performing quantile regression whereby the neural networks are trained to output quantiles at each time step in the forecast window. Forecasts are then generated by drawing samples according to these quantiles.

For the 8 neural network models that we investigate, there are varying design choices made regarding the type of covariates that can be used. For this study, we employed 2 groups of models, one group used past covariates and the other used future covariates. For the models that accept past covariates (which included TCN, BlockRNN, NLinear, DLinear, Transformer and NBEATS), we used the other target variables recorded at that site as well as air temperature as covariates. For the models that only accept future covariates (which

included TFT and LSTM), we used the day of the year as the covariate. The historical time series were split into a training and a validation set, whereby the models were trained on time series from 2020 to 2023 and validated at 12 30-day non-overlapping intervals in 2023.

We provide fully reproducible code for fitting, scoring and visualizing the forecasts. All the machine learning forecasts were generated using the `darts` python library [**?**]. `darts` is similar to other libraries like `scikit-learn` in Python or `tidymodels` in R in that the library allows users to employ a variety of time series forecasting models without having to implement them.

3 Results

We examine the forecast skill of 8 neural network models (LSTM, BlockRNN, NBEATS, NLinear, DLinear, TFT, TCN and Transformer), 1 statistical model (AutoTheta) and 2 null models (naive persistence and climatology) on time series taken from the NEON Forecasting Challenge's Aquatics Theme. In addition to these individual models, we created an ensemble model that aggregates the forecasts taken from all the neural network models. Following recent work that has established that multi-model ensembles offer advantages over individual models for water temperature forecasting [**?**], we were inspired to investigate multi-model ensembles in this comparative study. For the neural network multi-model ensemble, all the NN models are represented equally, hence its name "Naive Ensemble".

In Table **??**, we present the mean CRPS and RMSE scores for the respective target variables. For dissolved oxygen (DO) and water temperature (WT), the models perform similarly: the neural network models outperform the AutoTheta, naive persistence and climatology model with few exceptions; and the naive ensemble model is the best performing model with respect to both CRPS and RMSE. Yet, with Chlorophyll-A, there are different patterns in model performance: while the neural network models still generally outperform the reference models in CRPS, the Naive Ensemble model is no longer the top performing model in either CRPS or RMSE; and, BlockRNN and the naive persistence model attain the best performance with respect to CRPS and RMSE, respectively.

By examining some of the individual forecasts (Figure **??**) used for the evaluations, it is possible to gain intuition for why the neural network models perform well across the target variables. The AutoTheta model produces forecasts that look similar to a linear regression based upon recently observed values. These relatively simple forecasts perform well for many evaluation intervals, but there a few cases when AutoTheta fails significantly, which negatively impacts the model's overall performance. For instance, after a peak of dissolved oxygen in the winter of 2024, the AutoTheta model forecasts that DO will continue to increase which is opposed to the trend that DO peaks in the winter and declines through the spring. Similarly with Chlorophyll-A, the AutoTheta model wrongly extrapolates that a spike in Chlorophyll-A will lead to even higher Chlorophyll-A concentrations instead of an immediate reversion to the non-bloom state.

Conversely, the neural network models are able to learn from historical trends. For instance, at the beginning of 2024, the neural network models have learned from the training data that DO peaks in the winter and declines through the spring, so the neural networks correctly predicts that there will be a decline in dissolved oxygen during this month. With Chlorophyll-A, however, we see that the neural network models fail to predict any of the spikes in concentration which originate from algae blooms; instead, the neural networks take a conservative approach, only predicting that the chla concentration will remain at non-bloom levels throughout the year.

These general patterns in model performance can also be seen in the skill score plots displayed in Figures **??**. The AutoTheta model tends to have longer tails, indicating that the AutoTheta forecasts are more likely to perform very well or very badly. Meanwhile, the neural network models generally have fewer forecasts that underperform relative to AutoTheta and less forecasts that are outlier outperformers, but the distributions of NN skill scores are translated towards outperforming. In these plots, we display scores according to water body type, and we did not see any significant differences in performance across these categories. This may be due to an underreporting of scores for lakes and non-wadeable rivers relative to the number of scores found for wadeable streams.

