scientific reports

Check for updates

OPEN

Forecasting water quality indices using generalized ridge model, regularized weighted kernel ridge model, and optimized multivariate variational mode decomposition

Marjan Kordani[1], Mohsen Bagheritabar[2], Iman Ahmadianfar[3,4]✉ & Arvin Samadi-Koucheksaraee[5]

Permeability index (PI) and magnesium absorption ratio (MAR) are both primary irrigation water quality indicators (IWQI) used to evaluate the efficacy of agricultural water supplies. This is considered a complex environmental issue to reliably forecast IWQI parameters without its appropriate time series and limited input sequences. Hence, this research develops an innovative hybrid intelligence framework for the first time to forecast the PI and MAR indices at the Karun River, Iran. The proposed framework includes a new hybrid machine learning (ML) model based on generalized ridge regression and kernel ridge regression with a regularized locally weighted (GRKR) method. This research developed an optimized multivariate variational mode decomposition (OMVMD) technique, optimized by the Runge-Kutta algorithm (RUN), to decompose the input variables. The light gradient boosting machine model (LGBM) is also implemented to select the influential input variables. The main contribution of the intelligence framework lies in developing a new hybrid ML model based on GRKR coupled with OMVMD. Five water quality parameters from the Karun River at two stations (Ahvaz and Molasani) over 40 years are used to forecast the PI and MAR indices monthly. Statistical metrics confirmed that the proposed OMVMD-GRKR model, concerning the best efficiency in the Ahvaz (R = 0.987, RMSE = 0.761, and U95% = 2.108) and Molasani (R = 0.963, RMSE = 1.379, and U95% = 3.828) stations, outperformed the OMVMD and simple-based methods such as ridge regression (Ridge), least squares support vector machine (LSSVM), deep random vector functional link (DRVFL), and deep extreme learning machine (DELM). For this reason, the suggested OMVMD-GRKR model serves as a valuable framework for predicting IWQI parameters.

Keywords Permeability index, Magnesium absorption ratio, Generalized ridge regression, Optimized MVMD, Kernel ridge regression, Water quality

Rivers are vital sources of water for human life, they are playing a crucial role in agriculture, electricity generation, domestic usage, and industry. However, their dynamic nature increases the risk of pollution due to variations in water flow and runoff from nearby areas. Additionally, poor waste disposal practices contribute to both physical and chemical contamination of river ecosystems. These factors exacerbate the susceptibility of rivers to ecological pollution[1,2]. Historically, water quality (WQ) problems have often received less attention in research discussions because of abundant water resources[3]. Nonetheless, as global water availability is expected to decrease due to climate change[4], there is an urgent need to address WQ to ensure a sufficient water supply for irrigation and other essential uses. Consequently, forecasting WQ facilitates the early identification of pollutants and supports cost-effective treatment strategies. forecasting WQ enables preventive measures to safeguard water

[1]Department of Hydrology and Water Resources, Shahid Chamran University of Ahvaz, Ahvaz, Iran. [2]Department of Electrical Engineering, University of Cincinnati, Cincinnati, OH 45221-0030, USA. [3]Department of Civil Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran. [4]New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Nasiriyah 64001, Iraq. [5]Department of Civil, Construction and Environmental Engineering (Dept 2470), North Dakota State University, P.O. Box 6050, Fargo, ND 58108-6050, USA. ✉email: i.ahmadianfar@bkatu.ac.ir; im.ahmadian@gmail.com

supplies, ensuring compliance with health rules and mitigating public health hazards. Therefore, forecasting WQ enhances resource utilization, lowers treatment costs, and ensures access to clean water, particularly by identifying pollutants in river systems.

Reusing water from agricultural drainage systems shows promise as a means to increase the availability of irrigation water in the face of constraints[5], including growing water shortages, runoff pollution, and conflicting demands from urban and industrial sectors. However, insufficient treatment of this drainage water have adverse effects on soil quality, agricultural growth, and irrigation infrastructure. There are questions about the efficacy of recycled water persist. Several metrics are employed to evaluate the quality of irrigation water. These metrics include residual sodium carbonate (RSC)[6], the sodium adsorption ratio (SAR)[7], permeability index (PI)[8,9], magnesium absorption ratio (MAR)[10], and potential of salinity (PS)[8]. Accurate estimating of them can be helpful for use in agricultural purposes. Thus, ongoing attempts exist to create non-physical methods for predicting such indexes[11,12]. Non-physical methods utilize computational models and algorithms, such as machine learning (ML) models, to analyze and predict phenomena without relying on direct physical measurements.

ML models offer several advantages, such as their ability to handle missing data, non-linear structure, big datasets of varying sizes, and complex events. ML techniques such as support vector machines (SVMs)[13], artificial neural networks (ANNs)[14], and least square SVMs (LSSVMs)[15] are among the most reliable computational approaches for rapid and practical WQ modeling[16,17]. Numerous studies have demonstrated the reliability and accuracy of these models. For example, the monthly sodium (Na^+) levels were predicted by two different models, namely, hybrid wavelet-linear genetic programming (WLGP) and wavelet-ANN (W-ANN) in Turkey[18]. The computed outcomes proved that the LGP model offers more accuracy compared to the ANN model. In another study, the Karun River's water quality index (WQI) was predicted by Ref.[19]. They employed advanced variable reduction methods, such as the forward selection M5 model tree (FS-M5), focusing on several WQ factors. Their finding showed that the river's WQ parameters were "Relatively Bad," with only a 19% chance of improvement. Nouraki et al. (2021) applied four ML models, namely, random forest (RF), multiple linear regression (MLR), support vector regression (SVR), and the M5P model tree to forecast total hardness (TH), total dissolved solids (TDS), and SAR in the Karun River from 1999 to 2019[20]. The study's results showed that, even with complicated pollution sources, such models could accurately predict WQ metrics.

In recent years, Ahmadianfar et al. (2022) proposed a new framework called the weighted exponential regression hybridized by gradient-based optimization (WER-GBO)[21] with the purpose of monthly sodium (Na^+) forecasts in Maroon River. When WER-GBO was compared to other methods, such as ANFIS, LSSVM, BLR, and RSR, WER-GBO showed better accuracy. Moreover, to forecast the irrigation WQI (IWQI) of Bahr El-Baqr, Egypt, the V-KELM-INFO model was created by Chen et al. (2024) to predict TDS[22]. They combined the weighted mean of vectors (INFO) algorithm, kernel extreme learning machine (KELM), and variational mode decomposition (VMD) with Boruta-XGBoost. This model greatly enhanced the accuracy of predictions made at the Iranian Idenak station compared to previous models based on metaheuristics.

The prediction of WQ indicators has been supported by a number of effective models in the work of prior scholars[19[,23,]24]. However, there is an urgent need for a high-performance, sophisticated framework because of complicated and non-linear datasets. The complexity of WQ data also is a common challenge for traditional ML models. Hence, new advanced ML and artificial intelligence methods offer tremendous potential for solving these challenges[21,25,26]. As an illustration, ridge regression, SVMs, hybrid models, deep learning methods, and others can better manage big datasets and non-linear connections. Optimization strategies also allow these models to maximize computing efficiency and forecast accuracy. Thus, to improve decision-making and long-term planning for water resources, it is essential to create and use these advanced models for WQ management.

The main shortcomings of the previous studies can be summarized as,

- Traditional ML models such as ANN, SVR, and MLR have been widely used; however, they exhibit limited effectiveness in capturing the complex, non-linear dynamics of water quality prediction.
- Decomposition methods like the VMD are highly sensitive to their control parameters. Most previous studies did not carefully optimize these settings, which could lead to less accurate results.
- Combining ML models and developing a robust and accurate framework is essential to solve complex forecasting problems. However, most previous water quality research relied on single machine learning models, which often limited the accuracy and comprehensiveness of predictions.

To address the above shortcomings, the main objective of this research is to develop a unique hybrid approach, namely the GRKR, for forecasting IWQI. The GRKR model is developed based on a combination of a generalized ridge regression, an efficient weighted least squares method with regularization, and kernel ridge regression with a wavelet kernel. The optimal control parameters for the proposed GRKR model are also determined using a mathematical optimization technique known as Runge-Kutta optimization (RUN). Decomposing the input variables through optimized multivariate variational mode decomposition (OMVMD) enhances prediction accuracy.

Consequently, the proposed intelligent framework significantly improves decision-making in WQ management by enabling water quality managers to take preventive actions through precise predictive insights. The framework integrates GRKR, OMVMD, LGBM, and RUN methods to provide reliable predictions, allowing decision-makers to make informed choices that lead to improved WQ outcomes. This research contributes to the academic understanding of ML applications in environmental science while also offering practical tools for professionals working in the field.

Material and method

Case study

The Karun River's watershed is situated in southwestern Iran. The Karun River with 950Km is the longest and most significant river in Iran. The Karun River's longitudinal and latitudinal coordinates are 48° 15′ to 52° 30′ E and 30° 17′ to 33° 49′ N, respectively. Figure 1 depicts the Karun River within the Khuzestan Plain, as well as the specific locations of the WQ monitoring units situated along the river. In recent years, the WQ of the Karun River has significantly declined because of agricultural activities, human consumption, and industrial processes. The lack of sewage networks, non-compliance with environmental regulations by major industries, and the direct discharge of wastewater into the river are the primary factors affecting the WQ in this river. As a result, forecasting WQ has become essential for monitoring and managing the river's environmental health, allowing for early problem detection, timely interventions, and informed decision-making to safeguard public health and preserve the ecosystem.

This study utilized monthly WQ data collected from two monitoring sites over a 40-year period. The time-series graph of WQ parameters is provided in Appendix A. The independent variables in this research include nine key WQ indicators: chloride (Cl), discharge (Q), sulfate (SO_4^{2-}), sodium (Na), magnesium (Mg), bicarbonate (HCO_3), calcium (Ca), electrical conductivity (EC), and total dissolved solids (TDS). The dependent variables also include two IWQI, namely: the magnesium absorption ratio (MAR) and the permeability index (PI), which are measured at the Ahvaz and Molasani stations along the river.

Table 1 presents the dataset's statistical summary, including the mean (Mean), maximum (Max), minimum (Min), and median (Med) values. It also reports skewness (Skew), standard deviation (SD), and kurtosis (Kur). Skewness and kurtosis are critical statistical measures for WQ evaluation, as they provide insights into data distribution. Skewness indicates the asymmetry of the data, showing whether pollutant concentrations tend to exceed or fall below the mean. Kurtosis quantifies the "tailedness" of the distribution, emphasizing the presence of outliers and the potential risks associated with extreme values. Therefore, these statistical factors enhance the understanding of data variability and trends, supporting informed decision-making and efficient WQ management strategies.

Concerning PI, the variable measures how water affects soil aeration and water infiltration, which are vital for plant growth. In fact, PI is used to evaluate the long-term impact of irrigation water on soil structure and permeability. High PI levels in water can hinder soil aeration and water penetration, negatively affecting plant development. When determining whether or not irrigation water is suitable for agricultural use, the MAR

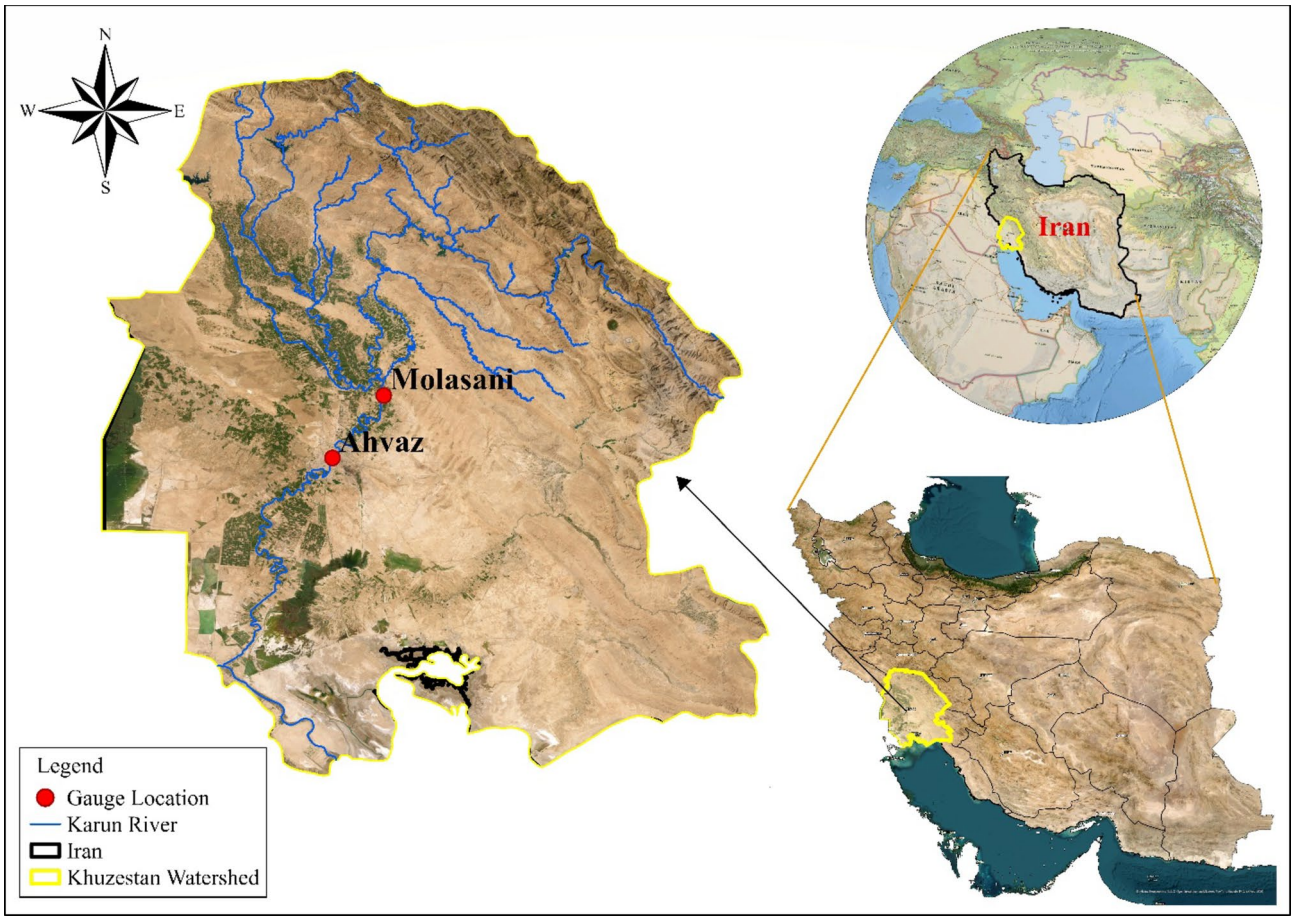

Fig. 1. Karun River location and selected two stations (Ahvaz and Molasani)[27].

