Consulting Project: Water Quality

Predicting Water Quality of the Iller in Kempten

Department of Statistics and Data Science
Ludwig-Maximilians-Universität München

Davit Martirosyan, Benjamin Dornow

Munich, June 4th, 2024

Supervised by Prof. Dr. Helmut Küchenhoff and Henri Funk

Abstract

This study aims to predict water quality indicators for the Iller River in Kempten, Germany, addressing an essential aspect of environmental sustainability within the broader context of climate change. By leveraging data collected from the Bavarian Hydrological Service and the Institute of Geology at LMU Munich, we employed advanced time-series forecasting models, including the Facebook Prophet model, to predict key water quality indicators such as water temperature and dissolved oxygen levels. Our methodology involved thorough data preprocessing, including MICE imputation for handling missing values, and feature importance analysis to identify critical predictors. The models demonstrated high predictive accuracy, with the FB Prophet model outperforming a baseline linear regression model. The results of this study provide valuable insights for water resource management and policy-making, highlighting the necessity of adaptive strategies to mitigate the adverse effects of climate change on water ecosystems. Future research should focus on expanding the dataset and exploring additional water quality indicators to further enhance predictive capabilities and address the complexities of water quality dynamics.

Contents

1 Introduction

Climate change represents one of the most pressing challenges of the 21st century, with profound implications for ecosystems, human health, and economic stability. The accelerating pace of global warming, driven primarily by anthropogenic emissions of greenhouse gases, has led to a cascade of environmental impacts. These include more frequent and severe weather events, rising sea levels, and shifting climatic patterns, all of which threaten biodiversity and disrupt human societies (IPCC, 2021). The urgency to mitigate and adapt to these changes cannot be overstated, as the window for effective action narrows rapidly.

Within the broader context of climate change, the issue of water resources stands out due to its critical importance for sustaining life and enabling economic activities. Water systems, including rivers, lakes, and aquifers, are directly influenced by changing climate conditions, which alter precipitation patterns, increase the frequency of droughts, and intensify the hydrological cycle (Arnell and Gosling, 2016). These changes impact water availability, quality, and distribution, posing significant challenges for water management and necessitating adaptive strategies to ensure sustainable use and equitable access.

Water quality, in particular, emerges as a crucial yet often underresearched aspect of the broader water issue. While substantial efforts have been directed towards understanding water quantity and availability, the quality of water remains inadequately addressed in many contexts. A high quality of water is essential for maintaining the ecological integrity of aquatic ecosystems and for human uses including drinking water supply, agriculture, and industry. However, the dynamic and multifaceted nature of water quality, influenced by both climatic and anthropogenic factors, complicates its assessment and management (Mateo-Sagasta et al., 2017). This gap in research is especially pronounced by the comparatively little knowledge about appropriate indicators for water quality.

Predicting water quality indicators is thus vital for informing conservation efforts, policy-making, and adaptive management strategies. In order to slightly fill this research gap, this paper addresses the prediction of water quality indicators based on data from the river Iller in the German Allgäu. By addressing this underexplored area, it is possible to better understand the interplay between climate change and water quality, ultimately

contributing to more resilient and sustainable water resource management in the face of ongoing environmental change.

2 Literature Review

Whilst the impact of climate change on water supply has been discussed thoroughly for many years now, the topic of water quality has only received little attention so far. However, a conclusion shared by several papers on this issue is that an increased water temperature in rivers and lakes is one of the most instantaneous reactions to climate change (Whitehead et al., 2009, Harvey et al., 2013, Ahmed et al., 2020).

Whitehead et al. (2009) were able to show that the surface temperature in various English bodies of water has already risen by up to 0.6 °C, depending on the region. Among other things, the water temperature influences the speed of chemical reactions and bacterial processes as well as the amount of dissolved oxygen. Further rises in temperature may therefore result in far-reaching and possibly irreversible consequences.

Water temperature also regulates a wide range of biological processes in a river system. Temperature influences spawning periods, growth rates and mortality rates of a river's aquatic inhabitants. A majority of those tend to be cold-blooded organisms that can only flourish when water temperatures are within a specific preferred thermal range. Therefore, a shift out of that range can adversely affect the overall health of the aquatic ecosystem. It has already been shown that high water temperatures lead to an increase of the mortality rates of a number of fish species (Harvey et al., 2013).

Furthermore, climate change does not only pose a threat towards the health of the ecosystem in rivers and lakes, but also to human health. Changes in water temperature play a key role in spreading waterborne and water-related diseases. As an example, the re-emergence of Cholera is a direct effect of climate change. The impact of rising water temperatures varies depending on regional factors. Therefore, rising water temperatures pose an especially huge challenge for developing countries in areas where water-related diseases already pose a considerable challenge (Ahmed et al., 2020). This problem is exacerbated by the fact that 80% of the sewage traceable to humans is discharged into rivers

and oceans without any treatment. This results in massive environmental pollution and a plethora of diseases. Furthermore, 80% of diseases and 50% of child deaths worldwide can be linked with poor water quality (Lin et al., 2022).
Similarly, a decrease in the quality of drinkable water will be especially hard-hitting for countries already dealing with those exact quality issues, such as developing Asian countries (Ahmed et al., 2020).
Finally, a rise in water temperature might very well lead to a decrease in overall drinking water production. In 2022, 27% of the global population were not able to regularly use a safely managed drinking-water service (WHO, 2023). The variations of temperature occurring during climate change can make bodies of water more volatile for contamination and therefore not only worsen the quality, but also the quantity of drinking water. This poses a direct thread for the global drinking water supply.

3 Data Collection

The datasets used for this project consists of two parts. The first dataset was scraped from the website of the Gewässerkundlicher Dienst Bayern (Bavarian Hydrological Service). The second dataset was provided by the project partners of the Institute of Geology of the LMU Munich.
The Gewässerkundlicher Dienst Bayern provides hydrological monitoring data in graphs and tables as well as detailed information on the monitoring sites on its homepage. One of these measuring stations is the Kempten/Iller monitoring station operated by the Wasserwirtschaftsamt Kempten (Gewässerkundlicher Dienst Bayern, 2024). In order to automatically import the data from the measuring station in Kempten/Iller, a data collection pipeline was created. The script iteratively accessed the table for each potential target variable and extracted relevant information using Python's *Requests* package. The package *Selenium* was not necessary as the data that was needed to be scraped from the website was not being dynamically loaded. Provided the website's underlying HTML code has not been altered, the code from the data collection pipeline is entirely reproducible.
The second dataset was provided by the project partners from the Institute of Geology

at LMU Munich. The dataset contains the variables river discharge, radiation, air temperature, relative humidity and precipitation. All five variables contain multiple daily observations recorded at three-hour intervals (eight observations per day) from February 1, 1980 to October 31, 2010. The files containing the data were provided in .txt format without any spacing or formatting. Therefore, the dataset was read as .txt using the *Pandas* package and merged with the scraped dataset. Finally, the merged dataset was saved as a .csv file.

4 Data Preprocessing

Although we merged the datasets, two urgent issues had to be addressed before further steps could be taken. Firstly, it was necessary to achieve a uniform time interval between observations. Furthermore, some variables exhibited missing values, as can be seen in Figure 1.

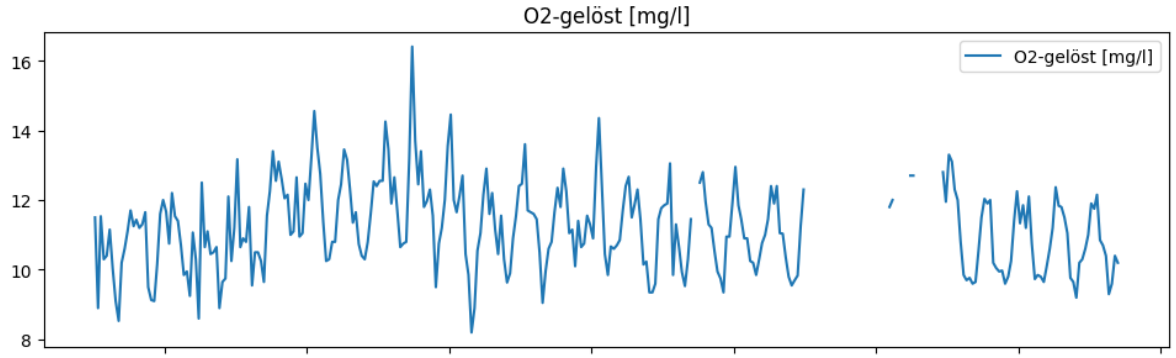

Figure 1: Line plot of the variable diluted O2 with missing values

In order to obtain a uniform time interval, all variables contained in the dataset were aggregated down to one observation per month. Subsequently, the problem of missing values was approached. The gaps in the period between 15.02.2002 and 15.10.2005 were filled by simulating data using the MICEforest package. The use of MICE (Multiple Imputations by Chained Equations) made it possible to obtain a complete dataset without any gaps by filling in missing values from random draws over non-missing data (PyPI, 2024). The resulting data can be seen in Figure 2.

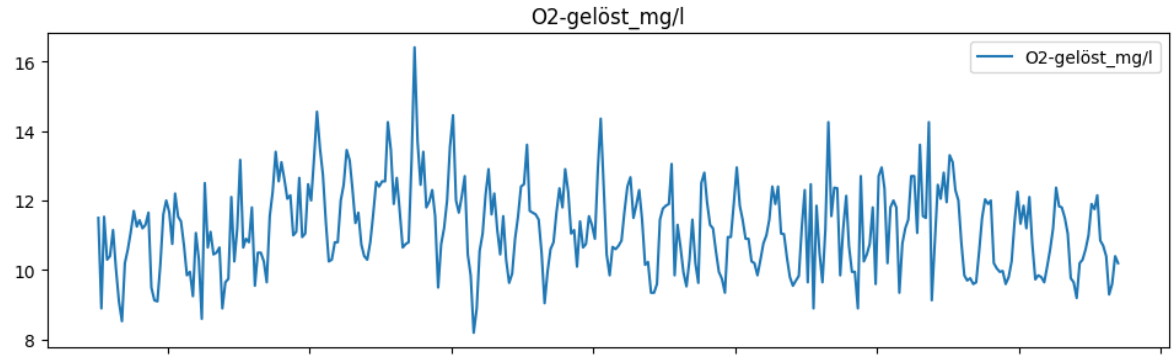

Figure 2: Line plot of the variable diluted O2 with imputed values

The MICE imputation was conducted for every variable obtained from the data collection pipeline. The plots of the variables with imputed values can be viewed in the appendix. In Figure 3, we can observe the distributions of our covariates.