When examining the performance of the models within a forecast horizon, additional nuances emerge, confusing the perspective that neural network models are the best performing model class. In Figure **??**, the AutoTheta model is consistently the best performing model for short time horizons across target variables, but by the end of the 30-day horizon, the AutoTheta model is the worst perfoming model universally. Meanwhile the neural network models underperform AutoTheta during the early stages of the forecast window (generally, when $t < 5$), but their forecast skill declines less rapidly than the skill score of AutoTheta. So while, the neural networks generally outperform AutoTheta and the null models according to coarse scoring metrics like mean CRPS and RMSE, AutoTheta is the best performer in short horizons.

4 Discussion

Over the last decade, machine learning has gained significant prominence after a spate of high profile successes based on neural network models, some of these successes include beating the master champion of Go, accurately predicting protein structures and the advent of large language models like ChatGPT [**?**, **?**, **?**]. During this time, researchers in computer science have developed new neural network models for time series forecasting with promises of significant accuracy improvements over previous methods [**?**]. Simultaneously during the last 10 years, there has been a growing focus in limnology towards using neural networks for water quality forecasting, but this focus has primarily centered on older neural network architectures and there has been little exploration of newer methods. This study aims to address this research gap by performing a

(a) Dissolved Oxygen ($\mathrm{mg\,L^{-1}}$)

Model	Mean CRPS	Mean RMSE
Climatology	0.62	0.92
Naive Persistence		1.55
AutoTheta	0.67	1.02
BlockRNN	0.52	0.80
DLinear	0.52	0.81
NBEATS	0.51	0.80
NLinear	0.52	0.80
NaiveEnsemble	**0.47**	**0.74**
RNN	0.60	0.92
TCN	0.62	0.98
TFT	0.52	0.80
Transformer	0.51	0.80

(b) Water Temperature (°C)

Model	Mean CRPS	Mean RMSE
Climatology	1.46	2.19
Naive Persistence		6.05
AutoTheta	1.55	2.35
BlockRNN	1.45	2.26
DLinear	1.35	2.08
NBEATS	1.36	2.10
NLinear	1.33	2.06
NaiveEnsemble	**1.18**	**1.80**
RNN	1.32	2.03
TCN	1.89	3.02
TFT	1.19	1.86
Transformer	1.32	2.05

(c) Chlorophyll-a ($\mathrm{mg\,L^{-1}}$)

Model	Mean CRPS	Mean RMSE
Climatology	4.83	7.50
Naive Persistence		**4.86**
AutoTheta	4.20	6.31
BlockRNN	**3.14**	5.16
DLinear	3.47	5.48
NBEATS	3.78	5.95
NLinear	3.67	5.63
NaiveEnsemble	3.44	5.50
RNN	4.15	6.40
TCN	3.72	5.77
TFT	3.67	5.87
Transformer	4.02	6.35

Table 1: Mean CRPS and RMSE for dissolved oxygen, water temperature and chlorophyll-a forecasts. The neural network-based models generally outperform the statistical benchmark model, AutoTheta, as well as the null models, naive persistence and climatology, across the target variables. For dissolved oxygen and water temperature, the NN ensemble model is the best performing model with respect to CRPS and RMSE. With chlorophyll-a, BlockRNN attains the best CRPS score, and naive persistence attains the best RMSE.

