



Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-arid regions of India

Title	Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid and semi-arid regions of India
Author(s)	Khan, Imran;Nizam, Sarwar;Bamal, Apoorva;Sajib, Abdul Majed;Diganta, Mir Talas Mahammad;Shaida, Mohd Azfar;Ashekuzzaman, S.M.;Nash, Stephen;Olbert, Agnieszka I.;Uddin, Md Galal
Publication Date	2025-05-03
Publisher	Elsevier
Repository DOI	https://doi.org/10.1016/j.clet.2025.100984

1
2
3
4 **1 Optimized intelligent learning for groundwater quality prediction in diverse aquifers of arid**
5 **2 and semi-arid regions of India**
6
7 3

8 4 Imran Khan^{a*}, Sarwar Nizam^b, Apoorva Bamal^{c,d,e,f}, Abdul Majed Sajib^{c,d,e,f}, Mir Talas
9
10 5 Mahammad Diganta^{c,d,e,f}, Mohd Azfar Shaida^g, S. M. Ashekuzzaman^h, Stephen Nash^{c,d,e},
11
12 6 Agnieszka I. Olbert^{c,d,e,f}, Md Galal Uddin^{c,d,e,f,h**}
13

14 7 ^a Department of Geology, Aligarh Muslim University, Aligarh-202002, India

15 8 ^b German Research Centre for Geosciences GFZ, 14473 Potsdam, Germany

16 9 ^c School of Engineering, University of Galway, Ireland

17 10 ^d Ryan Institute, University of Galway, Ireland

18 11 ^e MaREI Research Centre, University of Galway, Ireland

19 12 ^f Eco-HydroInformatics Research Group (EHIRG), Civil Engineering, University of Galway,
20 13 Ireland

21 14 ^g Department of Industrial Chemistry, Aligarh Muslim University, Aligarh-202002, India

22 15 ^h Department of Civil, Structural and Environmental Engineering, and Sustainable Infrastructure
23 16 Research & Innovation Group, Munster Technological University, Cork, Ireland

24 17
25 18 * Corresponding author 1: Imran Khan, DS Kothari Post-doctoral Fellow, Department of Geology,
26 19 Aligarh Muslim University, Aligarh-202002, India. Email: imranalig.iitk@gmail.com

27 20 ** Corresponding author 2: Md Galal Uddin, Post-doctoral researcher, Civil Engineering, College
28 21 of Science and Engineering, University of Galway, Ireland. Email:
29 22 mdgalal.uddin@universityofgalway.ie
30 23

1
2
3
4 **24 Abstract**

5
6 25 Ensuring access to safe, affordable drinking water while implementing sustainable management
7
8 26 practices is vital for achieving the United Nations' Sustainable Development Goals-2030. Accurate
9
10 27 groundwater (GW) quality assessment plays a crucial role in enhancing water management
11
12 28 strategies. This study evaluates GW resources across the diverse aquifer systems of arid and semi-
13
14 29 arid regions of northwest India using the recently developed Root Mean Squared-Water Quality
15
16 30 Index (RMS-WQI) model, optimized with machine learning (ML) techniques. A total of 772 GW
17
18 31 samples from 36 districts of state Rajasthan were analyzed for 16 water quality (WQ)
19
20 32 indicators/parameters, including pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS),
21
22 33 major cations (Ca^{2+} , Mg^{2+} , Na^+ , K^+), anions (Cl^- , CO_3^{2-} , HCO_3^- , SO_4^{2-} , NO_3^- , F^- , PO_4^{3-}),
23
24 34 Alkalinity (ALK), and Total Hardness (TH). The results indicate slightly alkaline GW (average
25
26 35 pH 7.9), with elevated concentrations of Na^+ , Cl^- , SO_4^{2-} and NO_3^- exceeding Bureau of Indian
27
28 36 Standards (BIS). This study employs the eXtreme Gradient Boosting (XGB) algorithm,
29
30 37 demonstrating strong predictive capabilities within the RMS-WQI model across diverse aquifers
31
32 38 of Rajasthan. This marks the first application of RMS-WQI at a state-wide scale in India. Model
33
34 39 performance assessment indicated groundwater quality ranging from 'fair' to 'marginal', generally
35
36 40 meeting BIS standards, with high sensitivity and low uncertainty. Statistical metrics (Root Mean
37
38 41 Square Error-RMSE, Mean Squared Error-MSE, Mean Absolute Error-MAE, and Percentage of
39
40 42 Absolute Bias Error-PABE) validated the model's efficiency, with minimal error and high
41
42 43 sensitivity. Optimization using "Optuna" further enhanced model performance, confirmed by
43
44 44 Tukey's Honest Significant Difference (HSD) test. Sensitivity analysis demonstrated robust
45
46 45 goodness-of-fit, while uncertainty analysis indicated minimal discrepancies, with overall
47
48 46 uncertainty below 2%. Spatial analysis revealed varying WQ across districts, ranging from
49
50 47 marginal to poor, while efficiency metrics demonstrated the model's effectiveness in providing
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 48 accurate assessments. The configured WQI model could substantially contribute to informing
5
6
7 49 aquatic managers and strategic planners for sustainable water resource management and policy
8
9 50 development aimed at enhancing GW quality.

10
11
12 51 **Keywords:** Hydrochemistry; Machine learning; eXtreme Gradient Boosting; Optuna; Water
13
14 52 quality index (WQI); Rajasthan.

15
16 53
17
18 54 **1. Introduction**

19
20
21 55 Groundwater (GW), an essential component of the earth's hydrological cycle, holds
22
23 56 paramount importance in sustaining ecosystems, facilitating agricultural activities, and meeting
24
25 57 domestic and industrial water requirements globally. Consequently, routine monitoring of GW
26
27 58 quality becomes imperative to address pertinent environmental concerns effectively (Boo et al.,
28
29 59 2024; Khan and Umar, 2024a, 2019; Nizam et al., 2022b; Srinivas et al., 2015; WHO, 2017).
30
31 60 Certain countries, particularly developed nations, have established robust management strategies
32
33 61 and action plans to uphold WQ standards. Conversely, developing nations continue to struggle
34
35 62 with formulating effective water management frameworks personalized to their specific needs and
36
37 63 challenges (Manzar et al., 2022; Dimple et al., 2023; Pandey et al., 2023; Bahrami et al., 2024;
38
39 64 Jahan et al., 2025). The escalating pressures of population growth, urbanization, and climate
40
41 65 variations have presented substantial obstacles to the sustainable management of GW reservoirs
42
43 66 (Samani, 2021; Singha et al., 2021; Tripathy and Mishra, 2024; Ahmad et al., 2025). Amidst these
44
45 67 complexities, the integration of machine learning (ML) techniques presents a paradigm shift in
46
47 68 water resource management, offering numerous advantages over traditional modeling approaches.
48
49 69 ML algorithms excel in handling large and intricate datasets, discerning nonlinear relationships,
50
51 70 and adapting to evolving environmental dynamics. In response to these challenges, the integration
52
53 71 of ML techniques into GW management has emerged as a promising avenue for effective
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 72 monitoring and remediation (Manzar et al., 2022; Uddin et al., 2022; 2024b; Sajib et al., 2023;
5
6
7 73 2025; Bamal et al., 2024; Khan and Umar, 2024a).

8
9 74 Water quality (WQ) monitoring programs traditionally rely on measuring physical,
10
11 75 chemical, and biological indicators/parameters to ascertain compliance with established guidelines
12
13 76 (Nizam et al., 2022b; 2022c; Haidery et al., 2024; Jhan et al., 2025). However, conventional
14
15 77 monitoring methods are often outdated, resource-intensive, and yield voluminous data that pose
16
17 78 challenges in interpretation (Dai et al., 2007; Dai and Samper, 2004; Zhu et al., 2023). In recent
18
19 79 years, Water Quality Index (WQI) models have gained importance for their simplified and
20
21 80 generalized approach to water resource management (Patel et al., 2022; Rajkumar et al., 2022;
22
23 81 Chidiac et al., 2023; Sadeghi-Lari et al., 2024). These models translate diverse WQ indicators into
24
25 82 standardized values, offering a comprehensive assessment of the GW of a given region. Despite
26
27 83 their popularity, traditional WQI models have been criticized for their classification schemes,
28
29 84 unpredictability, and lack of reliability (Khan and Umar, 2024b). To address these shortcomings,
30
31 85 novel and reliable WQI models have been developed to solve these problems. Examples of these
32
33 86 models are the Weighted Quadratic Mean Water Quality Index (WQM-WQI), and the Unweighted
34
35 87 Root Mean Square Water Quality Index (RMS-WQI) which have been proposed by Uddin et al.
36
37 88 (2022a), showcasing significant improvements in model accuracy and reduction of uncertainty
38
39 89 (Gani et al., 2023; Sajib et al., 2023; 2024; Uddin et al., 2024a). ML techniques, notably feature
40
41 90 selection algorithms, have emerged as precise tools for identifying crucial variables within
42
43 91 datasets. ML algorithms such as support vector machine (SVM), Gaussian process regression
44
45 92 (GPR), and random forest (RF) have been instrumental in feature selection and predicting WQI
46
47 93 scores with notable success. Additionally, ML models have found application in river WQ
48
49 94 pollution modeling, further accentuating their versatility and efficacy in water resource
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

95 management. It is also evident from the global research studies that ML technologies have been
96 utilized for WQ prediction using reliable WQIs in many countries such as India (Talukdar et al.,
97 2023), Bangladesh (Nath et al., 2023), Ireland (Uddin et al., 2023a), China (Ding et al., 2025), etc.
98 This also proves the ability of different ML models to test the reliability and sensitivity of WQIs
99 such as RMS-WQI across different study areas. While ML offers promising advancements in GW
100 quality assessment, it has notable limitations. Feature selection challenges may introduce irrelevant
101 parameters, reducing reliability, while data variability affects model generalization across regions
102 and timeframes. Model interpretability remains a concern, as many ML models act as "black
103 boxes," limiting insight into prediction factors (Tripathy and Mishra, 2024). Additionally,
104 hyperparameter optimization is complex and resource-intensive, requiring careful tuning to
105 balance accuracy and robustness (Uddin et al., 2024b). These challenges highlight the need for
106 continuous refinement and cautious application of ML in GW studies.

107 The existing literature highlights the varying performance of different ML models and their
108 corresponding accuracy levels in GW management for sustainable development worldwide
109 (Haggerty et al., 2023; Kumar et al., 2023; Boo et al., 2024; Sajib et al., 2024; Tripathy and Mishra,
110 2024; Uddin et al., 2024a; 2024b). However, despite the utilization of numerous ML and artificial
111 intelligence (AI) techniques for predicting the WQI, there remains a notable gap in understanding
112 the factors influencing the effectiveness of these predictions. Specifically, the impact of
113 nonlinearities and stochastic WQ characters on the accuracy of WQI predictions remains
114 inadequately explored for many regions such as the Rajasthan state of India. Rajasthan, India's
115 largest state, is mainly arid and semi-arid, with GW supplying around 90 percent of drinking water
116 and 70 percent of irrigation needs (Pandey et al., 2023). Additionally, the Central Ground Water
117 Board (CGWB) in India reported the contamination of GW in Rajasthan in 2021 that rendered

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

118 unsuitable for drinking and other domestic purposes (Jalan et al., 2023; Kumar and Mishra, 2024;
119 Pandey et al., 2023). Given the current water conditions in Rajasthan, enforcing a comprehensive
120 monitoring and management model is crucial to ensure sustainable water resources and safeguard
121 public health (Sunita and Ghosh, 2024).

122 The existing methodologies for assessing WQ lack over and/or under-estimation in weight
123 assignment and, therefore, contain erroneous WQI values and are limited in robustness in decision-
124 making. Therefore, Rajasthan has been selected as the study area to evaluate the effectiveness of
125 the RMS-WQI model in assessing WQ and to validate the accuracy of ML/AI models in predicting
126 water conditions. The RMS-WQI model was selected for its higher reliability, reduced uncertainty,
127 and improved classification, overcoming the limitations of traditional WQI models by providing a
128 balanced, data-driven approach with enhanced accuracy and adaptability through ML integration
129 (Mohammadpour et al., 2024; Nong et al., 2025; Sajib et al., 2023, 2025; Uddin et al., 2024a,
130 2024b). Hence, this study has been conducted with the aims: (i) to assess the accuracy and efficacy
131 of distinct ML/AI models in predicting the WQ considering the hydrogeological characteristics of
132 the study area, (ii) to investigate the influence of nonlinearities and stochastic WQ characteristics
133 on the performance of ML/AI models, (iii) to develop an advanced approach for WQ assessment
134 that incorporates impartial weight calculation and enhances flexibility and robustness in decision-
135 making policies. By addressing these objectives, this study not only demonstrates the value of
136 utilizing statistical and ML/AI techniques in assessing GW quality but also provides a crucial
137 approach for enhancing GW resource management world. Nonetheless, findings can provide
138 essential information for better policy making and sustainable GW development in line with SDG-
139 6.

