Dissolved oxygen modelling of the Yamuna River using different ANFIS models

Sameer Arora* and Ashok K. Keshari
Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
*Corresponding author. E-mail: sameer_arora01@yahoo.co.in

ABSTRACT

Dissolved oxygen (DO) is one of the prime parameters for assessing the water quality of any stream. Thus, the accurate estimation of DO is necessary to evolve measures for maintaining the riverine ecosystem and designing appropriate water quality improvement plans. Machine learning techniques are becoming valuable tools for the prediction and simulation of water quality parameters. A study has been performed in the Delhi stretch of the Yamuna River, India, and physiochemical parameters were examined for 5 years to simulate the DO using various machine learning techniques. Simulation and prediction competencies of adaptive neuro fuzzy inference system–grid partitioning (ANFIS–GP) and subtractive clustering (ANFIS–SC) were performed on high dimensional river characteristics. Four different models (M1, M2, M3 and M4) were developed using different combination of input parameters to predict DO. Results obtained from the models were evaluated using root mean square error and coefficient of determination (R²) to identify the appropriate combination of parameters to simulate the DO. Results suggest that both types of ANFIS models work adequately and accurately predict the DO; however, ANFIS–GP outperforms the ANFIS–SC. M4 generated R² of 0.953 from ANFIS–GP compared to 0.911 from ANFIS–SC.

Key words: dissolved oxygen, grid partitioning, subtractive clustering, water quality, Yamuna River

HIGHLIGHTS

• ANFIS models were designed to predict the DO of urban steam.
• Fuzzy logic allows the classification, data mining, interpretation and optimization of time series data.
• Simulation was performed using ANFIS-grid partitioning (ANFIS–GP) and ANFIS–subtractive clustering (ANFIS–SC).
• The extensive formulation of the rule base helps identify vital parameters and improves the accuracy of the model.
INTRODUCTION

The production and consumption of dissolved oxygen (DO) in rivers are dynamic and complex (Zahraeifard & Deng 2012). DO remains in water as free oxygen and its concentration varies due to diffusion. The concentration of DO in water depends on several sources, sinks and solubility rates. The atmosphere is the most significant external source of oxygen to stream, and photosynthesis plays a significant role as an internal source of oxygen (Lyons et al. 2014). Photosynthesis contributes more oxygen to water because the oxygen generated from the algae contains pure oxygen, whereas the atmospheric diffusion contains only 20% oxygen in overall gas transfer at the air–water interface (Holtgrieve et al. 2010). All the microorganisms, aquatic plants and aquatic animals consume oxygen through respiration, known as sinks and remain active throughout the day and night. In contrast, photosynthesis generates oxygen only during the daytime, and the algae act as both sources and the sink of oxygen (Arora & Keshari 2021b). Another critical factor is solubility, which depends on the water pressure, temperature and salinity. The increase in pressure increases the solubility of gas, whereas higher salinity and temperature reduces the solubility rate (Cox 2003; Verberk et al. 2011). A healthy riverine ecosystem maintains a synchronization in the sources and sinks of oxygen; however, several factors affect the DO concentration in the river along with the depth of the water body. Mathematically, the concentration of DO can be expressed as:

\[ \text{DO} = \text{DO}_{so}S - \text{DO}_{si} \]  

(1)

where \( \text{DO}_{so} \) is the source of DO, \( \text{DO}_{si} \) is the sink of DO, and S is solubility. A low concentration of DO in a river for a long duration increases the start of several environmental problems (Ay & Kisi 2012). The river system’s biota is affected if the oxygen content falls below 30% of the saturation limit. The variation in DO concentration occurs rapidly based on flow available in rivers, velocity, turbulence, the number of organics and atmospheric reactions involved in the riverine system (Cox 2003; Quick et al. 2019). Anthropogenic activities are becoming the significant sinks of oxygen that consume the available DO through partially or untreated wastewater from domestic, industrial, commercial and agricultural sectors (Arora & Keshari 2021b). It is mandatory to maintain equilibrium between sources and sinks for the aquatic ecosystem’s sustainability.

The assessment of DO variation for heavily polluted rivers based on statistical methods is not the appropriate approach nowadays due to complex and nonlinear water quality parameters (Cox 2003; Parmar & Keshari 2012; Arora & Keshari 2021b).
Various researchers have used machine learning techniques such as artificial neural networks (ANN) and adaptive neuro-fuzzy inference system (ANFIS) to predict the variation, simulate and forecast the water quality parameters (Singh et al. 2009; Chen & Liu 2014; Ay & Kisi 2017; Tiwari et al. 2018; Shah et al. 2021, Alsulaili & Refaie 2021).