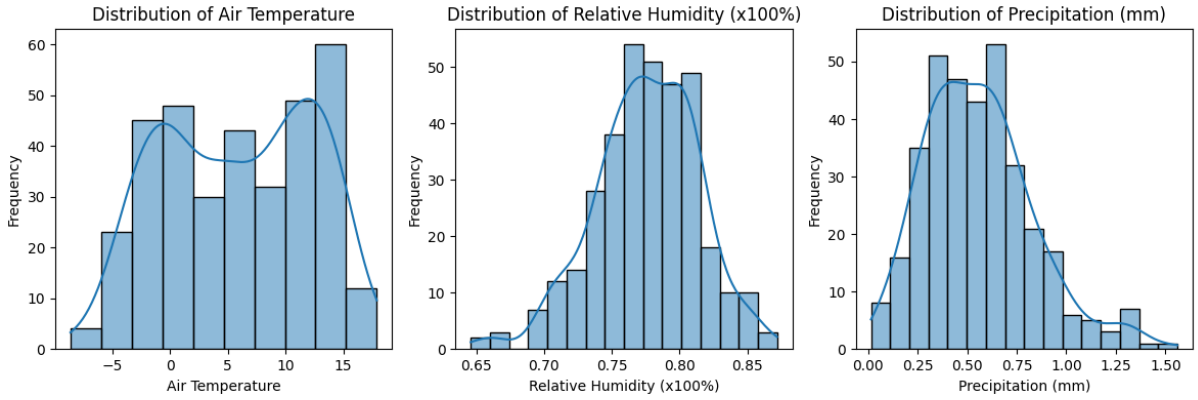

Figure 3: Distributions of the features air temperature, relative humidity and precipitation

The distribution of *air temperature* exhibits two peaks. This property suggests that the variable is subject to seasonal effects. With the exception of the two peaks, the distribution is reminiscent of a normal distribution. *Relative humidity* follows a near-normal distribution centered around 75%. The distribution of *precipitation* is slightly right-skewed, which indicates the presence of more low precipitation events compared to high ones. However, the distribution of the variable still displays a normal distribution, albeit slightly shifted to the left.

The multimodal distribution of *radiation* visible in Figure 4 reflects the varying levels of solar radiation throughout different times and conditions. Its distribution seems to be more or less uniform. The distribution of *river discharge* appears to be right-skewed.

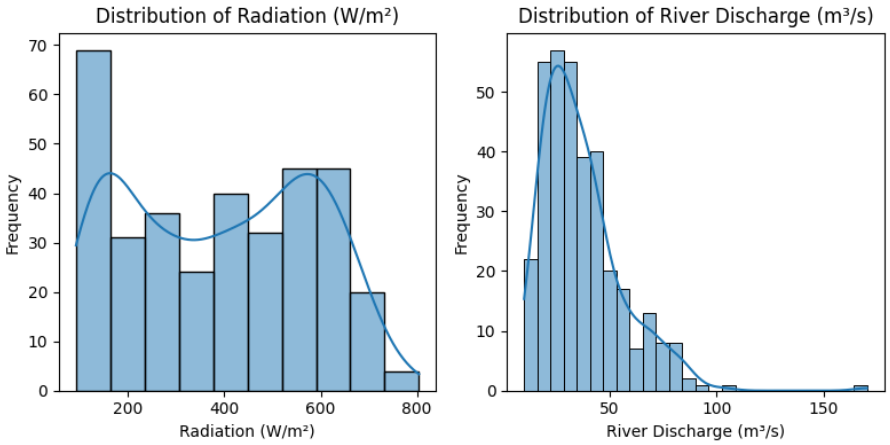

Figure 4: Distributions of the features radiation and river discharge

Overall, the distributions of the variables provided by the project partners resemble a normal distribution to a certain extent in most cases. This is advantageous for several reasons.

Many predictive models, especially linear regression, assume that the features and residuals are normally distributed. Normality in the features helps meet these assumptions, leading to more reliable parameter estimates. Furthermore, normal distributions facilitate more accurate inference. This ensures that confidence intervals and significance tests are valid, enhancing the reliability of the model's predictions. Since normal distributions are symmetric and feature fewer outliers, bias and variance in the model can be reduced. This results in more stable and generalizable predictions. Finally, features that follow a normal distribution are easier to interpret and analyze.

The distributions of the scraped variables, including water temperature and diluted O2, can be found in the appendix.

5 Modeling

Once the data had been prepared for further processing, a suitable model needed to be set up in order to predict the selected quality indicators. For this purpose, the characteristics of the data and the objective of the project had to be taken into account.

5.1 Model Selection

The prepared and preprocessed data consists of time series data with monthly observations. Furthermore, strong seasonal effects occur in some of the variables due to annual fluctuations. Based on research, a distinct trend can also be assumed in some variables (e.g. generally rising water temperatures due to global warming). Finally, occasionally changing growth rates can also be assumed due to increased water temperature fluctuations, for example. A suitable model for this task is the Facebook Prophet model.
The Facebook Prophet model (FB Prophet model) was developed by Facebook for forecasting time series data in the context of predicting stock market and housing prices.
Formally, the model can be described by the following components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

where:

- $y(t)$ is the time series,
- $g(t)$ is the trend function which models non-periodic changes in the value of the time series,
- $s(t)$ represents the seasonality component, capturing periodic changes,
- $h(t)$ denotes the effects of holidays or special events that occur at irregular intervals,
- ϵ_t is the error term which accounts for any idiosyncratic noise in the observations.

The model is known for its effectiveness and ease of use in handling data with strong seasonal effects, missing data and shifts in the trend. In order to model a trend, it automatically fits a piecewise linear or logistic growth curve trend to univariate time series data. This allows the estimation of changes in the trend component as well as shifts at points it identifies as change points. Furthermore, the model is able to handle rolling means, lags and other features (Facebook, 2024b).
Seasonality components, such as daily or yearly seasonality, are estimated using Fourier series. This allows the model to decompose complex periodic signals into simpler compo-

nents, making it easier to analyze and understand the underlying patterns and characteristics of the time series data (Facebook, 2024d).

Furthermore, depending on the nature of the components, FB Prophet allows these components to be modeled either additively or multiplicatively (Facebook, 2024a). Adding additional regressors (continuous or categorical) is also possible and provides better results than tree-based models (Facebook, 2024e). Finally, the model exhibits an exceptionally fast training time.

Since the advantages of the model appear to fit the given dataset, the FB Prophet model was chosen for analysis. In order to optimize the performance of the FB Prophet model, hyperparameter tuning was employed for predicting the target variable. FB Prophet offers various parameters that can be adjusted to customize the model according to specific data characteristics and modeling objectives. In order to attain the best possible performance for the given target variable, it was especially important to tune the following three parameters:

The parameter *changepoint_prior_scale* controls the flexibility of the model by regulating the sensitivity to changes in the underlying trend. Adjusting this parameter allows for fine-tuning the model's ability to detect shifts or turning points in the data (Facebook, 2024c).

The parameter *seasonality_prior_scale* also influences the flexibility of the model, specifically regarding yearly seasonality patterns. By adjusting this parameter, the model's sensitivity to seasonal fluctuations can be customized, thereby improving its ability to capture and forecast seasonal trends (Facebook, 2024c).

The parameter *seasonality_mode* determines how seasonality affects the trend component of the model. It offers different modes to specify the relationship between trend and seasonality, allowing for greater flexibility in modeling complex seasonal patterns (Facebook, 2024c).

By tuning these parameters using cross-validation, the FB Prophet model can be optimized to provide accurate and reliable predictions for the target variable, enhancing its utility in various forecasting applications (Facebook, 2024c).

5.2 Model Definition

The baseline model is a slightly complex linear regression model. The targets used were the variables *Water Temperature* and *Diluted O2*, which are known from research as indicators of water quality. As features, the variables *Air Temperature* (in °C), *Relative Humidity* (in percentage points), *Precipitation* (in millimeters), *Radiation* (in W/m^2) and *River Discharge* (in m^3/s) provided by the project partners were included in the model as well as a time index.

In predictive modeling, it is common practice to lag data to enable the model to learn from past observations. This approach leverages temporal dependencies within the dataset, enhancing the model's ability to make accurate forecasts. However, for the baseline model, this conventional approach was not followed due to the inclusion of additional regressors. These regressors provided the model with supplementary information that compensated for the absence of lagged variables, thus maintaining the integrity of the predictions. By integrating these external predictors, we were able to capture the underlying patterns and trends without relying on lagged data, simplifying the model structure and potentially improving its robustness.

The model equations for the respective targets read as follows:

$$\begin{aligned}\hat{Y}_{\mathrm{WaterTemperature}} = \beta_0 + \beta_1 X_{\mathrm{AirTemperature}} + \beta_2 X_{\mathrm{RelativeHumidity}} + \beta_3 X_{\mathrm{Precipitation}} \\ + \beta_4 X_{\mathrm{Radiation}} + \beta_5 X_{\mathrm{RiverDischarge}} + \beta_6 X_{\mathrm{TimeIndex}}\end{aligned} \tag{2}$$

$$\begin{aligned}\hat{Y}_{\mathrm{DilutedO_2}} = \beta_0 + \beta_1 X_{\mathrm{AirTemperature}} + \beta_2 X_{\mathrm{RelativeHumidity}} + \beta_3 X_{\mathrm{Precipitation}} \\ + \beta_4 X_{\mathrm{Radiation}} + \beta_5 X_{\mathrm{RiverDischarge}} + \beta_6 X_{\mathrm{TimeIndex}}\end{aligned} \tag{3}$$

In addition, the FB Prophet model was set up. Unlike the baseline model, it can be passed additional regressors more directly. It includes the features provided by the project partners as well as aforementioned parameters specific to the model.