2. Study area description

The study region comprises the entire state of Rajasthan (India), which covers an area of approximately 3,42,239 km², situated between north latitudes 23°03' and 30°12' and east longitudes 69°30' and 78°17' (Fig. 1). As per the 2011 census, Rajasthan's population was 68.5 million, and recent estimates suggest it has exceeded 80 million (Census, 2011). The physiography and landforms in the region are influenced by geological formations and structures, shaped by the previous fluvial cycles of aggradation and degradation, as well as the ongoing desert cycles of erosion and deposition. The geological formations include rocks from the Delhi Supergroup, Aravalli Supergroup, and upper Vindhyan Supergroup, and their ages range from the Cambrian through the Jurassic, Cretaceous, and Tertiary (Roy and Jakhar, 2002). The state's southernmost part is occupied by a pile of Cretaceous basaltic flows (Deccan Traps). The soils are primarily sandy, sandy loam, clay loam to clay, and rocky or mountainous soils (Pandey and Bahadur, 2009). The major catchments are the Yamuna-Ganga in the northeast, the Mahi and Sabarmati in the southwest, and Chambal in the southeast. The majority of rivers in the state are monsoon-fed and ephemeral, with the exception of the Chambal and Mahi rivers.

Climatically, the study area used to witness three primary conventional seasons, which are as follows: (1) Pre-monsoon (March to June), (2) Monsoon season (June to September), and (3) Post-monsoon (October to February). During monsoon season, the state receives 90% of its rainfall from the southwest monsoon (CGWB, 2022). The state's annual rainfall varies greatly; however, it acts as a primary source of GW recharge. Few parts of the study area also have access to reservoir water (Bisalpur dam), which is primarily used for irrigation purposes.

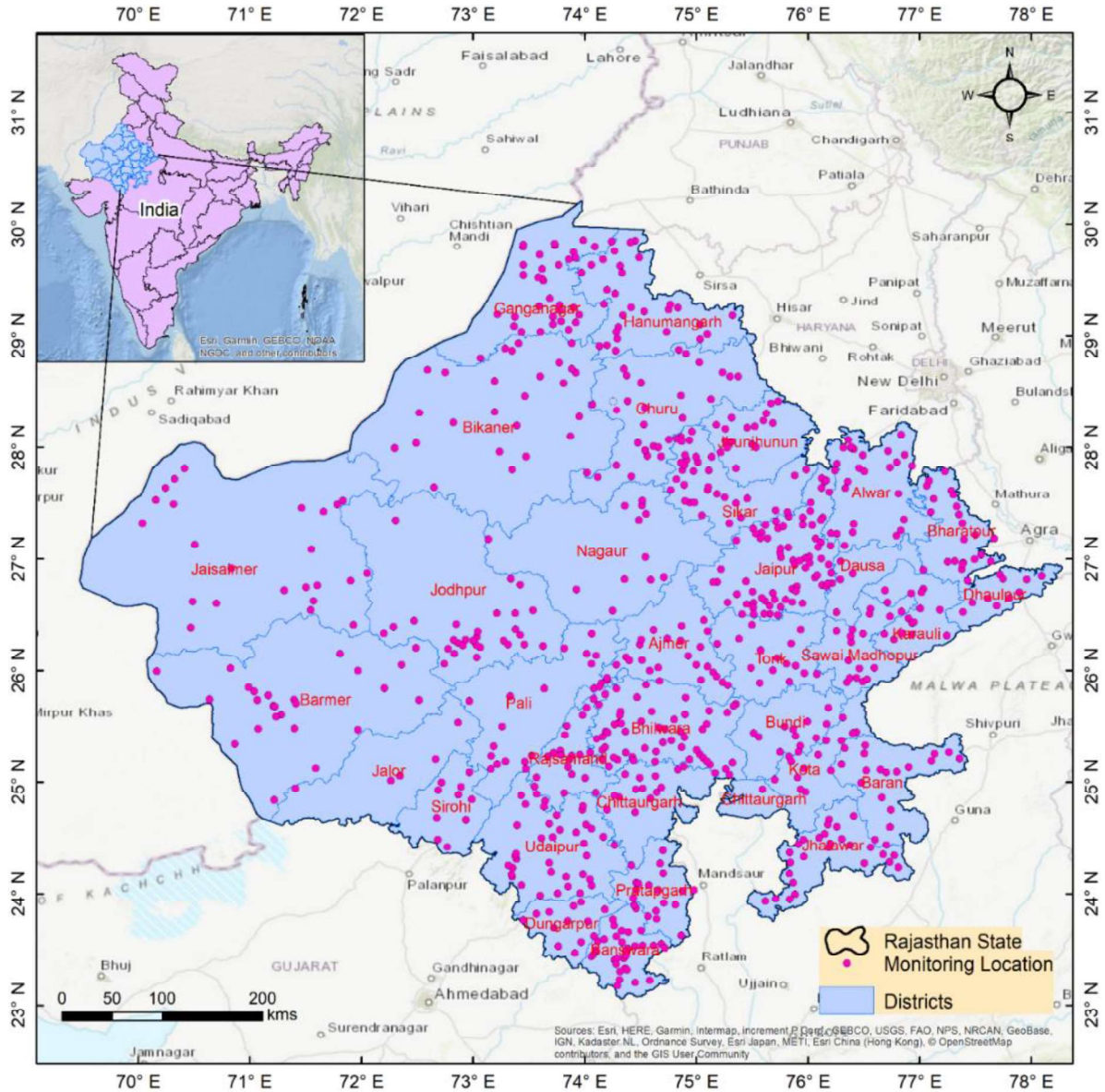


Fig. 1. Study area map showing GW sample locations.

In the study area, the GW occurrence and movement generally depend on the presence of geological formations, which are mainly divided into porous (consolidated or semi-consolidated) and fissured formations (CGWB, 2022). The consolidated formations (sand, gravel, and their admixtures) form a potential aquifer of the region. The semi-consolidated formations are mainly composed of siltstone, claystone, conglomerate, shale, sandstone, and limestone. Later, two formations form the main aquifers in a few districts of the state, e.g., Barmer, Jaisalmer, Bikaner,

1
2
3
4 169 and Jodhpur. As a hydrogeological unit, fissured formations cover around 32% of the state and are
5
6
7 170 broadly categorized into four units such as consolidated sedimentary rocks, igneous and
8
9 171 metamorphic rocks, and volcanic rocks, including Deccan Traps (CGWB, 2022). The depth to
10
11 172 water level varies from 20 and 40 m below ground level (bgl) in the north, east, and west parts of
12
13
14 173 the state. In the southern half of the state, it is recorded as <20 m bgl. Distinctly, shallow water
15
16 174 levels, <2 m bgl, are commonly recorded in the south-eastern part of the state (CGWB, 2022).

175 **3. Materials and methods**

176 **3.1. GW sample collection and analysis**

177 Geo-referenced GW chemistry data ($n = 772$) of Rajasthan state was gathered from surveys
178 conducted nationwide by the Central Groundwater Board (CGWB) India in 2021 (CGWB, 2022).
179 The selection of GW sampling sites was guided by the established protocols of the CGWB (2022),
180 India, with a focus on ensuring spatial and hydrogeological representativeness across the study
181 domain (Fig. 1). Water sample collection was conducted in accordance with the guidelines outlined
182 in APHA (2017). To accurately document sampling locations, geographic coordinates were
183 recorded using a handheld Garmin GPS. Before collecting the samples, hand pumps were purged
184 for approximately 5–8 minutes to remove any standing water in pumping assembly. The water
185 was then collected in 1-liter high-density polyethylene (HDPE) bottles that had been thoroughly
186 pre-cleaned and dried to prevent contamination. The bottles were securely sealed to avoid leakage
187 during transport. All samples were stored at 4°C and analyzed within the recommended holding
188 times to maintain data accuracy and reliability. Subsequently, these samples were examined for
189 pH, total hardness (TH), electrical conductivity (EC), total dissolved solids (TDS), including major
190 cations (Ca^{2+} , Mg^{2+} , Na^+ , and K^+), and anions (CO_3^{2-} , HCO_3^- , SO_4^{2-} , NO_3^- , F^- , Cl^- and PO_4^{3-}).
191 The methodological approach is provided in Table 1. For more details refer to the published

1
2
3
4 192 Central Ground Water Year Book Rajasthan 2021–2022 (CGWB, 2022). Moreover, the cation-
5
6
7 193 anion balancing approach was used to confirm the validity and reliability of the hydro-chemical
8
9 194 dataset. Datasets containing a charge balance error within $\pm 10\%$ were considered for this study
10
11
12 195 (Zhang et al., 2022).

13
14
15 196 **Table 1.** Analytical methods adopted to determine the WQ.

WQ Indicators	Abbreviation	Unit	Instrument/Method used*
pH	pH	--	Portable pH meter (Hach sensION)
Electrical Conductivity	EC	$\mu\text{S}/\text{cm}$	Portable EC and TDS meter (Hach sensION +EC5)
Total Dissolved Solids	TDS	mg/L	Portable EC and TDS meter (Hach sensION +EC5)
Total Hardness as CaCO ₃	TH	mg/L	EDTA# (0.05 N) titrimetric method
Calcium	Ca ²⁺	mg/L	EDTA# (0.05 N) titrimetric method
Magnesium	Mg ²⁺	mg/L	EDTA# (0.05 N) titrimetric method
Sodium	Na ⁺	mg/L	Flame photometer (SYSTRONICS 128, India)
Potassium	K ⁺	mg/L	Flame photometer (SYSTRONICS 128, India)
Chloride	Cl ⁻	mg/L	Titration using 0.05 N AgNO ₃
Carbonate	CO ₃ ²⁻	mg/L	Titration with 0.01 N H ₂ SO ₄
Bicarbonate	HCO ₃ ⁻	mg/L	Titration with 0.01 N HCl
Sulphate	SO ₄ ²⁻	mg/L	UV Spectrophotometer (UV Spectrophotometer (UV-VIS 2000 Spectrophotometer, Labindia Analytical, India)
Nitrate	NO ₃ ⁻	mg/L	UV Spectrophotometer (UV Spectrophotometer (UV-VIS 2000 Spectrophotometer, Labindia Analytical, India)
Fluoride	F ⁻	mg/L	UV Spectrophotometer (UV Spectrophotometer (UV-VIS 2000 Spectrophotometer, Labindia Analytical, India)
Phosphate	PO ₄ ³⁻	mg/L	UV Spectrophotometer (UV Spectrophotometer (UV-VIS 2000 Spectrophotometer, Labindia Analytical, India)
Alkalinity	ALK		Titrimetric method (APHA, 2017)

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38 197 * Methods and approach adopted (APHA, 2017); #Ethylene Di-amine Tetra Acetic acid.
39

40 198 3.2. Computation of RMS-WQI scores

41
42
43 199 Index-based models are frequently utilized while incorporating GW hydro-chemical data
44
45 200 to assess the suitability for human consumption and irrigation, particularly concerning major ions
46
47 201 contamination in GW (Muniz et al., 2020; Wu et al., 2020; Zhang et al., 2020). The calculation of
48
49 202 WQI scores may sometimes yield erroneous results due to subjectivity being involved in the
50
51 203 calculation process (Uddin et al., 2023a). This subjectivity can arise from various sources,
52
53 204 including the selection of WQ indicators, the assignment of weights to these indicators, and the
54
55 205 determination of sub-index values (Uddin et al., 2022). Each of these steps involves subjective
56
57 206 judgment, which can introduce biases and inaccuracies into the final WQI calculation using
58
59
60
61
62
63
64
65

1
2
3
4 207 conventional models (Horton, 1965; Amiri et al., 2014; Khan and Umar, 2019; 2024a; Rajkumar
5
6 208 et al., 2020). Among traditional models, the RMS-WQI model proposed by Uddin et al. (2022)
7
8 209 stands out as a commonly applied tool in recent studies in terms of the model's reliability and
9
10
11 210 uncertainty (Sajib et al., 2023; Uddin et al., 2024a). For the purposes of the computation of WQI
12
13
14 211 scores, the research utilized the RMS-WQI model. This newly developed RMS-WQI model has
15
16 212 gained significant attention for assessing WQ conditions across diverse geographic regions (Gani
17
18 213 et al., 2023; Sajib et al., 2024). A thorough assessment of WQ is made possible by the four main
19
20
21 214 parts of the RMS-WQI model: the indicator selection process, sub-index function, aggregation
22
23
24 215 function, and classification scheme. Details of the process computing RMS-WQI are presented in
25
26 216 Fig. 2 and supplementary material as continuation of this section.

