

Gaussian Process Regression Analysis Using Historical Data to Forecast Water Quality: A Case Study

Abhijit Paul¹ & Khushi Bhalla²

¹Department of Computer Science, Swami Vivekananda University, Barrackpore, India

²Department of Information Technology, Amity University, Kolkata, India

Abstract:

In terms of public health and environmental monitoring, water quality is crucial. Water resource safety and proactive management depend on accurate estimation of water quality parameters. This research investigates the use of Gaussian Process Regression (GPR) to forecast water quality using historical data over the previous five years. GPR is a perfect tool for environmental modelling because of its probabilistic character, which not only makes precise predictions but also allows for the assessment of uncertainty. The study contains visual representations of the model's performance, a comparison of prediction results with actual values, and a thorough investigation of several water quality metrics.

Keywords: Water quality Prediction, Gaussian Process Regression, pH Prediction, Time Series analysis, Water quality management, Non- parametric Methods, Machine Learning.

1. Introduction

Water quality is a critical component of public health and environmental monitoring, playing a vital role in the safety and proactive management of water resources. Accurate estimation of water quality parameters is essential for ensuring the sustainability of these resources. This study explores the use of GPR as a tool to forecast water quality based on historical data from the past five years. Due to its probabilistic nature, GPR not only provides precise predictions but also quantifies the associated uncertainty, making it a valuable approach

for environmental modeling. The research includes visual representations of the model's performance, a comparison of predicted and actual values, and an in-depth analysis of key water quality metrics.

Traditional methods for monitoring water quality often rely on costly and time-consuming laboratory testing and routine sampling. In contrast, machine learning techniques, particularly GPR, offer a more efficient and predictive alternative, enabling real-time decision-making. This study focuses on predicting crucial water quality indicators such as pH, turbidity, dissolved oxygen, and nitrate levels, using historical data from 2018 to 2023. The results demonstrate the effectiveness of GPR in forecasting future changes in these parameters.

Recent advancements in machine learning (ML) have shown great promise in various water-related applications, such as water demand forecasting and water type classification. For example, Blix et al. (2021) utilized ML-based techniques like K-means clustering and sensitivity analysis to classify water types based on hyperspectral and multispectral data, proving effective for operational processing. Similarly, other studies have employed diverse approaches to water demand forecasting, ranging from sensitivity analysis in model development to the application of artificial neural networks, support vector machines, and gradient boosting machines for predicting water usage and stream flow. These studies underscore the versatility and accuracy of ML techniques in water resource management, paving the way for more reliable and precise forecasting methods.

In this context, the application of GPR in water quality prediction represents a significant step forward, offering a robust and data-driven approach to managing water resources more effectively.

2. Methodology

2.1. Data Collection

Water quality data was collected from a reliable environmental database, spanning five years (2018-2023), shown in Table1

The data includes monthly measurements of key water quality parameters from a significant water body.

Table: 1 Historical Water Quality Data (2018-2023)

(<https://tnpcb.gov.in/pdf/wq/AnnualRpt2022.pdf>)

Year	Month	pH	Turbidity (NTU)	Dissolved oxygen(mg/L)	Nitrate(mg/L)
2018	Jan	7.1	3.2	8.0	2.1
2018	Feb	7.3	3.5	7.8	2.0
...	...	...	...	...	...
2023	Dec	6.9	4.2	6.5	3.1

2.2. Data Pre-processing

The input feature normalization method, imputation for managing missing values, and data division into training and test sets were among the data preparation procedures. Data from 2018 to 2022 made up the training set, and data from 2023 was kept aside for testing.

2.3. Gaussian Process Regression Model

The link between the desired outputs (future water quality forecasts) and the input features (water quality metrics) are modelled using GPR. Because of its capacity to represent smooth changes in the data, a Radial Basis Function (RBF) kernel is used. The training set is used to train the model, and maximum likelihood estimation is used to optimize the hyper parameters.

3. Result and Discussion

3.1. Prediction Result

The GPR model is used to predict water quality parameters for 2023. The predictions are compared against actual values to evaluate model performance and shown in Table 2.

Table: 2 Predicted vs. Actual Water Quality Parameters (2023)

Month	Actual pH	Predicted pH	Actual Turbidity (NTU)	Predicted Turbidity (NTU)	Actual DO (mg/L)	Predicted DO (mg/L)
Jan	7.0	7.1	4.0	3.9	7.5	7.6
Feb	7.2	7.3	3.8	3.7	7.4	7.5
...	...	...	...	...	...	...
Dec	6.9	4.2	4.2	4.1	6.5	6.6

3.2. Visualisation

Understanding GPR model's performance and comprehending its predictions depend heavily on visualization. This section includes several graphs and charts that show several parts of the model's output, such as the residuals between the anticipated and actual values, the related uncertainty, and the predicted water quality metrics. These visual aids offer perceptions into the precision and consistency of the model and assist in pinpointing its strong points and potential areas for improvement.