Fuzzy logic has several advantages in classification, data mining, interpretation, and optimization of time series data of various fields (Wijayasekara & Manic 2014; Tiwari et al. 2018). The fuzzy theory has been widely used to model the nonlinear behaviour for various hydrological applications (Altunkaynak et al. 2005; Keskin et al. 2006; Chang et al. 2015; Khan & Valeo 2015; Ay & Kisi 2017; Arora & Keshari 2021a). The fuzzy system can remove the uncertainties from the data and develop the model structure through the rule-based system (Huang et al. 2010; Shah et al. 2021). Altunkaynak et al. (2005) used the Takagi–Sugeno fuzzy logic approach to model fluctuations in DO at Golden Horn, Turkey, and compared the results with autoregressive moving average (ARMA) models. The results reveal that the fuzzy models are superior to ARMA in predicting DO fluctuations. Güldal & Tongal (2010) identified the variation in the water depth in the lake and compared the accuracy of recurrent neural networks (RNN), ANFIS and stochastic models using the coefficient of determination. They found that RNN and ANFIS performed better than stochastic models. Moosavi et al. (2013) compared different data-driven models to predict a reservoir’s groundwater level at two distinct basins. The researchers used ANN, ANFIS and ANN–ANFIS coupled models and found that the ANFIS and combinations of various models perform better than the ANN due to the errors involved in selecting the adequate number of neurons for the ANN model. ANFIS is also better than ANN because of its capability of analyzing uncertainties in input parameters. Parmar & Bhardwaj (2015) compared regression, ANN, Wavelet and ANFIS to predict chemical oxygen demand (COD) in the Yamuna River, India. They also compared the conventional techniques with the wavelet-coupled model. Khan & Valeo (2015) applied fuzzy regression and compared it with the Tanaka and Diamond method of fuzzy modelling to predict the DO and found that the ability to record water quality parameters’ uncertainty makes the fuzzy regression technique a substantial approach for predicting DO. Shah et al. (2021) compared the performance of various ANFIS models by varying the type and number of membership functions (MFs) to predict the electrical conductivity (EC) and total dissolved solids (TDS) in the Indus River. The ANFIS model was developed using three MFs where triangular and Gaussian MF type was used for EC and TDS, respectively. The study generated high coefficient of correlation (R) of 0.91 and 0.92 for EC and TDS, respectively, and revealed that with pre-processing of data and selecting appropriate parameters, ANFIS can simulate the water quality parameter with lesser intricacy than deterministic models.

The literature review reflects that the fuzzy modelling techniques can be applied to a wide area with high accuracy. However, detailed studies over the differences between the two approaches (subtractive clustering [SC] and grid partitioning [GP]) of fuzzy modelling are unavailable. The application of a correct approach for the prediction of the parameter can improve the simulation results significantly. The derivation of fuzzy models depends on linguistic terms designed via MFs and delivers input parameters to the optimization model (Cordón 2011). As the DO acts as the health indicator of the riverine system, the predominant tasks are the accurate prediction of DO for assessing the state of the water body, designing policies for the water resource management, and appropriate allocation of available water that keeps a sufficient amount of flow in rivers. In this study, the GP and SC approaches are applied to model the DO of the stream passing through a highly urbanized area, which receives a large volume of wastewater from domestic, industrial and agricultural sources through multiple drains.

**MATERIAL AND METHODOLOGY**

ANFIS models were developed for the simulation and prediction of DO. A hybrid algorithm combining the least-squares and gradient descent methods was used to conserve the search space and minimize the model’s operational time. The model’s structure is designed using GP and SC methods, and various combinations of input parameters are tested using both methods.

**Adaptive neuro-fuzzy inference system**

ANFIS is the combined structure of neural network and fuzzy logic. This composite structure allows neurons to record the input data and fuzzy rules to optimize the solution. The fuzzy sets in the model define the fuzzy rule base and make the ANFIS capable of simulating the nonlinear behaviour of input parameters. The rule base of the network increases with the number of input parameters. However, it also increases the computational time of the model (Chang & Chang 2006). The ANFIS structure uses five layers: input, fuzzification, normalization, defuzzification and output layers. The number of
input parameters is defined in the first layer. Fuzzification includes the distribution of MF to each input parameter and allocation of type of MF. For splitting the ANFIS input function, the model uses fuzzy MFs, which cover the input space and activate several local regions simultaneously using single input through overlapping. The number of MFs plays a vital role as the MFs control the partitioned input function’s resolution and the ANFIS model’s approximation.