27
28 217 The calculation of the sub-index function plays a crucial role in the computation of the
29
30
31 218 RMS-WQI score. It represents a fundamental stage in the process, as any misapplication or
32
33 219 improper execution of this function can potentially introduce ambiguity and uncertainty into the
34
35
36 220 results (Uddin et al., 2022; 2023b; Sajib et al., 2023). The sub-index approach serves as an
37
38 221 important step employed to convert the measured values of various WQ indicators into
39
40
41 222 dimensionless values. The significance of sub-index functions in WQI models has been
42
43 223 highlighted by researchers in various studies (Khouri and Bashir Al-Moufti, 2022; Uddin et al.,
44
45 224 2022). It is emphasized that the proper determination and application of these functions are critical
46
47
48 225 to ensure the accuracy and reliability of WQI assessments. Subsequently, the sub-indexing process
49
50
51 226 has become a focal point of research attention due to its implications for uncertainty management
52
53 227 and the overall WQ assessments. In the present study, the computation of sub-index scores was
54
55 228 conducted based on the guidelines provided by the Bureau of Indian Standards (BIS, 2012). These
56
57
58 229 guidelines serve as essential reference points for establishing the sub-index values corresponding

to individual measured WQ parameters. By adhering to recognized standards and guidelines, researchers can ensure consistency and validity in the sub-indexing process, thus enhancing the credibility and applicability of the derived RMS-WQI scores.

The following equation (Eq. 1) can be used to calculate the RMS-WQI scores for each sampling site (Uddin et al., 2022).

$$RMS - WQI = \sqrt{\frac{1}{n} \sum_{i=1}^n S_i^2} \quad \text{Eq. (1)}$$

where n is the number of WQ indicators, and S_i denotes the sub-index of the i^{th} WQ indicator.

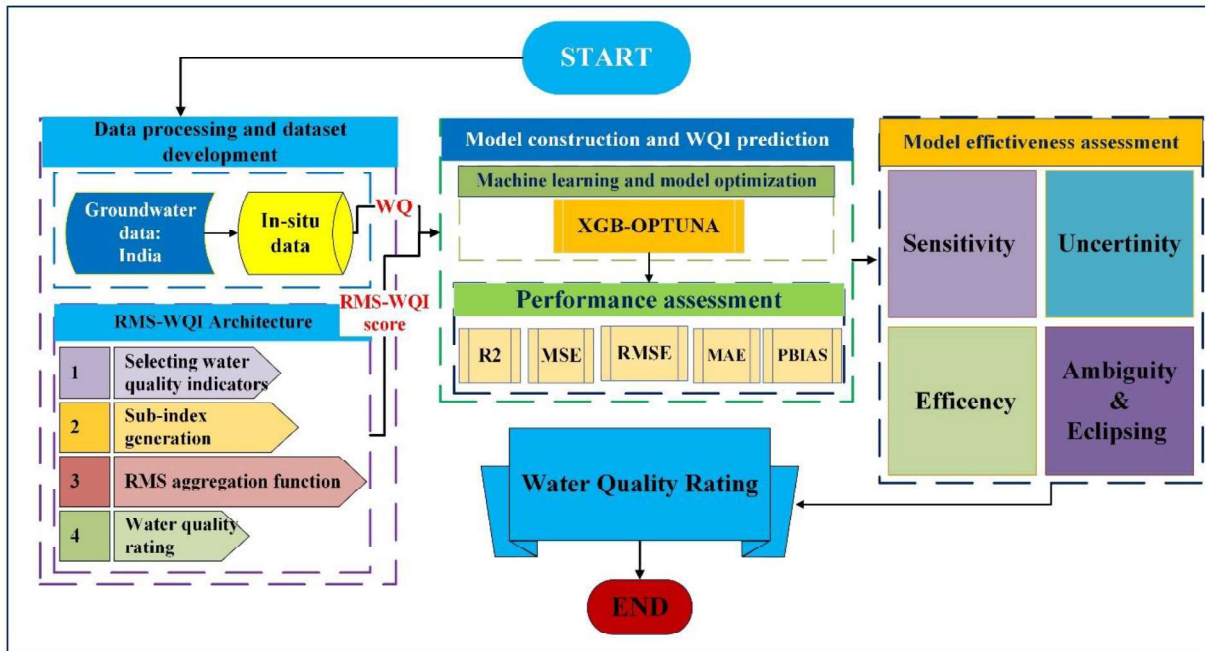


Fig. 2. A conceptual framework of this research.

3.2.1. Rating scheme for RMS-WQI model

After computing the RMS-WQI scores, they are deciphered into WQ ranks using various classification schemes. Detailed information about the development process of these classification schemes is provided by Uddin et al. (2022), while the validation of this approach is thoroughly discussed in Uddin et al. (2023a). This study utilized a modified classification scheme originally

244 proposed by Uddin et al. (2022). Table 2 shows the classification, which includes four WQ classes
 245 ranging from good to poor.

246 **Table 2.** Classification categories of the RMS-WQI model for GW according to Uddin et al. (2022).

GW quality	RMS-WQI score range	Explanation
Good	80 – 100	All WQ indicators meet the established standards, indicating that the water is suitable for all intended purposes. Consequently, the WQ is unlikely to have any adverse impact on human health.
Fair	50 – 79	A few WQ indicators exceed the established standards, making the water moderately vulnerable to human health risks. Although the immediate effects may be minimal, long-term use of such water can lead to various health issues. Therefore, it is essential to purify the water before use to prevent potential health risks.
Marginal	30 – 49	Majority of WQ indicators exceed the established standards, necessitating caution before using the water for any purpose. The WQ poses a significant risk to human health and should be treated with utmost care.
Poor	0 – 29	All WQ indicators exceed the established standards, rendering the water unsuitable for any purpose. The WQ can pose severe risk to human health and should be avoided entirely.

247

248 3.3. Machine Learning (ML) prediction approach

249 3.3.1. Data pre-processing

250 GW studies often involve complex and heterogeneous datasets, comprising diverse
 251 physicochemical parameters measured through various analytical techniques. The variability in
 252 measurement scales and units across these parameters can introduce inconsistencies, impacting the
 253 effectiveness of ML models in predicting WQ trends. Therefore, rigorous data preprocessing is a
 254 fundamental step to ensure data integrity, improve model reliability, and enhance predictive
 255 performance. A critical aspect of preprocessing is data standardization, which ensures that all input
 256 variables are transformed onto a common scale. Standardization is particularly important in ML
 257 applications, as many algorithms are sensitive to differences in magnitude and distribution of
 258 numerical features. Without proper scaling, models may assign undue importance to parameters
 259 with larger numerical ranges, leading to biased predictions. In ML-based hydrogeochemical
 260 modeling, z-score normalization is a widely adopted technique for standardizing data (Solanki et

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

261 al., 2015). This method transforms each variable by centering it around a mean of zero and scaling
262 it based on its standard deviation, thereby facilitating better model convergence. Proper
263 standardization minimizes training error rates and mitigates the risks of overfitting (where the
264 model learns noise rather than patterns) or underfitting (where the model fails to capture complex
265 relationships in the data). By ensuring uniformity in data representation, standardization enhances
266 the robustness and generalizability of ML models for WQ assessment (Solanki et al., 2015; Uddin
267 et al., 2022). In the present study, we employed a technique known as z-score normalization to
268 standardize the WQ data variables. For a detailed explanation, refer to the methodology of data
269 standardization elucidated by Uddin et al. (2022).

270 **3.3.2. Input preparation**

271 The standardized dataset was systematically partitioned into training and testing sets, with
272 80% (618 samples) allocated for training and 20% (154 samples) reserved for testing. This
273 partition ensured a balanced approach to model development and evaluation. The training set was
274 utilized to build and fine-tune applied ML model, while the testing set was employed to assess the
275 predictive performance and generalization ability. The model(s)'s performance was assessed
276 during both the testing and training stages. Therefore, the utilization of standardized data (z-score
277 normalized data), rigorous partitioning strategies (80% training and 20% testing), and thorough
278 performance evaluation (Uddin et al., 2023a) facilitated the creation of robust prediction models
279 capable of accurately capturing and predicting the target variable's behavior.

280 **3.3.3. Development of prediction model**

281 For the purposes of predicting the RMS-WQI score, the present study utilized the eXtreme
282 Gradient Boosting (XGB) ML algorithm (Fig. 2). The selection of the XGB algorithm is based on
283 its superior performance in earlier studies conducted by the authors in the realm of water research

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

284 (Uddin et al., 2023b, 2024a). Previously, the authors have conducted an extensive study where the
285 authors have compared the performance of eight ML algorithms (Random Forest-RF, Decision
286 Tree-DT, K- Nearest Neighbors-KNN, XGB, Extra Tree-ExT, Support Vector Machine-SVM,
287 Linear Regression-LR, and Gaussian Naïve Bayes-GNB) for predicting WQ state and the XGB
288 algorithm outperformed remaining seven algorithms (Uddin et al., 2023a). The RMS-WQI
289 framework was integrated with ML models to enhance the accuracy and adaptability of WQ
290 assessments. The XGB algorithm was used to predict RMS-WQI scores, exploiting optimized
291 hyperparameters to improve model performance. This integration allows for better generalization
292 of WQ patterns by minimizing errors and reducing the uncertainty commonly associated with
293 traditional WQI models. Additionally, a number of previous studies have also reported the robust
294 performance of XGB algorithm in several field of interest including hydrological modeling
295 (Budholiya et al., 2022; Ibrahim Ahmed Osman et al., 2021; Niazkar et al., 2024; Xiong et al.,
296 2023). Therefore, this research implemented the XGB algorithm following the approach as
297 outlined by Uddin et al. (2022). The XGB algorithm was developed using Python in the Google
298 CoLaboratory platform. Please refer to Uddin et al. (2022) and Uddin et al., (2024a) regarding the
299 details of the implementation process.

300 As mentioned in the studies by Faruq et al. (2025), Ibrahim Ahmed Osman et al. (2021),
301 Mohammadpour et al. (2024), Sajib et al. (2025), Uddin et al. (2024b) and Zhang et al. (2024),
302 XGB operates as an additive model, where its constituent sub-models, in this study, are regression trees.
303 The model's updates are sequentially dependent on the prediction outcomes of prior models. Initially, a base
304 model is trained on the data, and the residuals—representing the discrepancies between actual and predicted
305 values—are computed. In each iteration, an additional regression tree is introduced to the ensemble to
306 reduce these residuals. Notably, optimization is applied exclusively to the newly added trees at each update
307 step.

1
2
3
4 308 The mathematical representation of XGB is as follows:

$$5 \quad 6 \quad 7 \quad 309 \quad F_m(X_i) = F_{m-1}(X_i) + F_m(X_i) \quad \text{Eq. 2}$$

8
9
10 310 where m is the current step, $F_m(X_i)$ is the current step sub-model and $F_{m-1}(X_i)$ represents the
11
12 311 sub-models before step m .

13
14
15 312 Additionally, the object function of XGB is expressed as follows:

$$16 \quad 17 \quad 18 \quad 19 \quad 313 \quad Obj = \sum_{i=1}^N L[F_m(X_i), Y_i] + \sum_{j=1}^m \Omega(f_j) \quad \text{Eq. 3}$$

20
21
22 314 where Y_i presents the real value, $L[F_m(X_i), Y_i]$ depicts the loss function and $\Omega(f_j)$ represents the
23
24 315 regularization item, indicating the complexity of the sub model f .