Station	Variables	MAX	Mean	MIN	Median	SD	Skew	Kur
Ahvaz	SO4 (mg/L)	11.80	3.93	0.90	3.60	1.82	0.96	3.99
	Cl (mg/L)	17.80	6.87	0.50	6.70	2.73	0.97	4.66
	EC (μ S/cm)	3045.00	1362.70	128.00	1320.50	419.00	0.69	3.77
	TDS (mg/L)	1954.00	858.84	68.00	828.00	279.27	0.67	3.59
	Q (m^3/s)	4387.00	701.41	37.00	514.00	603.94	2.55	10.84
	Na (mg/L)	34.60	7.41	0.10	6.50	3.98	1.49	7.11
	Mg (mg/L)	7.80	2.55	0.20	2.30	1.07	1.19	4.94
	Ca (mg/L)	29.10	4.57	1.50	4.20	1.85	4.85	56.71
	HCO_3 (mg/L)	33.90	2.94	1.20	2.90	1.39	19.44	432.94
	PI	89.77	63.21	42.40	63.54	4.82	-0.64	6.83
	MAR	69.23	36.01	7.32	36.62	9.10	-0.06	3.78
Molasani	SO4 (mg/L)	9.50	3.69	1.10	3.30	1.60	0.79	3.16
	Cl (mg/L)	35.40	6.40	1.80	5.80	3.21	2.54	17.92
	EC (μ S/cm)	4510.00	1292.38	537.00	1234.00	442.56	1.59	9.29
	TDS (mg/L)	2490.00	813.84	278.00	771.00	280.03	1.24	6.29
	Q (m^3/s)	4656.00	684.36	0.00	504.00	596.46	2.67	11.82
	Na (mg/L)	23.80	8.03	0.33	7.15	3.85	1.05	3.97
	Mg (mg/L)	6.50	2.72	0.10	2.50	1.15	0.75	3.24
	Ca (mg/L)	15.60	4.77	0.80	4.30	1.75	1.38	6.08
	HCO_3 (mg/L)	4.90	2.92	0.20	2.90	0.54	-0.56	5.51
	PI	83.71	62.80	46.52	62.99	5.00	-0.19	3.82
	MAR	67.21	35.83	5.36	36.37	7.86	-0.30	3.71

Table 1. Statistical characteristics of Ahvaz and Molasani stations.

provides a useful metric to consider. An evaluation of the potential effects of magnesium content in water on soil and plant life is now part of the research process. To sustain fertile soil and achieve peak crop production, managing the MAR in irrigation water is critical. Here are the steps to compute the PI and MAR:

$$\mathrm{PI} = \left(\frac{[\mathrm{Na}] + \sqrt{[\mathrm{HCO3}]}}{[\mathrm{Ca}] + [\mathrm{Mg}] + [\mathrm{Na}]} \right) \times 100 \tag{1}$$

$$\mathrm{MAR} = \left(\frac{[\mathrm{Mg}]}{[\mathrm{Ca}] + [\mathrm{Mg}]} \right) \times 100 \tag{2}$$

PI and MAR were determined using parameters Ca, Na, and Mg, HCO3, which were not suitable to be employed as input parameters in ML models. Consequently, SO_4^{-2}, TDS, Q, EC, and Cl were the factors that were used as inputs.

Generalized ridge regression

The proposed generalized ridge regression (GRM) method is designed to resolve issues associated with multicollinearity and overfitting. GRM integrates ridge regression with generalized linear models (GLM)[28] to yield a powerful and flexible model. This hybrid approach updates the model coefficients continually through an iteratively reweighted least squares (IRLS) procedure. The IRLS chooses the appropriate link function, its derivative, and a variance function to be used in any given distribution namely: normal, binomial, Poisson, or gamma. The link function is used to compute the mean response at each iteration and the linear predictor. Then, it makes a weight matrix and a pseudo-response variable. The main formulas in GLM can be defined as,

$$\eta = X\alpha \tag{3}$$

where η represents the linear predictor for the observation in a GLM. X indicates the matrix of explanatory variables and α denotes the vector of coefficients.

The link function connects the expected outcome of the dependent variable with its linear component. The aforementioned function transforms a mean value for the response variables into a linear combination, even if they are not directly related to the predictors[28]. This transformation facilitates an efficient modeling using linear regression methods. The link function of GLM is expressed as,

$$f(\mu) = \eta \tag{4}$$

where μ represents the mean of the response variables. For different GLM distributions (such as normal, binomial, or Poisson), there are different link functions f according to μ and η appropriately. Various link

functions are used to connect the mean response (μ) to the linear predictor (η) in a GLM method. Each distribution, such as normal, binomial, or Poisson, requires a specific link function to accurately represent the relationship between the response variable and the predictors.

Based on Eq. (3), the coefficient α is formulated by using GLM approach. Consequently, α is obtained by the standard weighted least squares method. In order to minimize the weighted sum of squared residuals, the mentioned method optimizes the coefficients α . In this step, the residuals are weighted by the diagonal matrix w. This method is very helpful in the GLM because the distribution of response variable may not be normal and the variance might be non-constant across observations. Consequently, the weights w are adjusted to align with these model features, which after accurate approximations of α are obtained. The coefficient α can be calculated as,

$$\alpha = (X^T \times w \times X)^{-1} \times (X^T \times w \times z) \tag{5}$$

in which

$$w = diag\left(\frac{f'(\mu))}{Var(Y)}\right) \tag{5.1}$$

and

$$z = \eta + \frac{y - \mu}{f'(\mu))} \tag{5.2}$$

where $(f'(\mu))$ expresses the derivative of the link function f with respect to μ . $Var(Y)$ represents the variance function associated with the distribution of the response variable Y .

Based on the main formula for GLM (Eq. (3)), the GRM can be obtained by incorporating ridge regularization. We add the regularization term (λ_1) to the original equation (Eq. (4)). Therefore, the coefficient of GRM method is defined as,

$$\alpha = (X^T \times w \times X + \lambda_1 \times I)^{-1} \times (X^T \times w \times z) \tag{6}$$

where λ_1 indicates the regularization term. I denotes the unit matrix. To calculate the predicted value based on the GRM model ($\acute{y}_{GRM}$), Eq. (7) is employed.

$$\acute{y}_{GRM} = X\alpha \tag{7}$$

Kernel ridge regression with regularized locally weighted method

Kernel ridge regression (KRidge) is built on ridge regression (Ridge) by enhancing its efficiency, particularly in modeling non-linear relationships through kernel methods[29]. Ridge regression effectively addresses linear models and mitigates overfitting via a penalty term that shrinks coefficients; the KRidge enhances this approach by integrating kernel approaches, enabling the capturing of non-linear correlations in the data. This new feature allows the KRidge to model complicated datasets with comparability, in contrast to conventional ridge regression. Thus, the KRidge serves as a significant enhancement of the original methodology, especially for non-linear variables. Equation (8) is employed to achieve the predicted value ($\acute{y}_{KRidge}$).

$$\acute{y}_{KRidge} = X\beta \tag{8}$$

In which

$$\beta = (KrF + \lambda_2 I)^{-1} X^T \mathrm{y} \tag{9}$$

Where β expresses the regression coefficient of the KRdige, and λ_2 indicates the regularization factor of the KRidge model. KrF represents the kernel function. The wavelet kernel function was employed, formulated as,

$$KrF_{ik} = \cos\left(\rho \times \frac{-(x_i - x'_k)}{\nu}\right) \times exp\left(\frac{-||x_i - x'_k||^2}{4 \times \delta}\right) \tag{10}$$

where ρ , ν , and δ indicate the kernel function factors. The RUN optimization method was applied to derive optimal values for these factors.

This research used the regularized locally weighted (RLW) approach to compute new input variable coefficients, which improved the KRidge forecasting accuracy even more. The main formula of RLW can be formulated as,

$$\psi = (X^T \times \omega \times X + \lambda_3 \times I)^{-1} \times (X^T \times \omega \times y) \tag{11}$$

where ω denotes a kernel function based on the wavelet kernel. λ_3 indicates the regularization factor for the RLW method. The factor ψ is used to create a new kernel function based on the following formulas,

$$X_{new} = \psi X \tag{12}$$

$$KrF_{new} = KrF\left(X_{new}, X_{new,k}\right) \tag{13}$$

where KrF_{new} is a new version of KrF, which is calculated based on X_{new}.

Equation (9) can be reformulated as,

$$\beta_{new} = (KrF_{new} + \lambda_2 I)^{-1} X^T \mathrm{y} \tag{14}$$

and

$$\acute{y}_{KRidge} = X\beta_{new} \tag{15}$$

where β_{new} is a new version of β, which is calculated based on KrF_{new}.

The proposed hybrid model

A new hybrid regression model for IWQI forecasting was created in this study. It is based on two successful regression models, the GRM (sometimes called the GRKR model) and the KRidge (described above). The following relationship was defined,

$$\acute{y}_{GRKR} = a \times \acute{y}_{KRidge} + (1 - a) \times \acute{y}_{GRM} \tag{16}$$

where $\acute{y}_{GRKR}$ denotes the predicted value calculated based on $\acute{y}_{KRidge}$ and $\acute{y}_{GRM}$. a indicates a number in the interval of [0, 1]. a is achieved by optimization method (RUN). Figure 2 displays the schematic of the proposed GRKR model.

Runge-Kutta optimization (RUN)

RUN algorithm with two main operators, namely the Runge-Kutta search (RKS) engine and the enhance solution mechanism (ESM), is a mathematics-based optimization method[30]. The RUN algorithm was developed based on the Rung-Kutta method (RKM). The main components of the algorithm are described in the following sub-sections.

RKS operator

RUN algorithm uses the RKS operator to search globally and locally in the feasible space of each problem. The RKS's solution is calculated as follows,

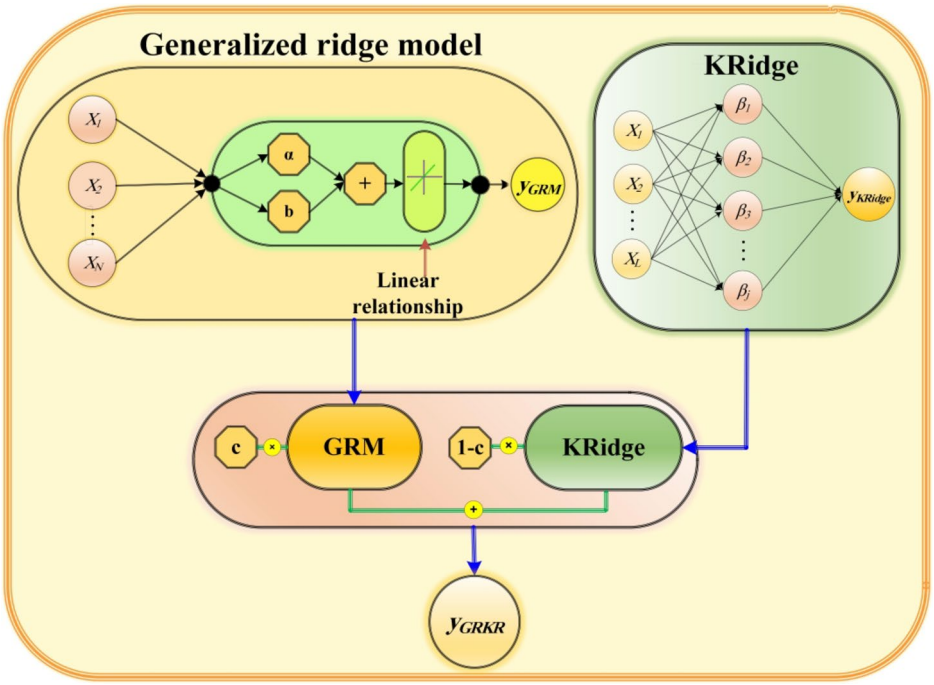

Fig. 2. Schematic of GRKR model.

$$x_{RKS} = \begin{cases} (x_a + \sigma_1 . A.c.x_a) + A.RKS + \phi\,.randn.(x_b - x_a) & if\ rand < 0.5 \\ (x_b + \sigma_1 . A.c.x_b) + A.RKS + \phi\,.\,randn.\,(x_{e1} - x_{e2}) & otherwise \end{cases} \quad (17)$$

where σ_1 indicates an integer number with the value of 1 or -1. c expresses a random value within the interval of [0, 2]. A is an adaptive factor. x_{e1} and x_{e2} express two random solutions ($e1 \neq e2 \neq n$), which are selected within the range of [1, *Np*]. *Np* indicates the number of populations and ϕ expresses a random number. RKS is defined based on the Runge-Kutta method. The RKS is thoroughly formulated in[30].

