A Comparison of Machine Learning 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)

April 16, 2024

Abstract

1.
2.
3.
4.

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

1 Introduction

The ability to accurately forecast water temperature, dissolved oxygen and chlorophyll concentration in freshwater systems is important to natural resource management. Since water temperature, dissolved oxygen and chlorophyll concentration effect many biological and physical processes in freshwater ecosystems [**?**, **?**, **?**], preemptive forecasts of these variables could allow managers to prepare for scenarios with far-reaching negative impacts. Limnologists have historically used empirical and process-based models to forecast water

quality metrics, but in recent years, researchers have shown that machine learning can outperform these traditionally used methods [**?**, **?**]. A shortcoming of the limnological studies that support machine learning is that they tend to focus on a handful of sites that are not necessarily representative of water systems on a broad geographic scale. This paper has the primary aim of making a comprehensive comparison of state-of-the-art machine learning methods by examining how these models perform 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 phenology 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 climatology model.

There are many neural network architectures that have not been explored for ecological time series forecasting. Past applications of machine learning to limnological time series forecasting have typically used the long-short-term-memory (LSTM) neural network which was first introduced in 1997 [**?**]. Over the nearly 30 years since LSTM's were introduced, there has been much development in the subfield of machine learning for time series forecasting, but recent studies examining machine learning for ecological time series forecasting have not explored more contemporary neural network architectures. This paper will compare 8 machine learning algorithms including LSTM's to ML methods that have been recently developed for time series forecasting like Temporal Fusion Transformer's and Temporal Convolutional Networks [**?**, **?**].

2 Materials and Methods

We employed ML models to generate water temperature, dissolved oxygen and chlorophyll-A concentration forecasts for 34 freshwater sites across North America. These sites consist of Lakes, Non-wadeable Rivers and Wadeable Streams, subtypes classified by NEON. 8 different ML models were trained for every site and target variable available in NEON Ecological Forecasting Challenge Aquatics theme. For the models that accept past covariates (**list these models**), 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 (**list these models**), 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 2018 to 2023 and validated at 12 30-day non-overlapping intervals in 2023.

We compared the ML models to a climatology and naive persistence null model. The climatology model generates forecasts by finding the daily mean and standard deviation of the target time series, and then drawing 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 window.

The imputation of missing data proved to be critical to the performance of the machine learning models. After experimenting with data filling methods that resulted in poor model performance, we settled on a imputation method inspired by the climatology model. Missing data is estimated using the daily historical mean when this statistic is available; and, for the case when there are no data collected for a day of the year, either the monthly, seasonal or global median was used with preference given to the median available from the shortest timescale.

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 `scikit-learn` in Python or `tidymodels` in R that allow users to easily employ time series forecasting models without having to implement them. For model comparison, we rely on the `CRPS` library in Python to efficiently compute the continuous ranked probability scores (CRPS). When comparing the probabilistic forecasts, we express the CRPS score in error-orientation – i.e., a lower CRPS indicates a smaller margin of error.

2.1 Machine Learning 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 theoretically, 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 ML 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 is simultaneously limiting as these models are not readily interpretable and generally require more data than mechanistic analogs []. For the management of critical resources like water, this lack of interpretability could be a deterrent to the use of ML methods, and there may not be enough data to accomodate ML approaches [].

The ML models are configured to generate probabilistic forecasts. The neural network architectures that are used in this study are deterministic, so they are not able to directly represent probability distributions. 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.

3 Results

4 Discussion

Over the last decade, ML has gained significant prominence after a spate of high profile successes like beating the master champion of Go, accurately predicting protein structures and the advent of large language models like ChatGPT [**?**, **?**, **?**]. During this time, researchers have developed many ML models for time series forecasting with promises of significant accuracy improvements over previous methods [**?**]. For this study, we selected 8 ML models that include s

Yet, these claims of improvement tend to have major limitations, notably that the studies examine a limited amount of time series and they do not compare to naive benchmarks [**?**].