25 26 27 316 **3.3.4. Model hyperparameter: Optuna**

28
29 317 A critical phase in developing ML models involves hyperparameter optimization, which
30
31 318 entails adjusting the hyperparameters to achieve optimal results (Elzain et al., 2023; R. Liu et al.,
32
33 319 2022). Hyperparameters are settings or configurations that are external to the model and are set
34
35 320 prior to the training process. They cannot be directly learned from the data but significantly affect
36
37 321 the performance and behavior of the model. The implementation of the XGB model adhered to the
38
39 322 methodology proposed in Uddin et al. (2024b), with hyper-parameter tuning performed using the
40
41 323 Optuna strategy (Fig. 2). Optuna, an open-source framework, is specifically designed to optimize
42
43 324 hyperparameters in ML and statistical models, making it particularly effective for improving the
44
45 325 predictive accuracy of GW quality assessments (Akiba et al., 2019). Most recently, the authors
46
47 326 have utilized five widely used hyperparameter optimization techniques including Grid Search,
48
49 327 Random Search, Bayesian Optimization, tree-based pipeline optimization tool-TPOT and Optuna
50
51 328 (Uddin et al., 2024a). This study revealed that ML models optimized with Optuna demonstrated
52
53 329 superior prediction capabilities for hydrological data (Uddin et al., 2024a). Furthermore, this
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

330 methodological approach ensured enhanced model reliability and accuracy, thereby providing
331 more precise insights into coastal GW quality of Bangladesh (Uddin et al., 2024a). Considering
332 Optuna’s consistent performance with hydrological data, this study utilized Optuna for optimizing
333 hyperparameters for the XGB model.

334 **3.3.5. ML model evaluation and Efficiency assessment**

335 Many prediction models use sensitivity analysis and other data processing and input
336 variable selection approaches to evaluate the predictive accuracy of implemented algorithms.
337 Statistical metrics like Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean
338 Absolute Error (MAE), Percentage Absolute Bias Error (PABE), and coefficient of determination
339 (R^2) have been computed (Uddin et al., 2022; Sajib et al., 2023). The mathematical formulas for
340 these models are provided in the supplementary information. These metrics are used to evaluate
341 the performance of the model during both the training and validation phases (Esmailbeiki et al.,
342 2020; Manzar et al., 2022; Rasool et al., 2022; Singha et al., 2021). The R^2 statistic measures the
343 percentage of the variance in the observed data explained by the model. RMSE, MSE, and MAE
344 metrics provides the discrepancies between predicted and observed values. Lower values of
345 RMSE, MSE, and MAE signify superior model performance. Ideally, for better model
346 performance, the values for RMSE, MSE, and MAE should be minimized (approaching zero). R^2
347 quantifies the extent to which the model can explain variance in the dependent variable. A higher
348 R^2 value indicates a better fit to the data, approaching 1. For detailed information regarding the
349 general methodological approach and mathematical equations, readers can refer to Uddin et al.
350 (2024a).

351 Model efficiency can be assessed based on different parameters of the user-defined interest,
352 such as the Nash-Sutcliffe efficiency (NSE) and the Model Efficiency Factor (MEF) (Clark et al.,

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

353 2021; McCuen et al., 2006; Sharif et al., 2022). Most of the recent model-based WQ prediction
354 research has applied these two metrics to improve the model results (McCuen et al., 2006; Sahour
355 et al., 2020; Che Nordin et al., 2021; Clark et al., 2021; Khosravi et al., 2023). The same metrics
356 were also used in the present study to evaluate the efficiency of the model deployed for spatial
357 WQ assessment. In this context, an NSE value close to one represents an excellent model, while a
358 smaller MEF value indicates a better (bias-free) model (Boo et al., 2024; Che Nordin et al., 2021;
359 McCuen et al., 2006; Sahour et al., 2020; Sajib et al., 2024).

360 **3.3.6. ML model sensitivity and uncertainty analysis**

361 Conducting sensitivity and uncertainty analysis in ML models for WQ assessment is
362 essential for evaluating the robustness and reliability of the predictions. Sensitivity analysis
363 involves systematically varying model parameters and inputs to identify the most influential
364 factors affecting the model's output, thereby ensuring consistency and dependability (Amaranto et
365 al., 2020; Manzar et al., 2022; Uddin et al., 2022; 2024; Talukdar et al., 2023). A sensitive WQ
366 model will respond appropriately to changes in input conditions, such as the addition or removal
367 of pollutants. The R^2 is a widely used tool to assess model sensitivity, with values ranging from 0
368 (indicating poor agreement) to 1 (indicating excellent agreement between the model's inputs and
369 outputs) (Uddin et al., 2024a). In this research, the methodology of Uddin et al. (2023a) was
370 employed to determine the R^2 (sensitivity) of the RMS-WQI model.

371 Uncertainty analysis, on the other hand, quantifies the degree of confidence in the model's
372 predictions by accounting for potential errors in data, model parameters, and structural
373 assumptions. Techniques such as Monte Carlo Simulations (MCS) and bootstrapping are
374 commonly employed, with MCS being the most widely used (Seifi et al., 2020; Xiong et al., 2022;
375 Sajib et al., 2023; 2024). MCS generates a distribution of possible outcomes, providing insights

1
2
3
4 376 into the range and likelihood of different WQ scenarios. Detailed information on model sensitivity
5
6
7 377 and uncertainty analysis has been provided by Uddin et al. (2024b). This study utilized the MCS
8
9 378 approach to estimate the model uncertainty by generating 2,50,000 samples.

12 379 **3.3.7. ML model ambiguity and eclipsing analysis**

15 380 Analyzing ambiguity and eclipsing in ML models for WQ prediction is crucial for
16
17 381 evaluating model performance and reliability. Previous studies have demonstrated that ambiguity
18
19 382 and model eclipsing issues in existing WQI models can introduce substantial uncertainty into the
20
21
22 383 evaluation process (Khouri and Bashar Al-Moufti, 2022; Uddin et al., 2022; Sajib et al., 2024).
23
24 384 Ambiguity arises from incorrect aggregation, while model eclipsing relates to the sub-index
25
26
27 385 functions of WQI models, both resulting from the overestimation or underestimation of these
28
29 386 functions (Uddin et al., 2022; 2023b). Ambiguity analysis focuses on the uncertainty in model
30
31
32 387 outputs due to inherent variability in data and model parameters, highlighting areas where
33
34 388 predictions may be less certain or stable. Eclipsing analysis, conversely, examines how dominant
35
36
37 389 features or patterns in the data can obscure the influence of other significant factors (Uddin et al.,
38
39 390 2023a). Together, these analyses enhance the understanding of ML model limitations and guide
40
41 391 improvements, ensuring more accurate and reliable WQ predictions. The methodology of Uddin
42
43
44 392 et al. (2022) was applied in this study to analyze ambiguity and eclipsing.

47 393 **4. Results and interpretations**

49 394 **4.1. Descriptive summary of the GW quality indicators**

52 395 Table 3 presents a detailed analysis of the groundwater chemical composition, comparing
53
54 396 WQ indicators with BIS (2012) drinking water guidelines. Fig. 3 complements this analysis with
55
56
57 397 visual representations in the form of box and whisker plots. Moreover, Pearson's correlation test
58
59 398 (at a 99% confidence level) revealed significant positive and negative correlations between the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

399 GW indicators (Fig. 4). Notably, pH, a fundamental measure of water acidity or alkalinity, exhibits
 400 a mean value of 7.9, falling within the range of 7 to 9, indicating a slightly alkaline nature. TH
 401 ranges from 15 to 4000 with a mean of 443 mg/L in studied samples. Approximately 79 % of the
 402 samples exceed the desirable limit of 200 mg/L, as specified by BIS (2012). TH is a parameter
 403 crucial for assessing water's propensity to form scale deposits and impact industrial processes and
 404 domestic appliances. Moreover, TH exhibits a strong positive correlation with electrical EC and
 405 TDS, both of which have correlation coefficients of 0.79. This suggests that as the hardness of the
 406 water increases, there is a corresponding increase in the conductivity and the total amount of
 407 dissolved solids, reflecting higher mineral content in the water. Furthermore, TH is highly
 408 correlated with Ca^{2+} and Mg^{2+} , with correlation coefficients of 0.90 and 0.97, respectively. This
 409 strong correlation is expected, as Ca^{2+} and Mg^{2+} ions are primary contributors to water hardness
 410 (Nisa and Umar, 2023; Haidery et al., 2024; Khan and Ayaz, 2024). EC ranges from 200 to 16560
 411 μ S/cm, with a mean of 2496 μ S/cm, while TDS vary from 130 to 10764 mg/L, with a mean of
 412 1623 mg/L. Approximately 85 % of the samples exceed the permissible drinking water levels for
 413 EC, while 84 % surpass the recommended levels for TDS. Elevated EC levels can adversely affect
 414 WQ and its suitability for various applications. The high TDS levels observed in the water samples
 415 may stem from dissolved salts resulting from sediment–water interaction, irrigation return flows,
 416 and potential evaporation processes (Merchán et al., 2020), impacting water taste, odor, and clarity,
 417 thereby posing health related challenges.

418 **Table 3.** Descriptive statistics of analyzed groundwater samples ($n = 772$) (CGWB, 2022). All concentrations are in
 419 mg/L except pH and EC (μ S/cm).

WQ Indicators	Min	Max	Mean	SD	BIS (2012)	Present Study			
					Standard Range*	# SEDL	% SEDL	# SEPL	% SEPL
pH	6.9	9.1	7.9	0.3	6.5–8.5	772	100	15	2
TH	15	4000	443	448	200–600	611	79	138	18
EC	200	16560	2496	2529	750–3000	654	85	193	25
TDS	130	10764	1623	1644	500–2000	647	84	185	24
Ca^{2+}	2	570	76	68	75–200	258	33	38	5

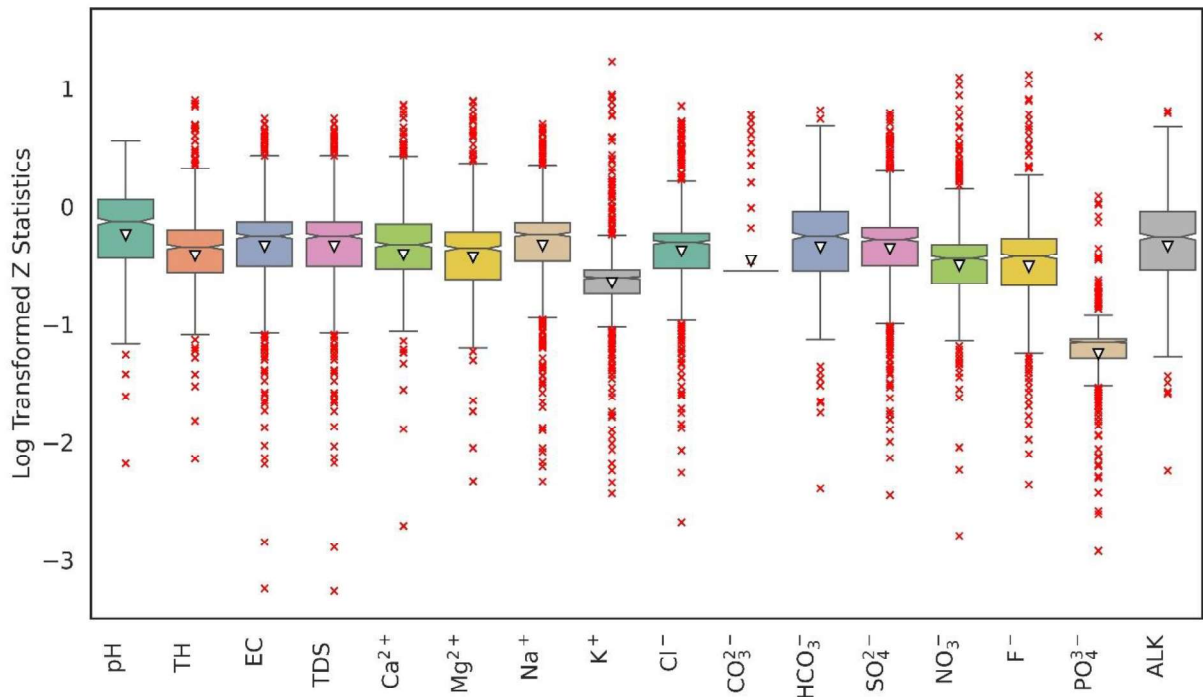
Mg ²⁺	2	642	62	74	30–100	482	62	115	15
Na ⁺	12	2714	384	465	200	395	51	—	—
K ⁺	0	843	16	49	12	170	22	—	—
Cl ⁻	14	5176	441	661	250	337	44	—	—
CO ₃ ²⁻	0	120	5	19	—			—	—
HCO ₃ ⁻	24	1915	410	230	200–600	684	89	113	15
SO ₄ ²⁻	1	2550	259	369	200–400	276	36	149	19
NO ₃ ⁻	0	1750	69	135	45	277	36	—	—
F ⁻	0	22	1	2	1–1.5	228	30	130	17
PO ₄ ³⁻	0	220	1	8	—			—	—
ALK	20	1570	345	190	200–600	623	81	62	8

420 ALK = Alkalinity; SD = standard deviation; BIS = Bureau of Indian standards; Range* = Lower value indicates
421 desirable limit (DL) while higher value refers to permissible limit (PL); # SEDL = number of samples exceeding
422 desirable limit; % SEDL = percentage of samples exceeding desirable limit; while # SEPL = number of samples
423 exceeding permissible limit; % SEDL = percentage of samples exceeding permissible limit.