6 Results

In this section, the performance of the proposed models is examined. Subsequently, the predictions of the models with respect to water temperature and saturated oxygen are compared with the respective original data. Afterwards, the overall performance of the models are assessed with regard to several performance indicators.

6.1 Feature Importance

The analysis section begins with a review of the feature importance graphs.

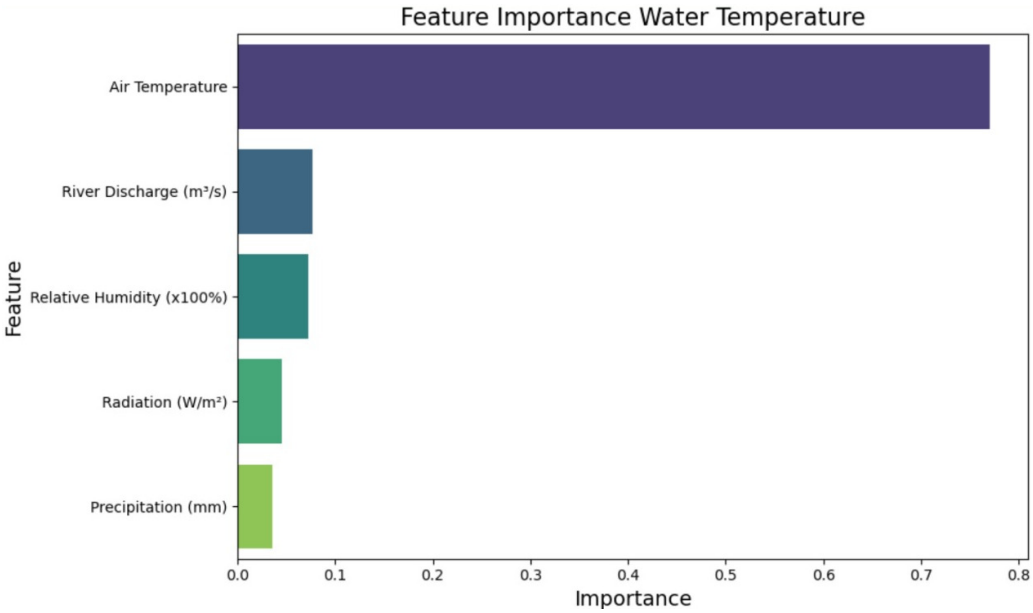

Figure 5: Feature Importance for the target water temperature

Figure 5 depicts the importance of the model features in predicting water temperature. *Air temperature* is the most influential predictor for water temperature, with a distinctly higher importance compared to other features. This underscores the well-known direct relationship between air and water temperatures, where changes in air temperature are strongly reflected in water temperature (Harvey et al., 2013).
Relative humidity is the third most important feature for predicting water temperature. This importance suggests that moisture content in the air may have a slight effect on the temperature in a body of water.

Radiation is the fourth most significant predictor. Solar radiation heats the water to a certain degree, thereby affecting its temperature (Boyd, 2020).

Precipitation has the least importance in predicting water temperature. While rainfall can cause slight cooling of the water body and affect its thermal properties (Croghan et al., 2018), its impact is less significant compared to the other features.

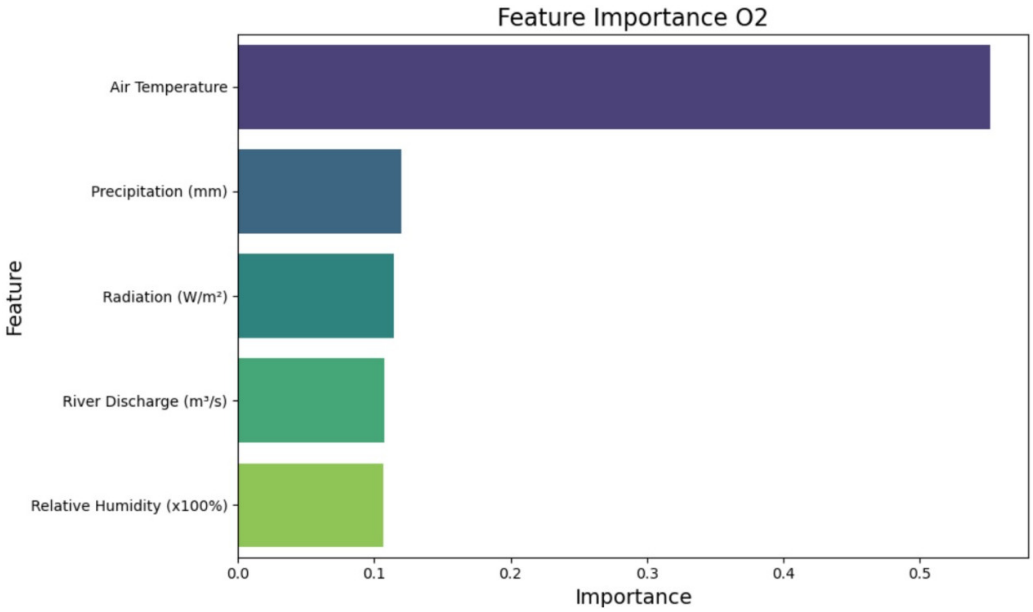

Figure 6: Feature Importance for the target diluted O2

Figure 6 illustrates the importance of the model features in predicting diluted oxygen (O2) levels in a body of water.

Similar to its role in predicting water temperatures, *air temperature* is the most crucial predictor for diluted O2 levels, with a predominant importance score. This finding is also supported by current research (Harvey et al., 2013).

Precipitation is the second most important feature, though its impact is considerably lower than air temperature. This suggests that rainfall slightly affects dissolved oxygen levels. The same applies to *Radiation* and *relative humidity*, indicating that while these three features may have some effect, it is not as significant in predicting diluted O2 levels compared as air temperature.

6.2 Predictions of dataset with prior MICE imputation

Subsequently, the prediction results of the models based on the dataset with prior MICE imputation are evaluated. The models with the target variable water temperature are considered first. For this purpose, the values predicted by the baseline model and the FB Prophet model are first compared with the actual values using a line plot.

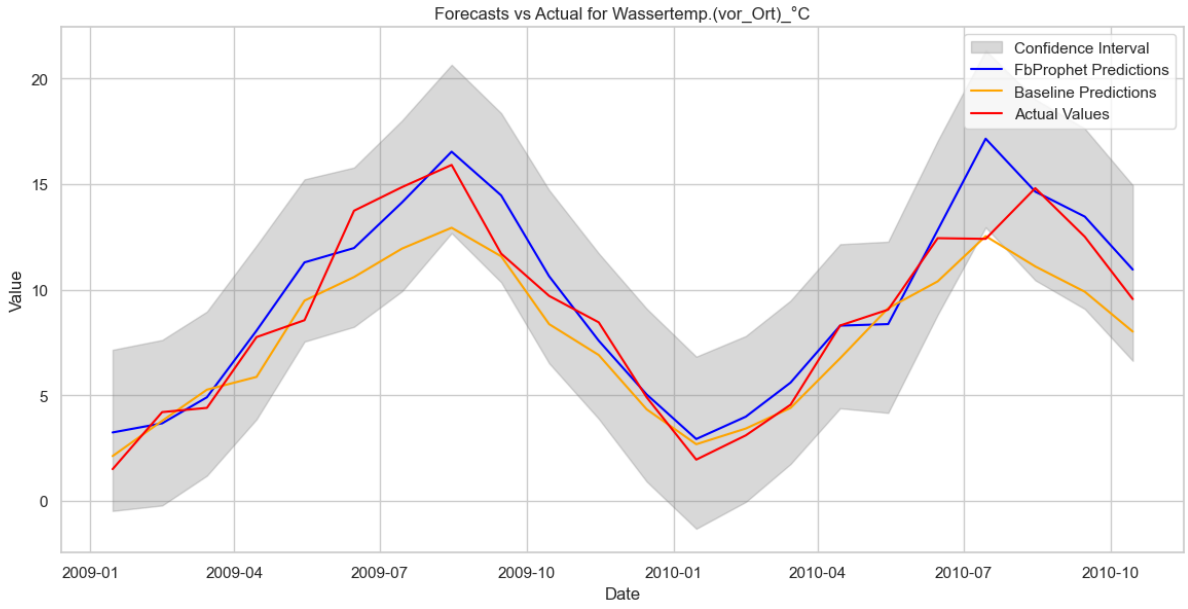

Figure 7: Forecasts vs Actual Values for the target water temperature

When comparing the line plots in Figure 7, it is immediately apparent that both the FB Prophet model (blue line) and the baseline model (yellow line) were able to predict values that are very close to the actual values (red line). It can also be seen that the FB Prophet model is closer to the true values than the baseline model in the majority of cases. It can therefore be assumed that the FB Prophet model is to be preferred over the baseline model.

Although the models perform quite well in most cases, there are areas where deviations from the true values are slightly higher. Particularly at peaks (both local minima and local maxima), the forecast values are somewhat further away from the actual values than at the majority of other segments in the model. This can also be seen from the wider confidence interval at peaks. The FB Prophet model tends to predict temperatures higher

than the actual values both at peaks and in valleys. The baseline model, on the other hand, predicts temperatures that are too low at the peaks, while values near the minimum are estimated to be higher than the true values. Overall, however, both models are able to make predictions close to the true values.