5 Acknowledgements

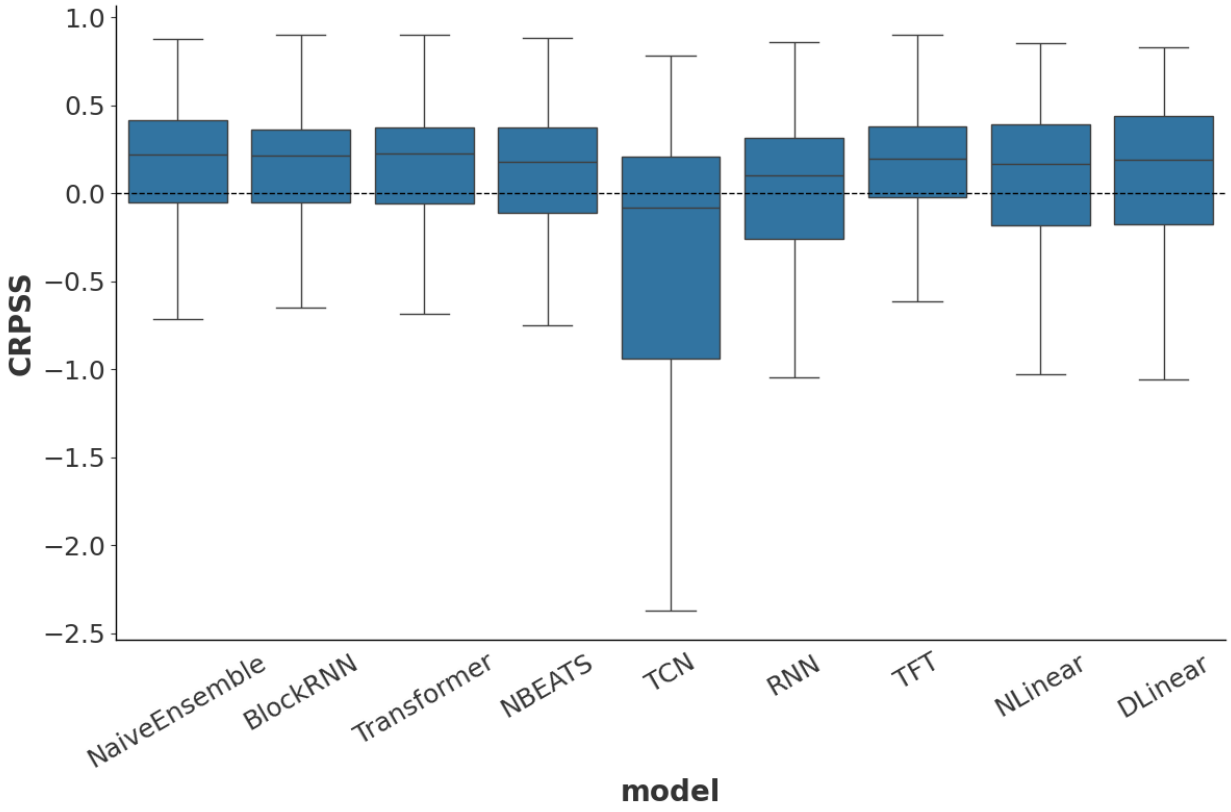
Conflicts of Interest

The authors declare there are no conflicts of interest.