ESM operator

In order to enhance the quality of solutions and escape from local optima solutions, the ESM operator is embedded in the RUN. The solution achieved by the ESM (x_{ESM}) is defined as,

$$\begin{aligned} &\boldsymbol{if}\ \ rand < 0.5 \\ &\quad \boldsymbol{if}\ \ \boldsymbol{\rho} < 1 \\ &\qquad x_{ESM} = x_{p1} + \sigma_2 . \boldsymbol{\rho} . |\,(x_p - x_m) + randn| \\ &\quad \boldsymbol{else} \\ &\qquad x_{ESM} = (x_{p2} - x_m) + \sigma_2 . \boldsymbol{\rho} . |\,(2.rand.x_p - x_m) + randn| \\ &\quad \boldsymbol{end} \\ &\boldsymbol{end} \end{aligned} \quad (18)$$

$$\varrho\ = rand\,(0, 2)\,.exp\left(-q.\left(\frac{Iter}{MIter}\right)\right) \quad (18.1)$$

$$x_m = \frac{x_{e1} + x_{e2} + x_{e3}}{3} \quad (18.2)$$

$$x_{p1} = \theta\ \times\ x_m + (1 - \theta\,) \times\ x_{bst} \quad (18.3)$$

where θ indicates a random value within the range of [0, 1]. q expresses a random value ($q = 5 \times\ rand$), and σ_2 denotes an integer number, equal to 1, 0, or -1.

x_{ESM} may not exhibit a superior objective function compared to solution x_n (a new solution). x_{p2} is generated to explore the possibility of obtaining more promising solution. The formulation of x_{p2} is as follows:

if *rand*< ϱ

$$x_{p2} = (x_{ESM} - rand.x_{ESM}) + A.(rand.x_{RKS} + (2.rand.x_{bst} - x_{ESM})) \quad (19)$$

end.

Feature selection

Overloading ML models with excessive parameters weakens their overall performance. Many techniques are employed for input selection, mainly emphasizing linear connections. They include auto-correlation, correlation, principal component analysis, and so on[31]. This research used the LGBM data filtering technique for improving accuracy of the potential model. This technique is a nonlinear way to select parameters for input. LGBM also is a well-known gradient-boosting ML approach that consistently delivers top-notch results across several domains[32]. LGBM trains decision trees using a histogram-based approach, which divides continuous information into bins to speed up training. This approach reduces data complexity leading to faster computation and lower memory usage while maintaining high accuracy during training and testing. To promote instances with more considerable gradients, LGBM prioritizes these and incorporates an autonomous feature-selection technique. This study employed LGBM to identify the optimal variables for input, to enhance the overall effectiveness of the model and reduce the dimensionality of the forecasting problem (specifically, the number of input variables that can complicate the prediction process).

Optimized multivariate variational mode decomposition (OMVMD)

Decomposition techniques break down complicated data into separate high- and low-frequency components, enhancing clarity and streamlining analysis. A notable multivariate variant of the VMD method is the MVMD[33]. The main parameters in MVMD are the total value of decomposition methods (K) and the quadratic penalty component (ϕ). The value of IMF and the bandwidth of IMF is determined by K and ϕ , respectively. When the K value is too high, mode aliasing occurs, and when it is too low, feature extraction is inadequate, and incomplete decomposition occurs. Therefore, the optimization of the *K* and ϕ variables was achieved in the current study by using the RUN optimization method. RUN mitigated the negative impacts of tuning parameter selection processes by making sure that the MVMD decomposition factors were adequately scaled.

Employing an adaptation function as the optimization criteria is necessary for optimizing the MVMD variables. Therefore, developing an adaptive function compatible with strain time series is of paramount importance. The envelope entropy quantifies the degree of sparsity in a signal, and its numerical value exhibits an inverse correlation with the periodic nature of the signal. To put it simply, as the signal's periodicity increases, the envelope entropy decreases. Because of the mentioned conditions, the objective function chosen for optimization is the minimal envelope entropy. This function aims to enhance the extraction of periodic features from the input parameter and increase the decomposition performance by optimizing the *K* and ϕ . Here is the formula for determining the envelope entropy[34]:

$$Objective\ Function = \frac{1}{K}\sum\nolimits_{k=1}^{K} EI\,(k) \tag{20}$$

in which

$$EI_{pd}\,(k) = -\sum\nolimits_{j=1}^{m}\ pd_i \log_2 pd_i \tag{21}$$

$$pd_i = es\,(j) \,/\, \sum\nolimits_{j=1}^{m}\ es\,(j) \tag{22}$$

$$es\,(j) = \sqrt{[x(j)]^2 + \{H\,[x(j)]\}^2} \tag{23}$$

The equation EI_{pd} expresses the envelope entropy of an IMF signal x, where m and pd denote the length of the signal and probability distribution, respectively. es and $H\,[*]$ define the series of envelope signals derived by Hilbert demodulation of the signal x and the Hilbert transform, respectively.

The following represent the precise phases of the signal decomposing process based on RUN-OMVMD presented in this study, as illustrated in the flow diagram (Fig. 3):

(1) Consider the input signal $x\,(t)$ and employ the OMVMD control variable pair [K, ϕ] as the two-dimensional population of the RUN. Establish the value interval for the control factor pair, with K having a range of [2,10] and ϕ having a range of [200,5000] options. Put the RUN algorithm's parameters, such as the population size Np and the maximum number of iterations $MIter$, into the initial state.

The limits for K [2, 10] in the MVMD decomposition method prevent excessive dimensionality, balancing complexity and manageability. The range for ϕ [200,5000] allows flexibility in capturing data characteristics while avoiding excessive smoothing. These parameters are informed by previous studies[35,36] to optimize performance and maintain interpretability.

(2) Use the initialized two-dimensional population [K, ϕ] as an input variable to decompose the input features in OMVMD. Compute the fitness value of each mode and choose the initial solution with the least fitness value.
(3) Evaluate the present iteration's fitness value in relation to the previous iteration. Make sure to update the answers by substituting the current iteration's fitness value in the prior iteration if the current iteration's value is low.
(4) Use Eqs. (17–19) to revise the best solution and existing solution locations.
(5) To achieve the ideal fitness value of RUN and the accompanying optimal parameters [K, ϕ], repeat steps 3 through 4, and continue loop iterations until the maximum number of iterations $MIter$ is achieved.
(6) Utilize OMVMD to decompose the input signal into many IMFs by setting an ideal value [K, ϕ] as the control factor.

Statistical metrics

This study employs seven statistical criteria for selecting the most effective ML models in predicting IWQI. A lower result for RMSE (root-mean-square error), which may range from 0 to ∞, indicates a better match. The correlation coefficient (R) ranges from −1 to 1, with 1 showing perfect correlation. The uncertainty coefficient ($U_{95\%}$) at a 95% confidence level also ranges from 0 to ∞, signifying uncertainty levels. Vicis symmetric distance

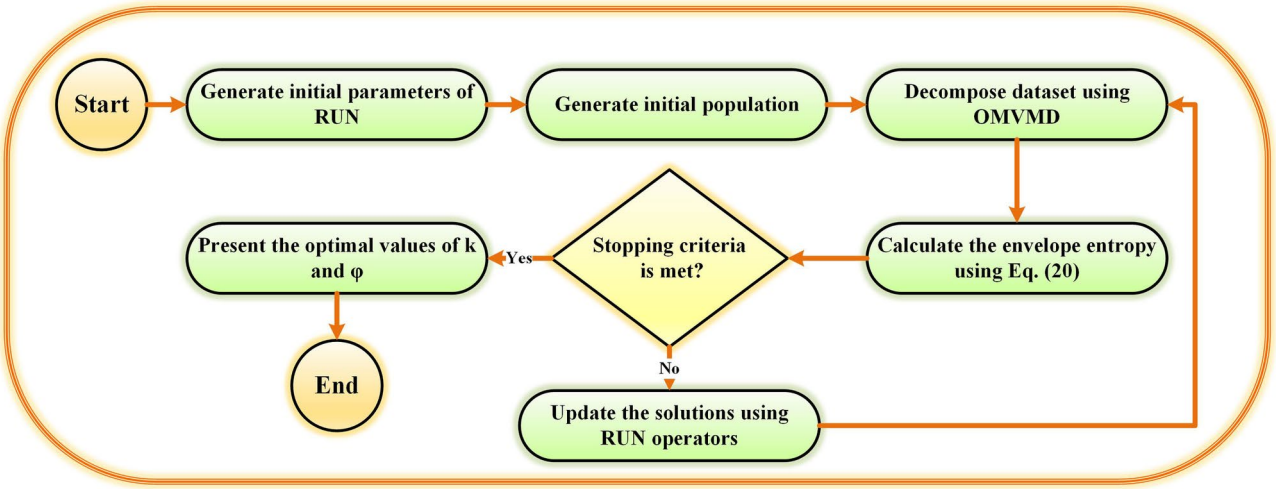

Fig. 3. Flowchart of optimized MVMD.

(VSD) ranges from 0 to ∞, where lower values suggest better similarity. The index of agreement (I_A) ranges from 0 to 1, with 1 indicating perfect agreement. Maximum absolute error (MaxAE) and mean absolute percentage error (MAPE) both range from 0 to ∞, with lower values showing better performance. Despite the fact that a few of these measures are associated with linear models, they are useful for evaluating non-linear models as well. They provide insight into the consistency and accuracy of predictions made using different methods. These metrics are formulated as,

$$R = \frac{\sum_{i=1}^{N}\left(IWQI_{Ms,i} - \bar{IWQI}_{Ms}\right) \times \left(IWQI_{F,i} - \bar{IWQI}_{F}\right)}{\sqrt{\sum_{i=1}^{N}(IWQI_{Ms,i} - \bar{IWQI}_{Ms})^{2} \times \sum_{i=1}^{N}(IWQI_{F,i} - \bar{IWQI}_{F})^{2}}} \tag{24}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{IWQI_{Ms,i} - IWQI_{F,i}}{IWQI_{Ms,i}}\right| \times 100 \tag{25}$$

$$MaxAE = max_{i=1,\ldots,N}\left|IWQI_{Ms,i} - IWQI_{F,i}\right| \tag{26}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(IWQI_{Ms,i} - IWQI_{F,i})^{2}} \tag{27}$$

$$IA = 1 - \frac{\sum_{i=1}^{N}\left(IWQI_{F,i} - IWQI_{Ms,i}\right)^{2}}{\sum_{i=1}^{N}\left(\left|\left(IWQI_{F,i} - \bar{IWQI}_{F}\right)\right| + \left|\left(IWQI_{Ms,i} - \bar{IWQI}_{Ms}\right)\right|\right)^{2}},\ 0 < IA \leq 1 \tag{28}$$

$$VSD = \sum_{i=1}^{N}\frac{(IWQI_{Ms,i} - IWQI_{F,i})^{2}}{\min(IWQI_{Ms,i}, IWQI_{F,i})} \tag{29}$$

$$U_{95\%} = 1.96\sqrt{SD^{2} + RMSE^{2}} \tag{30}$$

where $IWQI_{Ms,i}$ and $IWQI_{F,i}$ express the measured and forecasted *IWIQ* values, correspondingly. $\bar{IWQI}_{Ms}$ and $IWQI_{F}$ are the average values of the measured and forecasted IWQIs, respectively. *N* indicates the size of the dataset, and SD expresses the standard deviation of dataset.

Model development

The IWQI at two Iranian stations were predicted using several ML models in this research. A total of five ML methods, namely deep random vector functional link (DRVFL)[37], GRKR, LSSVM, deep ELM (DELM)[38], Ridge, and a stacking technique based on the regression tree model and generalized linear regression (GLM), were integrated into the proposed framework. They also incorporated the OMVMD decomposition method and the LGBM feature selection technique. The architecture that was built to anticipate IWQI is shown in Fig. 4. A three-month advance forecast of IWQI values (t + 3) was the aim of this research. The two primary steps in developing the model were:

Decomposition and feature selection approach

The MAR and PI parameters represented the output elements, whereas Q, SO_4^{-2}, Cl, TDS, and EC constituted the key input variables, as stated in Sect. 2.8. A total of 50 input parameters (5 × 10 = 50) were employed with a 10-time lag applied to each of the 5 input factors. The 10 lag-time of PI and MAR were also used as the input variables. Therefore, there were a total of 65 variables for input. The 65 variables were first deconstructed by the OMVMD approach to simplify the signals to capture different frequency patterns, isolating noise and significant trends. Simplifying signals can also be accomplished through OMVMD aimed at reducing complexity and enhancing clarity.

The mode decomposition factor (*k*) and the penalty parameter ϕ were the main adjustment parameters for the MVMD approach. They also were crucial for achieving promising accuracy. This goal was achieved by optimizing the MVMD using the RUN algorithm to determine the optimal values of *k* and ϕ at all the stations. Table 2 reports the ideal values of *k* and ϕ.

The OMVMD method used 65 input parameters to perform a simultaneous decomposition, resulting in the storage of IMFs that represented the independent variables' constituent components. The parameters for input were the initial set of 65-time delays, which improved accuracy. In the case of PI-A forecasting, for example, 65 × 8 IMFs were applied as input variables (a total of 520 variables). Then, the optimal input variables were specified using the LGBM, as mentioned in Sect. 2.5. The step towards dimensionality reduction in the input variables was the selection of the most significant features (35% of the total) for further analysis. According to this process, the procedure yielded the following feature counts: 73 for PI-A, 128 for MAR-A, 129 for PI-M, and 128 for MAR-M. The features used for the PI-A and MAR-A are shown in Fig. 5. It is worth mentioning that the important features for the PI-M and MAR-M are found in Appendix (B).

