Article

Not peer-reviewed version

Improving Inland Water Quality Predictions Using Machine Learning and Global Dissolved Oxygen Datasets

Ayush Prasad *, Snehal Verma , Mohammad Aatish Khan

Posted Date: 9 December 2024

doi: 10.20944/preprints202412.0671.v1

Keywords: Water quality; Dissolved Oxygen; Machine Learning

 doi:10.20944/preprints202412.0671.v1

Article

Improving Inland Water Quality Predictions Using Machine Learning and Global Dissolved Oxygen Datasets

Ayush Prasad *, Snehal Verma and Mohammad Aatish Khan

NatureDots, India

* Correspondence: ayush.prasad@naturedots.com

Abstract: Dissolved Oxygen (DO) is a crucial parameter for monitoring inland water systems' health, yet predicting its variability accurately is challenging due to spatial variability across different water bodies. Traditional process-based models lack sufficient accuracy, and while recent data-driven methods and deep learning have shown promise, they are typically site-specific and require retraining for each new location. This limitation is particularly problematic in areas lacking ground observations. In this work, we address these challenges by using deep learning models trained on global water quality datasets. These models were pretrained on global historical water quality observations and subsequently fine-tuned for specific regions. This approach demonstrated improved results compared to using models trained exclusively on regional data and offered a more robust solution for predicting Dissolved Oxygen levels in diverse inland water systems.

Keywords: water quality; dissolved oxygen; machine learning

1. Introduction

Predicting the variability of Dissolved Oxygen (DO) levels in inland water systems is a critical environmental challenge. DO is an essential parameter that reflects the health of aquatic ecosystems [1]. Its levels can be indicative of the overall quality of the water, impacting aquatic life and the broader environmental balance. For instance, in aquaculture, measuring DO plays a crucial role in determining the water quality and health of the water ecosystem for the fisheries-produce. Rapid or sudden fluctuations in DO over prolonged periods can lead to fishery mortality and impact other biodiversity health as well. Thus detecting timely risks on occasions where DO is at below danger level, depending on the type of water ecosystem and purpose is crucial. In the context of fisheries, this information enables fish farmers to take timely actions and save their fisheries stock, which is essential for their sustenance and food security [2]. By forecasting and monitoring the changes in DO parameters, fish farmers can allocate the right resources and take correct actions to derisk their production system. Similarly, we have observed how DO levels are critical for ensuring the productivity and health of water ecosystems, especially freshwater ecosystems such as wetlands, lakes and rivers. In the event of severe droughts or heat waves, before the impacts of water scarcity (quantity) come into play, the effect of changing water chemistry or quality can be observed. DO is a fundamental physical water quality parameter that indicates how much oxygen is available for all the chemical and biological processes to take place and for sustaining life in water. The mass mortality event of Freshwater river dolphins of the Amazon in 2023 where 120 critically endangered species of river dolphins along with other biodiversity perished is an active example [3]. The present approach of measuring DO manually is ineffective and very expensive However, accurately predicting DO variability is complex due to the diverse and dynamic nature of different water bodies. Traditional methods for predicting DO have primarily relied on process-based models. These models are rooted in the understanding of the physical, chemical, and biological processes that affect DO levels. While these models can provide insights into DO dynamics, they often lack the accuracy needed for precise predictions [4]. This shortcoming is primarily due to their inability to fully account for the spatial variability and complex interactions in natural water systems. In recent years, the rise of data-driven methods, particularly deep learning techniques, has offered new avenues for predicting DO levels. These approaches, which

 doi:10.20944/preprints202412.0671.v1

include algorithms like Random Forest, Long Short-Term Memory (LSTM) networks, and Recurrent Graph Convolutional Networks, have shown promising results [5–7]. They leverage large datasets to identify patterns and relationships that may not be apparent through traditional modelling. However, these methods are typically tailored to specific sites. This means they require extensive retraining and calibration when applied to new locations, which can be a significant hurdle, especially in regions lacking sufficient ground observations. In this work, we focus on using models that are trained on global historical water quality datasets. This pretraining allows the models to learn from a broad range of conditions and scenarios, capturing the diverse dynamics of various water systems. Once trained, the models are then fine-tuned with data specific to a given region. This approach, combining global learning with regional fine-tuning, presents an approach that is both robust and adaptable.