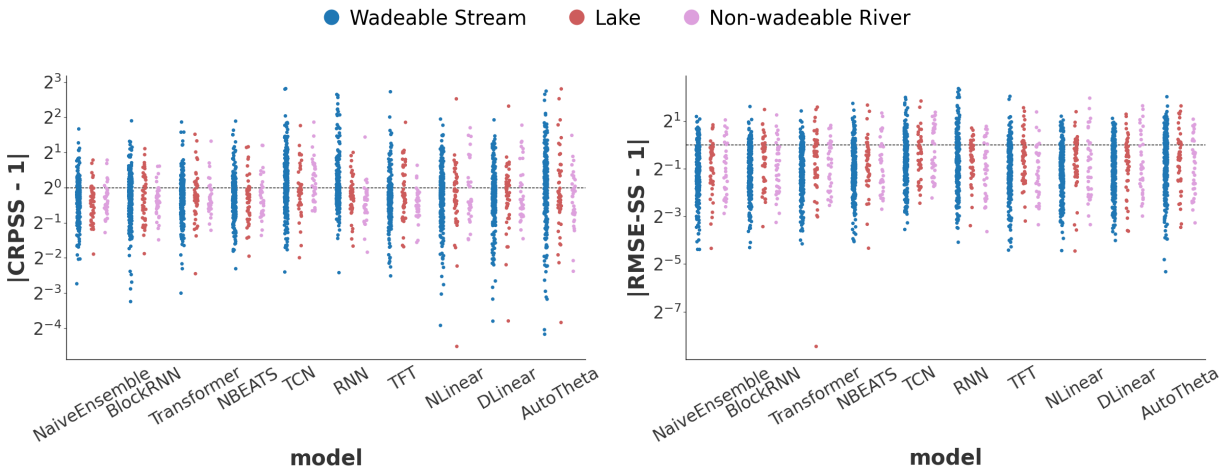

Figure 1: Skill score distributions for dissolved oxygen. The CRPSS plot measures the CRPS skill relative to the climatology model. The RMSE-SS plot measures the RMSE skill relative to the naive persistence model. A guideline is plotted at the threshold for underperformance: a score above this line denotes that a forecast is underperforming the reference model; a score below this line denotes that a forecast is outperforming the reference model. The AutoTheta model exhibits longer tails relative to most of the neural network models, indicating that AutoTheta has a larger amount of forecasts that do either very well or very poorly. The neural network models tend to have shorter tails in the underperforming region.

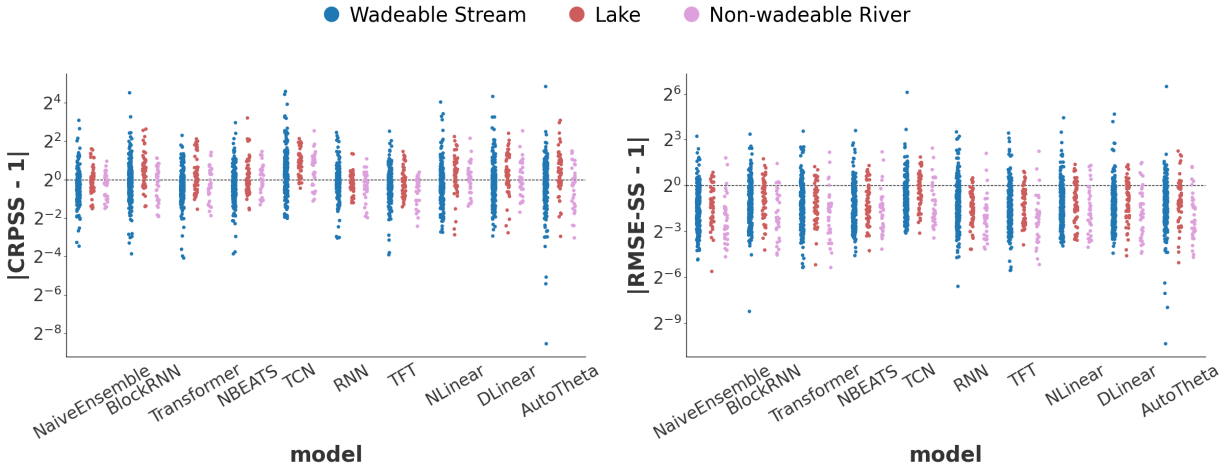

Figure 2: Skill score distributions for water temperature. AutoTheta does not have the long tails as seen in the skill score distributions for dissolved oxygen and chlorophyll-a, instead the neural network models have heavier tails in the outperforming region. This indicates that the neural network models are producing high accuracy forecasts more frequently than the AutoTheta model.