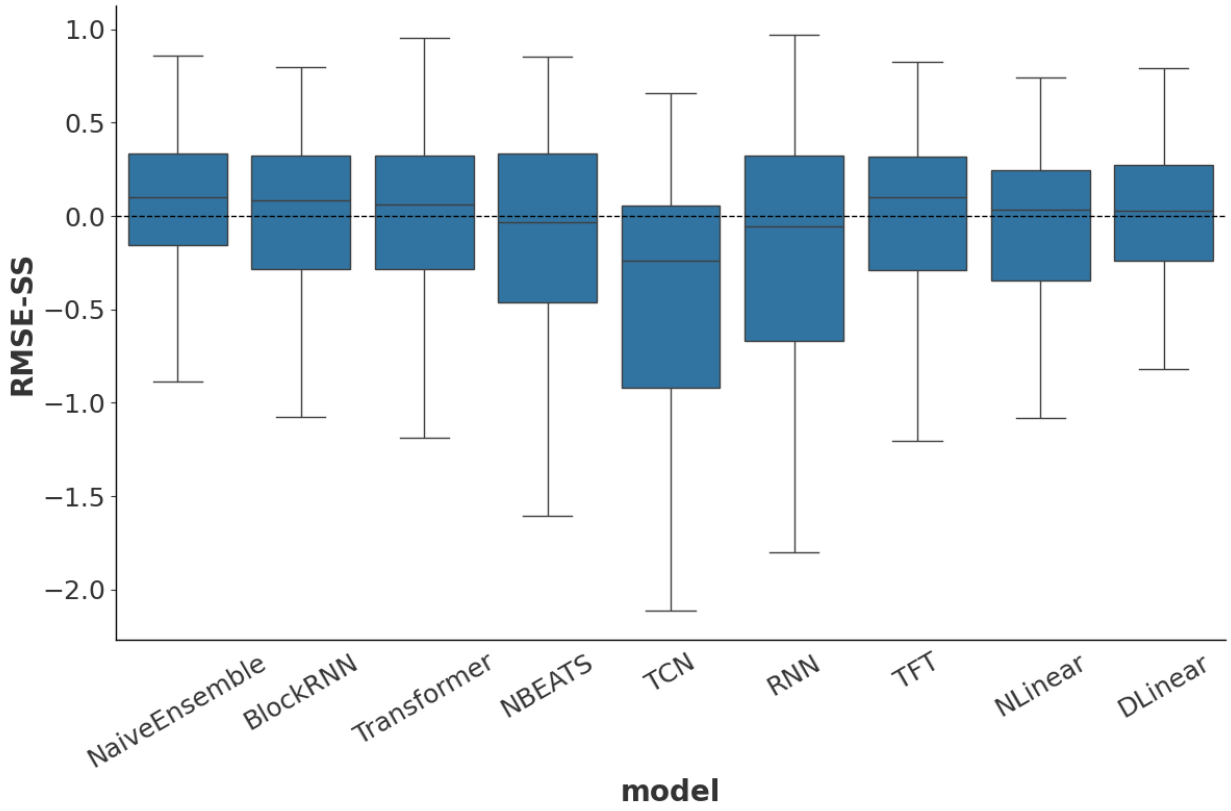
Authors' Contributions

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

Data Availability

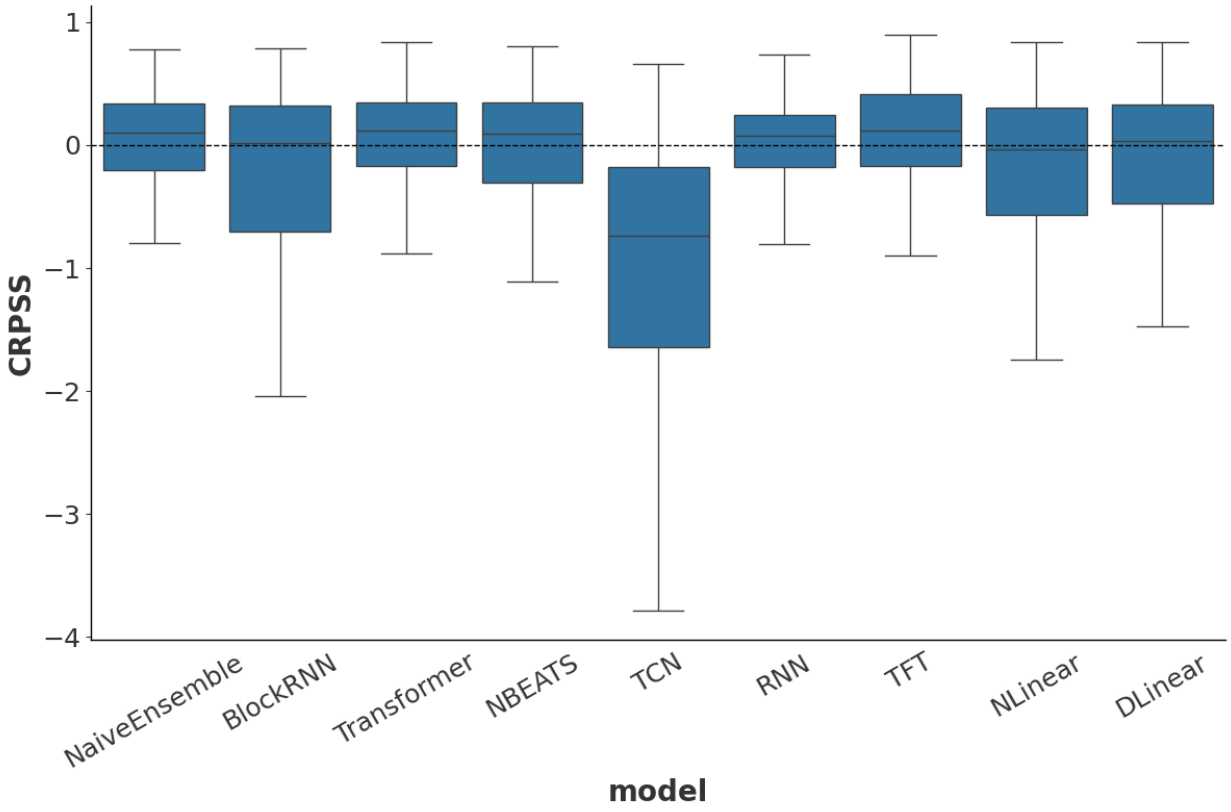


(a)

