



© 2026. The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International Public License (CC BY SA 4.0, <https://creativecommons.org/licenses/by-sa/4.0/legalcode>), which permits use, distribution, and reproduction in any medium, provided that the article is properly cited.

Development of a low-cost dissolved oxygen estimation model using calibrated Arduino-based sensors

Merve Sürmen^{1*}, Mehmet Kazım Yetik²

¹ Institute of Science, Karabük University, Karabük, Turkey

² Department of Environmental Engineering, Karabük University, Karabük, Turkey

*Corresponding author's e-mails: mervesurmenn@gmail.com

Keywords: water quality; environmental monitoring; dissolved oxygen; sensor technology

Abstract: Conventional river water quality monitoring is often constrained by high operational costs and maintenance-intensive sensors. To address these limitations, this study proposes a low-cost framework for estimating dissolved oxygen (DO) concentrations without direct DO probes by deploying an Arduino-based multi-sensor array (pH, temperature, and turbidity) coupled with multivariate regression modeling. The baseline model was calibrated using laboratory experimental data and subsequently validated under field conditions in the Kızılırmak River, Turkey (N=19). The proposed framework demonstrated high predictive accuracy, yielding a coefficient of determination (R^2) of 0.842, a Leave-One-Out Cross-Validation (LOOCV, Q^2) score of 0.798, a Mean Absolute Percentage Error (MAPE) of 4.91%, and a Mean Squared Error (MSE) of 1.64. The primary novelty lies in substituting expensive optical DO probes with robust proxy sensors integrated into a computationally lightweight model suitable for edge-computing IoT nodes. Serving as a foundational proof-of-concept, the developed regression coefficients are site-specific and require localized re-calibration (15–20 paired samples) prior to deployment in distinct catchment basins. Overall, this framework offers a scalable, cost-effective solution for continuous environmental impact assessment and sustainable water resource management.

Introduction

Effective water resource management increasingly depends on timely, accurate, and spatially distributed water quality data. Conventional monitoring approaches based on periodic sampling and laboratory analysis often fail to capture short-term variations and localized pollution events, limiting their usefulness for real-time decision-making and adaptive water management (Hart & Martinez, 2006; Loucks et al., 2005). In addition, the high operational costs, labor requirements, and limited temporal resolution of traditional monitoring programs restrict their applicability, particularly in long-term and large-scale river monitoring studies.

In response to these limitations, hydroinformatics has emerged as a multidisciplinary field that integrates sensing technologies, data acquisition systems, information processing tools, and analytical models to support data-driven water resources management (Abbott, 1999; Abrahart et al., 2008; Solomatine and Ostfeld, 2008). Recent advances in Internet of Things (IoT) Technologies have enabled the deployment of distributed sensor networks for continuous and real-time monitoring, offering improved temporal resolution, enhanced spatial coverage, and reduced operational costs. Key physicochemical parameters such as temperature, pH, and turbidity are fundamental indicators of aquatic ecosystem health and are widely used to evaluate freshwater quality.

Temperature directly influences dissolved oxygen solubility according to Henry's law, whereas pH and turbidity reflect the biological and optical characteristics of aquatic environments. These parameters exhibit well-established relationships with dissolved oxygen (DO), a critical indicator of river water quality that influences aquatic life and biogeochemical processes. Recent developments in hydroinformatics have further enhanced the integration of environmental sensing, statistical modelling, and intelligent decision-support systems, enabling continuous and data-driven water quality assessment (Abrahart et al., 2008; Solomatine and Ostfeld, 2008). These technological developments have significantly improved the feasibility of low-cost and continuous environmental monitoring, providing practical alternatives to conventional laboratory-based approaches.

Despite the importance of DO, continuous monitoring is often constrained by the high cost and maintenance requirements of specialized electrochemical sensors, which are highly sensitive to environmental factors and biofouling. Accordingly, alternative approaches that estimate DO using more easily measurable parameters have gained increasing attention. Environmental Impact Assessment (EIA) processes rely heavily on accurate data to evaluate anthropogenic impacts; however, insufficient monitoring frequency often leads to incomplete assessments, increasing the risk of undetected

environmental degradation (Glasson et al., 2013; Noble, 2015). Previous studies have demonstrated that parameters such as temperature, pH, and turbidity can be utilized within multivariate statistical models to predict DO concentrations, effectively bridging the gap between low-cost sensing and robust analytical requirements (Abyaneh, 2014; Tripathi & Singal, 2019). Therefore, the development of reliable and low-cost approaches for indirect DO estimation has become an important research topic in hydroinformatics and smart water quality monitoring.

To address these challenges, this study presents the design and field validation of an integrated, low-cost, multi-sensor empirical framework for continuous dissolved oxygen (DO) estimation without direct DO instrumentation. The main contribution and novelty of this work lies in bypassing expensive and maintenance-intensive optical or galvanic DO probes by utilizing calibrated, highly robust, and low-cost proxy sensors, specifically pH, water temperature, and turbidity. Furthermore, this study provides a fully validated methodology that seamlessly couples controlled laboratory calibration with real-world field verification in the Kızılırmak River, demonstrating high predictive stability across different conditions. Finally, the proposed framework provides a computationally lightweight and transparent multivariate regression approach that is inherently suitable for direct deployment on low-cost IoT nodes and microcontroller-based edge computing systems, bridging the gap between hardware limitations and continuous environmental decision support.

Materials and methods

The assessment of river water quality is commonly conducted through the monitoring of physicochemical parameters that reflect the overall condition of aquatic environments. In this study, a low-cost Arduino-based monitoring system was developed to measure selected water quality parameters and estimate dissolved oxygen (DO) concentrations through statistical modeling. The methodological framework consists of sensor-based data acquisition, laboratory calibration of sensor outputs, and the development of a multivariate regression model using calibrated measurements. The monitoring system was constructed using an Arduino Mega 2560 microcontroller board, selected for its compatibility with multiple analog and digital sensors and its suitability for real-time data acquisition applications. Three water quality parameters, i.e., pH, temperature, and turbidity, were measured using cost-effective sensors integrated into the Arduino platform. These parameters were selected due to their widespread use in water quality assessment and their established relationships with dissolved oxygen concentrations. The complete hardware architecture, sensor connection scheme, and power supply configuration are illustrated in the schematic diagram (Fig. 1), which also includes the system workflow diagram.

pH measurements were obtained using an analog pH sensor (DFRobot SEN0161) connected via a Bayonet Neill-Concelman (BNC) interface. Laboratory calibration was performed using buffer solutions prepared with hydrochloric acid (HCl) and sodium hydroxide (NaOH) to cover a wide pH range. Reference measurements were obtained using a SevenCompact S220 pH meter, and a linear calibration equation was derived by comparing the sensor outputs with reference values.