424
425 Among the major cations, Ca²⁺ concentrations vary from 2–570 mg/L, averaging 76 mg/L, with
426 33% of samples exceeding the desirable limit (75 mg/L), though only 5% surpass the permissible
427 limit (200 mg/L). Similarly, Mg²⁺ ranges from 2–642 mg/L (mean 62 mg/L), exceeding the
428 desirable limit (30 mg/L) in 62% of samples, while 15% surpass the permissible limit (100 mg/L).
429 Na⁺ concentrations (12–2714 mg/L, mean 384 mg/L) exceed the desirable limit (200 mg/L) in 51%
430 of samples, suggesting possible ion exchange and evaporative concentration processes (Kumar et
431 al., 2023). Among the anions, Cl⁻ levels vary widely (14–5176 mg/L) and exceed the desirable
432 limit (250 mg/L) in 44% of samples, indicating evaporative concentration and anthropogenic
433 inputs. HCO₃⁻, which ranges from 24–1915 mg/L (mean 410 mg/L), surpasses the desirable limit
434 (200 mg/L) in 89% of samples, reflecting carbonate weathering processes (Rahman et al., 2021).
435 SO₄²⁻ levels (1–2550 mg/L, mean 259 mg/L) exceed the desirable limit (200 mg/L) in 36% of
436 samples, while 13% of samples surpass the permissible limit (400 mg/L), potentially due to sulfide
437 oxidation and industrial effluents. NO₃⁻ concentrations (0–1750 mg/L, mean 69 mg/L) exceed the
438 desirable limit (45 mg/L) in 35% of samples, indicating potential agricultural and domestic
439 contamination. F⁻ levels (0–22 mg/L, mean 2 mg/L) exceed the desirable range (1 mg/L) in 30%
440 of samples, while 17% of samples exceed the maximum permissible limit (1.5 mg/L), indicating

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

441 geogenic fluoride enrichment in in studied samples. Alkalinity (ALK), varying from 20–1570
442 mg/L, with an average of 345 mg/L, exceeds the desirable limit (200 mg/L) in 81% of samples,
443 while 8% of samples surpass the permissible limit (600 mg/L), suggesting significant carbonate
444 dissolution in the study area.



445
446 **Fig. 3.** Summary of statistics for GW quality indicators.
447

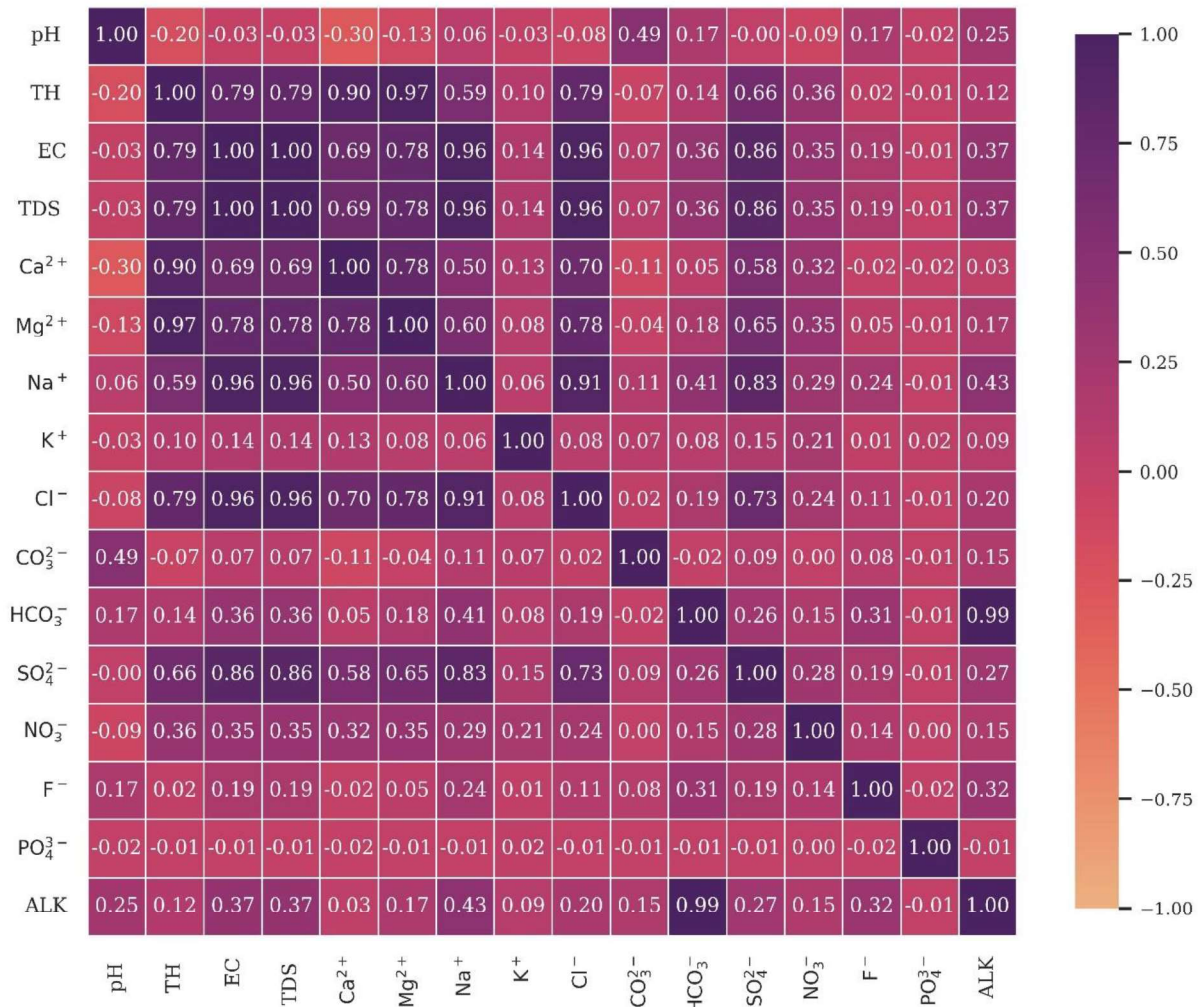


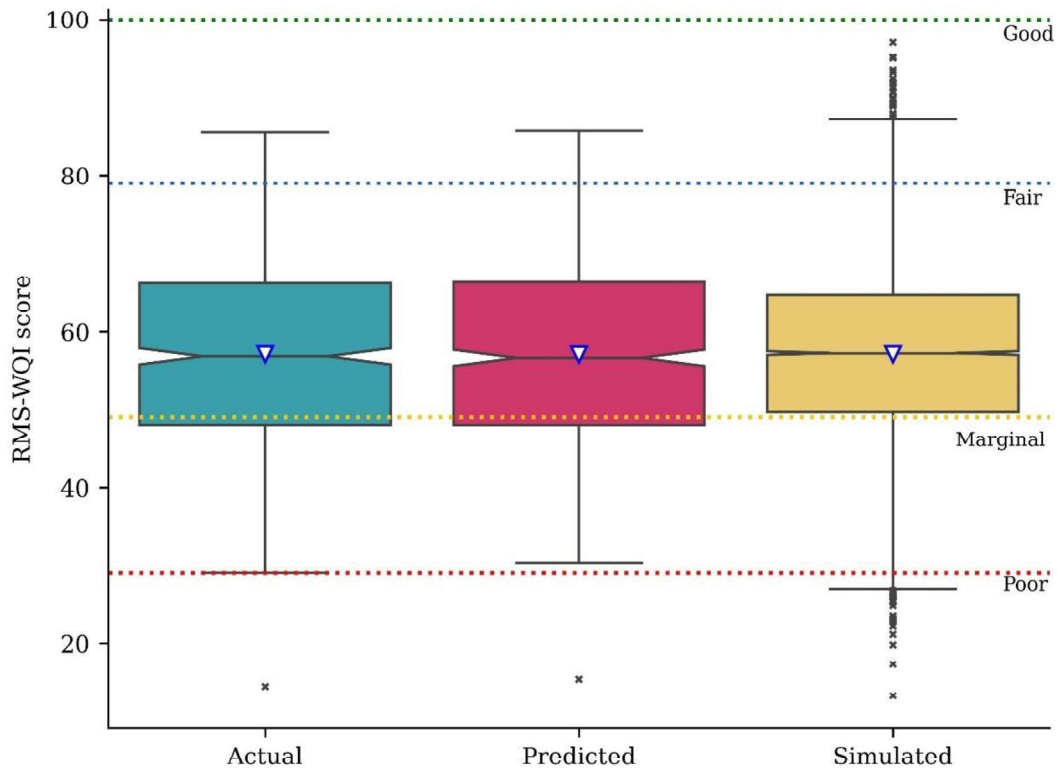
Fig. 4. Pearson's correlation between GW quality indicators.

4.2. Results of the RMS-WQI scores

As mentioned in the earlier section, the research utilized the RMS-WQI model for computing WQI scores and assessing the GW quality status of the study region. This model provides a quantitative measure of WQ, with RMS scores ranging from 0 to 100, where a score of 0 denotes severe contamination, and a score of 100 indicates pristine, uncontaminated water. This broad range allows for an assessment of WQ, classifying various categories of the RMS-WQI (Table 2). Table ST1 provides site-wise information of RMS-WQI scores. The statistical summary of the computed RMS-WQI scores for the study area is visually represented in Fig. 5, offering a

1
2
3
4 459 clear depiction of the GW quality based on this robust and objective index. The box plot in Fig. 5
5
6
7 460 compares the RMS-WQI scores for actual, predicted, and simulated GW quality data. Each box
8
9 461 plot displays the score distribution through their quartiles, highlighting the median, mean (white
10
11 462 dot), interquartile range, and potential outliers. The RMS-WQI scores varied between 14.5 to 85.5
12
13
14 463 with a mean of 57.1 ± 11.2 and a median score of 56.8 (Fig. 5). The simulated RMS-WQI score
15
16 464 also captured statistical attributes (mean: 56.0 ± 11.1 and median 56.1) which are consistent with
17
18 465 the actual RMS-WQI score for GW. The actual and predicted datasets exhibit similar central
19
20
21 466 tendencies, with medians and mean around 60, indicating that the prediction model accurately
22
23
24 467 estimates the observed GW quality. The consistency in variability, represented by the interquartile
25
26 468 range, between the actual and predicted data further suggests that the prediction model effectively
27
28
29 469 captures the inherent variability of GW quality. However, the simulated dataset shows a larger
30
31 470 spread and more outliers, reflecting greater variability and the presence of extreme values. This
32
33
34 471 suggests that the simulation model might introduce or capture additional variability not present in
35
36 472 the observed data. Overall, Fig. 5 indicates that the prediction model aligns closely with the actual
37
38
39 473 data in terms of central tendency and variability, demonstrating good XGB model performance in
40
41 474 predicting the IEWQI scores. Nevertheless, while the simulation model offers a broader
42
43 475 perspective on potential GW quality outcomes, it exhibits higher variability and more outliers.
44
45
46 476 Additionally, supplementary Fig. S2 to S5 illustrate the spatial distribution of the calculated RMS-
47
48 477 WQI scores at each monitoring site within the individual districts of Rajasthan state. These Fig.
49
50
51 478 S2 to S5 provide a detailed view of the GW quality across various locations, highlighting the
52
53 479 differences in RMS-WQI scores. These visualizations reveal that the RMS-WQI scores are
54
55
56 480 relatively high in the districts of Banswara, Chittorgarh, Pali, Tonk, and Udaipur, indicating better
57
58 481 WQ. Furthermore, the spatial distribution of WQ status throughout the study area is depicted in
59
60
61
62
63
64
65

1
2
3
4 482 Fig. S6 to S9. Despite the higher RMS-WQI scores, the overall WQ status in these districts ranges
5
6 483 from marginal to poor, as shown in Fig. S6 to S9. This comprehensive spatial analysis emphasizes
7
8
9 484 the variability in GW quality within Rajasthan and demarcates the areas where WQ improvements
10
11 485 are most needed.