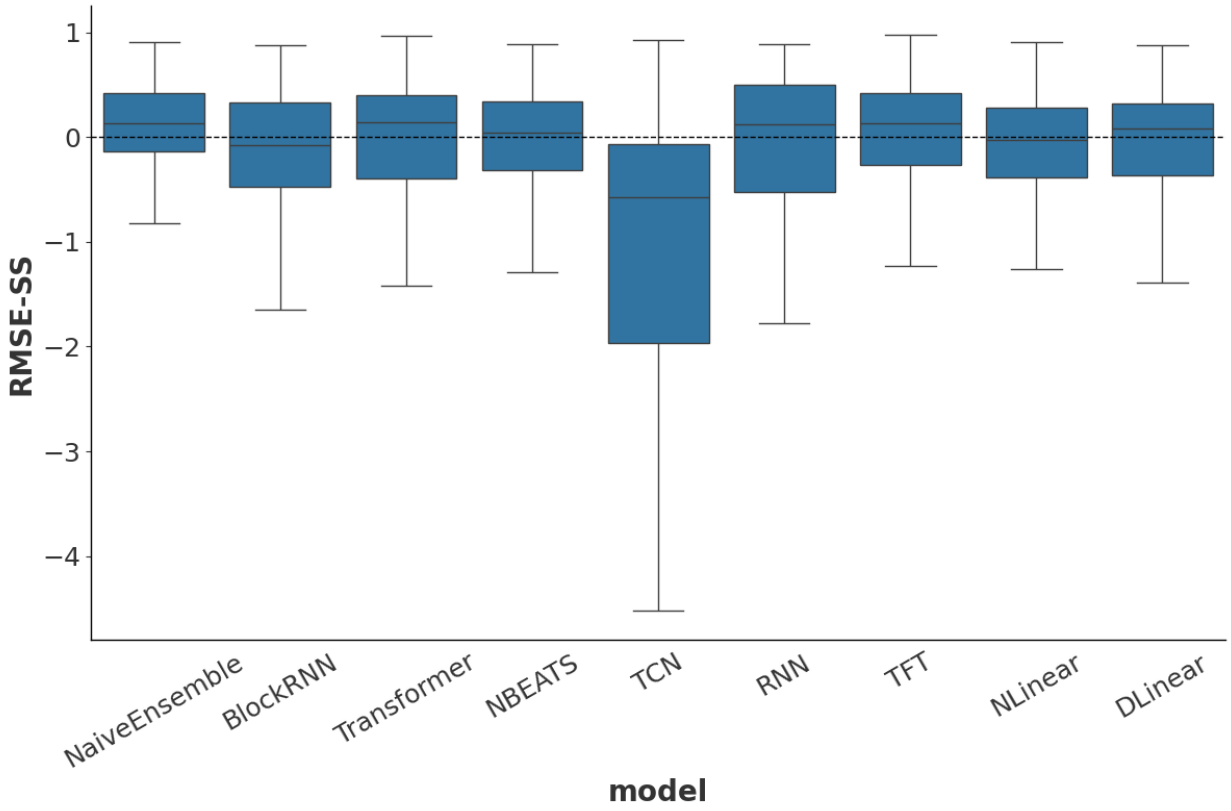


(b)

Figure 1: Comparison of