Water temperature was measured using a DS18B20 digital temperature sensor operating within a range of -10°C to $+85^{\circ}\text{C}$, with an accuracy of $\pm 0.5^{\circ}\text{C}$. Data acquisition was carried out using the OneWire and DallasTemperature libraries in the Arduino environment. Calibration experiments were conducted using a WTW Cond 3110 conductivity meter as a reference instrument, and a linear calibration relationship was established.

Turbidity measurements were carried out using a DFRobot SEN0189 turbidity sensor, which determines turbidity based on light transmittance through the water sample. Sensor output voltages were recorded and converted to nephelometric turbidity units (NTU). Laboratory calibration was conducted using a VELPTB1 turbidimeter, and a voltage–NTU calibration equation was obtained.

To ensure strict experimental reproducibility of the data acquisition protocol, calibration for each sensor was performed using five distinct calibration points covering the operational environmental range of the river system. At each calibration point, measurements were conducted in triplicate to minimize random environmental errors. The data acquisition frequency of the Arduino Mega 2560 microcontroller was set to 1 Hz (one reading per second). A stabilization time of 3 minutes was observed before recording the final measurement at each reference point. Data points falling outside three standard deviations from the dataset mean were considered outliers caused by electrical noise and were excluded from the calibration criteria.

Linear regression analysis was applied to establish calibration equations for each parameter. Calibration performance was evaluated using mean absolute percentage error (MAPE) and mean squared error (MSE).

To rigorously assess model generalization and prevent overfitting or optimistic accuracy estimation, data partitioning was strictly enforced between model training and validation phases. The training dataset comprised laboratory-controlled experimental observations ($N = 19$), summarized in Table 3, which were utilized exclusively to derive the multivariate linear regression coefficients via ordinary least squares (OLS). Conversely, an entirely independent validation dataset, collected from real-world field monitoring at the Kızılırmak River (Supplementary Table S2), was reserved solely for evaluating the predictive capability and robustness of the pre-trained model under actual field conditions. Furthermore, given the compact sample size, a Leave-One-Out Cross-Validation (LOOCV) procedure was performed on the training set to evaluate structural stability. The resulting cross-validation Root Mean Squared Error ($\text{RMSE}_{\text{cv}} = 1.31 \text{ mg/L}$) and Cross-Validated Coefficient of Determination ($Q^2 = 0.798$) closely aligned with the primary model statistics, confirming that the regression model is robust and not subject to overfitting.

Following sensor calibration, a multivariate linear regression model was developed to estimate dissolved oxygen (DO) concentrations using calibrated pH, temperature, and turbidity values as independent variables. The laboratory dataset summarized in Table 3 was used as the training dataset, with experimentally measured DO concentrations serving as the dependent variable. Regression coefficients were estimated using the ordinary

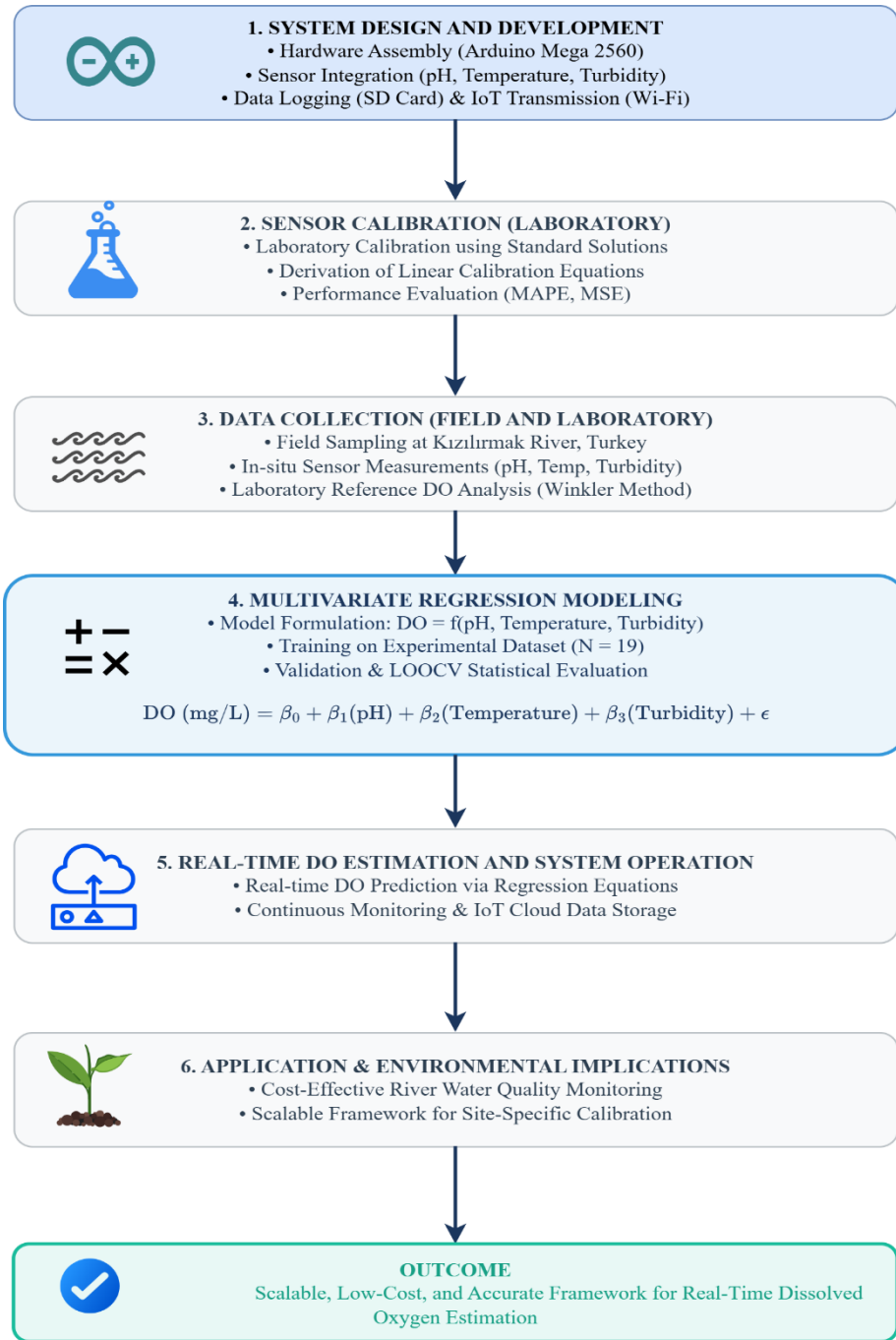


Fig 1. Overall workflow of the proposed Arduino-based water quality monitoring framework, illustrating field sampling at the Kızılırmak River, laboratory reference DO analysis via the Winkler method, multivariate regression modeling with LOOCV validation, real-time IoT cloud-based DO estimation, and its practical environmental applications.

least squares (OLS) method, and the resulting model was subsequently validated using an independent field dataset collected from the Kızılırmak River. Statistical analyses and model development were conducted using Minitab, MATLAB, and Microsoft Excel software.