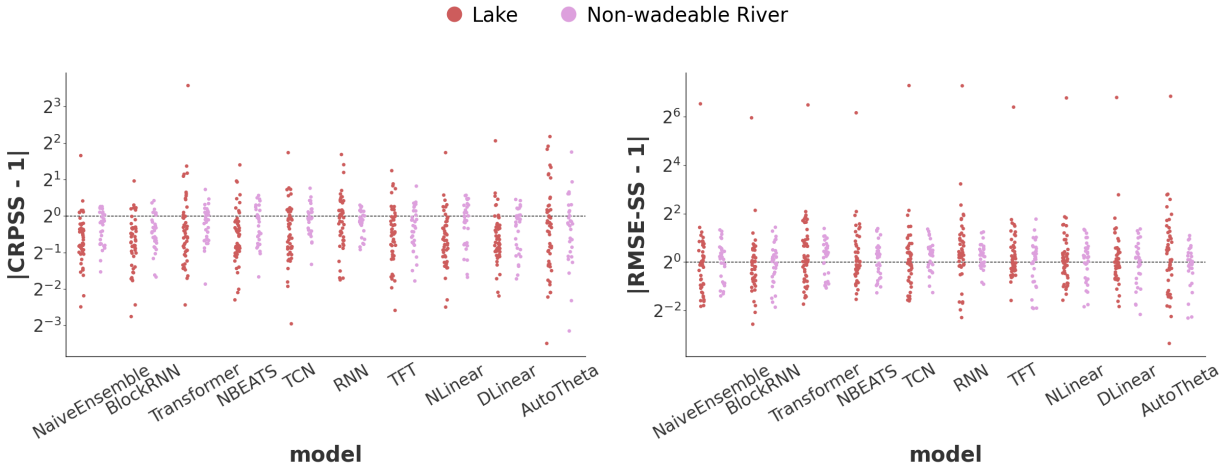

Figure 3: Skill score distributions for chlorophyll-a. The neural network models have shorter tails in the underperforming region relative to AutoTheta, indicating that the neural network models are producing a smaller amount of inaccurate forecasts. The naive persistence null model attained the lowest RMSE on chlorophyll-a, and this is apparent in the RMSE-SS plot as the models have skill score distributions that are generally centered above the underperformance threshold.

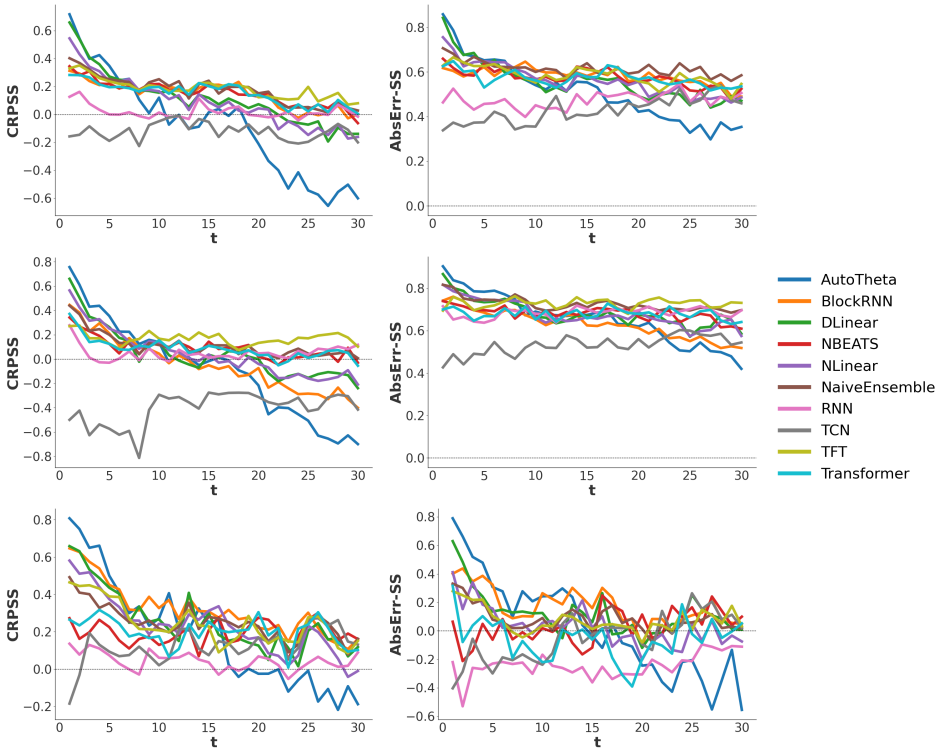

Figure 4: Mean skill scores within the 30-day forecast horizon. In the early phase of the forecast horizon ($t < 5$), the AutoTheta model is universally the top performing model. However, the performance of the AutoTheta model rapidly declines over the course of the 30-day horizon, concluding in AutoTheta being the worst performing model universally at the end of the horizon. The neural network models do not perform as well as AutoTheta at the beginning of the forecast horizon, but their performance declines less rapidly throughout the horizon.