If-then rule bases are formed based on the number and type of MF defined in the previous step. The fuzzification covers the input into breakable fuzzy sets, and defuzzification again converts the fuzzy sets into output after applying inference processes, normalization and optimization (Chang & Chang 2006). However, the fuzzy inference system (FIS) rule base can be altered by understanding the relationship between input parameters and reducing the computational time with optimized output. The alteration of the rule base and the modified structure of FIS make it worthwhile for wide application over neural networks (Arora & Keshari 2020). FIS is designed using Gaussian type MFs with a hybrid learning algorithm to optimize the model. FIS structure depends on the type and number of MF selected for modelling (Babuška & Verbruggen 2003; Sonmez et al. 2018). The overall architecture of the FIS model is shown in Figure 1.

**Grid partitioning**

GP is commonly used to design the FIS, a fuzzy clustering method when input variables are less. The input space is partitioned based on the minimum distance between each input variable divided into two member functions. The problem region is divided into sub-regions, and input space is further divided into sub-regions to refine the space depending on the type and number of MFs selected for designing the model. The partitioning method is preferred when the knowledge about the centre’s distribution is not adequate (Benmouiza & Cheknane 2019). The rule base of grid partitioned FIS is

![Figure 1 | ANFIS model structure.](http://iwaponline.com/wst/article-pdf/84/10-11/3359/967727/wst084103359.pdf)
defined as:

\[ x_1 = A_1^{k_1}, \ x_2 = A_2^{k_2}, \ldots , \ x_n = A_m^{k_m} \]  

\[ y_m = y^{k_1k_2 \ldots k_n}, \ k_i = 0, \ 1 \]  

where \( x \) is the input region varying from 1,2,...,\( m \) for \( m \)th sub-region and \( A_m^{k_i} \) is the fuzzy term. If \( k_i = 0 \), then \( A_m^{k_i} = a_i \), where \( a_i \) is the minimal value of the input parameter. If \( k_i = 1 \), then \( A_m^{k_i} = b_i \), where \( b_i \) is the maximal value of the input parameter, and both the values would be computed using the least square method. The input sub-region is divided into \( m \)th sub-regions, where \( x = x_1, x_2, x_3, \ldots, x_m \). The MF for the fuzzy term \( A_m^{k_i} \) would be:

\[ \mu^0_m(x_i) = \frac{b_m - x_i}{b_m - a_m} \] (for \( k_i = 0 \))

\[ \mu^1_m(x_i) = \frac{x_i - a_m}{b_m - a_m} \] (for \( k_i = 1 \))

where \( \mu^0_m(x_i) \) is the MF. The output (O) corresponding to \( m \)th sub-region is written as:

\[ O = \sum_{k_1k_2 \ldots k_n} A_m^{k_1k_2 \ldots k_n}(x_i)y^{k_1k_2 \ldots k_n} \]  

The sub-regions are divided on the maximum value of error from the training samples. Once the maximum errors of every sub-region are achieved, the region splits into two regions and the new approximation error is the minimum of the new sub-region. The sub-region splitting continues until the errors become constant in two regions. The sub-region into multiple regions is shown in Figure 2. The maximum error obtained from the sub-region at which split occur is written as:

\[ E_m = \frac{1}{N_m} \sum_{x_n} \left[ \frac{1}{2} (x_j - t_j)^2 \right] \]  

where \( E_m \) is the error obtained from \( m \)th sub-region from the \( N_m \) training samples and \( x_j \) and \( t_j \) are the output generated from the model and targeted, respectively, from \( j \)th training samples. The computational time in GP increases exponentially with the number of input parameters and MFs.

**Subtractive clustering**

In SC, the rule base formed is equivalent to the MF formed. In this method, each data point is considered the centre, and the importance of each centre is identified through the data point in the centre’s neighbourhood. The process runs through several iterations and allocates the centre by identifying the most influential centre with the highest number of data points in its surrounding. The radius of the cluster of points is identified using the centre of neighbouring points. The process repeats until

**Figure 2** | Grid formation of ANFIS–GP.
all the data points fall within the radius of every cluster. The potential of the data point is written as:

\[ P_i = \sum_{j=1}^{n} \exp \left( -\frac{||x_i - x_j^2||}{0.5r^2} \right) \]  

where \( P_i \) is the potential index of \( x_i \) data points and \( r \) is the radius where all the neighbourhood's data points fall. The second iteration is calculated as:

\[ P_i' = P_i - P_{c1} \exp \left( -\frac{||x_i - x_{c1}||^2}{0.5r_a^2} \right) \]  

where \( P_{c1} \) represents the potential of cluster 1 and \( r_a \) is the \( K_i r_b \), where \( K_i \) is the positive constant usually 1.5 and \( r_b \) is the neighbourhood radius. The process is repeated, and a cluster radius is recalculated until a sufficient number of clusters centres are not generated.