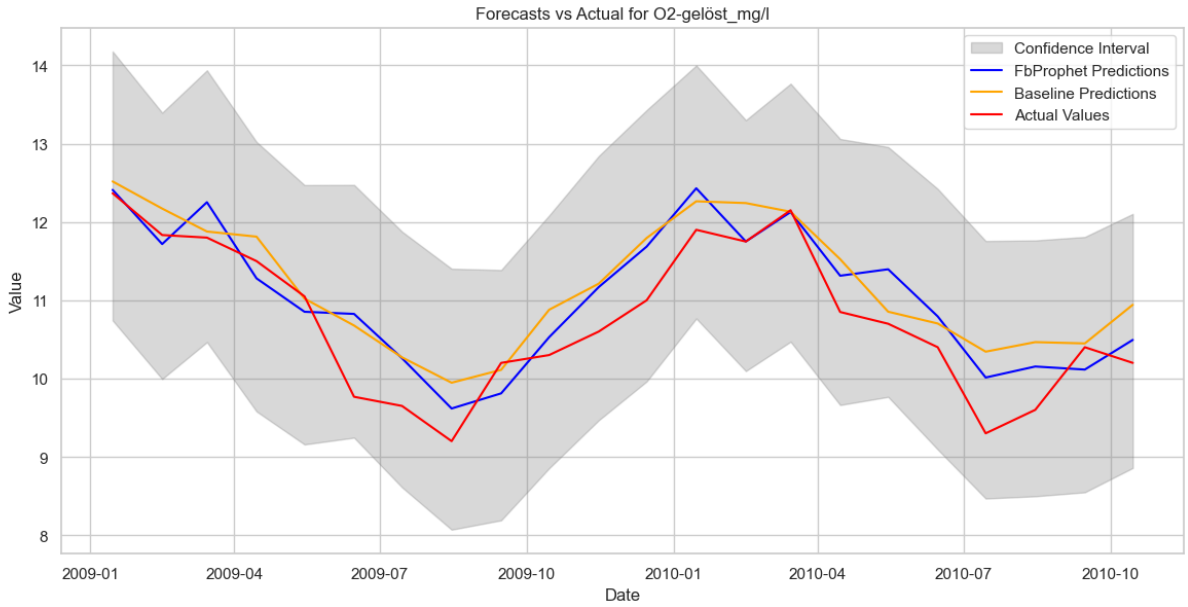

Figure 8: Forecasts vs Actual Values for the target diluted O2

The true values of the diluted oxygen seen in Figure 8 can also be predicted quite accurately by the two models. For this target variable, both the FB Prophet model as well as the baseline model display an observable tendency. Both models tend to overestimate the diluted oxygen. For most of the data points, both prediction graphs overestimate the actual values. These deviations are particularly noticeable with low target values.
Overall, both models perform well in predicting the target variable. However, when it comes to the performance at the peaks, the FB Prophet model performs better than the baseline model. However, as seen with the previous target variable, both models are able to make accurate predictions.
The previous visualizations indicate that the FB Prophet model performs slightly better than the baseline model. However, the difference between the two models has at times only been slightly noticeable. Considering the distribution of the predictions in relation to the actual distribution of the target, the performance difference between the FB Prophet

model and the baseline model becomes slightly more apparent.

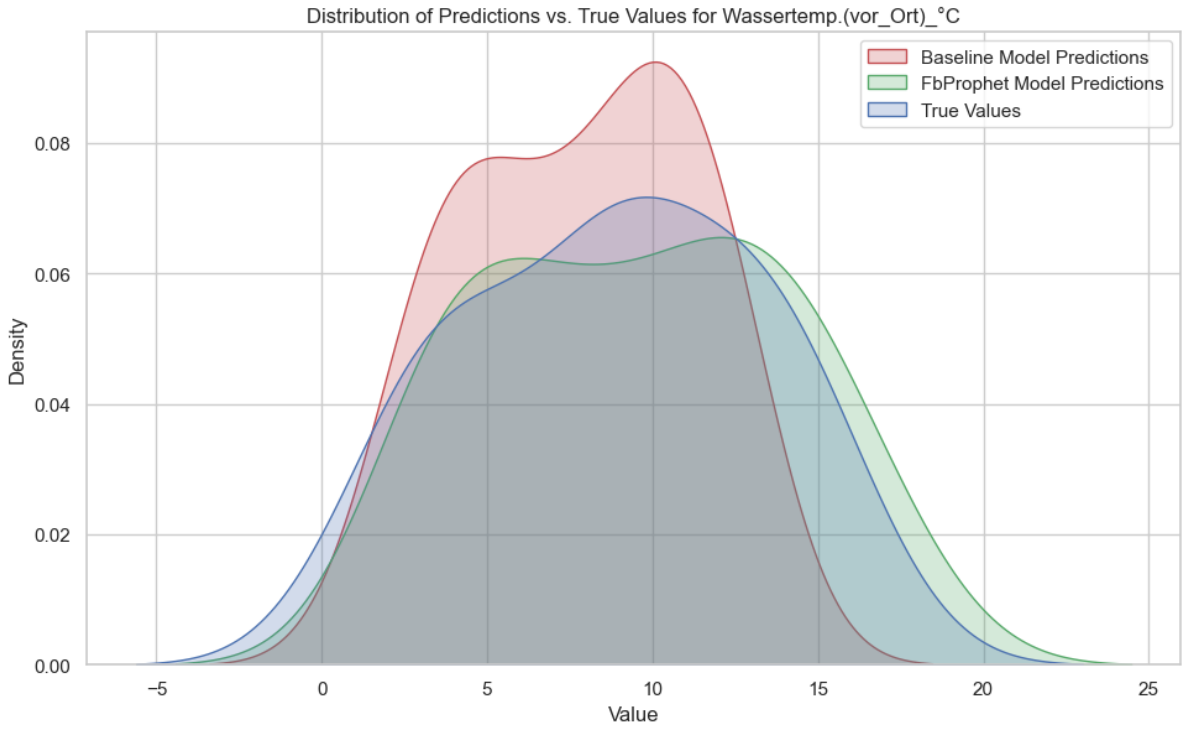

Figure 9: Distributions of Predictions vs Actual Values for the target water temperature

The density plot in Figure 9 compares the distribution of the values predicted by the FB Prophet model (green) and the baseline model (red) with the distribution of the actual water temperature values. As already indicated by the line plot, the distribution of the predicted values is similar to the distribution of the actual values. However, the difference between the FB Prophet model and the baseline model becomes more apparent. Whilst the density of baseline model predictions and the actual values for water temperatures between approximately -5 and 2 degrees Celsius appear quite similar, the density plots for temperatures of around 2 degrees Celsius or higher diverge more strongly. The difference is particularly pronounced for temperatures between around 4 and 12 degrees Celsius. Here, the baseline model has two peaks that cannot be found in this distribution shape of the original data. The tendency to predict overly high temperatures near the minima of the actual data and to predict overly low temperatures near the maxima of the actual data, which is also recognizable in the line plot, is also evident in this graph.

The density plot of the FB Prophet model, on the other hand, strongly resembles the

density plot of the true values across almost all the observed temperatures. The only striking deviation can be found in the range between around 5 and 13 degrees Celsius. The FB Prophet model also suggests two peaks here, however, these are considerably less severe compared to those of the baseline model and are therefore much closer to the distribution of the real values. It is also important to note that the FB Prophet model predicts high temperature values much more accurately than the baseline model. This quality is particularly beneficial for the FB Prophet model with regard to the expected trend of climate change leading to increased water temperatures over the upcoming years.

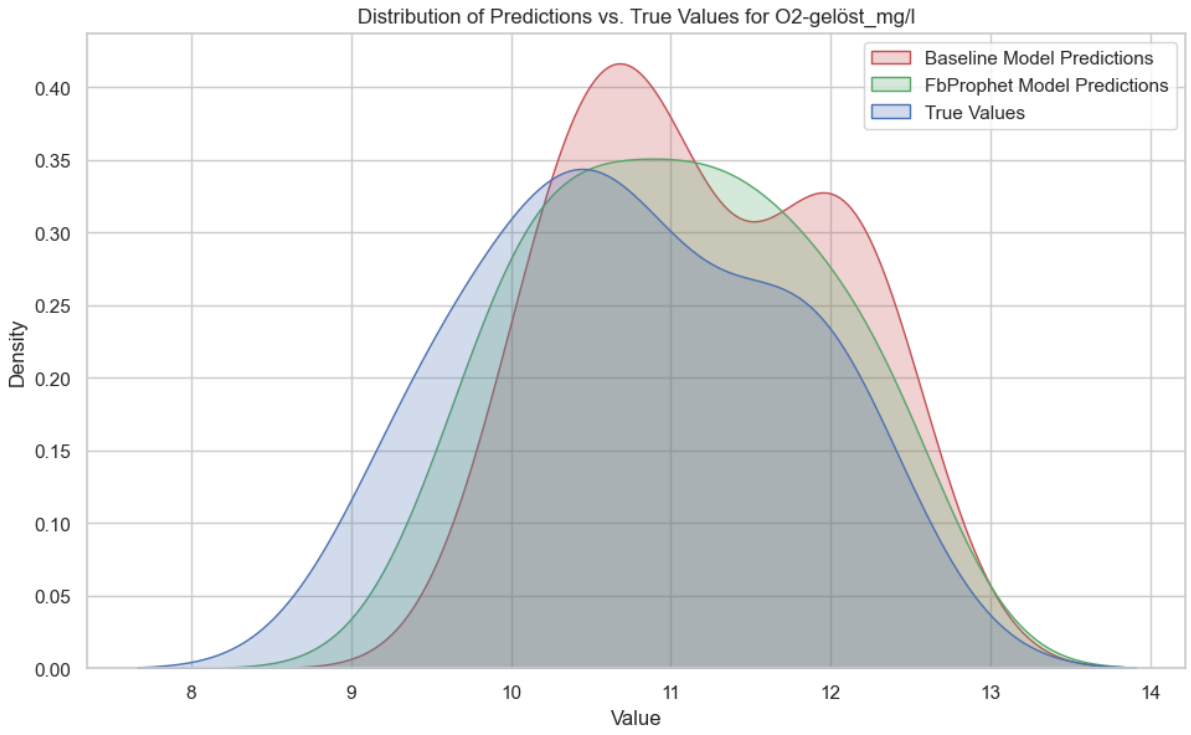

Figure 10: Distributions of Predictions vs Actual Values for the target diluted O2

Subsequently, the distributions of the values predicted by the FB Prophet model and baseline model are compared with the true values of diluted O2. Once again, initial findings from the line plot in Figure 10 prove to be accurate. The distribution of the predicted values of both the baseline model and the FB Prophet model is similar to the actual values. However, it is evident for this target variable as well that the baseline model deviates further from the true values than the FB prophet model. As already seen in the line plot, the baseline model predicts values that are higher than the actual values.