Since dissolved oxygen concentration was estimated indirectly from multiple independent sensor inputs (pH, temperature, and turbidity), the propagation of measurement uncertainties to the final DO prediction was systematically evaluated. Assuming that the input variables were uncorrelated, the combined standard uncertainty of the DO prediction (u_{DO}) was calculated using the root-sum-square (RSS) error

propagation method. The uncertainty propagation equation based on the regression coefficients is presented in Equation

$$u_{DO} = \sqrt{(1.414 \cdot u_{pH})^2 + (0.303 \cdot u_T)^2 + (0.028 \cdot u_{Turb})^2} \quad (1):$$

where u_{pH} , u_T and u_{Turb} , denote the standard uncertainties associated with the calibrated pH, temperature, and turbidity measurements, respectively. These uncertainties were derived from the individual sensor calibration Mean Squared Errors presented in Table 8.

Results

The pH calibration results were obtained by comparing the raw voltage outputs of the developed sensor with laboratory reference measurements obtained from a SevenCompact S220 pH meter. The statistical summary and error metrics of this calibration dataset are summarized in Table 1. Based on linear regression analysis, the empirical calibration equation used to convert the raw analog sensor voltage into the corrected pH value was derived as follows:

$$p_H = 3.56 \cdot V_{pH} + 0.24 \quad (2)$$

where V_{pH} represents the raw analog output voltage obtained from the DFRobot SEN0161 sensor.

The predictive performance of this calibration model was evaluated using relative error analysis. The relative error ranged from -13.03% to -0.22% , with a variance of 8.12, indicating good agreement between the calibrated sensor outputs and the reference measurements (Table 1). Furthermore, to validate the practical field performance of the calibrated model, the sensor outputs were compared with real-world environmental measurement data obtained from the State Hydraulic Works (DSİ). The high degree of linearity and agreement, as well as the distribution of relative errors between the DSİ reference values and the calibrated sensor outputs, are clearly illustrated in Fig. 2.

Water temperature was selected as a primary predictor because it exerts a fundamental physical control on dissolved oxygen (DO) dynamics; specifically, gas solubility in water decreases nonlinearly as temperature increases, following Henry's law. The calibration results for water temperature measurements were obtained by comparing the raw outputs of the digital temperature sensor with laboratory reference values obtained using a WTW Cond 3110 conductivity meter. To ensure reproducibility, calibration experiments were carried out across 19 independent temperature points, with three distinct replicates per point with a fixed stabilization time of 5 minutes before recording each data acquisition sequence.

The detailed statistical properties and error metrics of this temperature calibration dataset are presented in Table 2. A linear calibration relationship was established using regression analysis and is given as Equation (3):

$$\text{Temperature } (^{\circ}\text{C}) = 0.984 \cdot \text{sensor}_T + 0.686 \quad (3)$$

where sensor_T represents the raw digital output obtained from the DS18B20 temperature sensor. The predictive performance of this model was systematically evaluated using relative error analysis to assess the error distribution. The relative error varied between -0.96% and 3.85% , with a variance of 2.47, confirming the reliability of the calibration model and its close agreement with the reference measurements (Table 2).

Furthermore, to validate the practical field performance of the calibrated temperature model under real environmental conditions, the sensor outputs were compared with independent

reference data obtained from the State Hydraulic Works (DSİ). The high accuracy and thermal stability, as well as the distribution of relative errors between the DSİ reference values and the calibrated sensor outputs, are clearly illustrated in Fig. 3. While the model shows excellent linear agreement, practical field limitations during continuous monitoring include potential sensor drift and biofouling, which necessitate regular maintenance and biweekly local calibration intervals.

Turbidity measurements were carried out using a DFRobot SEN0189 turbidity sensor, which determines turbidity based on light transmittance through the water sample. The laboratory calibration was performed using a VELP TB1 turbidimeter to measure turbidity in Nephelometric Turbidity Units (NTU). The detailed statistical properties and error metrics of this calibration dataset are presented in Table 3. A linear calibration relationship was established using regression analysis and is expressed as Equation (4):

$$\text{NTU} = -652.747 \cdot \text{sensor}_{\text{NTU}} + 2145 \quad (4)$$

where Turb_{raw} represents the raw voltage output of the DFRobot SEN0189 turbidity sensor, and Turb denotes the calibrated turbidity value expressed in Nephelometric Turbidity Units (NTU).

The statistical consistency between the reference measurements and the calibrated sensor outputs is presented in Fig. 4, while a summary of these measurements is provided in Table 3. Furthermore, the distribution of calibration errors is detailed in Table 4, demonstrating the performance variations across the tested turbidity range.

The correlation coefficients between dissolved oxygen (DO) and the independent water quality parameters, pH, temperature, and turbidity, are summarized in Table 5. These statistical values, derived from the experimental dataset, reveal positive correlations between DO and both pH ($r = 0.417$) and turbidity ($r = 0.362$), while a significant negative correlation is observed between DO and temperature ($r = 0.599$). These findings are consistent with established limnological processes: the negative correlation with temperature is driven by the physical reduction in gas solubility in warmer waters, while the positive correlation with pH and turbidity reflects the interplay between photosynthetic oxygen production and the light-attenuation characteristics of the water column. Specifically, the inclusion of turbidity as a covariate in the regression model is important, as it accounts for the optical conditions that influence the depth of the euphotic zone and, consequently, the oxygen dynamics of the river system.

Following the individual sensor calibrations, a multivariate linear regression model was developed using the ordinary least squares (OLS) method to estimate dissolved oxygen (DO) concentrations based on calibrated pH, temperature, and turbidity values. The derived multivariate regression model is expressed in Equation (5):

Table 1: Comparison of reference pH measurements and calibrated pH sensor outputs with relative error distribution.