**MODEL DEVELOPMENT AND EVALUATION**

The sampling data is divided into two parts of 70:30 to design the model, where 70% of sampling data points are used for model training, and 30% are used for testing the model. Different sets of parameters are selected for designing the model. The four models are designed using temperature, biological oxygen demand (BOD), COD, conductivity and ammonia. The spatial and temporal analysis of data suggests that temperature, BOD, COD and ammonia produces the most significant impact on the variation in DO concentration, hence they are used for the model development (Arora & Keshari 2021b). The first model (M1) is developed considering temperature, BOD and COD as input parameters. The additional parameter selected in the second model (M2) is conductivity. The presence of ammonia reflects the generation of algae in the water, which acts as both the source and sink of oxygen. The third model (M3) is designed by combining the base parameters with ammonia.

The fourth model (M4) covers the combined effect of conductivity and ammonia over the base parameters to simulate the river’s DO. The GP and SC fuzzy clustering algorithms were used to design the model with similar input parameters. ANFIS input optimization technique was applied to identify the appropriate combination of input parameters. Four models were designed with different input parameters to observe the contribution of each parameter in affecting DO concentration. The spatial and temporal behaviour of the parameters and results of hierarchically aligned cluster analysis and principal component analysis (PCA) is used to select the appropriate parameters (Arora & Keshari 2021b). The input parameters selected to design the FIS models are shown in Table 1.

The coefficient of determination (\( R^2 \)) and root mean square error (RMSE) were evaluated to analyze the model’s performance. When the RMSE is closest to zero, it indicates the model is adequate, and when \( R^2 \) is closest to 1 it represents a better correlation between the observed and the predicted values obtained from the FIS model. Formulas used to identify the

| Model | Parameters | Output |
|-------|------------|--------|
| M1    | Temperature, BOD, COD | DO     |
| M2    | Temperature, BOD, COD, conductivity | DO     |
| M3    | Temperature, BOD, COD, ammonia | DO     |
| M4    | Temperature, BOD, COD, conductivity, ammonia | DO     |
performance of models are:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (p_i - o_i)^2}{\sum_{i=1}^{n} (p_i - \bar{o}_i)^2}$$

where $p_i$ is the predicted value of DO and $o_i$ is the observed value of DO and $\bar{o}_i$ is the mean observed value of DO.

**STUDY AREA**

Delhi is one of India’s largest and most dense cities, and all the wastewater generated from various sectors (domestic, commercial, industrial and agricultural) of Delhi joins the Yamuna River. Some part of wastewater before its confluence with the river passes through the treatment processes. However, the percentage of wastewater that is treated is too low compared to the untreated wastewater that is discharged into the drain through irregular means and subsequently joins the river. The Yamuna River travels about 375 km before reaching Delhi, and the flow of the river is obstructed at the Wazirabad barrage for water supply. The freshwater remains low throughout the year, and only wastewater from Delhi flows in the river except during the monsoon period. Monthly water samples for 5 years were collected from Nizamuddin, Delhi, which is located 16 km downstream of the Wazirabad barrage in Delhi. In between the Wazirabad barrage and Nizamuddin, the Yamuna River receives effluents from several drains, out of which maximum effluents are discharged by the Najafgarh drain, which contains 2.5 times more water than is available in the river (CPCB 2006). As the Najafgarh drain joins the river just after the Wazirabad barrage (0.5 km downstream), it causes maximum damage to the river’s water quality (CPCB 2006).

The untreated or partially treated wastewater discharge from 16 drains makes the Delhi stretch of the Yamuna River one of the most polluted sections of the river, with 22 km between the Wazirabad and Okhla barrages. The water quality falls into category E of the designated best-use water quality criteria of Indian water quality standards, which indicates that the river’s water is not fit for drinking, even after advance treatment (CPCB 2006). The BOD load increases in the river up to 80 tonnes/day after the Najafgarh drain’s confluence. The sampling location also receives the effluents from a thermal power plant on the river’s right bank. The flow of the river is obstructed by several structures in the study area, which includes six road bridges, two railway bridges and two metro railway bridges that causes silting near the bridge piers. The DO content of the river falls to zero in this stretch and causes a significant degradation of aquatic plants and animals. The generation of anaerobic conditions has also been observed in the river due to the decomposition of organic matter in the absence of oxygen.