This is particularly apparent at the lower end of the value range and at the top of the distribution. In addition, although the baseline model is again able to model the slightly indicated second peak, both peaks are much more conspicuous than in the actual data.
The FB Prophet model also tends to predict higher values of diluted O2 than those found in the data. However, its predictions deviate less strongly from the actual values than those of the baseline model across the entire value range. Only the structure with two peaks, which is slightly recognizable in the real values, is not properly captured by the FB Prophet model. Overall, it can be concluded that the FB Prophet model is also to be preferred to the baseline model based on the distribution of the predicted diluted O2 values.

Scores						
Target Variable	MAE Prophet	MSE Prophet	MAPE Prophet	MAE Baseline	MSE Baseline	MAPE Baseline
Diluted O2	0.406266	0.231250	3.929181	0.452975	0.304431	4.409691
WaterTemp	1.130095	2.420209	18.980118	1.370061	3.068575	16.435779

Table 1: Scores

Finally, the models were analyzed using a number of selected key performance indicators for the target variables diluted O2 and water temperature. The Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Mean Absolute Percentage Error (MAPE) were utilized for this purpose. The scores can be found in Table 1.
The previous findings from the analysis of the density plots can be confirmed for the target variable diluted O2. While the FB Prophet model has an MAE of 0.4063 and an MSE of 0.2313 for this variable, the baseline model has slightly higher values with an MAE of 0.4530 and an MSE of 0.3044. The lower values for the FB Prophet model suggest that it can predict the true values better than the baseline model. At 3.9292, the MAPE value of the FB Prophet model is also slightly lower than that of the baseline model (4.4097) and therefore also supports this assumption.
There is also a difference in the key figures for the target variable water temperature between the FB Prophet model and the baseline model. The MAE (1.3701) and the MSE (3.0686) of the baseline model are also slightly higher than the MAE (1.1301) and the MSE

(2.4202) of the FB prophet model. However, there is a noteworthy difference to the first target variable regarding the MAPE. Here, the baseline model (16.4368) performs better than the FB Prophet model (18.9801). Nevertheless, the chosen performance indicators suggest a slightly better performance of the FB Prophet model compared to the baseline model.

6.3 Predictions of dataset without prior MICE imputation

The same analysis was carried out for the dataset not filled out using MICE imputation. This time, the task of dealing with missing values was left to the FB Prophet model.

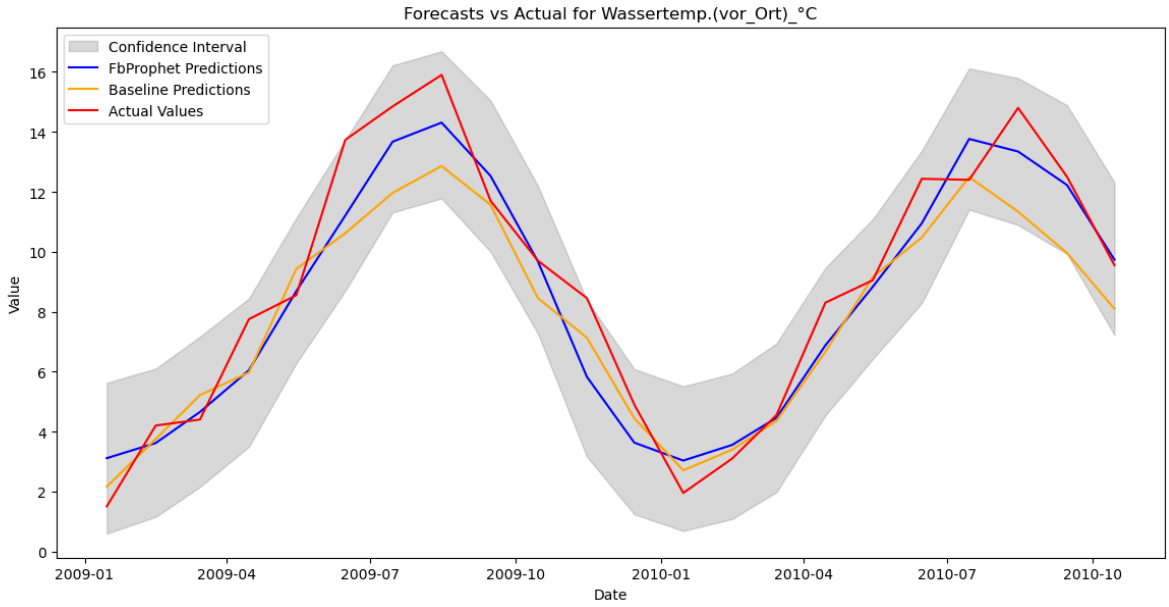

Figure 11: Forecasts vs Actual Values for the target water temperature without previous MICE imputation

Regarding the prediction of the water temperature, similar patterns can be recognized in Figure 11 as with the imputed data. Once again, both models are able to predict temperatures close to the actual values. However, it can be observed that the confidence interval for the predicted water temperatures without prior MICE imputation is slightly narrower and the predictions of both the FB Prophet model and the baseline model are slightly more accurate than it was the case for the predicted water temperatures from the dataset with prior MICE imputation.

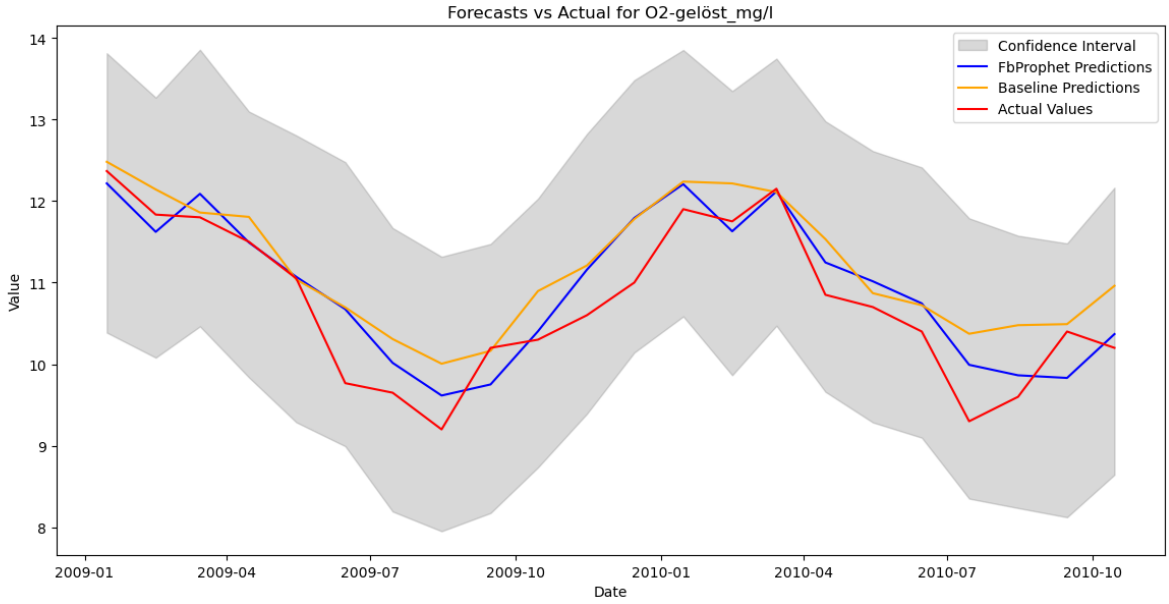

Figure 12: Forecasts vs Actual Values for the target diluted O2 without previous MICE imputation

For the prediction of the values for diluted O2, there are also only slight differences between the dataset with previous MICE imputation and the dataset without previous MICE imputation. In contrast to the predictions for water temperature, however, the predictions for diluted O2 have slightly worsened compared to the predictions based on the dataset with previous MICE imputation.

These conclusions are also supported by the comparisons of the densities between the model-based predictions and the real values as well as the metrics MAE, MSE and MAPE attached in the appendix.

6.4 Predictions of other potential targets

The baseline model and FB Prophet model were also used to predict the values of other potential targets. The figures discussed here, namely the comparisons between the forecasts and the actual values as well as the comparisons between the distributions of the forecasts and the actual values, were also created with the remaining scraped variables as targets. However, a brief examination of these graphs, which are shown in the appendix (see Figures 15, 16, 17, 18), quickly reveals that these other possible targets can hardly be used in a reasonable manner in the models at hand due to the strong divergence of

the forecasts from the actual values. For the remaining reasonably performing variables, no sufficient research based evidence could be found to suggest that they are suitable as indicators of water quality.

7 Discussion

Altogether, the models developed in this study provide reliable forecasts for water temperatures and diluted O2 levels based on the given data. The models' accuracy in predicting these key variables is highly encouraging and indicates a possible practical utility for water quality monitoring and management.

In particular, it is notable that the FB Prophet model proved to be an appropriate model for predicting the target variables. This is evident, on the one hand, in the aforementioned proximity of the predicted values to the actual values. Furthermore, this is expressed by the similarly strong and for the target water temperature marginally better performance of the FB Prophet model based on the dataset without prior MICE imputation of the missing data compared to the dataset supplemented by MICE imputation. This not only confirms the previously described strength of the FB Prophet model in dealing with missing values, but also suggests that this model might be suitable for similar future projects with partially incomplete datasets.

Furthermore, the tuning processes applied to the models resulted in clear improvements in the accuracy of predictions, particularly in the higher temperature range. This enhancement is significant given the challenges posed by due to climate change increasingly elevated water temperatures. The improved performance in this critical range highlights the model's robustness and its adaptability to varying temperature conditions, making it a valuable tool for anticipating and managing the impacts of higher water temperatures on water quality.

Finally, the code developed for this project is entirely reproducible, ensuring that the methodology and results can be consistently replicated and validated by other researchers. This property also allows others to build on the findings and to further refine the model.