Historical and Naive Oxygen

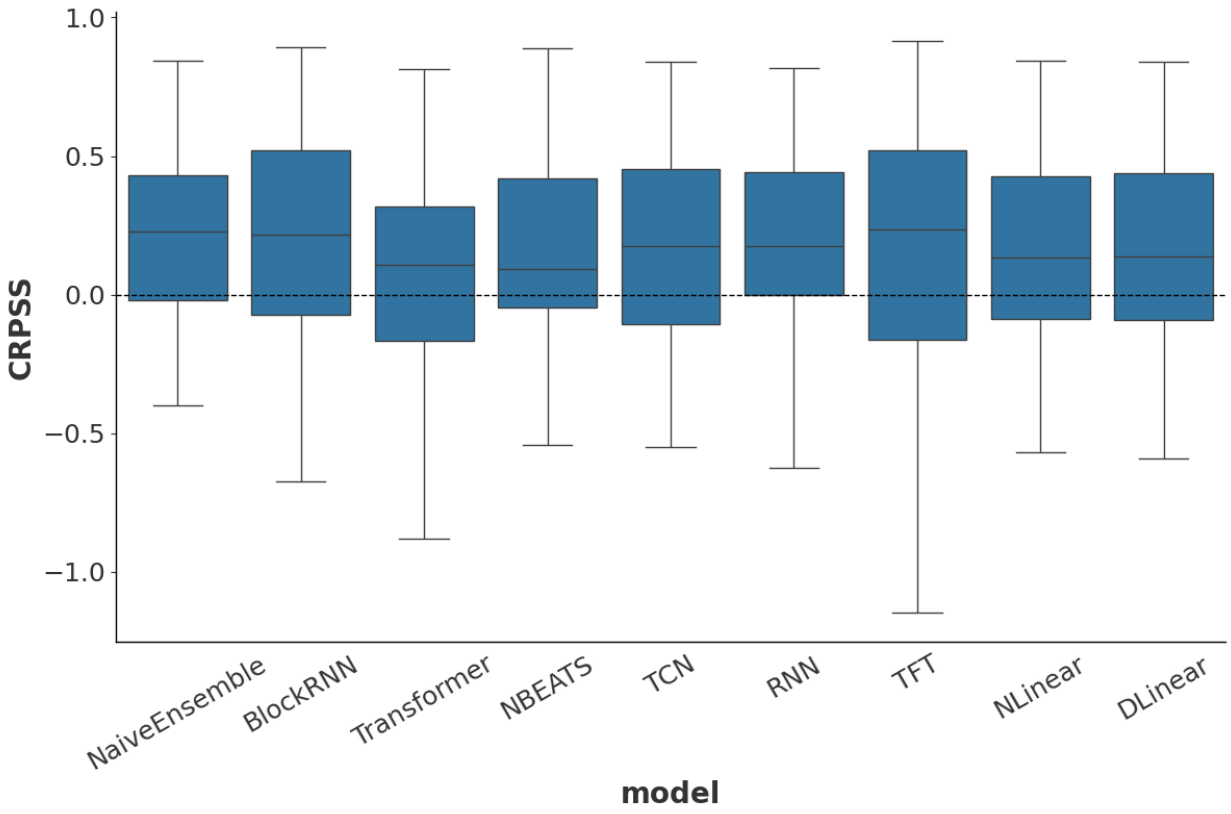


(a)