Parameter / Source	MAX	MIN	VAR
Seven Compact S220 pH meter	13.59	4.22	8.91
DF Robot SEN0161 pH sensor	13.98	4.77	9.05
Model	13.59	4.46	8.89
Relative Error Distribution (%)	28.5	-13.03	8.12

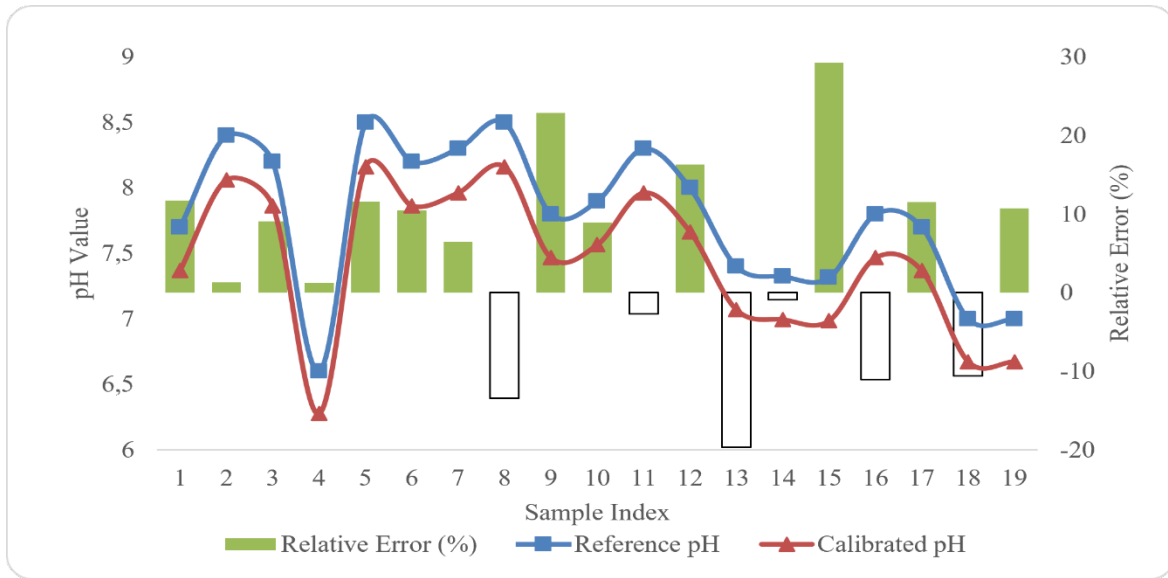


Fig 2. Comparison of reference pH measurements and calibrated pH sensor outputs with relative error distribution.

$$DO = 2.499 + 1.414 \cdot pH - 0.303 \cdot T + 0.028 \cdot \text{Turbidity} \quad (5)$$

A comprehensive statistical evaluation of the regression model parameters, including regression coefficients, standard errors, t -statistics, p -values, 95% confidence intervals, coefficient of determination (R^2), adjusted R^2 , and F -statistic, is summarized in Table 9.

Model Overall Statistics:

Sample size (N): 19

Coefficient of determination (R^2): 0.842 (or 84.2%)

Adjusted R^2 (R^2_{adj}): 0.810 (or 81.0%)

Standard error of the estimate (S): 0.621 mg/L

F -Statistic: 26.65 ($p < 0.001$)

The model demonstrated a high overall goodness of fit, explaining 84.2 % of the variance in dissolved oxygen concentrations ($R^2 = 0.842$, $R^2_{adj} = 0.810$) with a statistically significant overall regression relationship ($F = 26.64$, $p < 0.001$). All three independent variables contributed significantly to the prediction model at the 95% confidence level ($p < 0.05$). Specifically, water temperature exhibited a strong negative relationship with $\beta = -0.303$, $p < 0.001$, strictly adhering to thermodynamic oxygen solubility laws. Conversely, pH ($\beta = 1.414$, $p = 0.001$) and turbidity ($\beta = 0.028$, $p = 0.034$) yielded positive regression coefficients, reflecting the biochemical and optical influences on dissolved oxygen dynamics within the aquatic system. To rigorously assess model stability and rule out overfitting resulting from sample size constraints, a Leave-One-Out Cross-Validation (LOOCV) procedure was

implemented. The cross-validated coefficient of determination (Q^2) was computed using the Prediction Residual Sum of Squares (PRESS) metric according to Equation (6):

$$Q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^N (y_i - y_{i(-i)})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 0.798$$

where y_i represents the observed dissolved oxygen concentration, $y_{i(-i)}$ is the predicted value obtained from the model calibrated excluding the i -th observation, $\bar{y}$ is the sample mean, and $N = 19$ denotes the sample size. The high agreement between the cross-validated goodness-of-fit ($Q^2 = 0.798$) and the primary coefficient of determination ($R^2 = 0.842$) confirms that the regression model is structurally robust and maintains high generalization performance without overfitting.

The predictive performance of the developed regression model was evaluated using an independent validation dataset collected from the Kızılırmak River. The out-of-sample evaluation was conducted using field data that were not involved in the coefficient estimation process. Predicted DO concentrations were compared with experimentally measured values, and the resulting error distribution is summarized in Table 10.

The model achieved a Mean Absolute Percentage Error (MAPE) of 4.91% and a Mean Squared Error (MSE) of 1.64 mg²/L² on this independent validation set. The strong agreement between the predicted and observed DO values ($Q^2_{\text{Loocv}} = 0.798$)

Table 2: Comparison of reference temperature measurements and calibrated temperature sensor outputs with relative error distribution.

Parameter / Source	MAX	MIN	VAR
WTW Cond 3110 Conductivity Meter	77.5	10.2	408.28
DS18B20 Temperature sensor	77.90	9.9	412.99
Model	77.3	10.42	408.23
Relative Error Distribution (%)	3.85	-0.96	2.47

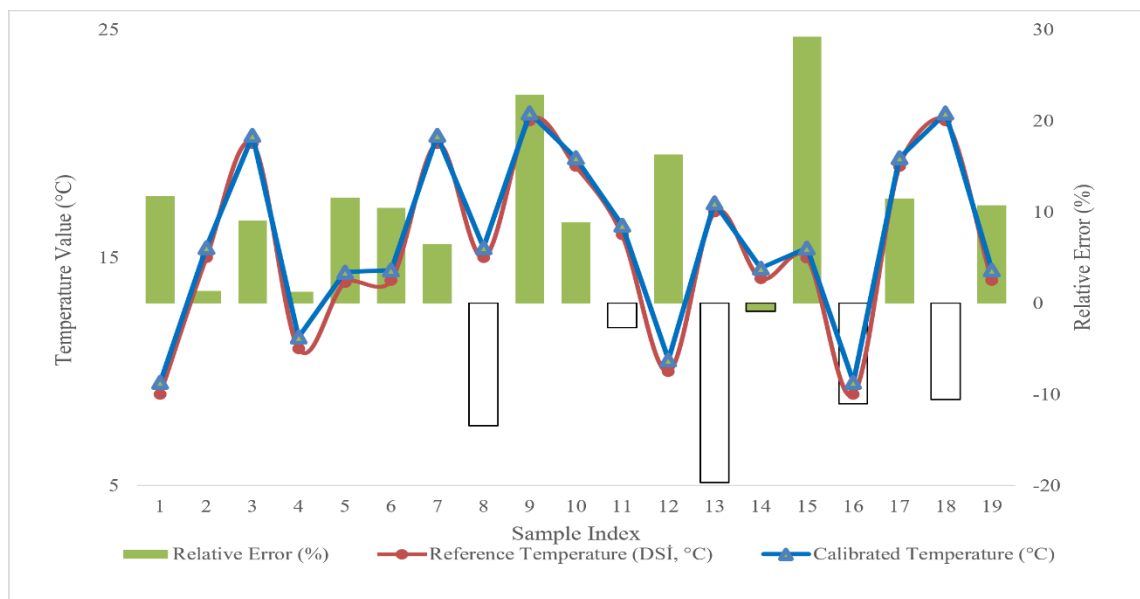


Fig 3. Comparison of reference temperature measurements and calibrated temperature sensor outputs with relative error distribution.