2. Data and Study Area

We constructed a global dataset that integrates historical dissolved oxygen (DO) data from inland water bodies with a variety of meteorological and topological parameters. Within the United States, our dataset includes information from 35 freshwater streams, sourced from the National Ecological Observatory Network (NEON). Additionally, we included data from the Finnish Environment Institute (SYKE) and the Swedish Meteorological and Hydrological Institute (SMHI) in Europe. Complementing these public datasets, we have also included proprietary data collected by NatureDots in India and the US. In addition to the in situ dissolved oxygen observations, we retrieved meteorological parameters listed below from the ERA5 reanalysis product and computed the Normalized Difference Chlorophyll Index (NDCI) and Normalized Difference Turbidity Index (NDTI) for the sites using Sentinel 2.

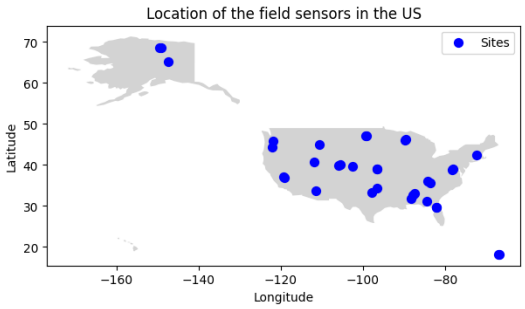

Figure 1. Location of NEON field sensors in the US

Table 1. Input Features for the Study

Source	Feature
ERA5	Max temperature
	Min temperature
	Precipitation sum
	Max wind speed
	Shortwave radiation sum
Sentinel 2	Chlorophyll Index
	Turbidity Index
Topographical Features	Slope
	Roughness

3. Methodology

3.1. Remote Sensing Metrics

We computed two remote sensing metrics, the Normalized Difference Chlorophyll Index (NDCI) and the Normalized Difference Turbidity Index (NDTI), using Sentinel-2 satellite imagery. These

 doi:10.20944/preprints202412.0671.v1

indices provide valuable information about the chlorophyll content and turbidity levels in water bodies, which can influence DO dynamics. The NDCI is calculated using the following formula:

$$NDCI = \frac{R_{705} - R_{665}}{R_{705} + R_{665}} \tag{1}$$

where R_{705} and R_{665} represent the reflectance values at wavelengths 705 nm and 665 nm, respectively. The NDTI is computed as follows:

$$NDTI = \frac{R_{1600} - R_{820}}{R_{1600} + R_{820}} \tag{2}$$

where R_{1600} and R_{820} denote the reflectance values at wavelengths 1600 nm and 820 nm, respectively. These remote sensing metrics were included as input features in our models to provide additional information about the water quality conditions at the study sites.

3.2. Model Selection and Pretraining

1. **Long Short-Term Memory (LSTM):** LSTM, a variant of recurrent neural networks (RNNs), its architecture enables it to capture long-term dependencies, making it exceptionally suitable for predicting variables like DO that are influenced by historical data. In our study, the LSTM model was trained to recognize and interpret complex temporal patterns in DO levels, utilizing sequences of past observations to enhance the accuracy of future predictions.
2. **Recurrent Graph Convolutional Network (RGCN):** RGCN, an advancement of graph convolutional networks (GCNs), incorporates a recurrent structure to effectively manage spatial data. This model is adept at understanding spatial relationships and interactions, an essential aspect for analyzing water systems with diverse geographical features. In our application, the RGCN was utilized to construct a spatial network of water bodies, with a 200 km threshold applied to define connections between sites. This approach enabled the model to account for spatial dependencies in predicting DO levels, offering insights into the spatial dynamics of aquatic ecosystems.
3. **Random Forest:** This ensemble learning method operates by constructing multiple decision trees during the training phase. Random Forest is known for its versatility in handling both numerical and categorical data, effectively dealing with non-linear relationships within datasets. In this study, Random Forest was employed to analyze complex interactions within the environmental data, leveraging its ensemble nature to improve predictive accuracy and robustness.