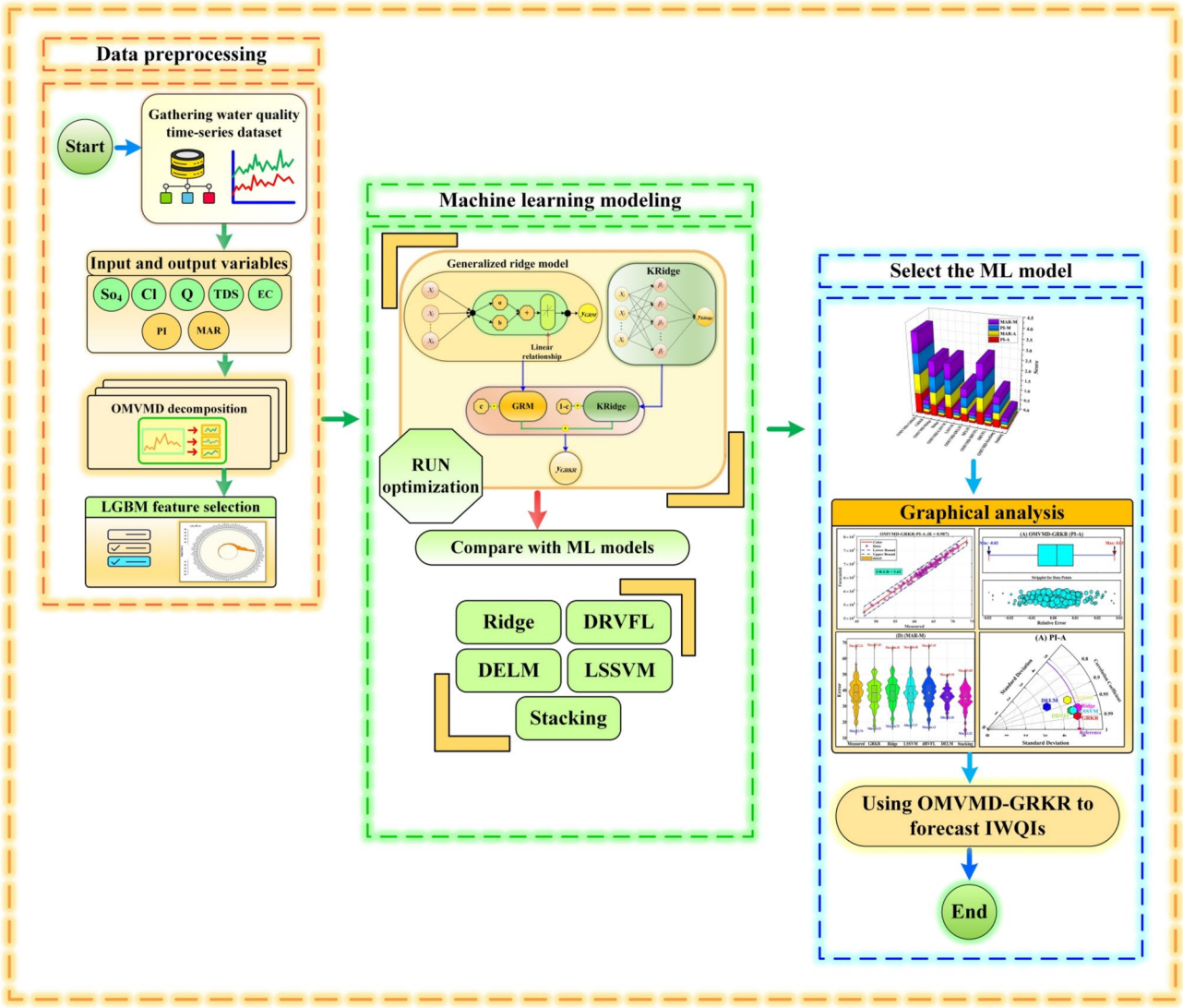

Fig. 4. Developed framework to forecast IWQIs using ML models.

Cases	k	ϕ
PI-A	8	1373
MAR-A	9	1190
PI-M	9	2603
MAR-M	8	2650

Table 2. Optimal values of MVMD parameters.

Model adjustment

Parameter tuning in ML algorithms is a significant part of model development for predicting. However, using solutions that are close to the optimal region during tuning decrease the precision of the method, which can lead to an inequitable evaluation of various prediction approaches. This occurs because models may appear less effective due to suboptimal parameter configurations. The RUN method was employed to optimize the main parameters of the GRKR method, ensuring a more thorough exploration of the parameter space and providing a fairer evaluation of forecasting techniques. The RUN method also helped to tune the tuning parameters for the other ML methods (DELM, LSSVM, Ridge, DRVFL, and Stacking). Tables 3 and 4 demonstrate the ideal values for the parameters relating to the simple-based and OMVMD-based ML methods, respectively.

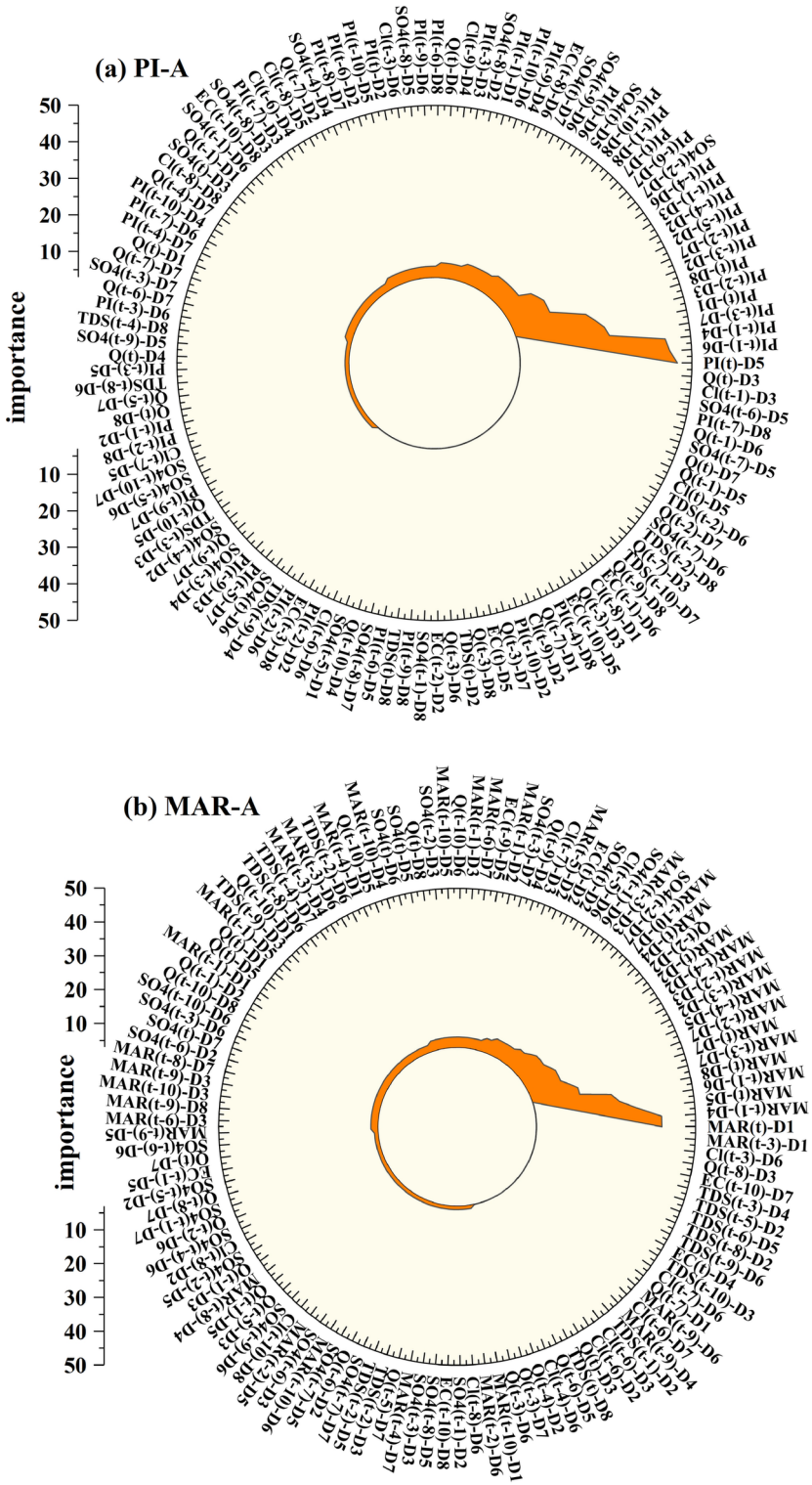

Fig. 5. Selected features for (**a**) PI-A and (**b**) MAR-A.

Results and discussion

Evaluate ML models using statistical metrics

Table 5 compares the performance of several models (simple-based and OMVMD-based) with R, RMSE, MAPE, IA, MaxAE, VSD, and $U_{95\%}$ to forecast the PI-A index. The study like previous studies are also evaluated over testing stages to better demonstrate the performance of ML models[21,22]. Almost all criteria demonstrated that the OMVMD-GRKR model performed well. When the predicted amounts compared with the observed values of the PI-A parameter, OMVMD-GRKR achieved the highest R values (0.987), the lowest RMSE (0.761) and

Cases	Models	Tuning parameter models
PI-A	GRKR	$\rho = 4.92E+04,\ \nu = 2.00E+06,\ \delta = 8.55E+05,\ \lambda_1 = 9.71E-01$ $\lambda_2 = 1.67E+00$
	DRVFL	LN*= 5, NeN* =20, SF* = 2, AF = sign, C = 0.023
	LSSVM	$gam = 1.00E-06,\ sigma = 33.30$
	Ridge	Ridge coefficient = 0.44
	DELM	NeN = [100, 20], AF = selu, regularization factor = 0.80
	Stacking	Number of trees = 105, learning rate = 0.045
MAR-A	GRKR	$\rho = 1.83E+07,\ \nu = 1.92E+08,\ \delta = 6.03E+07,\ \lambda_1 = 7.67E-01$ $\lambda_2 = 3.90E+02$
	DRVFL	LN*= 6, NeN* =15, SF* = 2, AF = sign, C = 1.00E-4
	LSSVM	$gam = 0.011,\ sigma = 103$
	Ridge	Ridge coefficient = 0.51
	DELM	NeN = [100, 24], AF = relu, regularization factor = 0.016
	Stacking	Number of trees = 120, learning rate = 0.31
PI-M	GRKR	$\rho = 6.71E+05,\ \nu = 1.75E+08,\ \delta = 2.00E+08,\ \lambda_1 = 1.02$ $\lambda_2 = 20.3$
	DRVFL	LN*= 4, NeN* =10, SF* = 2, AF = sign, C = 1.00E-04
	LSSVM	$gam = 9.20E+8,\ sigma = 9.82E+8$
	Ridge	Ridge coefficient = 25.3
	DELM	NeN = [40, 40], AF = selu, regularization factor = 1.00E-06
	Stacking	Number of trees = 100, learning rate = 1.00E-05
MAR-M	GRKR	$\rho = 8.54E+08,\ \nu = 1.72E+10,\ \delta = 1.20E+10,\ \lambda_1 = 1.1$ $\lambda_2 = 1.45E+00$
	DRVFL	LN*= 5, NeN* =20, SF* = 2, AF = sign, C = 5.00E-04
	LSSVM	$gam = 1.2,\ sigma = 2.20E+4$
	Ridge	Ridge coefficient = 502
	DELM	NeN = [100, 100], AF = selu, regularization factor = 1.00E-04
	Stacking	Number of trees = 132, learning rate = 0.052

Table 3. Control parameter values of simple-based ML models for four cases. LN*: layer number, NeN*: neuron number, SF**: scaling factor, AF*: activation function.

MAPE (0.990). The I_A value, which was near to 1, highlighted even more the strong predictive capacity of the OMVMD-GRKR model. To demonstrate minor prediction errors and variability, OMVMD-GRKR reported quite low MaxAE, VSD, and $U_{95\%}$ values compared to other models. The OMVMD-GRKR model's low values suggest that the proposed model outperforms other models in accuracy and reliability. On the other hand, simple-based models like GRKR, Ridge, LSSVM, DELM, DRVFL, and Stacking showed much higher error metrics, such as RMSE, MAPE, MaxAE, $U_{95\%}$, and VSD, along with lower R values. OMVMD-GRKR stood out among the OMVMD-based models as it successfully obtained the best outcomes for forecasting the PI-A index.

The statistical results for the simple and OMVMD-based MAR-A forecasting models are shown in Table 6. With an R-value of 0.981, an RMSE of 1.862, and a MAPE of 3.945, the OMVMD-GRKR retained its superior performance for the test data. These results demonstrated that the model maintained its reliability and accuracy even when faced with unknown data (the testing dataset). The I_A stayed high at 0.990, which supported the model's accuracy even more. Simpler models such as GRKR, Ridge, LSSVM, DELM, DRVFL, and Stacking had far lower R values and much larger errors. Consequently, the OMVMD-GRKR model was the best option among the models assessed overall for predicting MAR-A because of minimal errors, strong correlation, and a reasonable agreement with accurate data.

Table 7 shows PI-M forecasting outcomes with all ML models. The OMVMD-GRKR model displayed the highest performance in predicting PI-M, With high prediction accuracy ($R = 0.963$), low RMSE (1.379), and low MAPE and MaxAE values. It also preserved a strong I_A (0.980), reflecting high model stability. Other OMVMD-based models, such as GRKR, Ridge, LSSVM, DELM, DRVFL, and Stacking, performed better than simpler models that did not use OMVMD, but none of them were as effective as OMVMD-GRKR. This demonstrated how the OMVMD approach, combined with GRKR, improves reliability and the accuracy of forecasts.

Table 8 shows the detailed statistical findings for MAR-M forecasting. The OMVMD-GRKR model outperformed the others in predicting MAR-M. Its high R and IA values, low RMSE, and MAPE demonstrate good alignment between predicted and actual values, and lower MaxAE and VSD values show better handling of severe errors. Though they were not as successful as the OMVMD-GRKR model, other OMVMD-based models also outperformed their simpler equivalents. These analyses show that OMVMD-based methods, especially when coupled with the GRKR model, were the most successful in predicting the IWQI with high accuracy.