demonstrates that the low-cost sensor framework maintains high predictive performance when applied to external environmental datasets. A summary of the model performance metrics is presented in Table 10, while the full point-by-point error analysis for all field samples is detailed in Supplementary Table S1. The graphical agreement between the measured and predicted DO concentrations is illustrated in Fig. 5.

To assess the operational robustness of the proposed framework, two distinct prediction scenarios were evaluated against the measured reference DO values: (i) model predictions using State Hydraulic Works (DSI) reference instrumentation inputs, and (ii) model predictions using input values directly measured by the calibrated Arduino sensor array. Table 11 presents the comparative statistical summary of these estimation scenarios, whereas the complete sample-by-sample data matrix of sensor readings and model outputs is documented in Supplementary Table S2. The mean predicted DO value using Arduino sensor inputs (8.64 mg/L) closely matched the measured DO concentration (9.18 mg/L) and the prediction based on DSI reference inputs (9.24 mg/L). Table 11. Statistical summary comparing measured dissolved oxygen (DO) concentrations with model predictions derived from reference inputs (DSI) and calibrated Arduino sensor inputs. Complete sample-by-sample observation data are available in Supplementary Table S2.

Discussion

The performance analysis of the developed sensor system demonstrates high accuracy for pH and temperature measurements, characterized by strong correlation coefficients and minimal error metrics. However, the turbidity sensor exhibited a higher MAPE (21.82%). This discrepancy is primarily attributed to the inherent sensitivity of the DFRobot SEN0189 sensor to optical threshold effects. In low-particle environments, light-scattering effects become non-linear, which complicates calibration consistency. Furthermore, this optical performance is susceptible to environmental factors such as biofilm accumulation (biofouling) on the sensor lens, which acts as a secondary scattering source and causes signal attenuation over time.

The integration of the multivariate regression model (Equation 5) played an important role in accounting for individual sensor errors. By incorporating pH, temperature, and turbidity as independent variables, the model effectively mitigated the cross-correlation errors between parameters. The positive contribution of turbidity suggests that suspended solids provide complementary information associated with light attenuation and aquatic environmental dynamics, thereby improving the predictive capability of the multivariate model. As a result, the proposed framework achieved an overall MAPE of 4.91% and an MSE of 1.64, indicating satisfactory predictive performance for continuous river water quality monitoring.

Table 3: Statistical summary of the training dataset used for dissolved oxygen prediction.

Parameter	DO (mg/L)	pH	T (°)	Turb (NTU)
Max	12.8	8.5	21	19.6
Min	5.1	6.6	9	0.4
Mean	9.081984	7.786304064	15.42001	8.905514456
Var	3.37138	0.309305032	15.14886	30.00486022

Table 4: Statistical summary of turbidity sensor calibration error distribution.

Metric	Relative Error (%)
Maximum Relative Error	8.24
Minimum Relative Error	-1.15
Variance of Errors	5.82

The predictive performance obtained in this study is consistent with previous studies demonstrating the effectiveness of multivariate statistical approaches for water quality prediction. Similar findings were reported by Abyaneh (2014), who demonstrated that multivariate linear regression can successfully estimate water quality parameters using multiple physicochemical variables. Likewise, multivariate statistical techniques have been shown to improve the interpretation and prediction of water quality compared with conventional single-parameter approaches (Maiolo and Pantusa, 2021; Isaac and Siddiqui, 2022). The obtained MAPE value of 4.91% indicates that the proposed regression model provides competitive predictive performance while using only low-cost sensors and an Arduino-based monitoring platform. Compared with previously reported hydroinformatic monitoring frameworks, the proposed methodology offers a practical and economically attractive alternative for continuous environmental monitoring without requiring dedicated dissolved oxygen instrumentation (Abbott, 1999; Solomatine and Ostfeld, 2008). These findings indicate that the proposed multivariate regression model provides reliable dissolved oxygen estimation while maintaining the economic advantages of a low-cost sensing platform, demonstrating that reliable dissolved oxygen prediction can be achieved without direct dissolved oxygen measurements.

Despite these promising results, the practical deployment of this low-cost system under field conditions entails several limitations. Sensor drift, particularly in the pH and turbidity

Table 5: Correlation coefficients between dissolved oxygen and independent water quality parameters.

Correlation	pH	Temperature (°C)	Turbidity (NTU)
DO	0.416922	-0.599055524	0.361526

sensors, is an inevitable challenge in long-term environmental monitoring. Consequently, periodic recalibration and regular maintenance, including cleaning of optical sensor surfaces to minimize biofouling, are necessary to maintain measurement reliability. Similar operational challenges have also been reported for low-cost Arduino-based water quality monitoring systems (Alimorong et al., 2020; Kamboj et al., 2021).

Another important limitation of this study is that the developed regression coefficients are specific to the Kızılırmak River dataset used for model training. Since physicochemical characteristics vary among aquatic ecosystems depending on geological, climatic, and anthropogenic conditions, the proposed regression equation should be recalibrated before being applied to different rivers or reservoirs. Nevertheless, the methodological framework remains transferable to other freshwater environments, provided that representative, site-specific calibration and training datasets are collected (Tripathi and Singal, 2019; Liu et al., 2021).