40 486
41 487
42
43 488
44
45 489
46
47 490
48
49 491
50
51
52 492
53
54 493
55
56 494
57
58
59 495
60
61
62
63
64
65

Fig. 5. Comparison of actual and predicted RMS-WQI Scores.

4.3. Prediction results of RMS-WQI model

4.3.1. ML Model evaluation results

The current study utilized eXtreme Gradient Boosting (XGB) ML algorithm and Optuna as model hyperparameter to forecast the RMS-WQI score, addressing the complexities involved in the intricate computation and sub-indexing process of WQI assessment (Lap et al., 2023; Sajib et al., 2023; 2024; Uddin et al., 2024). Accordingly, various statistical measures were adopted to evaluate model performance as detailed in methodological section 3.3.5. The model performance

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

496 matrix during the training and testing phase has been illustrated in Fig. 6. A similar approach has
497 been applied in earlier studies to identify optimal predictive models (Singha et al., 2021; Uddin et
498 al., 2022, 2023a; Manzar et al., 2022). By using these indicators, researchers can identify areas
499 where prediction models require improvement and determine the optimal strategies to enhance
500 model performance. RMSE, indicating the average magnitude of prediction errors, demonstrated
501 superior model performance with lower values (Fig. 6). The model achieved an RMSE of 1.02
502 during the training phase and 0.57 during the testing phase, accentuating its improved accuracy.
503 Similarly, MSE, which measures the average of the squared errors, yielded favorable results with
504 values of 1.05 and 0.32 for the training and testing sets, respectively, suggesting strong
505 generalization capability. MAE which is less affected by outliers compared to RMSE and MSE,
506 further validated the model's effectiveness with values of 0.67 for the training phase and 0.36 for
507 the testing phase, indicating superior performance. Additionally, PABE which quantifies bias in
508 percentage terms showed reduced bias in test predictions compared to training, with values of 1.19
509 and 0.64 for the training and testing phase, respectively. Moreover, Fig. 6 illustrates the model's
510 superior performance on test data across all four metrics, indicating robust generalization and
511 predictive capabilities of the RMS-WQI. This highlights the efficacy of the selected ML
512 algorithms in accurately predicting the RMS-WQI score, thereby enabling precise WQ assessment
513 with minimal error rates.

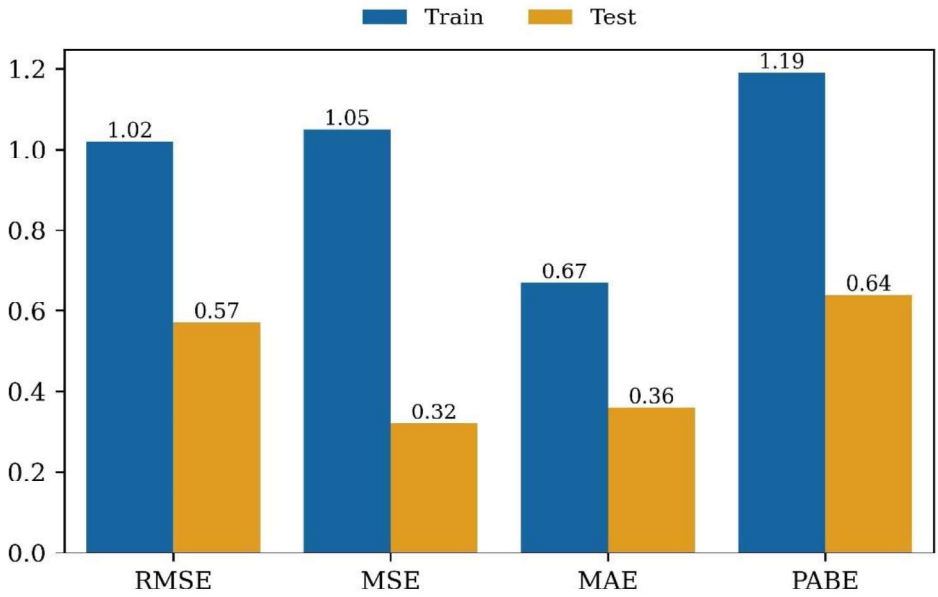


Fig. 6. Performance metrics of the model during training and testing phases.

4.3.2. Effects of Optuna hyperparameters on the prediction capabilities

The most commonly used ML optimization approaches include Bayesian, Optuna optimization, random search, and grid search. In this study, hyperparameters were tuned using an Optuna optimization technique to identify the best-fitting model (Akiba et al., 2019). Table 4 lists the optimal hyperparameters and their corresponding values for the XGB–Optuna model, representing the most effective combination for predicting GW quality. The use of Optuna for hyperparameter optimization significantly enhances the prediction capabilities of models used for GW quality assessment. By systematically searching for the best combination of hyperparameters, Optuna ensures that the model is neither overfitting nor underfitting the data. This leads to more accurate and reliable predictions. Previous studies on WQ prediction have shown that Optuna optimization yields impressive results (Krishnaraj and Honnasiddaiah, 2022; Uddin et al., 2024b).

1
2
3
4 **529** **Table 4.** Optimized values for the XGB model.
5

Model attributes	Optimized value
booster	Gbtree
lambda	0.090
alpha	0.632
max_depth	3
n_estimators	293
learning_rate	0.077
subsample	0.869
colsample_bytree	0.766
min_child_weight	1
Optimizer	Optuna

17
18 **530**

19
20 **531** **4.3.3. Comparison between actual and predicted WQI score**
21

22
23 **532** The pairwise comparison of the actual, predicted, and simulated RMS-WQI scores with a
24
25 **533** 95% confidence interval (CI) was conducted for statistical validation (Fig. 7). Tukey's Honest
26
27 **534** Significant Difference (HSD) test, a widely used advanced statistical method for multiple
28
29
30 **535** comparisons, was employed for this analysis (see Uddin et al., 2023a for further details). This
31
32 **536** method allows for the intercomparison of all pairs of models using the mean statistical metric.
33
34
35 **537** Both the simulation and prediction models effectively estimate RMS-WQI scores, as illustrated in
36
37 **538** Fig. 7, with mean differences being nearly equal across all comparisons. The narrower CI for the
38
39
40 **539** predicted scores suggests that the predictions are slightly more precise than the simulated results.
41
42 **540** This implies that the predictive model may offer more accurate and consistent estimates. Despite
43
44 **541** the minor variations in scores between the simulated and predicted results, the simulation model
45
46
47 **542** still proves to be a reliable alternative to direct forecasts. However, the slightly broader CI in the
48
49
50 **543** simulated comparisons indicates that there is room for improvement in the accuracy and
51
52 **544** consistency of the simulation model. Overall, Fig. 7 demonstrates that the predictive model
53
54 **545** accurately predicts WQ indices, while the simulation model performs well but with slightly more
55
56
57 **546** variation. Although there is a noticeable variation between the models' computed RMS-WQI
58
59 **547** scores (Fig. 7), the primary goal of the RMS-WQI model is to rank WQ using a specific
60
61
62
63
64
65

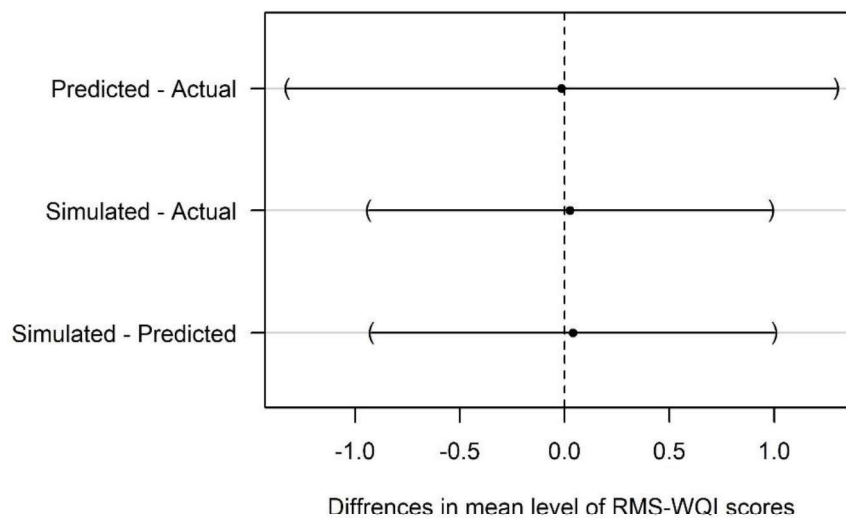
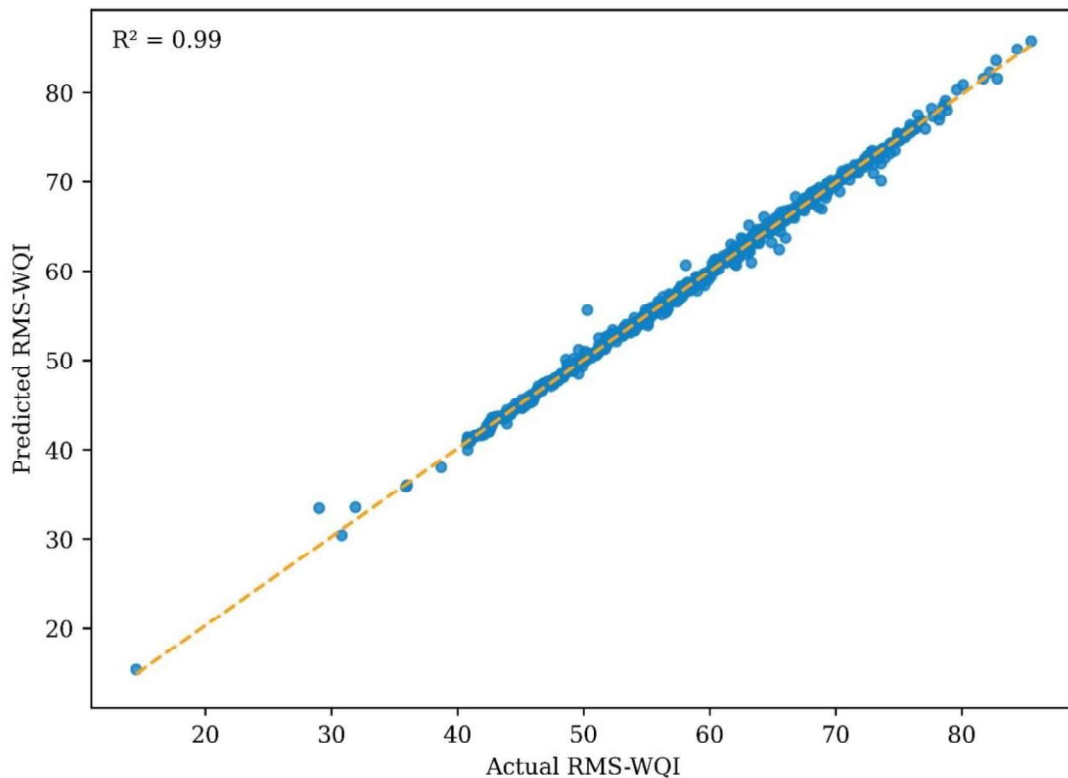


Fig. 7. Pair-wise comparison between the actual, predicted, and simulated RMS-WQI scores with 95% CI from Tukey's HSD.