Figure 6 presents the ARAS (additive ratio assessment (ARAS))[39] scores for all ML models. The ARAS method is a multi-criteria decision-making (MCDM) method, which is used to evaluate and prioritize several options based on multiple criteria[40]. The ARAS technique evaluates and ranks OMVMD- and simple-based

Cases	Models	Tuning parameter models
PI-A	GRKR	$\rho = 1.07E+04,\ \nu = 2.00E+05,\ \delta = 1.05E+05,\ \lambda_1 = 7.31E-01$ $\lambda_2 = 1.90E-02$
	DRVFL	LN*= 10, NeN* =45, SF* = 2, AF = sign, C = 0.15
	LSSVM	$gam = 401,\ sigma = 2010.50$
	Ridge	Ridge coefficient = 0.025
	DELM	NeN = [1500, 1100], AF = selu, regularization factor = 0.01
	Stacking	Number of trees = 150, learning rate = 0.5
MAR-A	GRKR	$\rho = 2.66E-03,\ \nu = 1.67E+05,\ \delta = 4.34E+04,\ \lambda_1 = 7.67E-01$ $\lambda_2 = 7.47E+01$
	DRVFL	LN*= 8, NeN* =52, SF* = 2, AF = sign, C = 0.05
	LSSVM	$gam = 4.50,\ sigma = 7.38$
	Ridge	Ridge coefficient = 0.021
	DELM	NeN = [4800, 4800], AF = relu, regularization factor = 0.001
	Stacking	Number of trees = 1000, learning rate = 0.11
PI-M	GRKR	$\rho = 3.67E+07,\ \nu = 1.26E+08,\ \delta = 8.73E+06,\ \lambda_1 = 2.08E-02$ $\lambda_2 = 2.65$
	DRVFL	LN*= 5, NeN* =100, SF* = 2, AF = sign, C = 0.11
	LSSVM	$gam = 105,\ sigma = 3000.45$
	Ridge	Ridge coefficient = 5.1
	DELM	NeN = [6100, 6100], AF = selu, regularization factor = 8E-04
	Stacking	Number of trees = 700, learning rate = 0.41
MAR-M	GRKR	$\rho = 3.96E+05,\ \nu = 1.68E+08,\ \delta = 6.76E+06,\ \lambda_1 = 7.08E-01$ $\lambda_2 = 1.38E-01$
	DRVFL	LN*= 8, NeN* =62, SF* = 2, AF = sign, C = 0.05
	LSSVM	$gam = 182,\ sigma = 2000.30$
	Ridge	Ridge coefficient = 30
	DELM	NeN = [2500, 2500], AF = selu, regularization factor = 0.001
	Stacking	Number of trees = 130, learning rate = 0.75

Table 4. Control parameter values of OMVMD-based ML models for four cases. LN*: layer number, NeN*: neuron number, SF**: scaling factor, AF*: activation function.

models with considering seven performance criteria (R, RMSE, MAPE, IA, MaxAE, VSD, and $U_{95\%}$). The models' relative efficacy is graphically shown in a cumulative bar graph that compares their overall performance. Each bar represents the entire ARAS score, and higher cumulative scores suggest better performance. The goal of this research is to identify the model type that performs best overall. The cumulative scores for OMVMD-GRKR, OMVMD-Ridge, OMVMD-LSSVM, OMVMD-DELM, and OMVMD-Stacking were 2.998, 1.977, 1.935, 1.198, 2.143, and 1.126, respectively. Consequently, the OMVMD-GRKR model achieved the best ARAS score to forecast the IWQI.

Evaluate ML models using scatter plot

An essential tool for evaluating model performance in regression analysis is the scatterplot of predicted values compared to actual values. A well-fitting linear correlation model is represented by points along a diagonal line. large variations also refer to predicting trends or biases. This graph offers valuable insights for assessing the reliability and accuracy of the model. Figure 7 shows the scatter plot representing all models at the Ahvaz station. The best ML method in Fig. 7 is the one that exhibits the least disparity between the lower and upper bounds (UB-LB) of the data samples. It can be seen that the OMVMD-GRKR model for PI-A had the lowest (UB-LB) of 3.62, demonstrating its high precision in predicting compared to the other ML methods. The LSSVM and DRVFL methods had a (UB-LB) value of 4.92, which was somewhat lower than the DELM model's (UB-LB) value of 5.83. The Ridge and Stacking models exhibited larger (UB-LB) values of 7.27 and 13.17, respectively, suggesting less accurate projections for the PI-A index. For MAR-A forecasting, the UB-LB values for OMVMD-GRKR, OMVMD-DELM, OMVMD-DERVFL, OMVMD-LSSVM, OMVMD-stacking, and OMVMD-Ridge were 12.38, 14.44, 14.93, 14.98, 18.11, and 21.32, respectively. These findings showed that the OMVMD-GRKR model could achieve the smallest UB-LB, resulting in the best model to forecast the MAR-A and PI-A index.

Scatter plots of PI-M and MAR-M is presented in Appendix C. In the case of PI-M and MAR-M forecasting, the proposed OMVMD-GRKR indicated the best performance compared with the other models. The OMVMD-GRKR was able to yield the smallest values of UB-LB for PI-M (6.41) and MAR-M (13.06), indicating high accuracy and reliability compared to the other models.

Evaluate ML models using density plot

ML models' relative error distribution for PI-A and MAR-A forecasting is shown in Fig. 8. A mix of box and strip graphs is used to display the relative error distribution of all the ML models. The figure illustrates that the relative

Model	Metric	*R*	RMSE	MAPE	I_A	MaxAE	VSD	$U_{95\%}$
OMVMD-GRKR	Train	0.989	0.668	0.791	0.995	3.250	2.333	1.853
	Test	0.987	0.761	0.990	0.994	1.801	1.312	2.108
GRKR	Train	0.229	4.487	5.416	0.288	18.583	121.331	12.447
	Test	0.317	4.555	5.683	0.312	14.802	51.610	12.649
OMVMD-Ridge	Train	0.987	0.736	0.912	0.994	2.780	2.849	2.040
	Test	0.969	1.197	1.403	0.984	4.460	3.362	3.322
Ridge	Train	0.401	4.224	5.107	0.515	18.013	106.590	11.718
	Test	0.171	5.439	6.970	0.439	16.802	71.462	15.099
OMVMD-LSSVM	Train	0.986	0.782	0.956	0.992	3.420	3.242	2.170
	Test	0.971	1.167	1.416	0.984	4.460	3.110	3.216
LSSVM	Train	0.375	4.276	5.179	0.494	18.121	109.205	11.862
	Test	0.192	5.175	6.656	0.452	14.425	65.089	14.370
OMVMD-DELM	Train	0.957	1.536	1.926	0.965	5.867	13.215	4.260
	Test	0.929	2.085	2.561	0.932	7.183	10.520	5.790
DELM	Train	0.421	4.182	5.049	0.530	17.872	104.520	11.602
	Test	0.167	5.748	7.332	0.435	19.143	79.231	15.933
OMVMD-DRVFL	Train	0.976	1.025	1.265	0.987	3.473	5.647	2.844
	Test	0.973	1.115	1.426	0.985	3.118	2.828	3.096
DRVFL	Train	0.248	4.499	5.404	0.400	18.660	121.880	12.479
	Test	0.242	4.852	6.142	0.469	14.330	57.998	13.463
OMVMD-Stacking	Train	0.992	0.579	0.729	0.996	1.750	1.750	1.607
	Test	0.934	1.730	1.923	0.964	9.160	7.108	4.777
Stacking	Train	0.257	4.917	5.953	0.327	20.350	61.779	13.589
	Test	0.135	4.752	5.931	0.239	15.370	56.111	13.196

Table 5. Statistical metrics achieved by ML models for PI-A case.

Model	Metric	*R*	RMSE	MAPE	I_A	MaxAE	VSD	$U_{95\%}$
OMVMD-GRKR	Train	0.995	0.884	2.224	0.997	3.519	8.070	2.453
	Test	0.981	1.862	3.945	0.990	7.974	15.141	5.171
GRKR	Train	0.660	6.380	15.985	0.758	24.761	575.220	17.695
	Test	0.442	8.713	21.979	0.609	29.841	439.040	24.119
OMVMD-Ridge	Train	0.994	0.901	2.135	0.997	3.580	8.034	2.499
	Test	0.967	2.529	4.623	0.983	12.040	34.012	7.018
Ridge	Train	0.642	6.502	16.536	0.752	24.999	601.495	18.035
	Test	0.262	10.312	26.662	0.509	31.289	636.788	28.225
OMVMD-LSSVM	Train	0.983	1.621	3.852	0.990	8.410	27.374	4.495
	Test	0.961	2.665	6.034	0.978	7.790	31.192	7.396
LSSVM	Train	0.585	7.084	17.948	0.587	27.424	731.331	19.650
	Test	0.270	9.193	23.410	0.329	31.895	511.943	25.489
OMVMD-DELM	Train	0.995	1.157	3.111	0.995	3.890	17.144	2.999
	Test	0.920	4.968	10.834	0.899	19.440	108.370	13.222
DELM	Train	0.537	7.150	18.040	0.646	28.595	721.910	19.834
	Test	0.230	9.727	23.586	0.442	34.500	524.854	26.537
OMVMD-DRVFL	Train	0.985	1.478	3.516	0.992	9.533	22.282	4.099
	Test	0.968	2.420	5.292	0.984	8.476	25.948	6.719
DRVFL	Train	0.588	6.860	17.271	0.701	26.092	671.925	19.029
	Test	0.272	10.524	27.815	0.509	34.485	694.293	28.220
OMVMD-stacking	Train	0.994	0.950	2.558	0.997	3.140	11.439	2.632
	Test	0.895	4.623	11.118	0.919	15.530	104.424	12.685
Stacking	Train	0.666	6.321	16.208	0.773	22.930	554.675	17.533
	Test	0.281	9.846	25.685	0.509	32.800	586.958	27.136

Table 6. Statistical metrics achieved by ML models for MAR-A case.

Model	Metric	*R*	RMSE	MAPE	I_A	MaxAE	VSD	$U_{95\%}$
OMVMD-GRKR	Train	0.968	1.237	1.565	0.983	3.505	8.273	3.429
	Test	0.963	1.379	1.704	0.980	4.140	4.314	3.828
GRKR	Train	0.202	4.799	5.978	0.247	21.487	130.222	13.312
	Test	0.362	4.847	6.015	0.327	15.074	58.227	13.421
OMVMD-Ridge	Train	0.968	1.245	1.591	0.983	3.713	8.318	3.453
	Test	0.952	1.675	2.025	0.971	8.517	6.580	4.494
Ridge	Train	0.454	4.373	5.386	0.547	18.461	106.832	12.1319
	Test	0.125	6.105	7.703	0.441	14.495	91.163	15.733
OMVMD-LSSVM	Train	0.965	1.325	1.705	0.979	3.859	9.503	3.675
	Test	0.937	1.863	2.384	0.962	6.139	8.111	5.086
LSSVM	Train	0.513	4.205	5.319	0.627	17.521	98.391	11.665
	Test	0.087	6.414	8.034	0.424	15.425	101.146	16.658
OMVMD-DELM	Train	0.989	0.926	1.232	0.990	2.567	4.748	2.490
	Test	0.885	2.762	3.367	0.896	10.004	17.738	7.396
DELM	Train	0.247	4.916	6.254	0.113	20.979	137.548	13.613
	Test	0.082	5.184	6.544	0.217	16.219	66.203	14.274
OMVMD-DRVFL	Train	0.964	1.312	1.678	0.980	3.839	9.297	3.640
	Test	0.949	1.630	1.950	0.972	7.350	6.196	4.508
DRVFL	Train	0.356	4.602	5.717	0.502	18.616	119.329	12.764
	Test	0.222	5.052	6.222	0.385	18.127	63.725	13.982
OMVMD-Stacking	Train	0.999	0.237	0.301	0.999	0.694	0.300	0.657
	Test	0.787	3.559	4.279	0.850	10.429	29.382	9.328
Stacking	Train	0.508	4.225	5.326	0.605	17.738	99.440	11.721
	Test	0.100	6.186	7.770	0.428	14.296	93.446	16.108

Table 7. Statistical metrics achieved by ML models for PI-M case.

Model	Metric	*R*	RMSE	MAPE	I_A	MaxAE	VSD	$U_{95\%}$
OMVMD-GRKR	Train	0.981	1.442	3.444	0.990	5.023	21.873	4.000
	Test	0.964	2.293	5.331	0.982	7.896	22.821	6.366
GRKR	Train	0.565	6.144	16.548	0.671	25.414	562.533	17.041
	Test	0.388	8.016	19.044	0.534	26.002	313.627	22.256
OMVMD-Ridge	Train	0.966	1.953	4.833	0.981	6.695	44.606	5.418
	Test	0.947	2.893	6.751	0.970	11.081	36.825	7.866
Ridge	Train	0.591	6.014	16.086	0.698	24.431	541.240	16.682
	Test	0.112	9.626	23.625	0.426	36.006	508.858	26.368
OMVMD-LSSVM	Train	0.981	1.490	3.702	0.989	5.439	26.301	4.132
	Test	0.948	2.765	6.528	0.972	9.643	33.866	7.669
LSSVM	Train	0.438	7.221	19.869	0.148	28.550	813.038	20.030
	Test	0.176	8.874	20.710	0.288	31.773	385.385	24.251
OMVMD-DELM	Train	0.974	2.001	5.462	0.979	9.740	64.911	5.429
	Test	0.877	5.204	11.220	0.858	18.030	117.530	13.888
DELM	Train	0.459	7.254	20.325	0.192	29.518	840.909	20.067
	Test	0.158	8.802	20.664	0.257	31.281	382.005	24.167
OMVMD-DRVFL	Train	0.969	1.847	4.520	0.983	6.570	38.723	5.123
	Test	0.950	2.775	6.433	0.973	9.497	33.432	7.613
DRVFL	Train	0.565	6.149	16.374	0.695	25.266	558.116	17.056
	Test	0.170	9.354	22.987	0.451	35.876	477.016	25.686
OMVMD-stacking	Train	0.988	1.141	2.835	0.994	3.484	13.778	3.165
	Test	0.777	6.211	13.427	0.846	16.322	173.539	16.268
Stacking	Train	0.489	6.492	17.479	0.604	24.865	651.163	18.008
	Test	0.198	8.682	19.942	0.395	30.661	384.000	23.990

Table 8. Statistical metrics achieved by ML models for MAR-M case.