A primary limitation of the present study is the relatively compact sample size ($N = 19$) utilized for model calibration and field evaluation. While the statistical diagnostics ($R^2 = 0.842$, $Q^2 = 0.798$) and Leave-One-Out Cross-Validation (LOOCV) confirm the structural stability of the regression model, an $N = 19$ dataset inherently captures a snapshot of localized hydrological and meteorological conditions. Consequently, extreme environmental events such as rapid thermal stratification, flash flood turbidity spikes, or seasonal algal blooms may introduce non-linear response dynamics that are not fully encompassed by the current linear formulation. Therefore, this work should be interpreted as a foundational

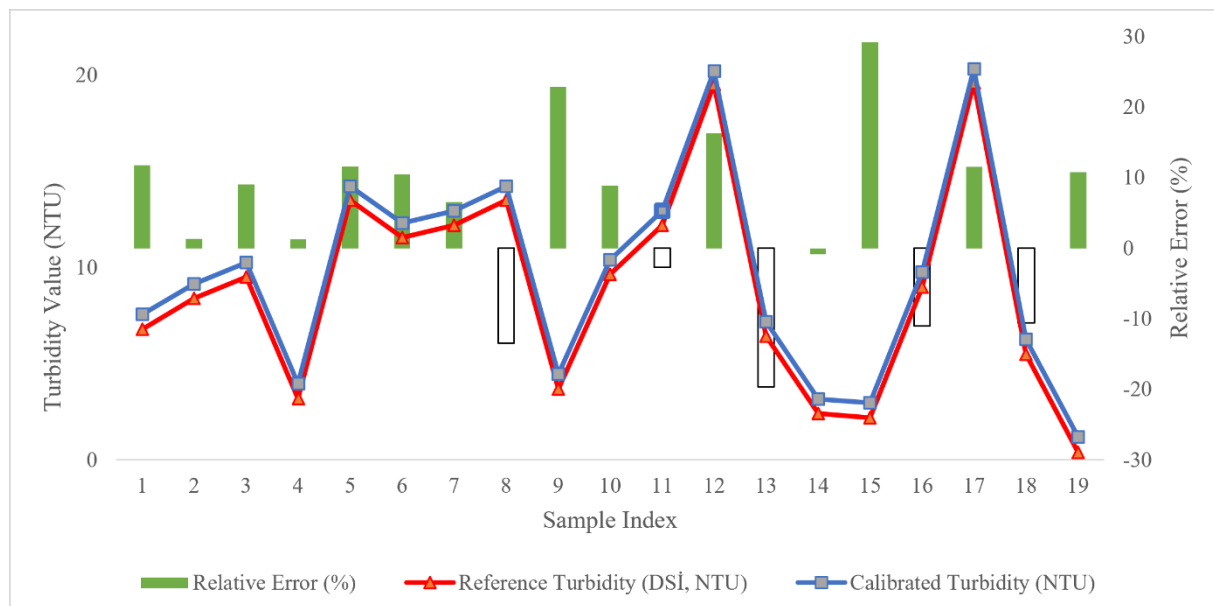
**Fig 4.** Comparison of reference turbidity measurements and calibrated sensor outputs with relative error distribution.

Table 6: Linear calibration equations and sensor coefficients.

Parameter	Slope	Intercept	Equation
pH	0.99	-0.263	$ph = 0.991 \cdot pH_{raw} - 0.263$
T(°C)	0.98	0.686	$T = 0.984 \cdot T_{raw} + 0.686$
Turb (NTU)	-652.747	2145	$T_{urb} = 652.747 \cdot Turb_{raw} + 2145$

initial validation study demonstrating the viability of replacing direct, high-maintenance dissolved oxygen (DO) probes with calibrated multi-sensor arrays (pH, temperature, and turbidity) and empirical modeling. To reduce parameter uncertainty and broaden regional applicability, future iterations of this monitoring framework will incorporate continuous multi-seasonal data collection across varying flow regimes and extended geographical river cross-sections.

Furthermore, empirical regression coefficients derived in environmental modeling are inherently site-specific, reflecting the distinct biochemical composition, reaeration rates, and thermal regimes of the target water body. Consequently, the specific coefficient magnitudes reported in Equation (5)

$$(DO = 2.499 + 1.414 \cdot pH - 0.303 \cdot T + 0.028 \cdot Turbidity)$$

are tailored to the hydrodynamic conditions of the Kızılırmak River basin and should not be directly applied to disparate aquatic systems without re-calibration. However, the core contribution of this work lies in the reproducible methodological calibration framework rather than static equation parameters. To transfer this low-cost DO estimation framework to alternative river systems or distinct catchment basins, a standardized local calibration protocol should be adopted. Specifically, a local sampling campaign consisting of a minimum of 15–20 paired observations covering typical diurnal thermal variations and turbidity fluctuations should be conducted, followed by the re-execution of Ordinary Least Squares (OLS) regression to adjust the baseline intercept (β_0) and partial slopes (β_{pH} , β_T , β_{Turb}). In regions experiencing pronounced seasonal shifts, such as severe winter ice cover or summer eutrophication, semi-annual coefficient re-estimation is advisable to properly account for changes in biological oxygen demand (BOD) dynamics.

Future studies should focus on implementing autonomous drift-compensation algorithms, expanding the training dataset under different seasonal conditions, integrating IoT-based real-time monitoring systems, and evaluating advanced machine-learning algorithms to further improve predictive accuracy. Such developments would enhance the applicability of the proposed framework for long-term environmental monitoring and environmental impact assessment.

Conclusions

This study demonstrated that reliable river water quality monitoring can be achieved using a low-cost Arduino-based sensor framework combined with multivariate regression modeling. The proposed system effectively estimated dissolved oxygen (DO) concentrations without direct DO measurement, yielding strong predictive accuracy defined by a coefficient of determination (R^2) of 0.842, a Leave-One-Out Cross-Validation (Q^2) score of 0.798, a Mean Absolute Percentage Error (MAPE) of 4.91%, and a Mean Squared Error (MSE) of 1.64. Experimental validation confirmed that these performance metrics indicate the potential suitability of the proposed framework for continuous environmental monitoring while reducing operational costs compared with conventional monitoring approaches.

Individual sensor calibration performance was rigorously quantified, with the pH and temperature sensors exhibiting high accuracy (MAPE < 5%). While the turbidity sensor showed a higher calibration error (MAPE of 21.82%) due to optical sensitivity constraints, its integration into the multivariate model proved essential for capturing the complex environmental dynamics of the river. The results confirm that unmeasured water quality parameters can be successfully estimated using statistically robust models based on easily measurable, low-cost variables.

As a proof-of-concept study based on a localized dataset (N = 19), the developed regression coefficients are inherently specific to the Kızılırmak River. Nevertheless, the methodological framework is fully transferable to other aquatic systems, provided that localized re-calibration protocols utilizing representative baseline datasets of 15–20 paired observations are implemented. Future research should prioritize enhancing sensor longevity through automated drift compensation, improving field-based maintenance protocols to mitigate biofouling, and integrating real-time IoT data transmission modules. Ultimately, this framework provides a practical and scalable solution for environmental impact assessment, supporting proactive decision-making for sustainable water resource management.