8 Limitations and Future Work

Despite these encouraging conclusions, several limitations must be acknowledged.
Even with the aforementioned improvements, the prediction accuracy of the model still declines slightly when forecasting water temperatures at the peaks. Especially high water temperatures are affected by this slightly lower prediction accuracy. This is particularly concerning in the context of climate change, which is expected to lead to an increase in average water temperatures. As global temperatures rise, the model may experience more significant deviations in its predictions, potentially reducing its reliability and usefulness over time. Future work should focus on enhancing the model's robustness to accommodate these higher temperature ranges and mitigate the impact of climate-induced variations.
Furthermore, the data given for this project allowed for the identification of only two appropriate targets. This limitation stems from the restricted number of variables available within the dataset, as well as the potential infeasibility of identifying additional predictors. Should the former be the reason for the limited amount of suitable targets, the limited scope of the dataset constrains the model's capacity to account for other potentially influential factors that could improve prediction accuracy. Expanding the dataset to include a broader range of variables or employing more comprehensive data collection methods could help in identifying further predictors, thus enhancing the model's performance and applicability.

9 Conclusions

In conclusion, the findings of this project contribute significantly to our understanding of water quality prediction and monitoring. By validating previous research findings, we confirm the importance of water temperature and dissolved oxygen as key indicators of water quality that are well suited to forecasting. This validation enhances our ability to predict and assess water quality, providing valuable insights for environmental management and conservation efforts.

Moreover, the mitigation of one of the project's limitations, namely the challenge of accurately predicting water temperatures in higher temperature ranges, through hyperparameter tuning underscores the potential for further improvement in predictive modeling. This improvement does not only enhance the accuracy and reliability of the predictions conducted during the project, but also highlights the importance of fine-tuning methodologies to address challenges within the domain of water quality prediction.

However, the presence of only a small number of useful predictive targets also highlights the need for further research and broader data collection efforts. The limited scope of available targets underscores the complexity of water quality dynamics and emphasizes the necessity of comprehensive, more granular datasets and multifaceted approaches to fully capture and understand the factors influencing water quality.

Even with a more detailed data collection, it remains questionable whether a higher number of indicators of water quality can be derived. One of the most significant factors influencing water quality, usually in a negative way, is agriculture. A large proportion of environmental pollution that cannot be traced back to a specific cause, known as Nonpoint Source Pollution (NPS), originates from agriculture. NPS, and thus also agriculture, is a leading remaining cause of water quality problems around the world, it is known that these pollutants have harmful effects on drinking water supplies, recreation, fisheries and wildlife (Environmental Protection Agency, 2024a). Agricultural operations can considerably effect water quality, i.e. due to the extent of farm activities on the landscape, the soil-disturbing nature of those activities or pollutants from agricultural operations entering groundwater and degrading sources of drinking water (Environmental Protection Agency, 2024b). Furthermore, for as much as agriculture contributes to these changes, it also

impacted by the problem as a major user of water resources in many regions. Therefore agriculture is therefore both a generator of this locally highly variable source of pollution as well as a victim of these partly self-inflicted wounds (OECD, n.D.). This powerful dual role considerably complicates the research for further reliable indicators of water quality. Despite these difficulties, water quality research should by no means be neglected. On the contrary, methodological difficulties should encourage us to continue exploring the topic of water quality, which might play a greater role in the future than many of us would prefer.