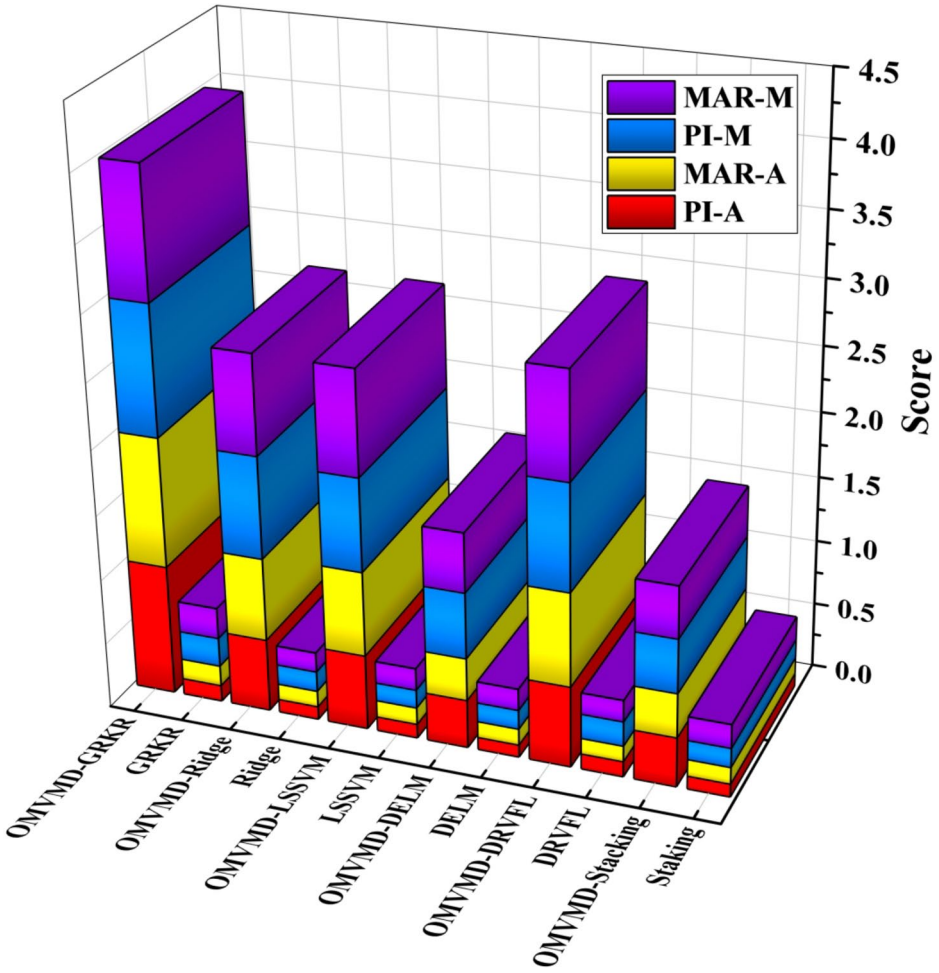

Fig. 6. ARAS score calculated for ML models.

errors of the OMVMD-GRKR model for PI-A forecasting were tightly distributed around zero, with the range of [-0.03, 0.03]. The predicted values of PI-A were quite accurate and consistent, as seen by the tightly packed data points around the zero-error line in the strip plot. For the OMVMD-Ridge, OMVMD-DRVFL, OMVMD-Stacking, OMVMD-DELM, and OMVMD-LSSVM models, the ranges of relative error were [-0.05, 0.07], [-0.05, 0.04], [-0.10, 0.14], [-0.15, 0.09], and [-0.04, 0.05], respectively, indicating a broader range of relative errors. The proposed OMVMD-GRKR model was able to achieve the smallest range of relative error ([-0.15, 0.23]) compared with the other models for the MAR-A forecasting.

Appendix D shows the relative error distribution of PI-M and MAR-M forecasting. The figure clearly indicates that the OMVMD-GRKR method outperformed the others, demonstrating the smallest range of relative error for PI-M ([-0.07, 0.04]) and MAR-M ([-0.28, 0.22]) compared to the other ML methods.

Evaluate ML models using violin plot

The violin plot (in Fig. 9) shows the distribution of PI-A, MAR-A, PI-M, and MAR-M for all ML models. The mentioned plots present a comparison between a measured reference and six prediction models (GRKR, Ridge, dRVF1, Stacking, DELM, and LSSVM) for the distribution of all IWQI (PI-A, MAR-A, PI-M, and MAR-M) values. Each plot shows the IWQI range and density; the median and the interquartile range are shown by a white dot and box. The PI-A value distribution for the GRKR model showed a high degree of resemblance by being somewhat near to the measured values. For measured, GRKR, Ridge, DRVFL, Stacking, DELM, LSSVM, PI-A ranges were [46.91, 73.70], [46.80, 73.60], [46.77, 73.28], [49.13, 72.86], [48.57, 73.23], [53.34, 70.03], and [47.84, 72.84]. Both showed the lowest minimum PI-A values (Measured = 46.91 and GRKR = 46.80). In contrast, other models such as DELM had higher minimum values (53.34) and lower maximum values (70.03). The GRKR's efficiency in accurately reproducing the measured data is shown by its general near match to the PI-A value distribution.

When it came to the other IWQI, such as MAR-A, PI-M, and MAR-M, the GRKR model consistently performed better. Predictions of IWQI made by the model showed remarkable agreement with the actual data. The ranges of suggested model (MAR-A (GRKR = [10.75, 63.33]), PI-M (GRKR = [47.16, 73.02]), and MAR-M (GRKR = [15.23, 67.83])) were somewhat comparable to the observed values (MAR-A = [11.29, 65.35], PI-M

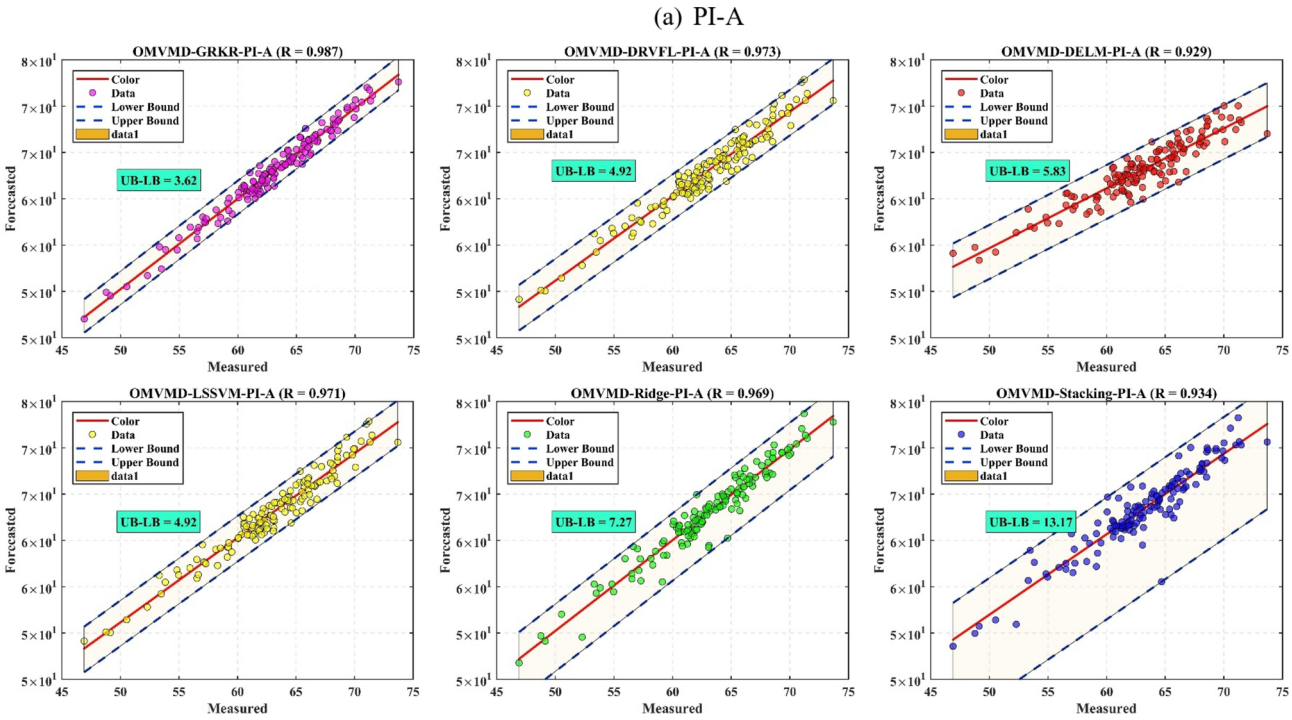

Fig. 7. Scatter plot achieved using ML models for (**a**) PI-A and (**b**) MAR-A.

= [46.52, 72.77], and MAR-M = [13.74, 67.21]). This made the GRKR model a useful ML model for forecasting IWQI as it showed its accuracy and dependability in estimating these parameters.

Evaluate ML models using residual density distribution

Figure 10 illustrates the residual density distribution across all IWQI (PI-A, MAR-A, PI-M, and MAR-M) for several predictive models (GRKR, Ridge, dRVFI, Stacking, DELM, and LSSVM). The colorful curves represent the residual distributions of the various models. Vertical dashes show all of the models' 95% confidence intervals. The residuals are spread uniformly about the average with varying degrees of dispersion and concentration since the distributions often have a central tendency around zero. the figures showed that the GRKR model with the narrowest and tallest distributions exhibited a higher density of residuals near zero for all IWQI (namely: (a) PI-A, (b) MAR-A, (c) PI-M, and (d) MAR-M). This model was able to provide more accurate and error-free predictions of the IWQI. Conversely, models with broader distributions (e.g., DELM and LSSVM) exhibited a higher variability in their prediction errors.

Evaluate ML models using Taylor plot

Figure 11 shows a comparison of six ML models' Taylor diagrams with a reference model. These models are DELM, DRVFI, Stacking, Ridge, LSSVM, and GRKR. The position of each model on the plot was estimated by the standard deviation and correlation coefficient of its data. the horizontal axis shows the standard deviation, and the radial lines show the correlation coefficient. The GRKR model was closest to the reference point for all IWQI (PI-A, MAR-A, PI-M, and MAR-M) comparing the position of ML models in the Taylor diagram. The closeness level between GRKR and the reference point indicated that GRKR had the most similarity to it. The results proved that when it came to predicting the reference dataset, the GRKR model was the most effective.

Conclusion

A complementary hybrid intelligence framework was developed for the first time to forecast monthly PI and MAR indices in Ahvaz and Molasani stations, Khouzestan province, Iran. The framework comprised a new hybrid ML model based on generalized ridge regression and kernel ridge regression with the regularized locally weighted method called the GRKR model. Furthermore, an optimized MVMD method was developed to decompose the input variables. Finally, the LGBM model was considered to select the most important features. Two stations' PI and MAR indices were predicted using the suggested framework: PI-A and MAR-A for the Ahvaz station and PI-M and MAR-M for the Molsani station. A novel hybrid ML model (GRKR) and an optimized MVMD based on the RUN optimization technique were developed as the key contributions of this study. Notably, the best control parameters for ML models were derived using the RUN optimization technique.