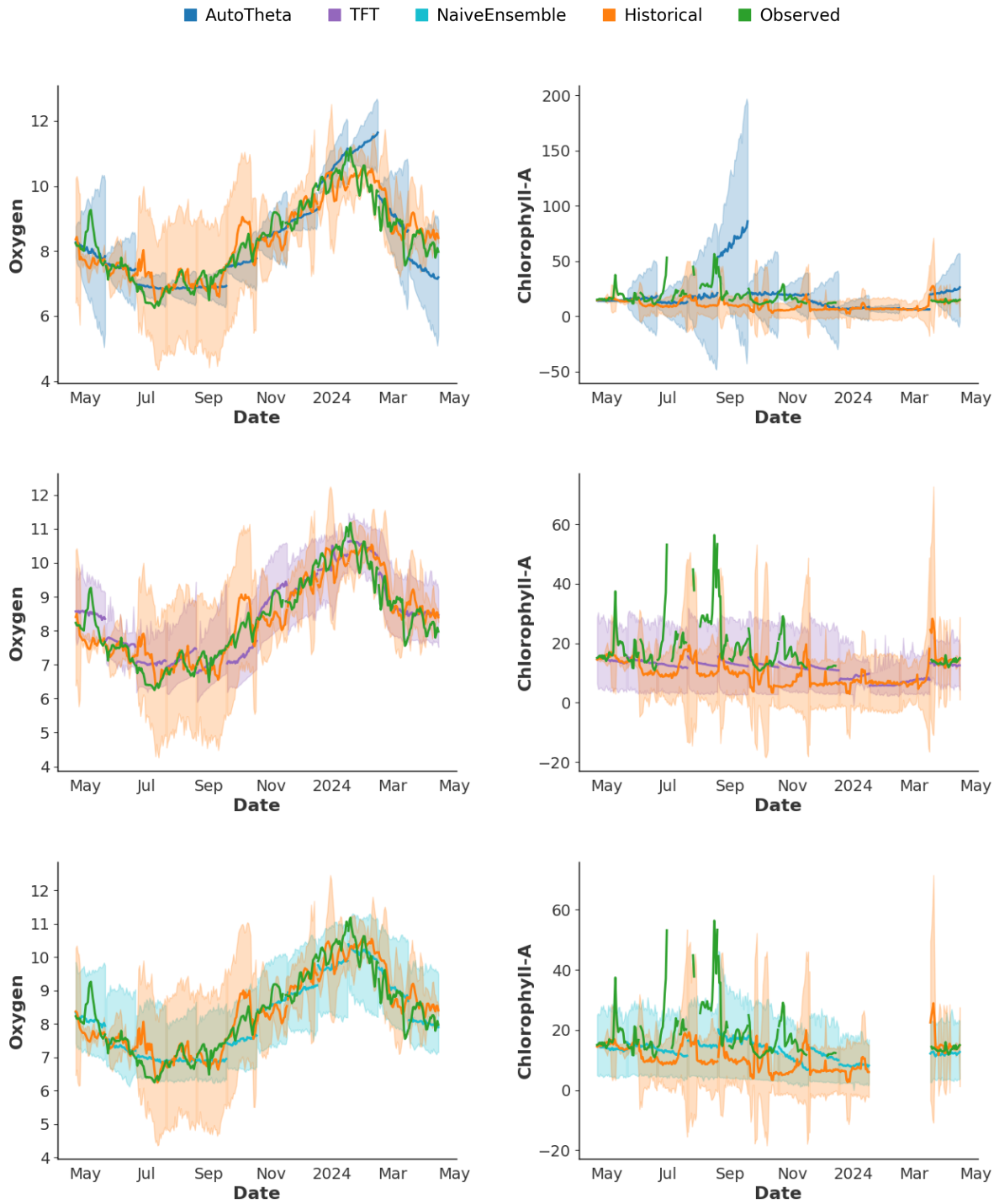

Figure 5: Individual forecasts of dissolved oxygen and chlorophyll-a. The AutoTheta model produces accurate forecasts most of the year, especially for dissolved oxygen; but, there are some intervals where AutoTheta produces extremely inaccurate forecasts. After peaks in dissolved oxygen and chlorophyll-a, the AutoTheta model wrongly predicts that the values will continue to increase. Conversely, the neural network-based forecasts from the TFT and the Naive Ensemble model have learned from the training data that high levels of dissolved oxygen in the winter are followed by a steady decline throughout the spring. With chlorophyll-a, the neural networks take a conservative strategy where they predict that chlorophyll-a will remain in a non-bloom state, and these models never take a risk in forecasting that a bloom will occur.

comprehensive comparison of neural network models, including models taken from more recent developments in computer science.