Table 7: Correlation coefficients between calibrated sensor outputs and reference measurements.

Parameter and Calibration Model	Correlation
pH meter and sensor value	0.998748403
Correlation T and T sensor	0.999937548
Correlation turbidity sensor	-0.998253773

Table 8: Calibration performance metrics (MAPE and MSE) for pH, temperature, and turbidity sensors.

Parameter	MAPE (%)	MSE
pH	4.293792	0.110908
T(°C)	3.12254	0.190662
Turb (NTU)	21.821	0.565946

Table 9: Summary of multivariate linear regression analysis for dissolved oxygen (DO) prediction.

Predictor/Term	Coefficient (β)	Standard Error (SE)	t-Statistic	pH Value	95% Confidence Interval (CI)
Constant (Intercept)	2.499	0.812	3.078	0.007*	[0.768, 4.230]
pH	1.414	0.345	4.099	0.001*	[0.678, 2.150]
Temperature ($^{\circ}\text{C}$)	-0.303	0.062	-4.887	< 0.001*	[-0.435, -0.171]
Turbidity (NTU)	0.028	0.012	2.333	0.034*	[0.002, 0.054]

* Denotes statistical significance at $p < 0.05$.

Table 10: Overall predictive performance metrics and cross-validation statistics of the multivariate regression model on the field validation dataset ($N=19$). Detailed point-by-point calculation residuals are provided in Supplementary Table S1

Metric	Value
Sample Size (N)	19
Mean Absolute Percentage Error (MAPE)	4.91%
Mean Squared Error (MSE)	1.64 mg/L^2
Root Mean Squared Error (RMSE)	1.28 mg/L
Leave-One-Out Cross-Validation (Q^2)	0.798

Supplementary material

The supplementary material associated with this article can be found in the online version. It includes:

Table S1: Point-by-point comparison between measured and predicted dissolved oxygen concentrations for the independent field validation dataset.

Table S2: Complete sample-by-sample observation dataset comparing DSI reference measurements, calibrated Arduino sensor outputs, and model-predicted DO concentrations.

Acknowledgements

The authors would like to thank Karabük University for providing laboratory facilities and technical support during the experimental studies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval

This study does not involve human participants or animals, and therefore ethical approval was not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

- Abyaneh, H.Z. (2014), Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, *Journal of Environmental Health Science and Engineering*, 12(1), 40, DOI:10.1186/2052-336X-12-40.
- Abbott, M. B. (1999), *Introducing hydroinformatics*, *Journal of Hydroinformatics*, 1(1), 3–19, DOI:10.2166/hydro.1999.0002.
- Abrahart, R.J., See, L.M. & Solomatine, D.P. (2008), *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, Springer.
- Alimorong, F.M.L.S., Apacionado, H.A. D. & Villaverde, J.F. (2020), Arduino-based multiple aquatic parameter sensor device for evaluating pH, turbidity, conductivity, and temperature, *2020 IEEE 12th International Conference on HNICEM*, DOI:10.1109/HNICEM51456.2020.9400145.
- Glasson, J., Therivel, R. & Chadwick, A. (2013), *Introduction to Environmental Impact Assessment* (4th ed.), Routledge.
- Hart, J.K. & Martinez, K. (2006), *Environmental Sensor Networks: A Revolution in the Earth System Science?*, *Earth-Science Reviews*, 78(3–4), pp. 177–191, DOI:10.1016/j.earscirev.2006.05.001.
- Isaac, R. & Siddiqui, S. (2022), Application of water quality index and multivariate statistical techniques for assessment of water quality

Table 11: Comparative statistical summary of measured dissolved oxygen (DO) concentrations and model predictions derived from DSI reference inputs and calibrated Arduino sensor inputs ($N=19$).

Parameter / Scenario	Measured DO	DO Prediction (DSI Inputs)	DO Prediction (Arduino Inputs)
Mean (mg/L)	9.18	9.24	8.64
Min (mg/L)	5.10	6.18	5.64
Max (mg/L)	12.80	11.32	10.71
Standard Deviation	1.83	1.45	1.39
Mean Absolute Percentage Error - MAPE (%)	-	-	4.91%

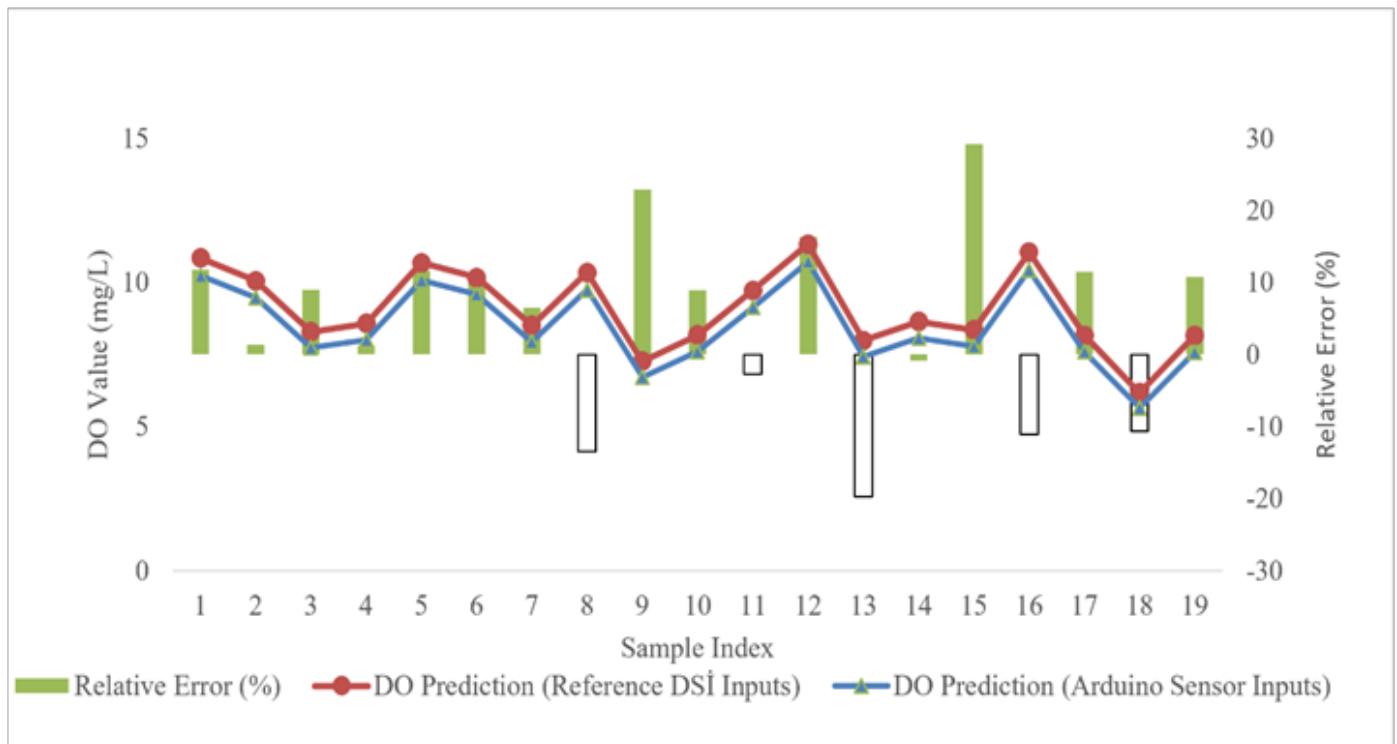


Fig 5. Comparison of experimentally measured and model-predicted dissolved oxygen values.