The decomposed input variables were utilized in the GRKR model to design OMVMD-GRKR to forecast monthly PI and MAR. In this work, OMVMD-Ridge, OMVMD-DRVFL, OMVMD-DELM, OMVMD-LSSVM, and OMVMD-Stacking models were developed to assess the forecasting precision versus the OMVMD-GRKR model utilizing seven statistical metrics (R, RMSE, MAPE, IA, MaxAE, VSD, and U95%). The findings

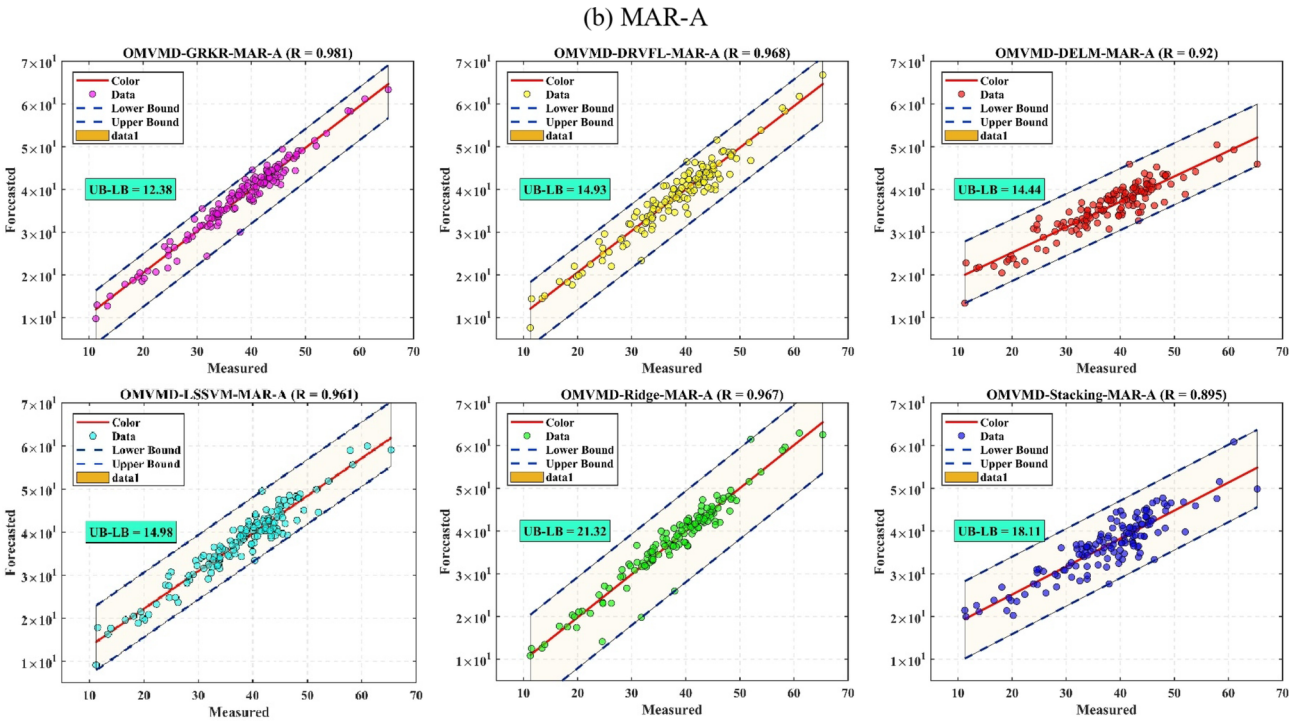

Figure 7. (continued)

showed that, when the suggested OMVMD-GRKR compared to other OMVMD-based models, it had better performance accuracy for both stations in predicting PI and MAR indices. The OMVMD-GRKR model could achieve the highest degree of accuracy in terms of ($R = 0.987$, RMSE = 0.761, IA = 0.994, VSD = 1.312, and U95% = 2.108) to forecast the PI index at the Ahvaz station and ($R = 0.963$, RMSE = 1.379, IA = 0.980, VSD = 4.314, and U95% = 3.828) to forecast the PI index at the Molasani station.

The results show that the OMVMD-GRKR and OMVMD-Stacking models were the most and least accurate for predicting PI and MAR at two stations, respectively. The simple GRKR, Ridge, LSSVM, DELM, and Stacking models also indicated lower accuracy as compared to the OMVMD-based methods. Therefore, the OMVMD-GRKR model could yield superior accuracy compared to other OMVMD-based models to forecast PI and MAR indices. The suggested hybrid intelligence framework OMVMD-GRKR can effectively be used in the future to address climate change, sustainable energy, environmental and agricultural fields, and renewable energy.

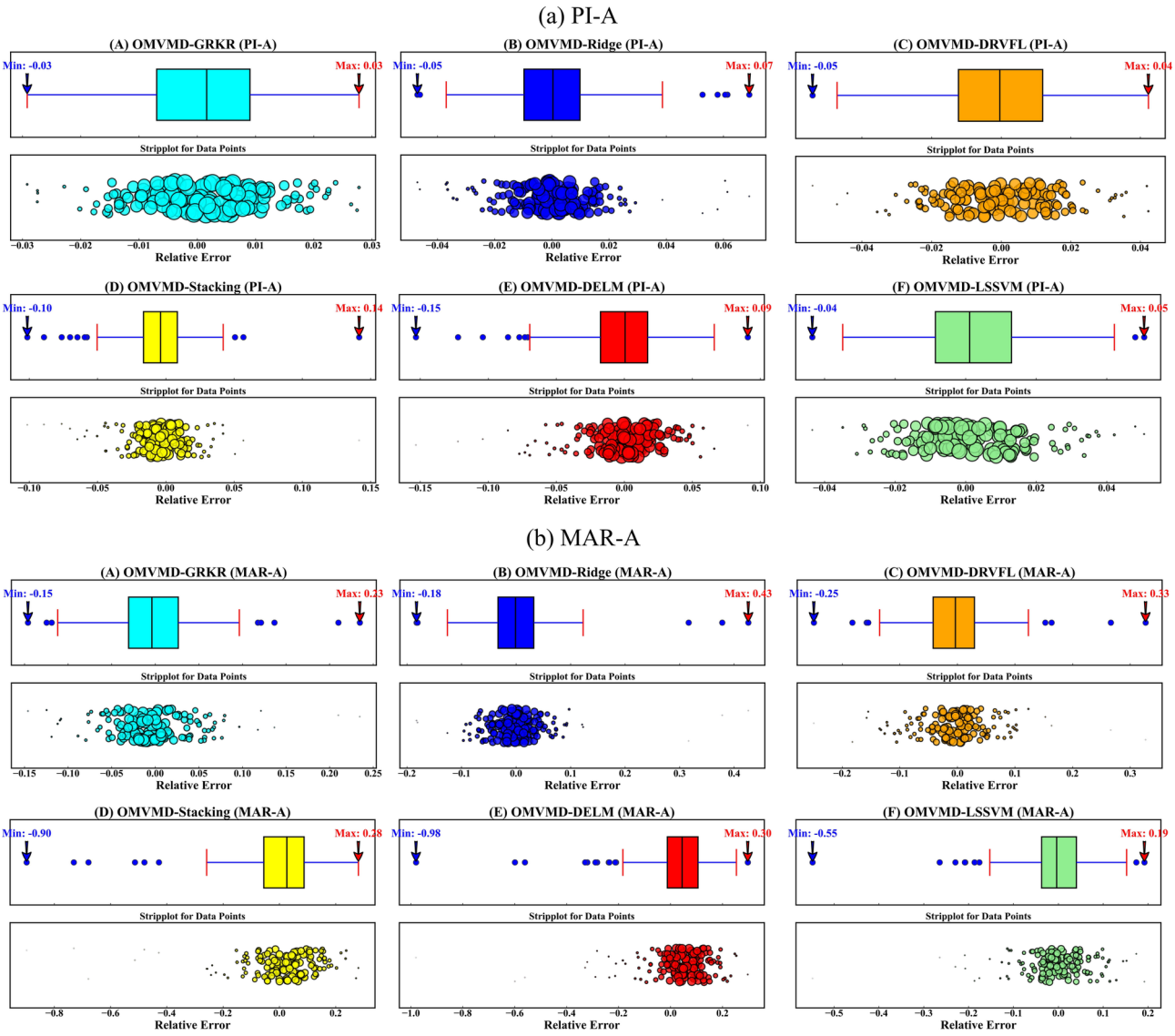

Fig. 8. Relative error distribution of ML models for (**a**) PI-A and (**b**) MAR-A.

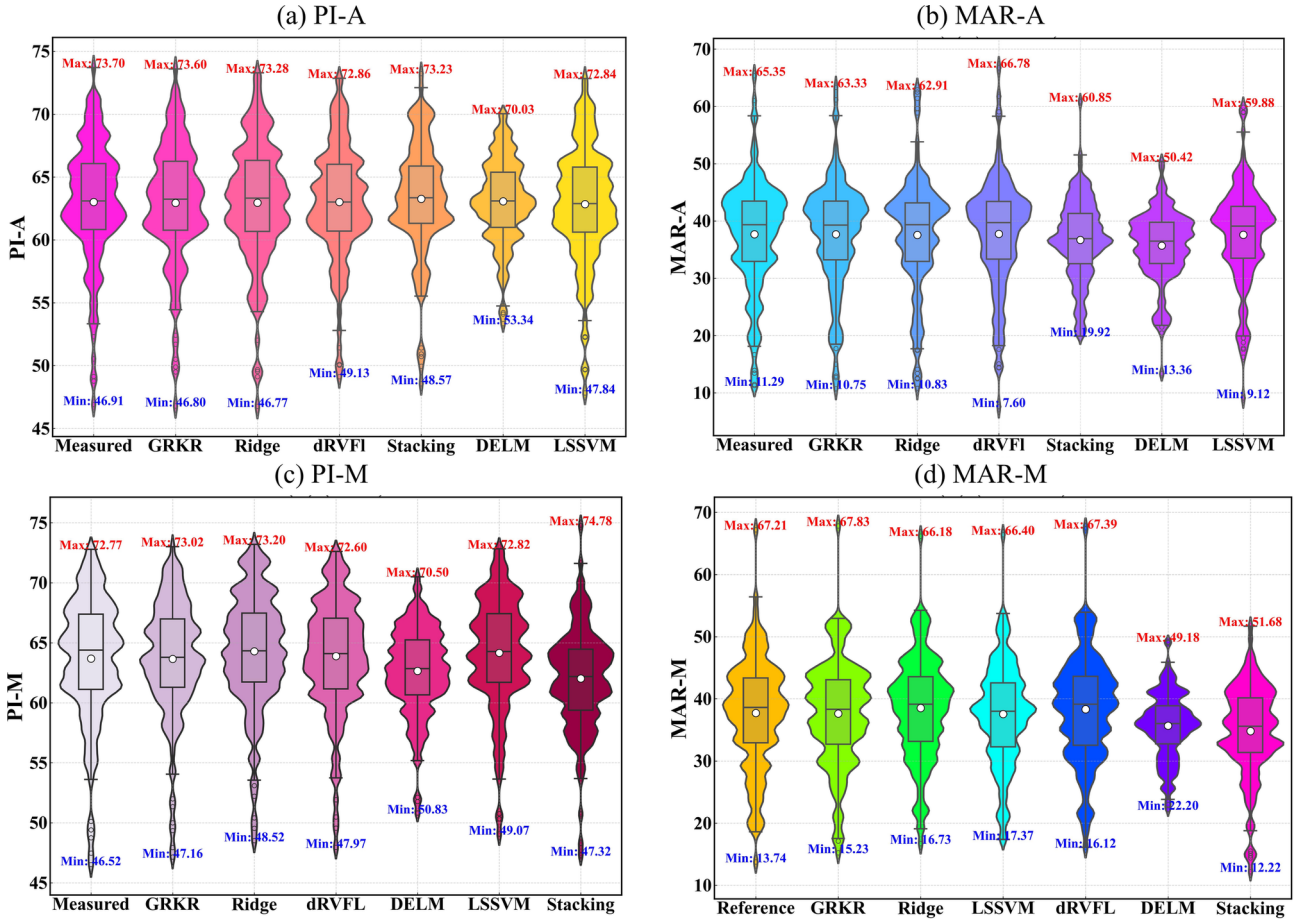

Fig. 9. Violin plot the distribution of (**a**) PI-A, (**b**) MAR-A, (**c**) PI-M, and (**d**) MAR-M for ML models.

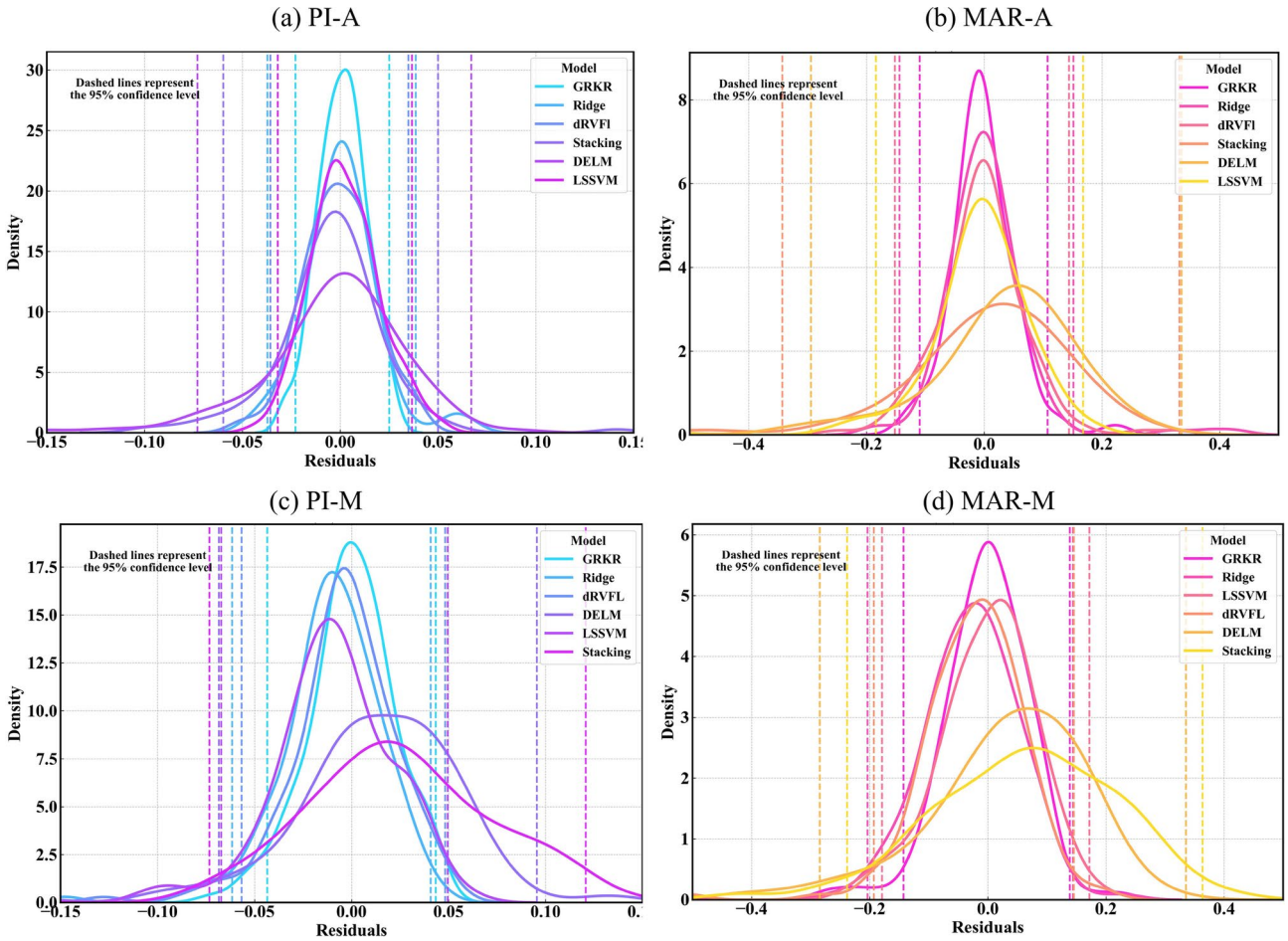

Fig. 10. Residual density distribution of ML models for (**a**) PI-A, (**b**) MAR-A, (**c**) PI-M, and (**d**) MAR-M.