Each model was pretrained on a global historical water quality dataset. This comprehensive dataset included a wide range of environmental conditions, providing a rich base for the models to learn diverse water system dynamics.

3.3. Fine-Tuning for Regional Specificity

Following pretraining, the models were fine-tuned using data specific to individual regions. This was done by calibrating each model with localized datasets, to adapt them to the unique environmental and topographical characteristics of each study area.

3.4. Cross-Validation Technique

For validating our models, we employed the Leave-One-Out Cross-Validation (LOOCV) method. In LOOCV, one data point is used as the validation set while the rest of the data is used for training. This process is repeated such that each data point is used once as the validation set. LOOCV is particularly effective in our study as it maximizes both the training and testing data, ensuring a thorough evaluation of the model's performance across all available data.

 doi:10.20944/preprints202412.0671.v1

4. Results

We evaluated the models' performance after fine-tuning them on new sites, using only 20 percent of the data compared to the amount used during pretraining. This assessment was executed within a methodical leave-one-out cross-validation (LOOCV) framework, where each test site had to be predicted independently, ensuring a comprehensive and unbiased evaluation of the model's predictive power. Table 2 shows the average Root Mean Square Error (RMSE) values obtained during the LOOCV process, serving as a quantitative measure of each model's prediction accuracy. Among the models, the Long Short-Term Memory (LSTM) network performed the best, achieving the lowest average RMSE of 1.14. This result underscores LSTM's robustness in handling the variability of environmental data across different geographical locations. However, it is important to note that the performance of LSTM was not uniform across all sites. The RMSE for LSTM varied from as low as 0.4 to as high as 2.1 in certain locations. This variation suggests that some sites posed more challenges than others, potentially due to factors such as complex water dynamics, varying data quality, or specific environmental conditions at these locations. These findings highlight the need for site-specific considerations in model deployment and the potential benefits of further tailoring machine learning models to address the unique characteristics of individual water bodies.

Table 2. Model Performance Comparison

Model	Average RMSE
LSTM	1.14
Random Forest	2.7
RGCN	1.2

5. Conclusion

In this study, we developed and evaluated machine learning models for predicting Dissolved Oxygen (DO) levels in inland water systems, demonstrating enhanced performance over traditional methods. The LSTM model showed the highest accuracy, as evidenced by its low average RMSE, effectively capturing the temporal patterns in DO levels. However, the variation in model performance across different sites highlights the impact of local environmental conditions and the need for customizing models to these specific contexts. The Recurrent Graph Convolutional Networks also proved effective in capturing spatial relationships, further enhancing our understanding of DO dynamics. Our approach combines pretraining on a global dataset with regional data fine-tuning, offering a more accurate and adaptable solution for environmental monitoring, particularly beneficial in areas with limited direct in-situ observations. In future, expanding the dataset and integrating more ecological variables, will further enhance the models' predictive capabilities and generalizability. The availability of open data from various government and research networks is crucial in this endeavour, as it allows for a more comprehensive and geographically diverse dataset. Additionally, further development of the models' spatial capabilities could provide deeper insights into the complex interactions within aquatic ecosystems. Such advancements will not only improve DO predictions but also contribute significantly to the sustainable management of water resources.