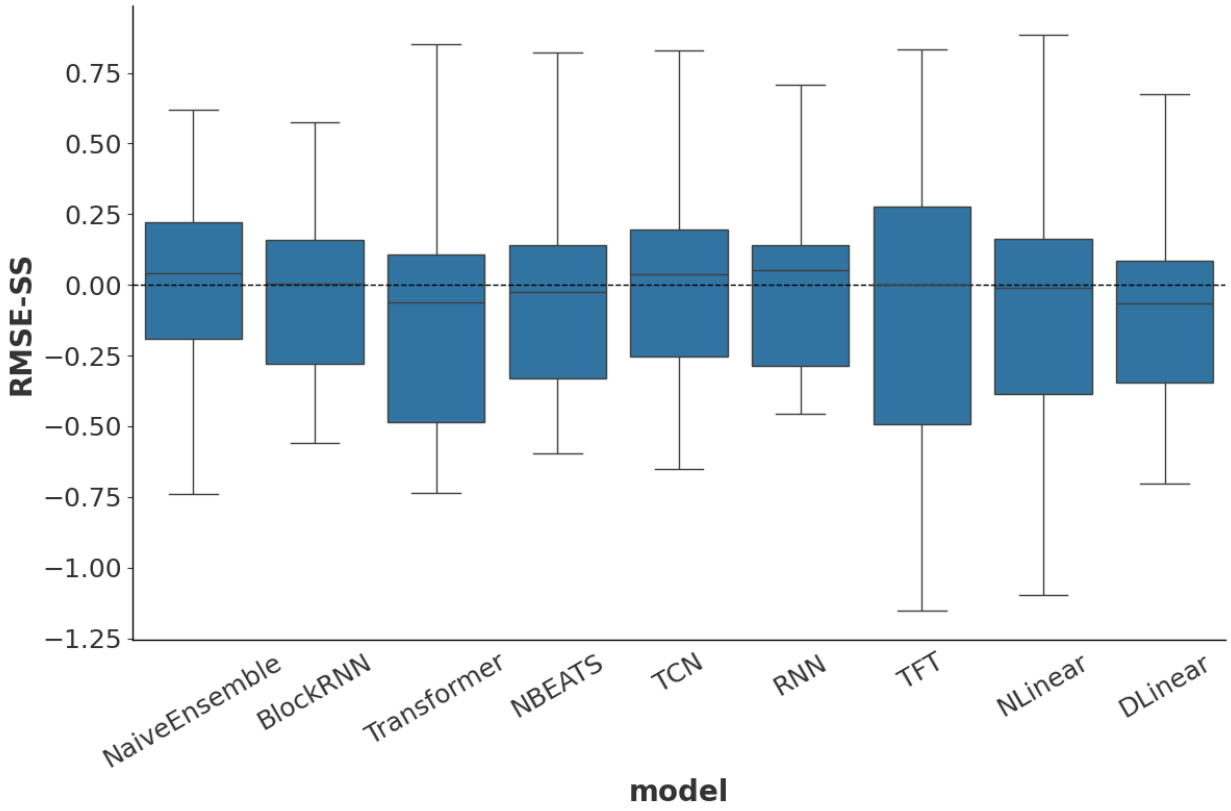


(b)

Figure 2: Comparison of Historical and Naive temperature



(a)



(b)