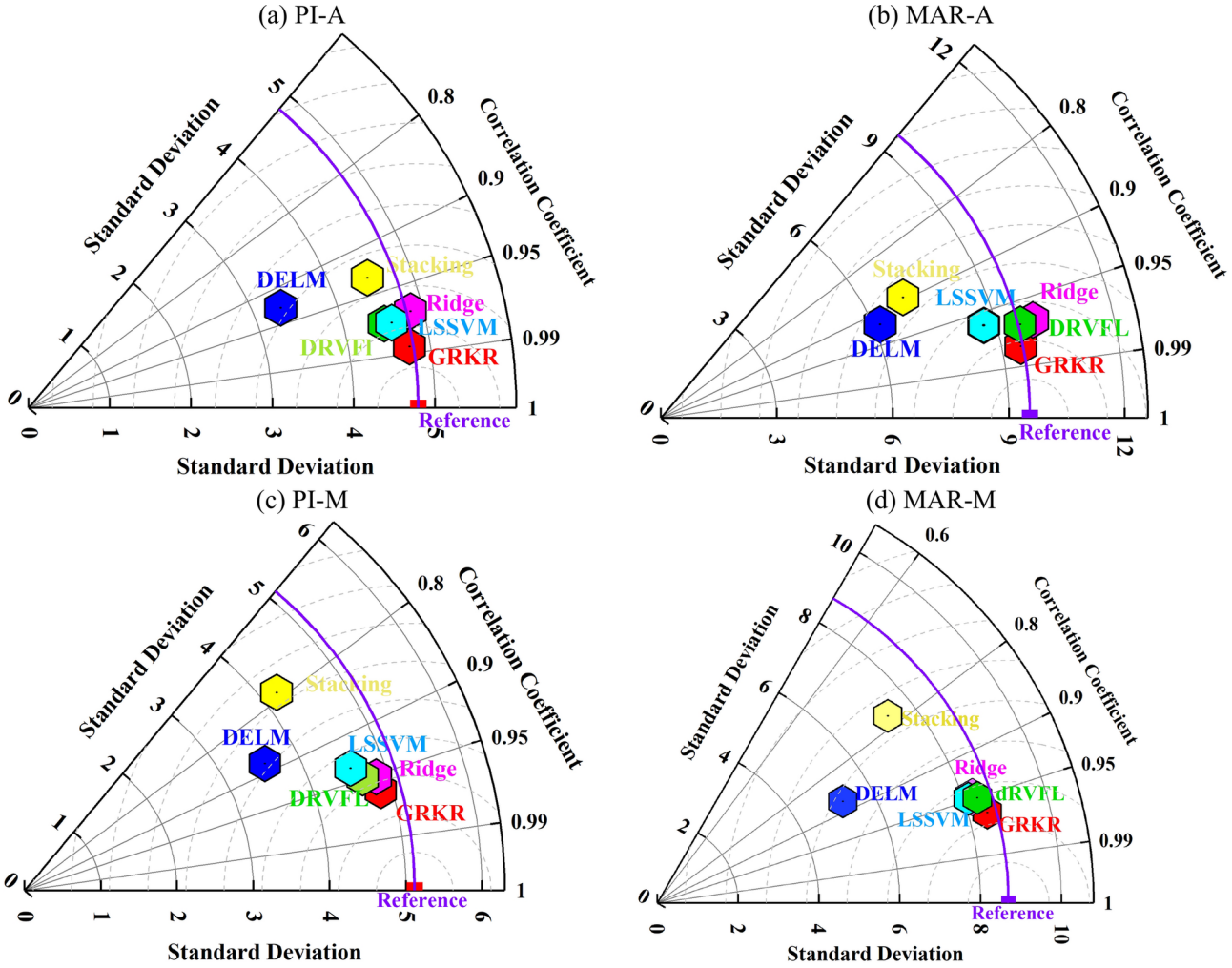

Fig. 11. Taylor diagram of (**a**) PI-A, (**b**) MAR-A, (**c**) PI-M, and (**d**) MAR-M for all ML models.

Data availability

Data availabilityThe data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 7 July 2024; Accepted: 18 April 2025
Published online: 10 May 2025

References

1. Zahedi, S. Modification of expected conflicts between drinking water quality index and irrigation water quality index in water quality ranking of shared extraction wells using multi criteria decision making techniques. *Ecol. Indic.* **83**, 368–379. https://doi.org/10.1016/j.ecolind.2017.08.017 (2017).
2. Chang, F. J., Tsai, Y. H., Chen, P. A., Coynel, A. & Vachaud, G. Modeling water quality in an urban river using hydrological factors–Data driven approaches. *J. Environ. Manage.* **151**, 87–96. https://doi.org/10.1016/j.jenvman.2014.12.014 (2015).
3. Ayres, R. S. & Cat, D. W. W. Water Quality for Agriculture, FAO, Irrigation and drainage paper (1985).
4. Delpla, I., Jung, A. V., Baures, E., Clement, M. & Thomas, O. Impacts of climate change on surface water quality in relation to drinking water production. *Environ. Int.* **35**, 1225–1233. https://doi.org/10.1016/j.envint.2009.07.001 (2009).
5. Assar, W. et al. Effect of water shortage and pollution of irrigation water on water reuse for irrigation in the nile Delta. *J. Irrig. Drain. Eng.* **146**, 5019013. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001439 (2020).
6. Richards, L. A. *Diagnosis and Improvement of Saline and Alkali Soils* (US Government Printing Office, 1954).
7. Ayers, R. S. & Westcot, D. W. *Water Quality for Agriculture*vol. 29 (Food and Agriculture Organization of the United Nations Rome, 1985).
8. Doneen, L. D. *Notes on Water Quality in Agriculture* (Department of Water Science and Engineering, University of California, 1964).

9. Gholami, S. & Srikantaswamy, S. Analysis of agricultural impact on the cauvery river water around KRS dam. *World Appl. Sci. J.* **6**, 1157–1169 (2009).
10. El Bilali, A., Taleb, A. & Brouziyne, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agric. Water Manag.* **245**, 106625. https://doi.org/10.1016/j.agwat.2020.106625 (2021).
11. Yaseen, Z. M. et al. Hybrid adaptive neuro-fuzzy models for water quality index Estimation. *Water Resour. Manag.* **32**, 2227–2245. https://doi.org/10.1007/s11269-018-1915-7 (2018).
12. Jin, T., Cai, S., Jiang, D. & Liu, J. A data-driven model for real-time water quality prediction and early warning by an integration method. *Environ. Sci. Pollut Res.* **26**, 30374–30385. https://doi.org/10.1007/s11356-019-06049-2 (2019).
13. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **13**, 18–28. https://doi.org/10.1109/5254.708428 (1998).
14. Hassoun, M. H. *Fundamentals of Artificial Neural Networks* (MIT Press, 1995).
15. Suykens, J. A. K. & Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **9**, 293–300. https://doi.org/10.1023/A:1018628609742 (1999).
16. Asadollahfardi, G., Taklify, A. & Ghanbari, A. Application of artificial neural network to predict TDS in Talkheh Rud river. *J. Irrig. Drain. Eng.* **138**, 363–370. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000402 (2012).
17. Bozorg-Haddad, O., Soleimani, S. & Loáiciga, H. A. Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming. *J. Environ. Eng.* **143**, 4017021. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001217 (2017).
18. Ravansalar, M., Rajaee, T. & Zounemat-Kermani, M. A wavelet–linear genetic programming model for sodium (Na+) concentration forecasting in rivers. *J. Hydrol.* **537**, 398–407. https://doi.org/10.1016/j.jhydrol.2016.03.062 (2016).
19. Najafzadeh, M., Homaei, F. & Farhadi, H. Reliability assessment of water quality index based on guidelines of National sanitation foundation in natural streams: integration of remote sensing and data-driven models. *Artif. Intell. Rev.* **54**, 4619–4651. https://doi.org/10.1007/s10462-021-10007-1 (2021).
20. Nouraki, A., Alavi, M., Golabi, M. & Albaji, M. Prediction of water quality parameters using machine learning models: A case study of the Karun river, Iran. *Environ. Sci. Pollut Res.* **28**, 57060–57072. https://doi.org/10.1007/s11356-021-14560-8 (2021).
21. Ahmadianfar, I., Shirvani-Hosseini, S., Samadi-Koucheksaraee, A. & Yaseen, Z. M. Surface water sodium (Na+) concentration prediction using hybrid weighted exponential regression model with gradient-based optimization. *Environ. Sci. Pollut Res.*, 1–26. https://doi.org/10.1007/s11356-022-19300-0 (2022).
22. Chen, H., Ahmadianfar, I., Liang, G. & Heidari, A. A. Robust kernel extreme learning machines with weighted mean of vectors and variational mode decomposition for forecasting total dissolved solids. *Eng. Appl. Artif. Intell.* **133**, 108587. https://doi.org/10.1016/j.engappai.2024.108587 (2024).
23. Asadollah, S. B. H. S., Sharafati, A., Motta, D. & Yaseen, Z. M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **9**, 104599. https://doi.org/10.1016/j.jece.2020.104599 (2021).
24. Rezaie-Balf, M. et al. Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: comparative assessment of a noise suppression hybridization approach. *J. Clean. Prod.* **271**, 122576. https://doi.org/10.1016/j.jclepro.2020.122576 (2020).
25. Hunter, J. M. et al. Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. *Hydrol. Earth Syst. Sci.* **22**, 2987–3006. https://doi.org/10.5194/hess-22-2987-2018 (2018).
26. Deng, W., Wang, G. & Zhang, X. A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting. *Chemom Intell. Lab. Syst.* **149**, 39–49. https://doi.org/10.1016/j.chemolab.2015.09.017 (2015).
27. Johnston, K., Ver Hoef, J. M., Krivoruchko, K. & Lucas, N. *Using ArcGIS Geostatistical Analyst*vol. 380 (Esri Redlands, 2001).
28. Nelder, J. A. & Wedderburn, R. W. M. Generalized linear models. *J. R Stat. Soc. Ser. Stat. Soc.* **135**, 370–384. https://doi.org/10.2307/2344614 (1972).
29. Saunders, C. & Gammerman, A. Ridge Regression Learning Algorithm in Dual Variables. In *15th International Conference on Machine Learning (ICML '98) (01/01/98)* (1998).
30. Ahmadianfar, I., Heidari, A. A., Gandomi, A. H., Chu, X. & Chen, H. RUN beyond the metaphor: an efficient optimization algorithm based on runge Kutta method. *Expert Syst. Appl.* **181**, 115079. https://doi.org/10.1016/j.eswa.2021.115079 (2021).
31. Hadi, S. J. & Tombul, M. Forecasting daily streamflow for basins with different physical characteristics through Data-Driven methods. *Water Resour. Manag.* **32**, 3405–3422. https://doi.org/10.1007/s11269-018-1998-1 (2018).
32. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process.Syst.* **30** (2017).
33. ur Rehman, N. & Aftab, H. Multivariate variational mode decomposition. *IEEE Trans. Signal. Process.* **67**, 6039–6052. https://doi.org/10.1109/TSP.2019.2951223 (2019).
34. Gray, R. M. *Entropy and Information Theory* (Springer Science & Business Media, 2011).
35. Zhang, J. et al. FBG strain monitoring data denoising in wind turbine blades based on parameter-optimized variational mode decomposition method. *Opt. Fiber Technol.* **81**, 103527. https://doi.org/10.1016/j.yofte.2023.103527 (2023).
36. Rahul, S. & Sunitha, R. Dominant electromechanical Oscillation mode identification using modified variational mode decomposition. *Arab. J. Sci. Eng.* **46**, 10007–10021. https://doi.org/10.1007/s13369-021-05818-x (2021).
37. Shi, Q., Katuwal, R., Suganthan, P. N. & Tanveer, M. Random vector functional link neural network based ensemble deep learning. *Pattern Recognit.* **117**, 107978. https://doi.org/10.1016/j.patcog.2021.107978 (2021).
38. Fayaz, M. & Kim, D. A prediction methodology of energy consumption based on deep extreme learning machine and comparative analysis in residential buildings. *Electronics* **7**, 222. https://doi.org/10.3390/electronics7100222 (2018).
39. Zavadskas, E. K. & Turskis, Z. A new additive ratio assessment (ARAS) method in multicriteria decision-making. *Technol. Econ. Dev. Econ.* **16**, 159–172. https://doi.org/10.3846/tede.2010.10 (2010).
40. Fandel, G. & Gal, T. *Multiple Criteria Decision Making Theory and Application: Proceedings of the Third Conference Hagen/Königswinter, West Germany, August 20–24*, vol. 177 (Springer Science & Business Media, 2012).

Acknowledgements

The authors would like to thank the reviewers and editors for their thorough and insightful feedback, which has contributed significantly to enhancing the quality of this manuscript.

Author contributions

Marjan Kordani: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Investigation, Supervision.Mohsen Bagheritabar: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Investigation, Supervision.Iman Ahmadianfar: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Software, Visualization, Investigation.Arvin Samadi koucheksaraee: Formal analysis, Software, Visualization, Writing-original draft, Software, Investigation.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-99341-w.

Correspondence and requests for materials should be addressed to I.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.