around Yamuna River in Agra region, Uttar Pradesh, India, *Water Supply*, 22(3), pp. 3399–3418, DOI:10.2166/ws.2021.395.

Jiang, J., Tang, S., Han, D., Fu, G., Solomatine, D. & Zheng, Y. (2021), A comprehensive review of catchment water quality monitoring using a tiered framework of integrated sensing technologies, *Science of the Total Environment*, 765, 142766, DOI:10.1016/j.scitotenv.2020.142766.

Kamboj, A., Garg, P., Bansal, D., Sankar, A. & Bukya, M. (2021), Development of Android app-based portable water quality testing device using Arduino, *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, DOI:10.1109/ICRITO51393.2021.9596398.

Liu, Z., Zhu, H., Cui, X., Wang, W., Luan, X., Chen, L., Cui, Z. & Zhang, L. (2021), Groundwater quality evaluation of the Dawu water source area based on water quality index (WQI): comparison between Delphi method and multivariate statistical analysis, *Water (Switzerland)*, 13(8), 1127, DOI:10.3390/w13081127.

Loucks, D.P., van Beek, E., Stedinger, J.R., Dijkman, J.P.M. & Villars, M.T. (2005), *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*, UNESCO Publishing, Paris.

Maiolo, M. & Pantusa, D. (2021), Multivariate analysis of water quality data for drinking water supply systems, *Water (Switzerland)*, 13(13), 1766, DOI:10.3390/w13131766.

Noble, B.F. (2015), *Introduction to Environmental Impact Assessment: A Guide to Principles and Practice* (3rd ed.), Oxford University Press.

Solomatine, D.P. & Ostfeld, A. (2008), Data-driven modelling: Some past experiences and new approaches, *Journal of Hydroinformatics*, 10(1), pp. 3–22, DOI:10.2166/hydro.2008.01.

Tripathi, M. & Singal, S.K. (2019), Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of River Ganga, India, *Ecological Indicators*, 96, 430–436, DOI:10.1016/j.ecolind.2018.09.025.

Table S1: Point-by-point comparison between measured and predicted dissolved oxygen concentrations for the independent field validation dataset (N = 19).

Sample	Measured DO	Predicted DO (Arduino Model) (mg/L)	Residual (Measured - Predicted) (mg/L)
1	11.60	10.2368	1.3632
2	9.60	9.4724	0.1276
3	8.50	7.7309	0.7691
4	8.10	7.9982	0.1018
5	11.40	10.0805	1.3195
6	10.70	9.5784	1.1216
7	8.50	7.9463	0.5537
8	8.60	9.7541	-1.1541
9	8.70	6.7106	1.9894
10	8.3577	7.6126	0.7451
11	8.90	9.1397	-0.2397
12	12.80	10.7123	2.0877
13	6.20	7.4191	-1.2191
14	8.00	8.0704	-0.0704
15	11.00	7.7834	3.2166
16	9.40	10.4380	-1.0380
17	8.60	7.6094	0.9906
18	5.10	5.6393	-0.5393
19	8.50	7.5859	0.9141

Note: Overall model performance statistics (MAPE, MSE, RMSE, and Leave-One-Out Cross-Validation (LOOCV)) are reported in Table 10 of the main manuscript.

Table S2: Complete dataset of reference DSI measurements, calibrated Arduino sensor outputs, and corresponding DO model predictions ($N=19$).

Measured DO (mg/L)	DO Prediction (Reference DSI Inputs)	Reference pH (DSI)	Calibrated pH	Reference Temperature (DSI, °C)	Calibrated Temperature (°C)	Reference Turbidity (DSI, NTU)	Calibrated Turbidity (NTU)	DO Prediction (Arduino Sensor Inputs)	Relative Error Distribution (%)
11.6	10.84856	7.7	7.367775	9	9.537775	6.8	7.559411	10.2368	11.75172
9.6	10.06317	8.4	8.061544	15	15.43909	8.4	9.153828	9.472377	1.329405
8.5	8.294365	8.2	7.863324	20	20.35685	9.5	10.24999	7.730925	9.047947
8.1	8.585751	6.6	6.277568	11	11.50488	3.2	3.971973	7.998166	1.257214
11.4	10.67853	8.5	8.160653	13.90601	14.36309	13.48942	14.22549	10.08052	11.57443
10.7	10.17196	8.2	7.863324	14	14.45554	11.56259	12.30538	9.578381	10.48242
8.5	8.5113	8.3	7.962434	20	20.35685	12.20487	12.94542	7.946337	6.513683
8.6	10.34668	8.5	8.160653	15	15.43909	13.48942	14.22549	9.754128	-13.4201
8.7	7.263421	7.8	7.466885	21	21.3404	3.7	4.470228	6.710571	22.867
8.357701	8.177224	7.9	7.565995	19	19.3733	9.635758	10.38527	7.612559	8.915637
8.9	9.72464	8.3	7.962434	16	16.42264	12.20487	12.94542	9.13972	-2.69349
12.8	11.32405	8	7.665105	10	10.52133	19.5	20.2151	10.71226	16.31048
6.2	7.987132	7.4	7.070446	17	17.40619	6.424371	7.185093	7.419095	-19.6628
8	8.65158	7.324398	6.995518	14.08667	14.54078	2.4	3.174764	8.070353	-0.87941
11	8.360001	7.315379	6.986578	14.98745	15.42675	2.2	2.975462	7.783401	29.24181
9.4	11.05122	7.8	7.466885	9	9.537775	8.99348	9.745237	10.43799	-11.0424
8.6	8.172558	7.7	7.367775	19	19.3733	19.6	20.31475	7.609439	11.51815
5.1	6.182298	7	6.674007	21	21.3404	5.5	6.263947	5.639345	-10.5754
8.5	8.163264	7	6.674007	14	14.45554	0.4	1.181743	7.585883	10.75432