A Appendix

A.1 Plotted Variables based on the dataset with prior MICE imputation

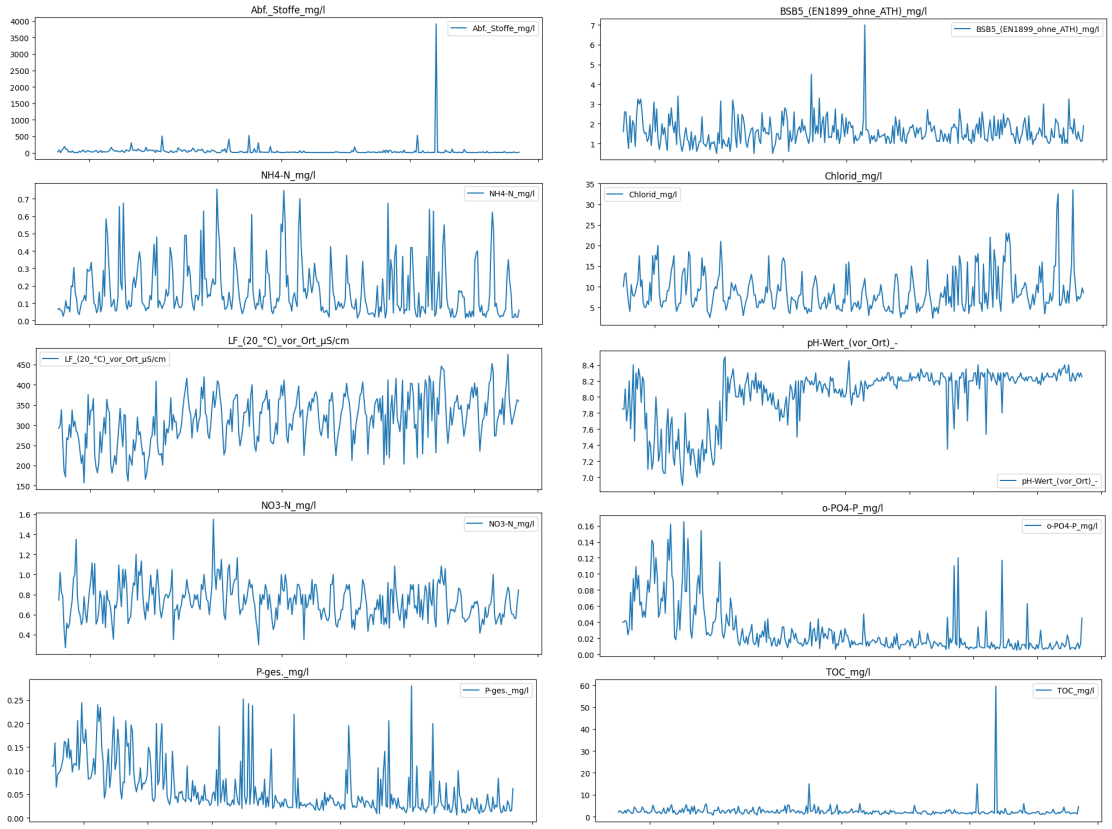

Figure 13: Plotted Variables based on the dataset with prior MICE imputation

A.2 Plotted Variable Distributions of the features

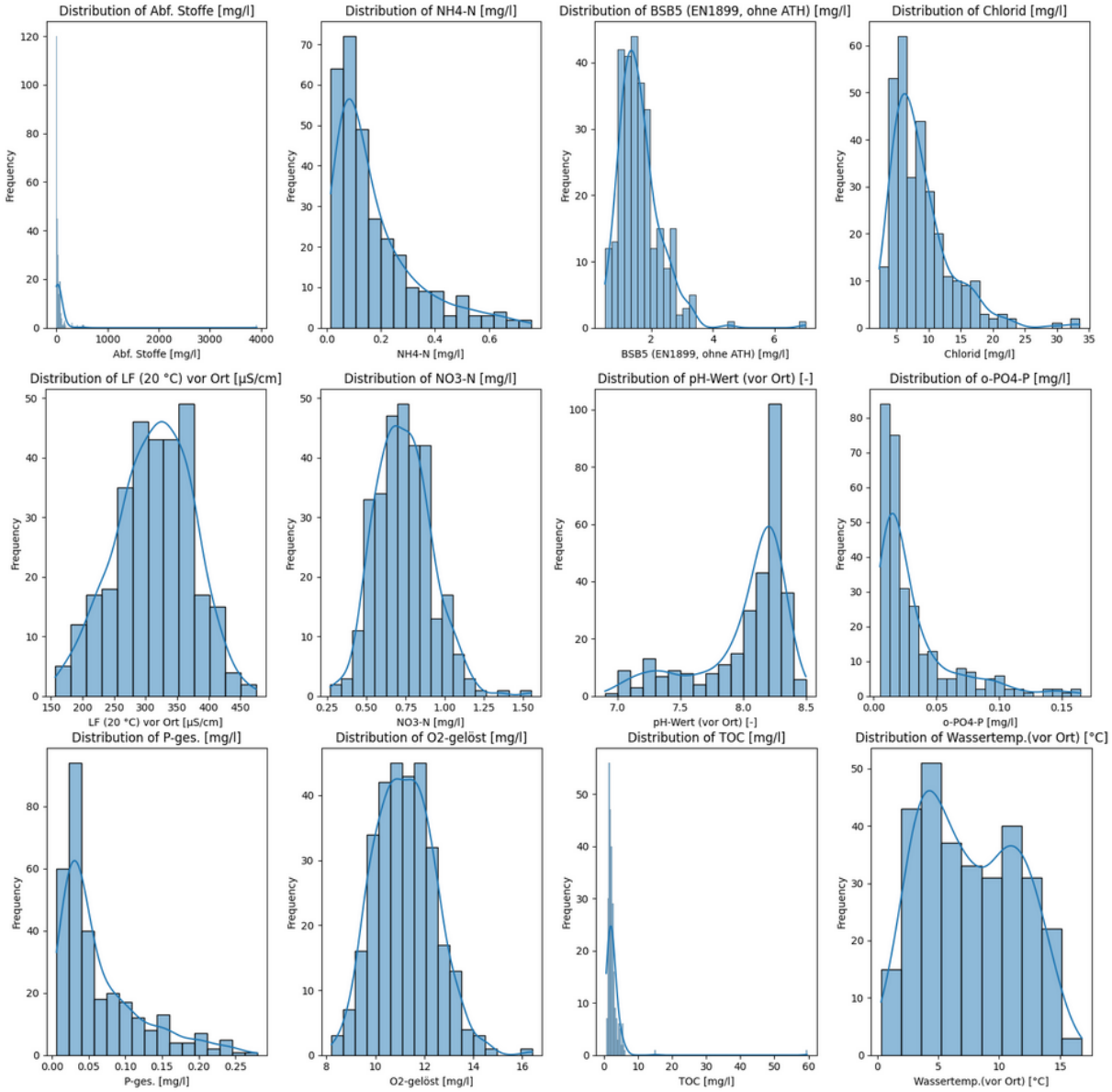

Figure 14: Plotted Variable Distributions of the features

A.3 Distributions vs Predictions

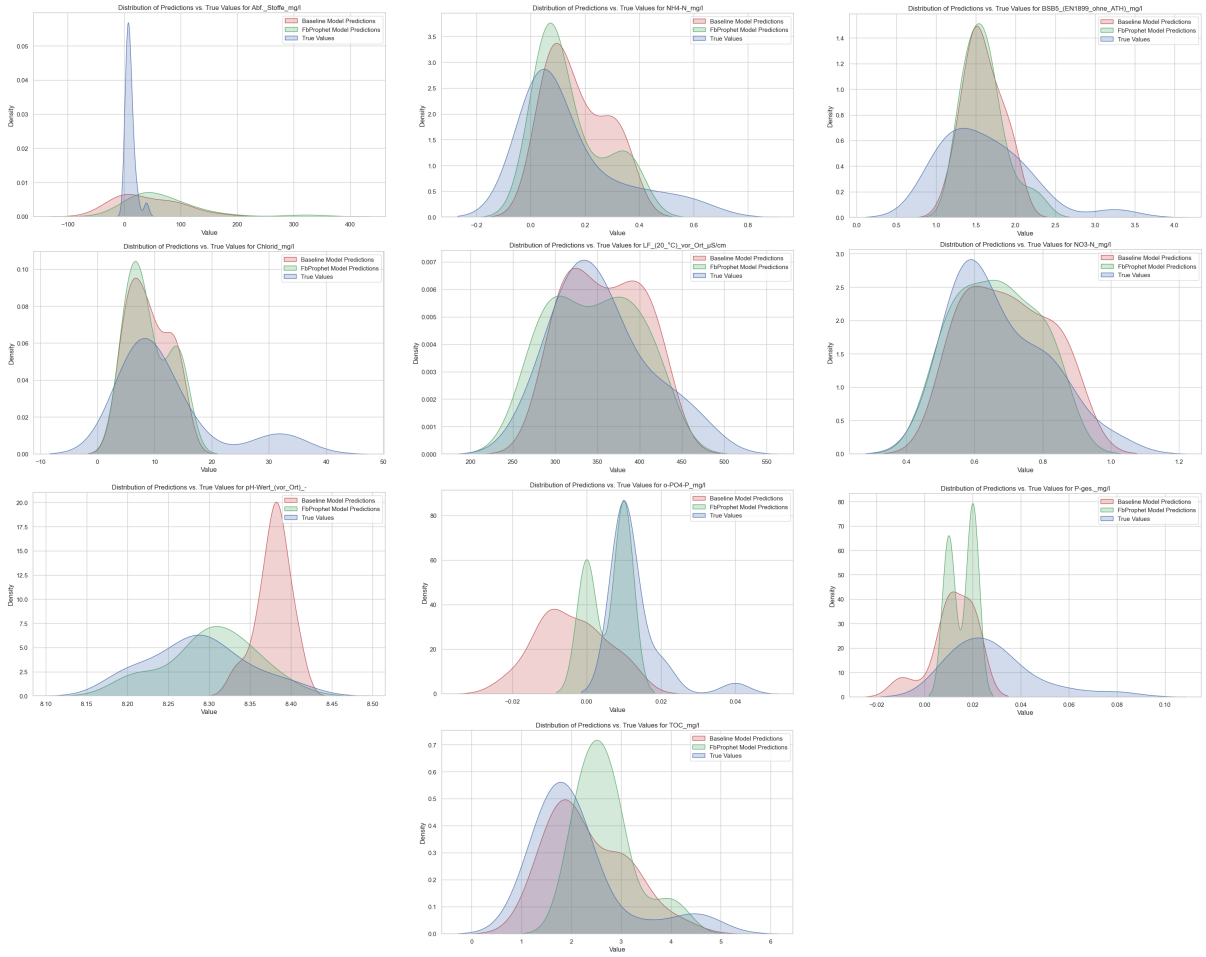

Figure 15: Distributions vs Predictions based on dataset with prior MICE imputation

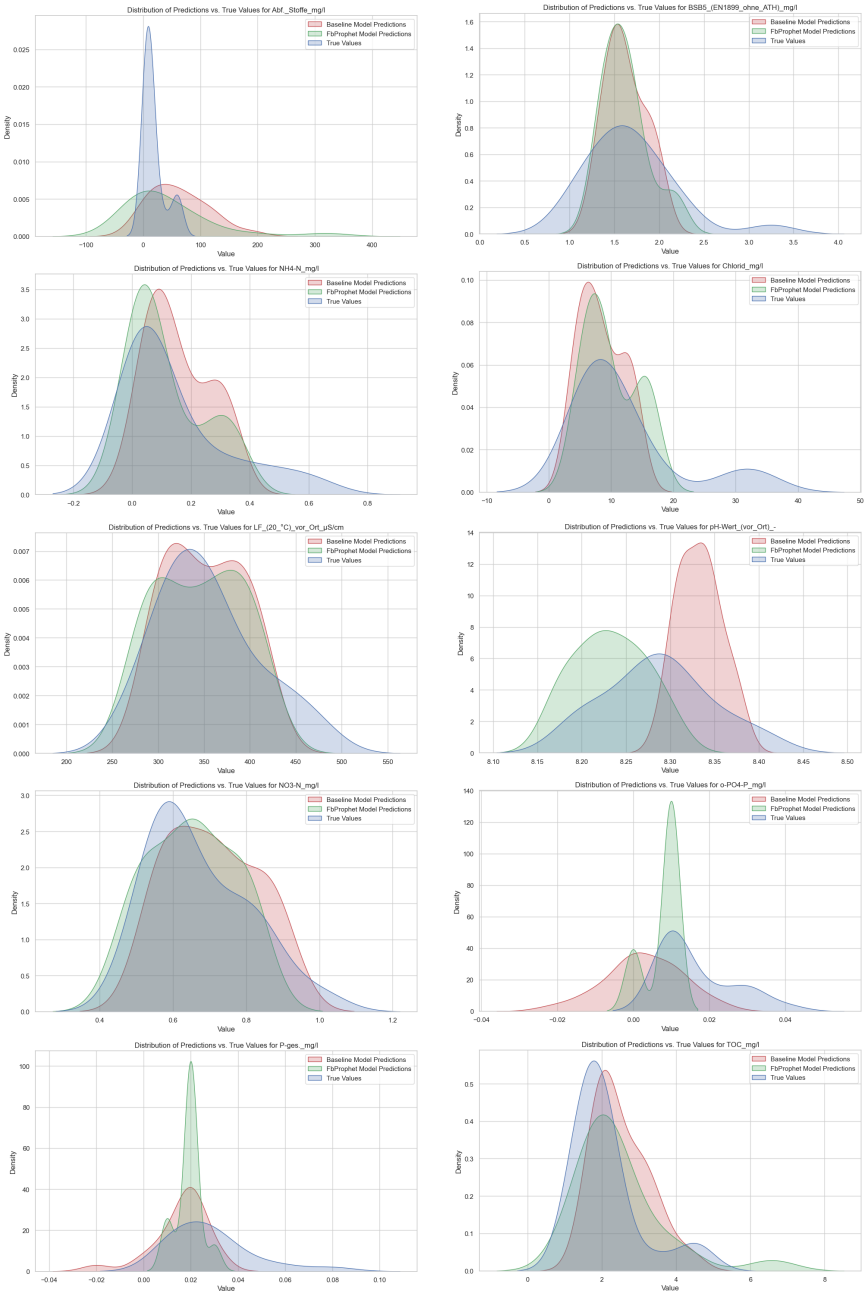

Figure 16: Distributions vs Predictions based on dataset without prior MICE imputation