Figure 3: Comparison of Historical and Naive chla

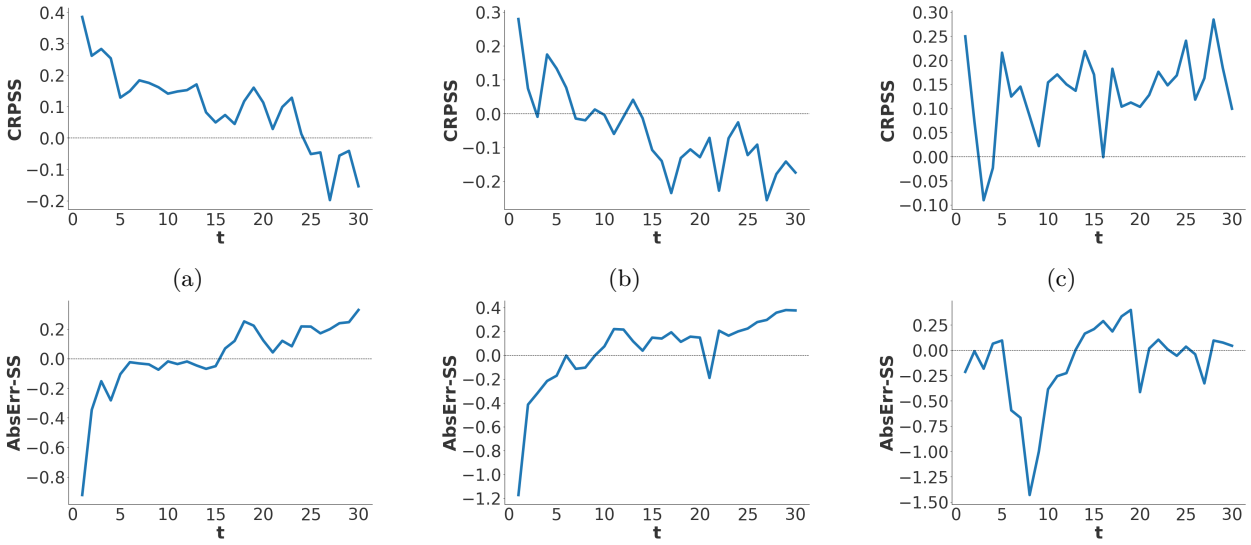
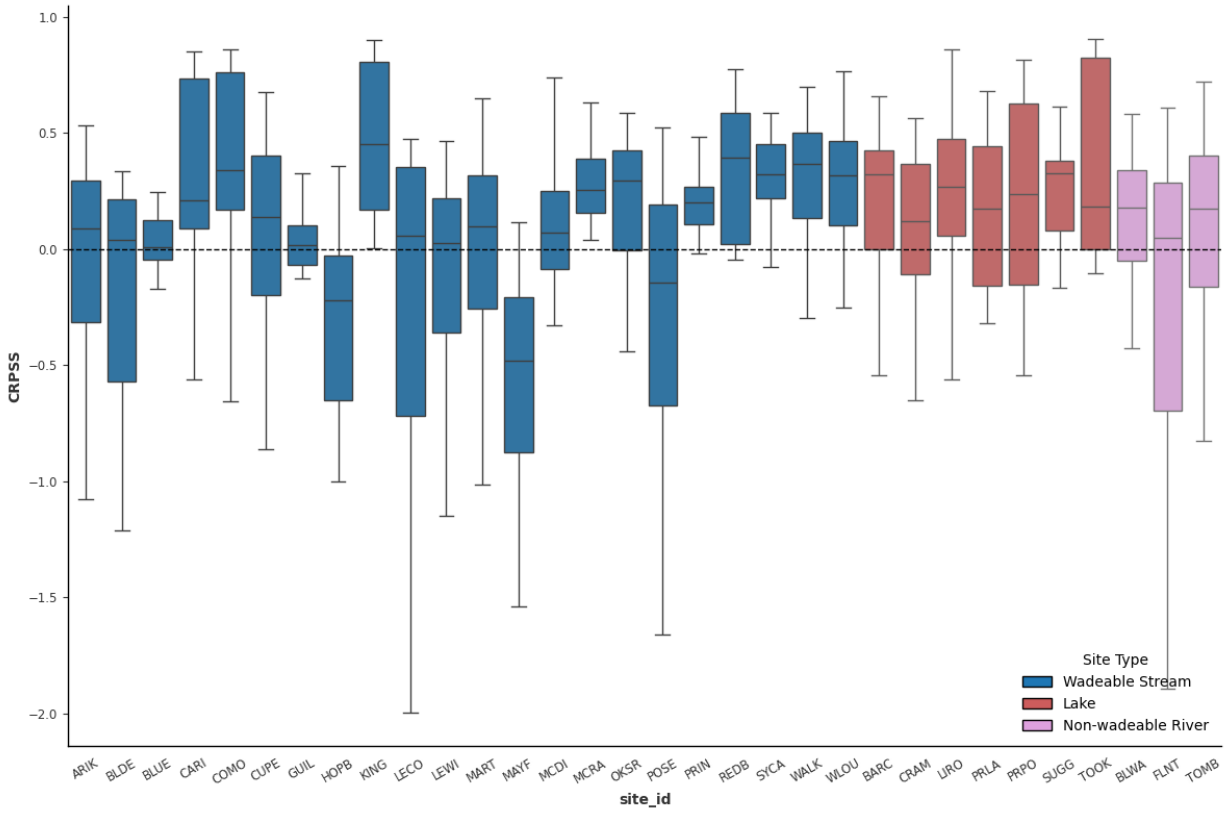