 doi:10.20944/preprints202412.0671.v1

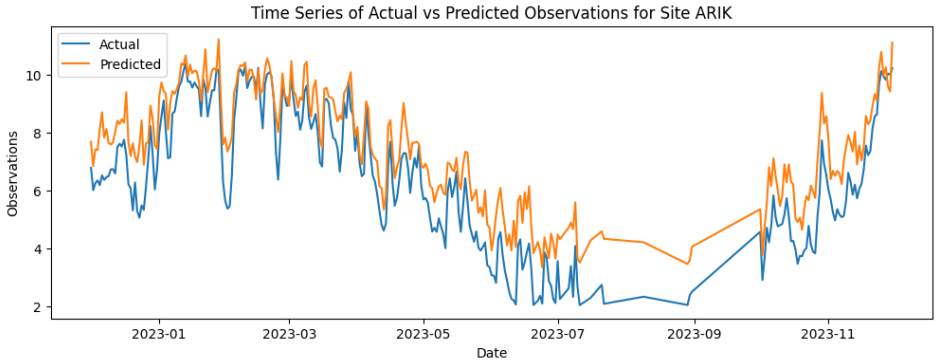

Figure 2. Figure showing the time series plots for actual vs predicted DO using LSTM model for Arikaree River (ARIK) in Northeastern Colorado, US

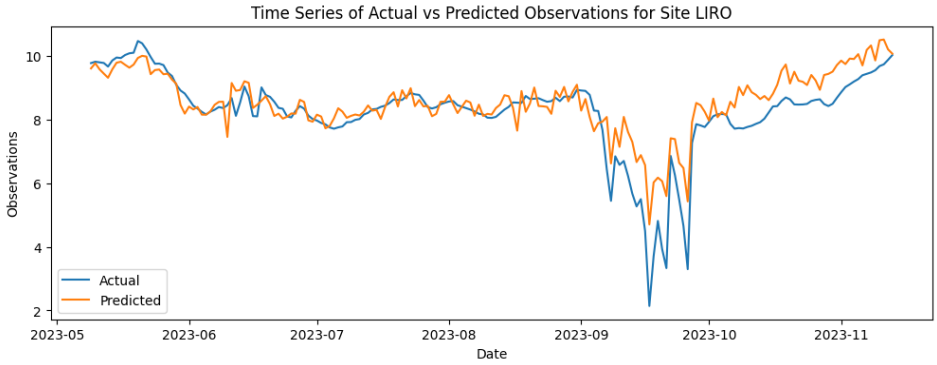

Figure 3. Figure showing the time series plots for actual vs predicted DO using LSTM model for the Little Rock Lake (LIRO) in north-central Wisconsin, US

References

1. Saari, G.N.; Wang, Z.; Brooks, B.W. Revisiting inland hypoxia: diverse exceedances of dissolved oxygen thresholds for freshwater aquatic life. *Environmental Science and Pollution Research* **2018**, *25*, 3139–3150.
2. Mallya, Y.J. The effects of dissolved oxygen on fish growth in aquaculture. *The United Nations University Fisheries Training Programme, Final Project* **2007**.
3. Kelly, B. Death of dolphins in Amazon linked to severe drought, heat. https://www.reuters.com/business/environment/mass-death-amazon-river-dolphins-linked-severe-drought-heat-2023-10-02/, 2023. [Accessed 05-01-2024].
4. Kannel, P.R.; Kanel, S.R.; Lee, S.; Lee, Y.S.; Gan, T.Y. A review of public domain water quality models for simulating dissolved oxygen in rivers and streams. *Environmental Modeling & Assessment* **2011**, *16*, 183–204.
5. Zhi, W.; Feng, D.; Tsai, W.P.; Sterle, G.; Harpold, A.; Shen, C.; Li, L. From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? *Environmental science & technology* **2021**, *55*, 2357–2368.
6. Ziyad Sami, B.F.; Latif, S.D.; Ahmed, A.N.; Chow, M.F.; Murti, M.A.; Suhendi, A.; Ziyad Sami, B.H.; Wong, J.K.; Birima, A.H.; El-Shafie, A. Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan. *Scientific Reports* **2022**, *12*, 3649.
7. Chatziantoniou, A.; Spondylidis, S.C.; Stavrakidis-Zachou, O.; Papandroulakis, N.; Topouzelis, K. Dissolved oxygen estimation in aquaculture sites using remote sensing and machine learning. *Remote Sensing Applications: Society and Environment* **2022**, *28*, 100865.