A.4 Forecasts vs Actuals

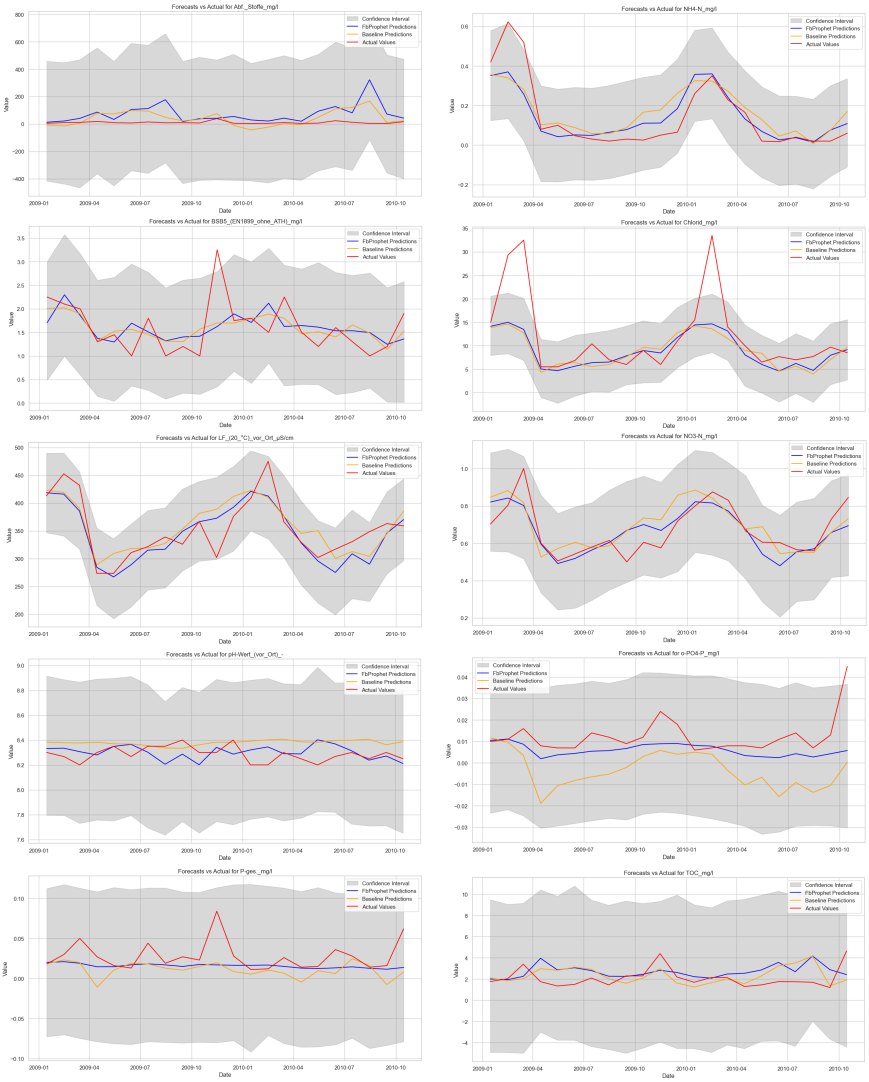

Figure 17: Forecasts vs Actuals based on dataset with prior MICE imputation

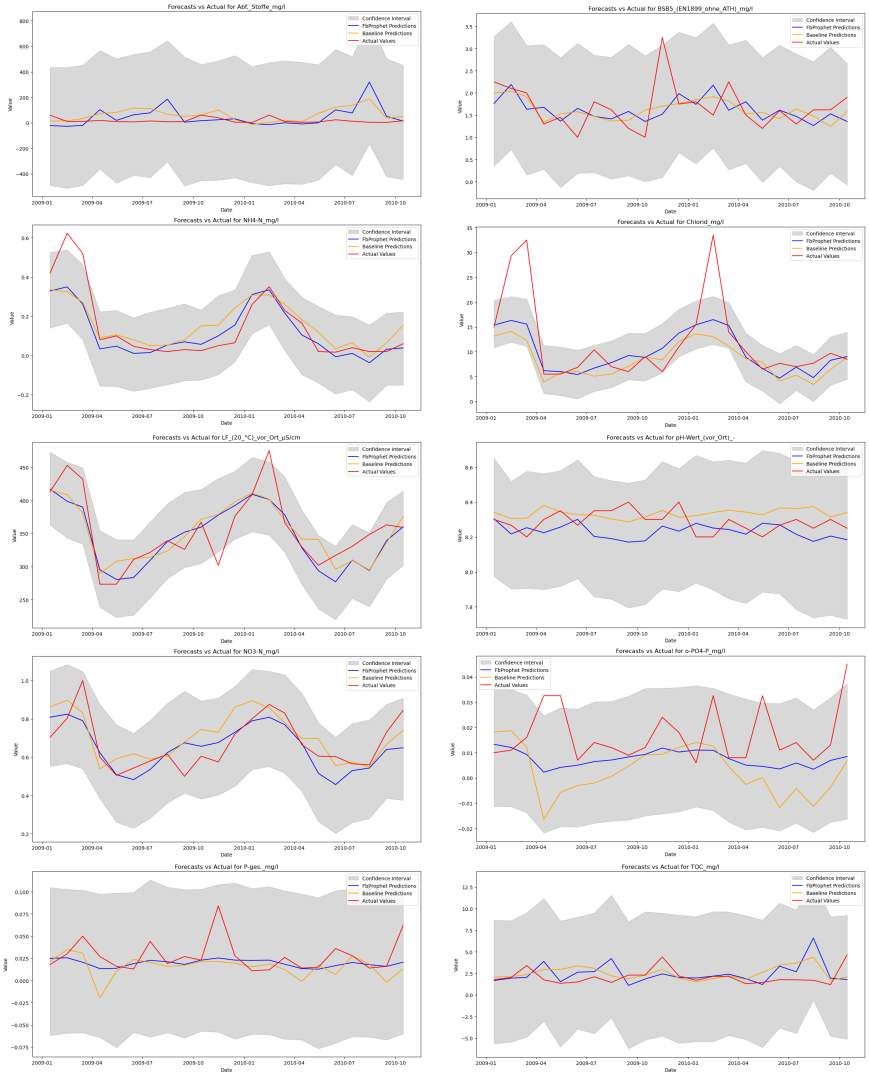

Figure 18: Forecasts vs Actuals based on dataset without prior MICE imputation

A.5 Full Scores

Scores based on dataset with prior MICE imputation						
Target Variable	MAE Prophet	MSE Prophet	MAPE Prophet	MAE Baseline	MSE Baseline	MAPE Baseline
Diluted O2	0.406	0.231	3.929	0.453	0.304	4.410
WaterTemp	1.130	2.420	18.980	1.370	3.069	16.436
AbfStoffe	61.838	8604.346	1104.822	44.187	0.304	4.401
NH4-N	0.062	0.009	92.766	0.080	0.012	145.960
BSB5	0.367	0.251	23.356	0.304	0.198	19.854
Chlorid	3.579	44.568	22.287	4.077	49.752	27.216
LF(20°C)	23.168	942.353	6.491	27.204	1140.283	7.802
NO3-N	0.063	0.007	9.192	0.081	0.009	12.091
pH-value	0.072	0.008	0.875	0.101	0.013	1.222
o-PO4-P	0.007	0.000	47.305	0.016	0.000	143.553
P-ges	0.014	0.000	37.930	0.018	0.001	58.782
TOC	1.054	1.675	58.181	0.967	1.468	48.514

Table 2: Scores based on dataset with prior MICE imputation

Scores based on dataset without prior MICE imputation						
Target Variable	MAE Prophet	MSE Prophet	MAPE Prophet	MAE Baseline	MSE Baseline	MAPE Baseline
Diluted O2	0.340	0.173	3.285	0.456	0.315	4.453
WaterTemp	1.017	1.599	17.280	1.331	2.903	16.176
AbfStoffe	56.978	7978.364	848.564	55.465	5054.899	732.450
NH4-N	0.061	0.008	85.273	0.074	0.011	128.686
BSB5	0.379	0.267	22.363	0.303	0.195	18.639
Chlorid	3.448	37.525	23.655	4.209	52.817	27.601
LF(20°C)	24.289	1093.558	6.757	25.698	1132.949	7.298
NO3-N	0.070	0.009	10.312	0.082	0.010	12.396
pH-value	0.082	0.010	0.983	0.073	0.007	0.878
o-PO4-P	0.010	0.000	49.677	0.016	0.000	101.931
P-ges	0.012	0.000	37.428	0.016	0.001	52.724
TOC	1.113	2.620	56.661	0.978	1.581	51.420

Table 3: Scores based on dataset without prior MICE imputation

B Electronic appendix

Data, code and figures are provided in electronic form, via our GitHub repository.

References

Ahmed, T., Zounemat-Kermani, M. and Scholz, M. (2020). Climate change, water quality and water-related challenges: A review with focus on pakistan, *International Journal of Environmental Research and Public Health* **17**(22): 8518. Accessed: June 4th, 2024.
URL: *https://doi.org/10.3390/ijerph17228518*

Arnell, N. W. and Gosling, S. N. (2016). The impacts of climate change on river flow regimes at the global scale, *Journal of Hydrology* **541**: 767–780.

Boyd, C. E. (2020). *Solar Radiation and Water Temperature*, Springer, Cham.
URL: *https://doi.org/10.1007/978-3-030-23335-8*$_2$

Croghan, D., Van Loon, A. F., Sadler, J. P., Bradley, C. and Hannah, D. M. (2018). Prediction of river temperature surges is dependent on precipitation method, *Hydrological Processes* . First published: 29textsuperscriptth October 2018.
URL: *https://doi.org/10.1002/hyp.13317*

Environmental Protection Agency (2024a). Basic Information about Nonpoint Source (NPS) Pollution, `https://www.epa.gov/nps/basic-information-about-nonpoint-source-nps-pollution`. Accessed: June 4th, 2024.

Environmental Protection Agency (2024b). Nonpoint Source: Agriculture, `https://www.epa.gov/nps/nonpoint-source-agriculture`. Accessed: June 4th, 2024.

Facebook (2024a). Multiplicative Seasonality, `https://facebook.github.io/prophet/docs/multiplicative_seasonality.html`. Accessed: June 4th, 2024.

Facebook (2024b). Prophet: Forecasting at Scale - Homepage, `https://facebook.github.io/prophet/`. Accessed: June 4th, 2024.

Facebook (2024c). Prophet: Forecasting at Scale - Hyperparameter Tuning, `https://facebook.github.io/prophet/docs/diagnostics.html#hyperparameter-tuning`. Accessed: June 4th, 2024.

Facebook (2024d). Prophet: Forecasting at Scale - Seasonality, Holiday Effects, and Regressors, `https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html`. Accessed: June 4th, 2024.

Facebook (2024e). Seasonality, Holiday Effects, and Regressors: Additional Regressors, `https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html#additional-regressors`. Accessed: June 4th, 2024.

Gewässerkundlicher Dienst Bayern (2024). Homepage Gewässerkundlicher Dienst Bayern, `https://www.gkd.bayern.de/`. Accessed: June 4th, 2024.

Harvey, R., Lye, L., Khan, A. and Paterson, R. (2013). The influence of air temperature on water temperature and the concentration of dissolved oxygen in newfoundland rivers, *Canadian Water Resources Journal* pp. 171–192. Published online: 23rd Jan 2013.
URL: *https://doi.org/10.4296/cwrj3602849*

IPCC (2021). *Climate Change 2021: The Physical Science Basis*, Cambridge University Press.

Lin, L., Yang, H. and Xu, X. (2022). Effects of water pollution on human health and disease heterogeneity: A review, *Frontiers in Environmental Science* **10**: 880246.

Mateo-Sagasta, J., Zadeh, S. M., Turral, H. and Burke, J. (eds) (2017). *Water pollution from agriculture: a global review*, FAO and IWMI.

OECD (n.D.). OECD - Water and agriculture, `https://www.oecd.org/agriculture/topics/water-and-agriculture/`. Accessed: June 4th, 2024.

PyPI (2024). MICEforest - Project Description, `https://pypi.org/project/miceforest/`. Accessed: June 4th, 2024.

Whitehead, P. G., Wilby, R. L., Battarbee, R. W., Kernan, M. and Wade, A. J. (2009). A review of the potential impacts of climate change on surface water quality, *Hydrological Sciences Journal* **54**(1): 101–123.
URL: *https://doi.org/10.1623/hysj.54.1.101*

WHO (2023). WHO - Drinking Water, https://www.who.int/en/news-room/fact-sheets/detail/drinking-water. Accessed: June 4th, 2024.