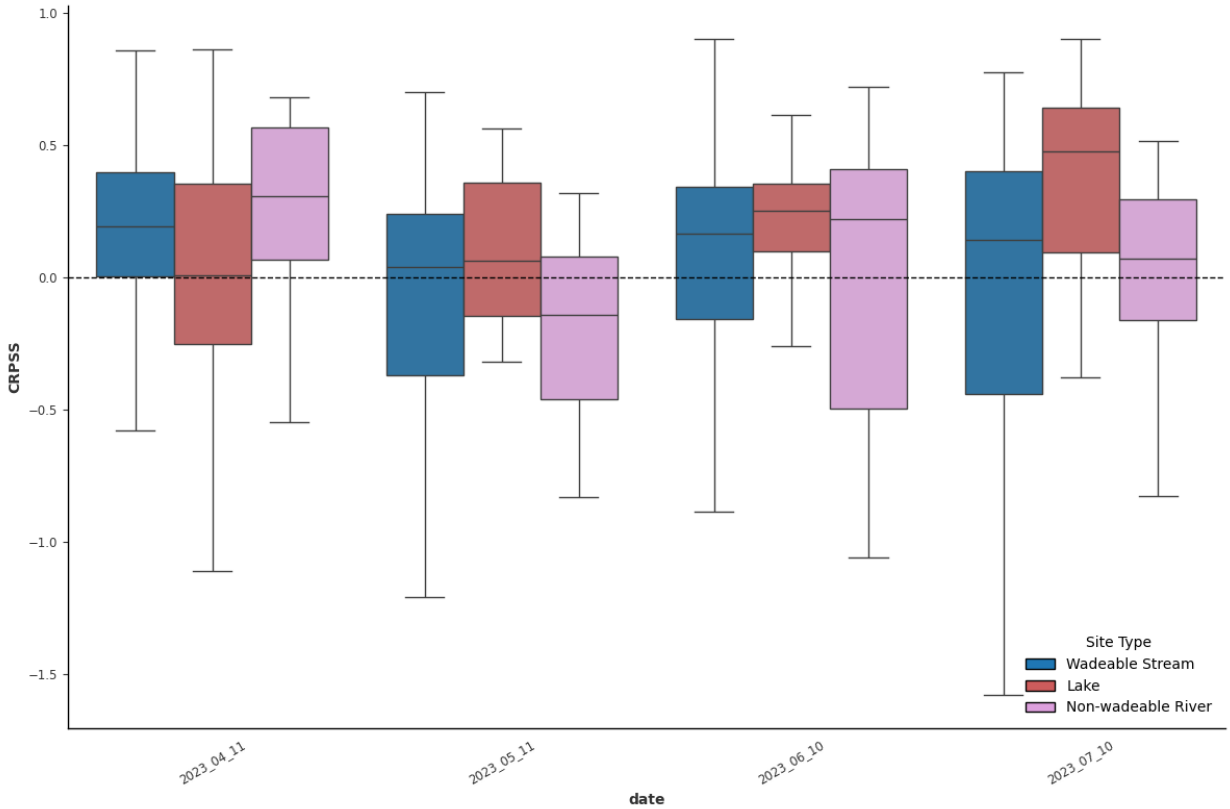
Figure 4: Oxygen, Temperature, Chlorophyll Intrawindow plots

6 References

References



(a)



(b)

Figure 5: Global performance on oxygen