4.4. Model sensitivity results

The sensitivity analysis results reveal a high level of agreement ($R^2 = 0.99$) between the model's input and output variables, as shown in Fig. 8. This outcome aligns with previous WQ evaluation studies (Singha et al., 2021; Parween et al., 2022; Uddin et al., 2022; Gani et al., 2023; Sajib et al., 2023). With R^2 result of 0.99, the model demonstrated that 99% of the variability in its predictions are accounted by independent variables. The residual of 1% in the model might be attributed to additional factors, which may include the existence of outliers in the dataset. Furthermore, the constructed model demonstrates a strong fit with the data, as evidenced by the random residual bar plot of anticipated RMS-WQI values (Fig. 8). This indicates that the model is highly accurate and could be effectively applied to GW quality assessments. Nevertheless, it is

1
2
3
4 567 worth mentioning that the developed model did not consider seasonal fluctuation of GW and
5
6
7 568 therefore, further refinement would be required with seasonal GW quality data for maintaining the
8
9 569 model's accuracy to assess GW resources. Additional statistical evaluations of the RMS-WQI
10
11 570 model were conducted at the district level, as shown in supplementary Fig. S10 to S12, to identify
12
13
14 571 any potential ambiguities in the model's local-level outputs. These findings suggest that the RMS-
15
16 572 WQI model is capable of accurately predicting GW quality in the study area, as the input factors
17
18
19 573 fully explain the variability in the output variables.
20

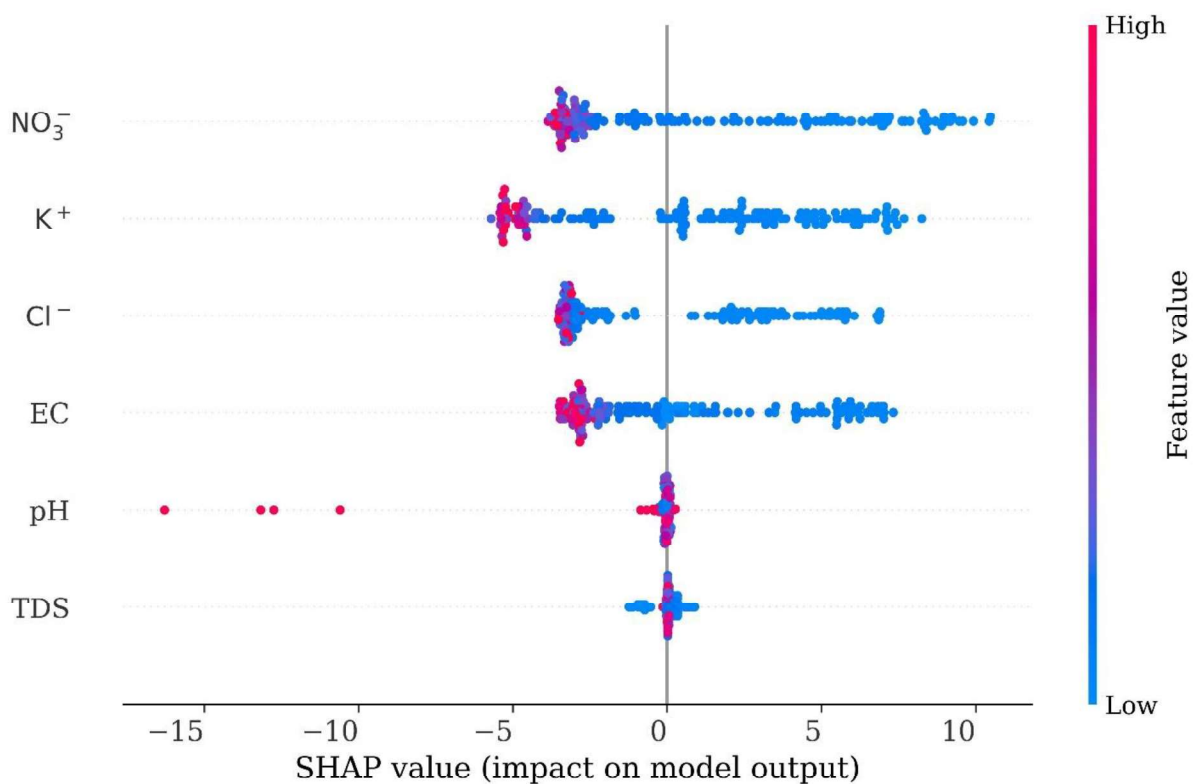


21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48 574
49 575
50 576
51 577
52
53 578
54
55
56 579
57
58 580
59
60
61
62
63
64
65

Fig. 8. Sensitivity analysis of the RMS-WQI model for predicting WQI scores.

In addition to R^2 , for the purposes of determining the WQ indicators' impact on the RMS-WQI model, the research utilized the SHapley Additive exPlanations (SHAP), because recently a range of water research studies have used this technique (El Bilali et al., 2023; Liu et al., 2022; Nallakaruppan et al., 2024; S. Wang et al., 2022; X. Wang et al., 2022). It is a powerful ML

1
 2
 3
 4 581 interpretability tool based on Shapley values from cooperative game theory technique that can be
 5
 6
 7 582 provided details model input impacts on output (Palar et al., 2023). Figure 9 shows various WQ
 8
 9 583 indicators impact on RMS-WQI model and their importance values. It can be seen from the figure,
 10
 11 584 the NO_3^- , K^+ , and Cl^- are most influential WQ indicators these are impacting the model output.
 12
 13
 14 585 Whereas the pH has significant negative impact on the model output, with consistently negative
 15
 16 586 SHAP values, suggesting that higher pH levels strongly decrease the predicted RMS-WQI score
 17
 18
 19 587 (Fig. 9). For instance, higher values of pH and NO_3^- tend to have a significant negative effect,
 20
 21 588 while variations in TDS and EC contribute to both positive and negative impacts. The results of
 22
 23
 24 589 the SHAP reveals that NO_3^- plays the most crucial role, followed by potassium (K^+) and chloride
 25
 26 590 (Cl^-), in predicting the RMS-WQI scores in groundwater.
 27
 28
 29 591
 30
 31
 32
 33
 34
 35
 36
 37
 38
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
 49
 50
 51
 52
 53
 54
 55
 56
 57



58 592
 59 593 **Fig. 9.** SHAP summary results of feature importance on RMS-WQI model.
 60
 61
 62
 63
 64
 65

4.5. Model uncertainty results

The probability density function (PDF) and cumulative distribution function (CDF) plots of the RMS-WQI model (Fig. 10) demonstrate a normal distribution among real, predicted, and simulated data, with no significant differences observed. Supplementary Table ST2 presents the standard uncertainty and expanded uncertainty for the RMS-WQI model as 77.05 ± 0.64 and 77.05 ± 1.30 , respectively. Standard uncertainty has been utilized in numerous studies to assess model uncertainty (Ding et al., 2023). Moreover, employing expanded uncertainty estimates enhances confidence levels in the uncertainty of measured outcomes (Carvalho et al., 2021). The overall uncertainty of the model is less than 2%, indicating that the RMS-WQI model is suitable for precise prediction. These uncertainty findings are consistent with previous studies (Sajib et al., 2024).

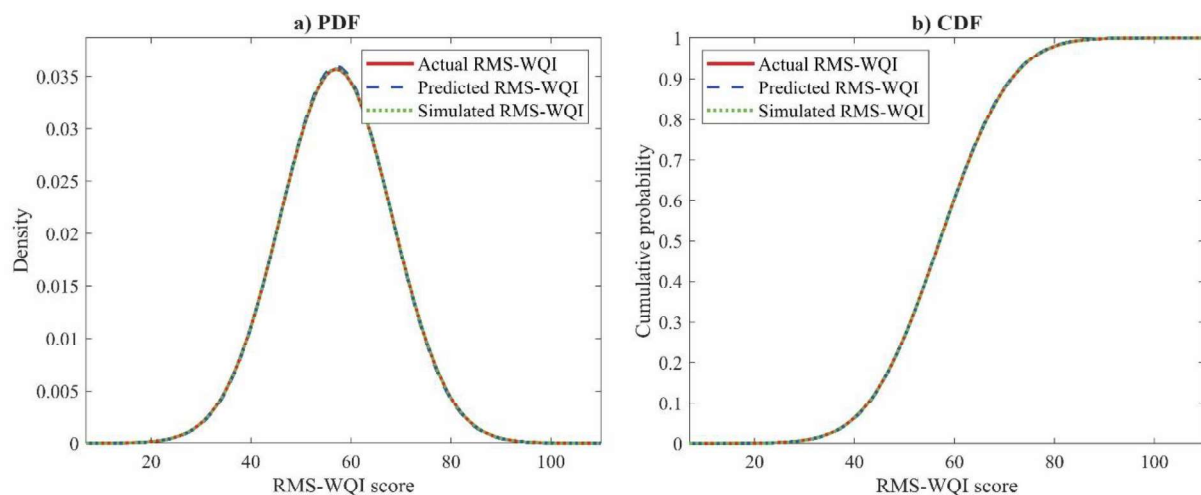


Fig. 10. Comparison of PDF and CDF plots for actual, predicted, and simulated RMS-WQI models.

4.6. Model eclipsing and ambiguity results

Research advancements in WQI studies (Mukate et al., 2019; Uddin et al., 2022; Chidiac et al., 2023; Gani et al., 2023; Haidery et al., 2024; Khan and Umar, 2024b) have underlined the noteworthy uncertainties introduced into the WQI computations owing to model eclipsing and ambiguity. The model ambiguity issues typically associated with the aggregation function and

1
2
3
4 611 ambiguity obscures the true WQ information used as model input (Uddin et al., 2022). On the other
5
6
7 612 hand, the aggregation function's tendency to overestimate the WQI score results in the occurrence
8
9 613 of the eclipsing issue and does not reflect actual WQ status (Uddin et al., 2022). Detailed
10
11 614 description of these two terms could be found in (Uddin et al., 2022). These model-associated
12
13
14 615 uncertainties can mislead the results and interpretation of the WQ, thereby hindering the overall
15
16 616 cost-effective and sustainable water resource management in a given ecosystem. To tackle such
17
18
19 617 problems, an effort was made utilizing an improved protocol (Uddin et al., 2022) to determine the
20
21 618 eclipsing and ambiguity in two WQI models at 772 sampling sites (Table 5). A detailed site
22
23
24 619 breakdown of these issues is graphically shown in the supplementary information as a prediction
25
26 620 error index (PREI) score (Fig. S13-S15). As a thumb rule, PREI scores with the least deviation
27
28
29 621 (i.e., closer to zero) result in better performance of the model (Parween et al., 2022; Uddin et al.,
30
31 622 2022; 2023a). The RMS-WQI output revealed that 25% of the total water sampling sites
32
33 623 encountered eclipsing issues, while half (54%) of the sites experienced ambiguity during the RMS-
34
35
36 624 WQI calculation (see Table 5). The WQ information from the RMS-WQI model, in general, is just
37
38 625 fine looking at such a large regional scale but could be effective with improvement in the
39
40
41 626 encountered issues for reliable and more accurate assessment of the WQ. Uddin et al. (2022)
42
43 627 suggested that proper computation of sub-index values with appropriate sub-index function and
44
45
46 628 threshold limit could potentially minimize modeling eclipsing and ambiguity issue. Nevertheless,
47
48 629 the RMS-WQI output of the current study is consistent with the previous findings in another part
49
50
51 630 of the region of the Indian subcontinent (Parween et al., 2022; Uddin et al., 2022;2023a; Sajib et
52
53 631 al., 2023).

54
55 632 **Table 5.** Model ambiguity and eclipsing results.