The sampling location is selected considering distance from the Wazirabad barrage, confluence of drains and time taken by the flow to provide sufficient mixing of wastewater with the river water to represent a homogeneous mixture. The river receives the effluent load from the right bank only and has heavy habitation close to the right bank, as shown in Figure 3. On the left bank, the river has a flood plain to cater for the excess water during a flood; however, the encroachment of the flood plain is another problem of the Yamuna River as it causes narrow channels of flow during the monsoon period and a river flow with low discharge during the rest of the year.

Monthly water samples were collected for 5 years (Arora & Keshari 2021b) and physiochemical analyses were carried out as per the standard methods (APHA 2005). In-situ measurements were performed for temperature, DO and pH of the water samples. The analysis of BOD, COD, ammonia, total Kjeldahl nitrogen (TKN) and conductivity were performed in the laboratory within 48 h of collection, after preserving the sampling with hydrochloric acid at temperature below 4 °C in dark.

**RESULTS AND DISCUSSION**

The experimental analysis shows significant variation in all water quality parameters except pH. During summer, the average temperature remains around 33 °C, which falls in winter to 11 °C. The DO remains zero throughout the year due to the low fresh water flow and confluence of wastewater drains at regular intervals. The BOD and COD were significantly higher than the discharge norms (CPCB 2006). Similarly, ammonia and TKN were also much higher throughout the year, except during the monsoon season. Delhi receives precipitation in the form of rainfall from July to September, and sufficient water flows
into the river during this period, reducing BOD levels to below 50 mg/L, and increasing DO to 2 mg/L. The minimum values observed for ammonia, TKN, BOD and COD are caused by the dispersion of impurities due to excess flow received in the monsoon season. This indicates the poor condition of the river ecosystem, but it is better than the river's non-monsoon state. The descriptive statistics of water quality parameters used for the study are shown in Table 2.

### Input optimization

The effectiveness of the models depends on the type and number of parameters selected for the simulation of output. BOD and COD were selected in every model and produced a direct and substantial impact on the DO concentration, whereas

### Table 2 | Descriptive Statistics of water quality parameter at Nizamuddin

| Parameters | Samples | Minimum | Maximum | Mean    | Standard deviation |
|------------|---------|---------|---------|---------|--------------------|
| Temperature| 60      | 11      | 33      | 24.20   | 5.915              |
| DO         | 60      | 0       | 5       | 0.59    | .993               |
| pH         | 60      | 7       | 8       | 7.53    | .256               |
| Ammonia    | 60      | 1       | 39      | 15.94   | 9.259              |
| TKN        | 59      | 2       | 47      | 21.13   | 10.916             |
| COD        | 60      | 7       | 161     | 69.65   | 32.530             |
| BOD        | 60      | 2       | 57      | 23.07   | 13.433             |
| Conductivity| 60    | 276     | 1752    | 1186.85 | 437.892            |

Figure 3 | Inset location of Nizamuddin sampling location in Delhi, India (the right part of the image was captured from Google Earth on 12 April 2021).
temperature affected the regeneration rate of DO from the atmosphere. The other parameters selected in M2 and M3 (i.e., ammonia and conductivity) were also found as principal factors in fluctuating DO values. The previous studies show that ammonia and TKN follow a similar pattern of variation and produced a similar degree of flux, and therefore only ammonia was selected for the model development.

**ANFIS model analysis**

The Takagi–Sugeno algorithm was used for the development of the model. Three MFs of Gaussian type were selected for each input parameter, and output is generated from the constant type MF. The number and type of MFs are selected based on the data distribution and computational time. The performance of GP depends on the fineness of grids and the SC calculations are proportional to data points. The model structure of GP and SC methods is shown in Table 3. In all the models based on GP, three MFs are used for each input, whereas the number of MF is generated automatically in SC, depending on the data points of each parameter and the number of input parameters in each model. In M1, there are three input parameters and five MFs are used; however, M2 and M3 have four input parameters in each model. Therefore, six MFs are used in both models. However, M4 contains five input parameters, and nine MFs are used to design the GP-based model.

Along with the increase in input parameters, the rule base also elevates exponentially in GP, whereas SC considers each data point as a centre of cluster. The computational time in GP increases exponentially with the number of input parameters and MFs. Therefore, the rule base for five input parameters and 3 MFs would become $5^3 = 243$, whereas, in SC the rule base remains low and consumes less computational space. The optimization was carried out for both partitioning methods using hybrid learning for all the models. Each model’s varying epochs were used until the observed error became constant or reduced to the minimum.