Instead of identifying that there is some new neural network model that performs exceptionally well, this study has affirmed the well-established forecasting axiom that ensemble forecasts are often more accurate and robust than forecasts from single models. We found that an ensemble model which aggregates the forecasts from the neural network models was the best performing model overall, evincing all the other models on the target variables of dissolved oxygen and water temperature, while also performing well for chlorophyll-a, attaining the second and third best CRPS and RMSE, respectively. This result affirms recent work from [**?**] which established that multi-model ensembles offer advantages over individual models for water temperature forecasting.

Yet, as is often the case when judging whether one method is superior to another, we have found that subtle changes to our criteria may have produced radically different results. For instance, if did not change any of the model training configurations but cut off the forecast horizon to 5 days, then the AutoTheta model would have been the best performing model since the AutoTheta model outperformed all the other models in short time horizons (see Figure **??**). But, it is important to qualify that if we changed the model training configurations to consider shorter forecasts horizons, this could result in the neural network models learning different behavior during training. How neural networks models perform with different forecast horizons warrants future research.

The neural network models struggled to forecast Chlorophyll-A, a result that is not surprising given that forecasting algae blooms is well established as a challenging prediction problem in limnology [**?**]. With Cholorophyll-A (chla), the neural network models displayed a tendency to take a conservative strategy, regularly predicting that the chla concentration would remain in a non-bloom state. It is possible that the neural network models settled on this conservative behavior because we did provide enough information to predict blooms; the neural networks were not provided with some of the typical covariates used in process-based models like Nitrogen and Phosphorous concentrations, and photosynthetic active radiation. It is also likely that this conservative strategy is an artifact of the length of the forecast horizon. Blooms are stochastic events, characterized by rapid fluctuations in chlorophyll-a. Predicting the exact time of such an event in a long term horizon will be more difficult than predicting the onset of a bloom in the near-term [**?**]. It seems reasonable that the neural networks would take a conservative approach to forecasting chla as making the wrong prediction for an algae bloom will be costly with respect to the model's objective.

In place of the one research question that we set out to answer, "How do neural network models perform for water quality forecasting?", many new questions and ideas have arose from this project. Outside of exploring how the performance of neural networks will vary with forecast horizon, we think that some other promising directions for exploration include cross-learning and hybrid models. Throughout this study, we

have trained models individually for each target time series; cross-learning presents a different approach where we would train one model across all the time series available for a given target variable. Through cross-learning, the model might be able to learn a pattern from one time series that will improve forecasts on a different time series; but, if the model is trained on a single time series, then the model may never learn this pattern. We also reckon that hybrid models may have superior forecasting performance to neural network or statistical models individually. We observed that AutoTheta were the best performing models early in the forecast horizon, and neural networks performed best later in the forecast horizon. This motivates the exploration of a hybrid model that employs AutoTheta at the beginning of the forecast horizon and neural network models later in the forecast horizon.

It is an exciting time to study water quality forecasting. There are large, publicly accessible water quality data sets that exist now, and there will be an increasing amount of limnological data in the future as more sensors will be deployed [**?**]. Neural network models offer promising advantages over alternative methods for limnological forecasting as these methods can learn complex patterns without being restrained by a need for domain knowledge. Furthermore, their performance has the potential to improve as more water quality data becomes available. Yet, they also present a range of problems like a lack of interpretability and generalizability that could restrain them from being useful decision support tools for the safety critical problem of water resource management. This study does not attempt to provide a definitive answer as to whether neural networks will be the best method for water quality forecasting as answering this questions will take numerous studies to reach a consensus on. Instead, we hope that this study will generate ideas and a collective momentum towards improving our general ability to forecast and manage water quality.

5 Acknowledgements

Conflicts of Interest

The authors declare there are no conflicts of interest.

Authors' Contributions

Marcus Lapeyrolerie and Carl Boettiger developed the code and wrote the manuscript.

Data Availability

6 References

References