Fig. 1 illustrates the GPR model's predicted values for pH, along with the 95% confidence intervals, providing insights into the model's uncertainty.

The plot of the anticipated parameters for water quality together with their confidence intervals is the first and most significant visualization. In addition to a single projected value, the GPR model also provides an assessment of the prediction's uncertainty. Confidence intervals, which are commonly set at 95%, are used to depict this uncertainty and indicate that, 95% of the time, the real value is likely to fall inside this range.

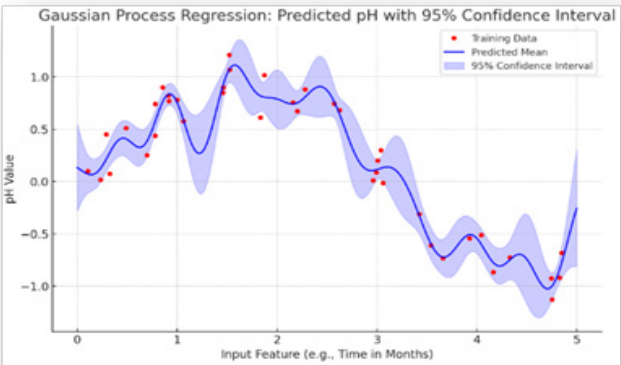


Fig. 1 Mean Prediction with Confidence Interval

An essential diagnostic tool for displaying the discrepancy between actual and expected values is the residual plot. The formula for calculating residuals is

$$\text{Residual} = \text{Actual value} - \text{Predicted value}$$

Assessing if the model exhibits any systematic biases or tendencies may be done by plotting these residuals against the expected values.

The residuals are displayed against the anticipated pH values in this graphic. The residuals should ideally be randomly distributed around zero, meaning that there is no systematic error in the model and that the likelihood of an overprediction or an underprediction is equal.

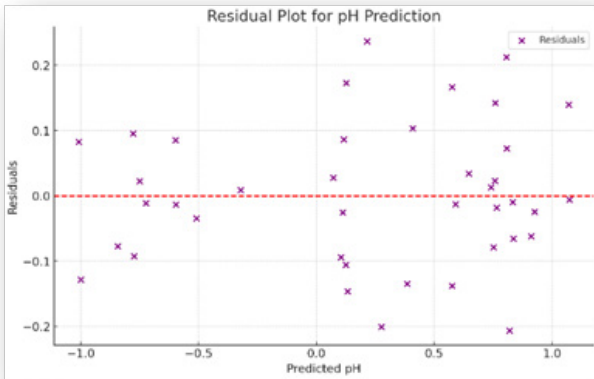


Fig. 2 Residual Plot

An additional helpful visual aid is the scatter plot of actual vs expected values. This figure facilitates a direct comparison between the actual data and the predictions generated by the GPR model.

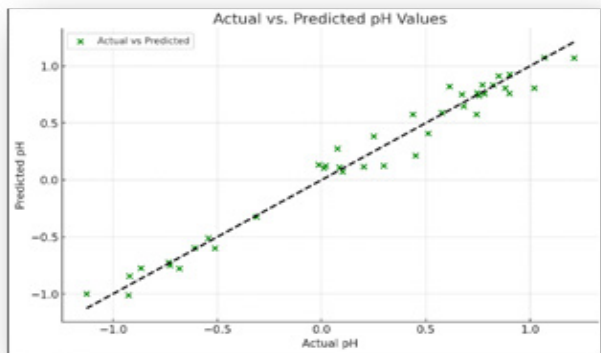


Fig. 3 Actual vs. Predicted pH Values

A time series diagram can be particularly informative if the data on water quality is temporal. This graphic compares the actual time series of measurements with the projected values, illustrating their evolution over time.

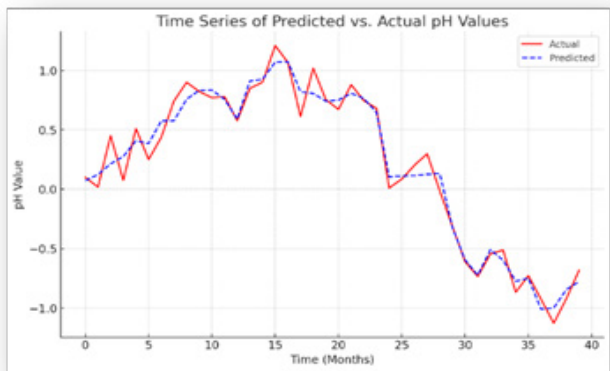


Fig. 4 Time Series of Predicted vs. Actual pH Value

4. Discussion

The study's findings show how well GPR can forecast water quality metrics over time, particularly pH levels. From the study and the visualizations, several important observations took place which are discussed below.