The performance of the models was evaluated using RMSE and $R^2$. Results of ANFIS–GP and ANFIS–SC, as shown in Table 4, indicate that both models produce suitable solutions for the prediction. The M1 of both ANFIS–GP and ANFIS–SC produces considerable but high RMSE compared to other models. It indicates that the input parameters used for modelling are insufficient to explain the phenomenon of DO variation in the river. However, $R^2$ of more than 0.75 indicates that input parameters are substantial factors that affect the variability in DO concentration. The performance of the M2 and M3

---

**Table 3 | ANFIS model structure**

|                        | Grid partitioning | Subtractive clustering |
|------------------------|-------------------|------------------------|
| Number of MF           | 3 for each input  | 5 for M1                |
|                        | in each model     | 6 for M2 and M3         |
|                        |                   | 9 for M4                |
| Type of MF             | Gaussian          | Gaussian                |
| Optimization model     | Hybrid learning   | Hybrid learning         |
| Fuzzy rules            | M1 – 27           | M1 – 5                 |
|                        | M2 – 81           | M2 – 6                 |
|                        | M3 – 81           | M3 – 6                 |
|                        | M4 – 243          | M4 – 9                 |

**Table 4 | Performance of developed models**

|                  | Grid partitioning | Subtractive clustering |
|------------------|-------------------|------------------------|
|                  | RMSE              | $R^2$                  | RMSE              | $R^2$                  |
| M1               | 0.642             | 0.758                  | 0.458             | 0.824                  |
| M2               | 0.181             | 0.908                  | 0.287             | 0.872                  |
| M3               | 0.308             | 0.861                  | 0.284             | 0.871                  |
| M4               | 0.049             | 0.953                  | 0.150             | 0.911                  |
models increases with the inclusion of conductivity and ammonia, respectively, compared to M1. However, the $R^2$ values indicate better performance of M2 compared to M3 in both the GP and SC methods. Results suggest that conductivity produces higher variation in DO compared to ammonia. Simultaneously, the combination of the conductivity, ammonia and other parameters is considered in M4 and it outperforms other models. The RMSE of M4 was only 0.049, and $R^2$ is 0.953.

**Figure 4** | Plot of observed and predicted DO concentration from four GP and SC models
for ANFIS–GP. The bi-plots between the observed and predicted DO from all the ANFIS models obtained from both partitioning methods are shown in Figure 4.

The performance of ANFIS–SC shows similar results for models. The highest RMSE is found in M1 and lowest in M4, which is similar to ANFIS–GP. The M2 and M3 deliver approximately similar results because both models contain a similar number of input parameters and MFs, which indicates that the output of ANFIS–SC essentially depends on the number of MFs rather than the characteristics of input parameters. The M4 model was observed with RMSE of 0.150, the lowest among all the ANFIS–SC models but higher than the M4 of ANFIS–GP, as shown in Table 4. The GP model does not use the coefficient and it relies on calculating the maximum number of rules based on the number of MF and input parameters. The mandatory optimization improves the results of the GP model for the larger number of parameters and subsequent rule bases, whereas the SC model works on good input–output relation and coefficient generated from the relation. Therefore, in the ANFIS–GP model that classifies data based on the rule base, the model performance improves with the size of the rule base and is not affected by the input–output relationship. It is evident from the results that the ANFIS–GP model outperforms the ANFIS–SC and could act as an effective tool for defining, planning and managing water quality parameters when input–output parameters are maintaining a regular variation and the input characteristics are significantly varied due to anthropogenic disturbances.

The results obtained from the present study are compared with recent studies performed to simulate the DO using various soft computing methods. Ay & Kisi (2017) developed several DO modelling techniques using different combinations of pH, EC, temperature and flow. The ANFIS models were selected for comparison with the present study, as shown in Table 5. Ay & Kisi (2017) used 2 MFs of triangular type to design the GP model. Stajkowski et al. (2020) modelled DO at two stations using normalized, standardized and spectral analysis of the ARIMA model. The normalized ARIMA model shows the highest accuracy with a configuration of 4,1,4, which states that the model is designed for fourth order of autoregressive, first order of differencing and fourth order of moving average model. In contrast, Abba et al. (2021) compared different types of neural network models with different combinations of input parameters. The emotional artificial neural network–genetic algorithm (EANN-GA) and neural network ensemble (NNE) produced the highest accuracy from the various models.

### CONCLUSION

The study was carried out for the simulation of DO using different ANFIS models. Monthly water quality data was collected from the Yamuna River for 5 years, and significant parameters were identified using spatial and temporal analysis, cluster analysis and PCA (published as a separate study). The simulation models were developed using two fuzzy clustering algorithms, i.e., GP and SC, and the results of both clustering algorithms were compared using R² and RMSE. The different combinations of input parameters were used to develop the models using the ANFIS algorithms (ANFIS–GP and ANFIS–SC), and the applicability of ANFIS models was tested using the water quality parameters of the Yamuna River. The M4 model of ANFIS–GP had the lowest RMSE and the maximum R² of 0.953 and contained three MFs of Gaussian type. However, all the models of the ANFIS–GP worked well over the ANFIS–SC and showed a good correlation with the observed values of DO.

BOD and COD are significant parameters that reflect the direct consumption of oxygen through biological and biochemical decomposition of organic matter, respectively. The study reveals that, other than BOD and COD, conductivity also significantly impacts DO and could affect the DO concentration compared to ammonia. However, both conductivity and ammonia are involved in the DO modelling and improve the model performance. It can be concluded that the appropriate

### Table 5 | Comparison of optimum model with recent studies of DO modelling

| Study            | Model                                           | RMSE | R²   |
|------------------|-------------------------------------------------|------|------|
| Ay & Kisi (2017) | ANFIS–SC                                        | 0.57 | 0.950|
|                  | ANFIS–GP                                        | 0.99 | 0.860|
| Stajkowski et al. (2020) | ARIMA (4,1,4)                              | 0.482| 0.944|
| Abba et al. (2021) | Emotional artificial neural network–genetic algorithm (EANN-GA) | 0.766| 0.934|
|                  | Neural network ensemble (NNE)                   | 0.758| 0.935|
| Present study (M4) | ANFIS–GP                                        | 0.049| 0.953|
selection of input parameter is the first and vital step for accurately predicting water quality parameters. Additionally, the assortment of specific model configurations may act as non-transparent, but will substantially improve the model performance. The extensive formulation of the rule base helps identify vital parameters and improves the model's accuracy. However, it is expected that the model's accuracy can be further improved with a larger data set that could train the model adequately without being bias towards larger data points of single variables.

CONFLICTS OF INTEREST/COMPETING INTERESTS
The authors declare that there is no conflict of interest.

FUNDING
Authors received no research grants or funding from any funding agencies to perform this study.

DATA AVAILABILITY STATEMENT
All relevant data are included in the paper or its Supplementary Information.

REFERENCES
Abba, S. I., Abdulkadir, R. A., Sammen, S. S., Usman, A. G., Meshram, S. G., Malik, A. & Shahid, S. 2021 Comparative implementation between neuro-emotional genetic algorithm and novel ensemble computing techniques for modelling dissolved oxygen concentration. Hydrological Sciences Journal 66 (10), 1584–1596.
Alsuwaili, A. & Refaie, A. 2021 Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. Water Supply 21 (5), 1861–1877.
Altunkaynak, A., Özger, M. & Çakmakci, M. 2005 Fuzzy logic modeling of the dissolved oxygen fluctuations in Golden Horn. Ecological Modelling 189 (5–4), 36–446.
APHA 2005 Standard Method to the Examination of Water and Wastewater, 21st edn. American Public Health Association, American Water-Works Association, Water Environment Federation, Washington, DC.
Arora, S. & Keshari, A. K. 2020 Monte Carlo simulation and Fuzzy modelling of river water quality for multiple reaches using QUAL2kw. Environmental Processes and Management: Tools and Practices 91 (3), 3–24.
Arora, S. & Keshari, A. K. 2021a ANFIS-ARIMA modelling for scheming re-aeration of hydrologically altered rivers. Journal of Hydrology 601, 126635.
Arora, S. & Keshari, A. K. 2021b Pattern recognition of water quality variance in Yamuna River (India) using hierarchical agglomerative cluster and principal component analyses. Environmental Monitoring and Assessment 193 (8), 1–18.
Ay, M. & Kisi, O. 2012 Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado. Journal of Environmental Engineering 138 (6), 654–662.
Ay, M. & Kisi, O. 2017 Estimation of dissolved oxygen by using neural networks and neuro fuzzy computing techniques. KSCE Journal of Civil Engineering 21 (5), 1651–1659.
Babuška, R. & Verbruggen, H. 2003 Neuro-fuzzy methods for nonlinear system identification. Annual Reviews in Control 27 (1), 73–85.
Benmouiza, K. & Cheknane, A. 2019 Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. Theoretical and Applied Climatology 137 (1), 31–43.
Chang, F. J. & Chang, Y. T. 2006 Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. Advances in Water Resources 29 (1), 1–10.
Chang, F. J., Tsai, Y. H., Chen, P. A., Coyenal, A. & Vachaud, G. 2015 Modeling water quality in urban water using hydrological factors–data driven approaches. Journal of Environmental Management 151, 87–96.
Chen, W. B. & Liu, W. C. 2014 Artificial neural network modeling of dissolved oxygen in reservoir. Environmental Monitoring and Assessment 186 (2), 1203–1217.
Cordón, O. 2011 A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: designing interpretable genetic fuzzy systems. International Journal of Approximate Reasoning 52 (6), 894–913.
Cox, B. A. 2005 A review of dissolved oxygen modelling techniques for lowland rivers. Science of the Total Environment 314, 303–334.
CPCB 2006 Water Quality Status of Yamuna River (1999–2005). Central Pollution Control Board, Ministry of Environment & Forests, Assessment and Development of River Basin Series: ADSORB/S41/2006-07.
Güldal, V. & Tongal, H. 2010 Comparison of recurrent neural network, adaptive neuro-fuzzy inference system and stochastic models in Eğirdir Lake level forecasting. Water Resources Management 24 (1), 105–128.
Holtgrieve, G. W., Schindler, D. E., Branch, T. A. & A’mar, Z. T. 2010 Simultaneous quantification of aquatic ecosystem metabolism and reaeration using a Bayesian statistical model of oxygen dynamics. Limnology and Oceanography 55 (3), 1047–1063.
Huang, Y., Chen, X., Li, Y. P., Huang, G. H. & Liu, T. 2010 A fuzzy-based simulation method for modelling hydrological processes under uncertainty. Hydrological Processes 24 (25), 3718–3732.
Keskin, M. E., Taylan, D. & Terzi, O. 2006 
Adaptive neural-based fuzzy inference system (ANFIS) approach for modelling hydrological time series. 
*Hydrological Sciences Journal* **51** (4), 588–598.