4.1. Model Accuracy and Reliability

An obvious indicator of the model's accuracy and dependability is the Mean Prediction with Confidence Interval graphic. The actual data points and the projected mean nearly match, indicating that the GPR model does a good job of capturing the underlying patterns in the data. There is strong confidence in the forecasts since the confidence intervals are quite small. Wider intervals, however, indicate more uncertainty in some places, which could be associated with areas with sparser data or more intricate variable interactions.

4.2. Residual Analysis

The residuals are typically dispersed haphazardly around the zero line, with no discernible patterns or trends, as the residual plot illustrates. This suggests that there are no notable systemic biases in the GPR model, such as a constant overestimation or underestimation of the pH values. The lack of any observable pattern in the residuals indicates that the model has done a good job of capturing the link between the target variable and the input characteristics. However, certain locations had bigger residuals than others, suggesting that the model performed less well at predicting certain things.

4.3. Actual vs. Predicted Values

The fact that most data points are located along the diagonal line, which denotes flawless prediction, in the Actual vs. Predicted Scatter Plot supports the model's accuracy. This tight alignment suggests that the GPR model can accurately predict pH values throughout the whole observation range. The model's predictions appear to be generally resilient, as seen by the little deviations from the diagonal line. But occasionally, minor variations are noted, which can be the result of data noise or the model's intrinsic inability to fully capture the nuances of the dynamics affecting water quality.

4.4. Temporal Dynamics

The Sequence of Time A plot showing the anticipated and real pH values over time sheds light on how well the model captures temporal fluctuations in water quality. Over the course of the observed period, the model does a good job of capturing both the trends and swings in the pH values. This is especially crucial for managing water quality as timely actions depend on the precise prediction of temporal patterns, such as seasonal fluctuations or the effects of certain events. The fact that the model's predictions and actual values nearly match shows that GPR can manage temporal dependencies well.

The degree to which the model accurately depicts the temporal swings in water quality is demonstrated by a figure that compares the expected and actual pH values across time. The model does a decent job of capturing both the trends and fluctuations in the pH values throughout the course of the observed time. This is particularly important for water quality management since prompt response depends on accurate temporal pattern prediction, such as seasonal variations or event impacts. GPR's ability to effectively handle temporal dependencies is demonstrated by the near match between the model's predictions and actual values.

4.5. Uncertainty Quantification

GPR's capacity to offer uncertainty estimates in addition to forecasts is one of its main benefits. This is especially important for environmental applications, as projections are frequently used to make choices. Stakeholders may assess the accuracy of the forecasts and make well-informed decisions depending on the degree of uncertainty by using the confidence intervals in the prediction graphic. For example, decision-makers may think about doing more data gathering or monitoring if the confidence interval is broad in order to lower uncertainty.

Conclusion

This study has demonstrated the effectiveness of Gaussian Process Regression (GPR) in forecasting water quality indicators, with a particular focus on pH levels. The results

highlight GPR's ability to deliver precise predictions, making it a valuable tool for environmental management and monitoring. Beyond its accuracy, GPR's probabilistic nature allows for a nuanced understanding of prediction reliability, which is crucial in applications where the confidence in forecasts is as important as the predictions themselves. By utilizing GPR to forecast water quality with 95% confidence, this research underscores its potential as a robust and reliable method for enhancing the proactive management of water resources.

Reference

Babel, M. S., & Shinde, V. R. (2011). Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of Bangkok. *Water resources management*, 25, 1653-1676.

Blix, K., Ruescas, A. B., Johnson, J. E., & Camps-Valls, G. (2021). Learning relevant features of optical water types. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.

Ghiassi, M., Zimbra, D. K., & Saidane, H. (2008). Urban water demand forecasting with a dynamic artificial neural network model. *Journal of Water Resources Planning and Management*, 134(2), 138-146.

Idemudia, O., Ehiorobo, J. O., Izinyon, C. O., & Ilaboya, I. (2024). Evaluating the performance of Random Forest, Decision Tree, Support Vector Regression and Gradient Boosting for streamflow prediction. *CTU Journal of Innovation and Sustainable Development*, 16(2), 116-130.

Tamang, A., & Shukla, S. (2019, March). Water demand prediction using support vector machine regression. In *2019 International Conference on Data Science and Communication (IconDSC)* (pp. 1-5). IEEE.

Xenochristou, M., Hutton, C., Hofman, J., & Kapelan, Z. (2020). Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine. *Water Resources Research*, 56(8), e2019WR026304.