Khan, U. T. & Valeo, C. 2015 
A new fuzzy linear regression approach for dissolved oxygen prediction. 
*Hydrological Sciences Journal* **60** (6), 1096–1119.

Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. 2014 
The rise of oxygen in Earth’s early ocean and atmosphere. 
*Nature* **506** (7488), 307–315.

Moosavi, V., Vafakhah, M., Shirmohammadi, B. & Behnia, N. 2013 
A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. 
*Water Resources Management* **27** (5), 1301–1321.

Parmar, K. S. & Bhargaw, R. 2015 
River water prediction modeling using neural networks, fuzzy and wavelet coupled model. 
*Water Resources Management* **29** (1), 17–33.

Parmar, D. L. & Keshari, A. K. 2012 
Sensitivity analysis of water quality for Delhi stretch of the River Yamuna, India. 
*Environmental Monitoring and Assessment* **184** (3), 1487–1508.

Quick, A. M., Reeder, W. J., Farrell, T. B., Tonina, D., Feris, K. P. & Benner, S. G. 2019 
Nitrous oxide from streams and rivers: a review of primary biogeochemical pathways and environmental variables. 
*Earth-science Reviews* **191**, 224–262.

Shah, M. I., Abunama, T., Javed, M. F., Bux, F., Aldrees, A., Tariq, M. A. U. R. & Mosavi, A. 2021 
Modeling surface water quality using the adaptive neuro-fuzzy inference system aided by input optimization. 
*Sustainability* **13** (8), 4576.

Singh, K. P., Basant, A., Malik, A. & Jain, G. 2009 
Artificial neural network modeling of the river water quality – a case study. 
*Ecological Modelling* **220** (6), 888–895.

Sonmez, A. Y., Kale, S., Ozdemir, R. C. & Kadak, A. E. 2018 
An adaptive neuro-fuzzy inference system (ANFIS) to predict of cadmium (Cd) concentrations in the Filos River, Turkey. 
*Turkish Journal of Fisheries and Aquatic Sciences* **18**, 1333–1343.

Stajkowski, S., Zeynoddin, M., Farghaly, H., Gharabaghi, B. & Bonakdari, H. 2020 
A methodology for forecasting dissolved oxygen in urban streams. 
*Water* **12** (9), 2568.

Tiwari, S., Babbar, R. & Kaur, G. 2018 
Performance evaluation of two ANFIS models for predicting water quality index of River Satluj (India). 
*Advances in Civil Engineering* **2018**, 1–10.

Verberk, W. C., Bilton, D. T., Calosi, P. & Spicer, J. I. 2011 
Oxygen supply in aquatic ectotherms: partial pressure and solubility together explain biodiversity and size patterns. 
*Ecology* **92** (8), 1565–1572.

Wijayasekara, D. & Manic, M. 2014 
Data driven fuzzy membership function generation for increased understandability. In: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Beijing, China, pp. 133–140.

Zahracifard, V. & Deng, Z. 2012 
VART model–based method for estimation of instream dissolved oxygen and reaeration coefficient. 
*Journal of Environmental Engineering* **138** (4), 518–524.

First received 29 June 2021; accepted in revised form 15 October 2021. Available online 27 October 2021