Attributes	Statistics
Eclipsing	190 (25%)
Ambiguity	416 (54%)
Total sites	772

4.7. Model efficiency analysis

The model efficiency results from the present study indicate that the majority of the districts investigated for WQ in Rajasthan exhibit bias-free outcomes with satisfactory model performance. This finding is consistent with similar studies conducted in other parts of the Indian subcontinent (Gani et al., 2023; Sajib et al., 2023). However, certain districts, specifically Bikaner, Gangangar, and Sirohi, demonstrate significantly lower NSE values, occasionally dropping below -5. Such low NSE values point to poor model performance in these regions, suggesting notable discrepancies between the model's predictions and the observed data (Fig. 11). Additionally, detailed results of the model efficiency, encompassing both NSE and MEF, are provided in Fig. S16 to S18. These results demonstrate that employing these approaches can effectively provide accurate WQI scores with minimal bias. Nevertheless, it is essential to also incorporate other parameters, such as the PREI score, which serves as an additional measure of model performance (Fig. 12). Integrating PREI scores helps in cross-checking any ambiguities and errors present in the model's RMS-WQI output. Moreover, point-wise model prediction error results for the study area are presented in Fig. S13 to S15. The model efficiency results are generally consistent with the PREI findings. However, exceptions exist in districts like Jhunjhunu, Naguar, and Hanumangarh (see Fig. 12), where significantly higher PREI scores were observed, contradicting the better model performance suggested by the NSE and MEF results. This discrepancy underscores the importance of using multiple evaluation metrics to achieve a robust and comprehensive assessment of WQ.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

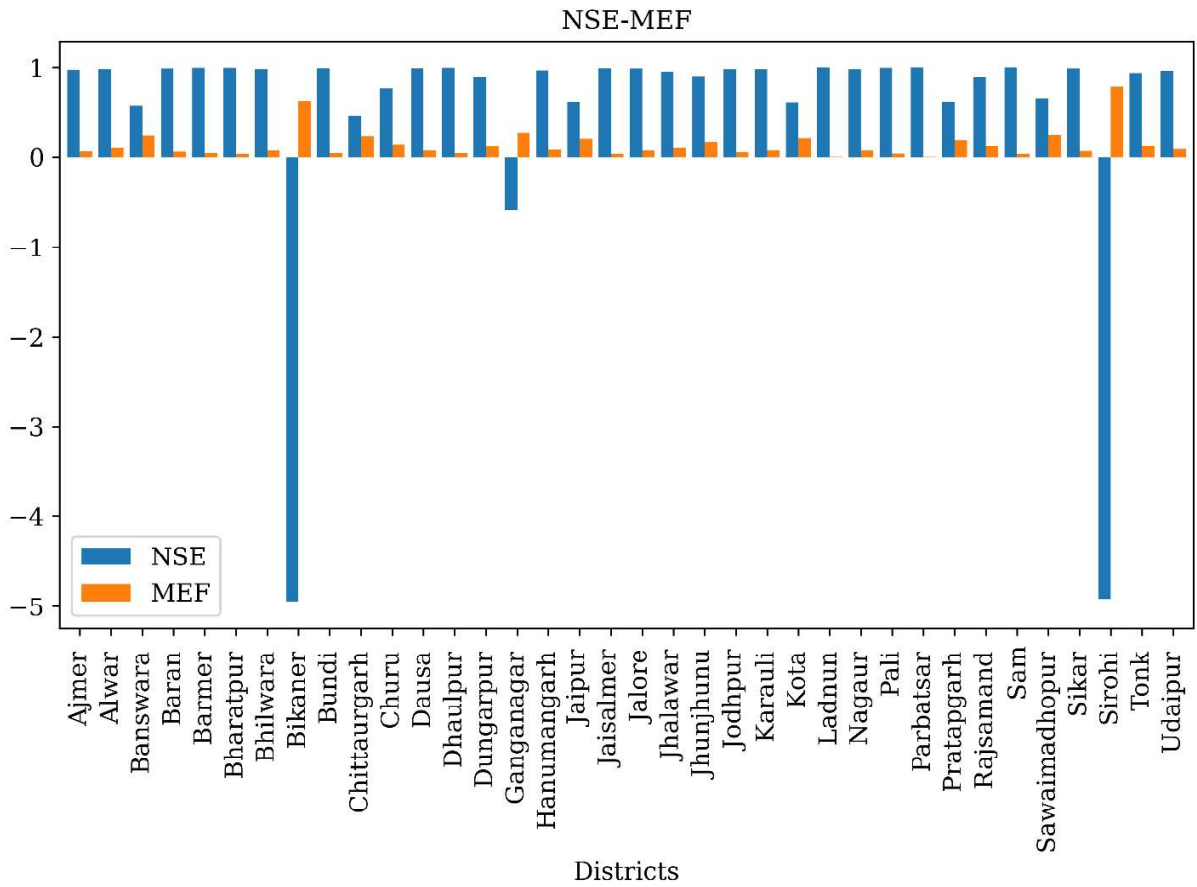


Fig. 11. Showing NSE and MEF for districts of Rajasthan.

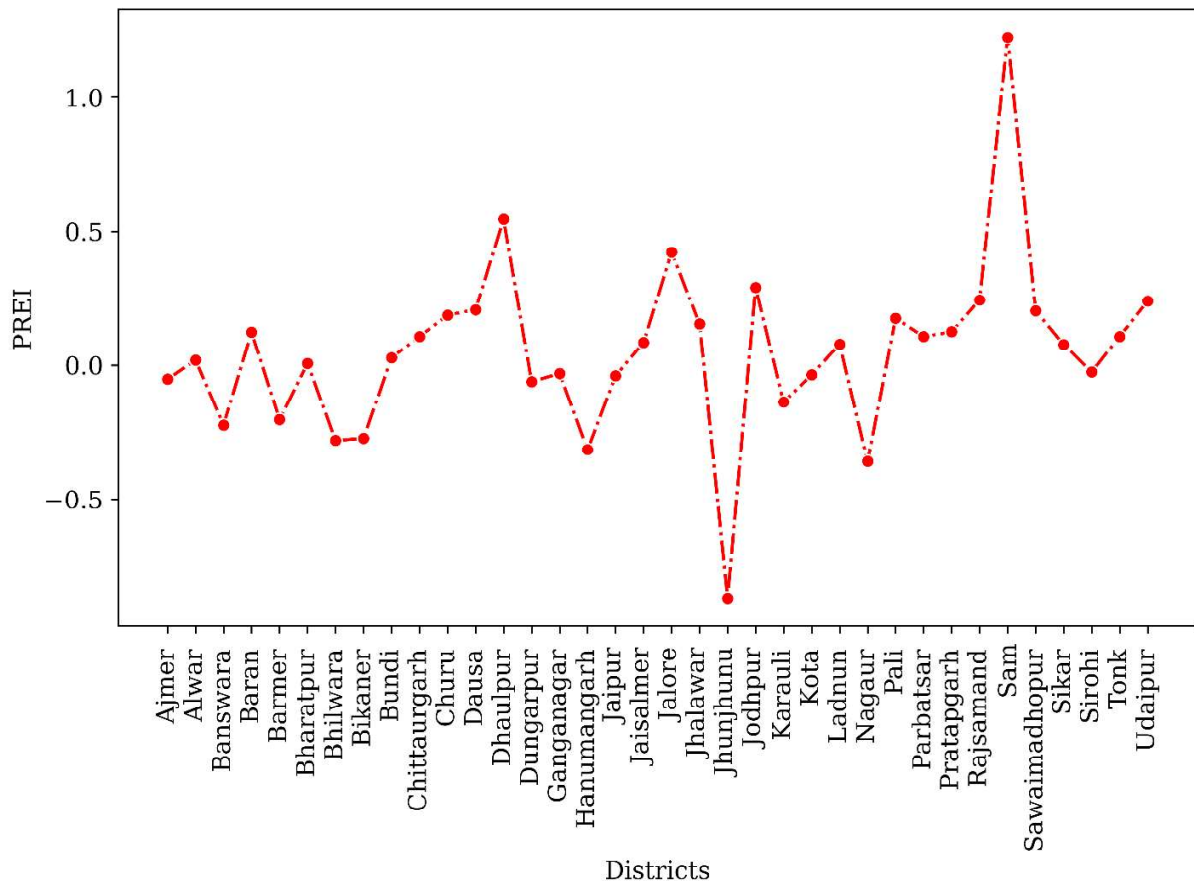


Fig. 12. Showing district-wise model prediction error results.

5. Assessing GW quality in Rajasthan

A comprehensive statistical summary of WQ monitoring across various districts in Rajasthan is presented in Fig. 13 and 14. These figures categorize the monitoring sites into four distinct WQ classification classes: Good, Fair, Marginal, and Poor, providing an insightful overview of the WQ status in the region. In Fig. 13, districts like Barmer and Bikaner display a notable presence of poor WQ sites, with Bikaner showing 8% marginal sites and Barmer with 16% marginal sites. The districts of Ajmer, Gangangar, and Hanumangarh also reflect significant concerns, with 10%, 12% and 10% of monitoring sites, respectively, falling into the marginal category. Ajmer, Alwar, and Banswara exhibit a relatively balanced distribution of WQ, with Ajmer showing 15% fair sites, Alwar having 24% fair and 1% marginal sites, and Banswara with

1
2
3
4 668 32% fair and 3% marginal sites. Bhilwara, on the other hand, shows a substantial proportion of
5
6
7 669 fair quality sites at 31%, yet a high percentage of marginal sites at 24%. Jaipur stands out with the
8
9 670 highest percentage of fair quality sites at 55%, indicating a predominance of satisfactory water
10
11 671 conditions. Fig. 14 extends this analysis to additional districts. Jodhpur and Sikar are highlighted
12
13
14 672 for their higher percentages of marginal sites at 18% and 9%, respectively, suggesting areas that
15
16 673 require attention to improve WQ. In contrast, districts like Jalore and Udaipur show predominantly
17
18 674 fair quality sites, with Udaipur at 24%. However, districts such as Jhunjhunu and Jhalawar indicate
19
20
21 675 a mix of WQ statuses, with Jhunjhunu also having a small percentage of poor sites (1%). In
22
23
24 676 addition, the spatial distribution of WQ status of selected districts is provided in supplementary
25
26 677 Figs. S6 to S9. As depicted in these figures, significant spatial variations in WQ are evident across
27
28
29 678 different locations. For instance, some regions exhibit consistently high levels of contaminants,
30
31 679 indicating poor WQ, while other areas demonstrate lower contaminant concentrations, suggesting
32
33
34 680 relatively good WQ. These variations may be attributed to several factors, including local
35
36 681 geological conditions, industrial activities, agricultural practices, and population density.
37
38 682 Therefore, the RMS-WQI model offers a valuable tool for monitoring and assessing WQ of a given
39
40
41 683 region.

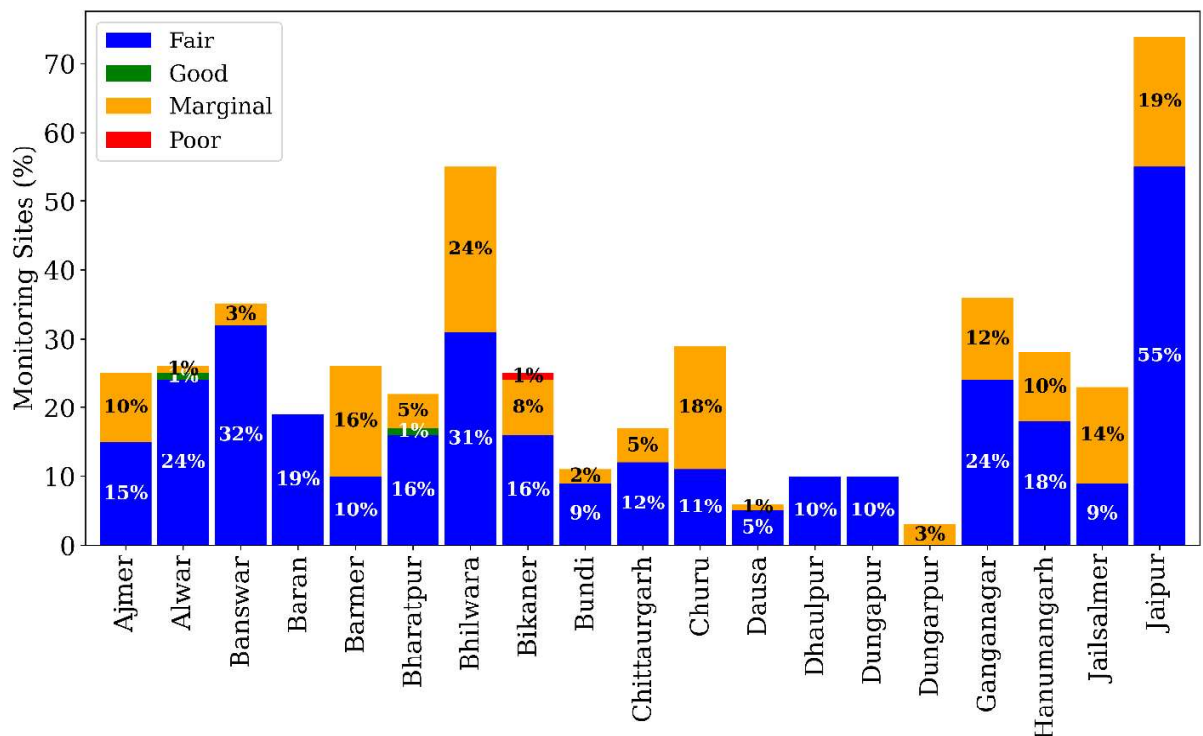


Fig. 13. Statistical summary of WQ status in different locations in the study area.

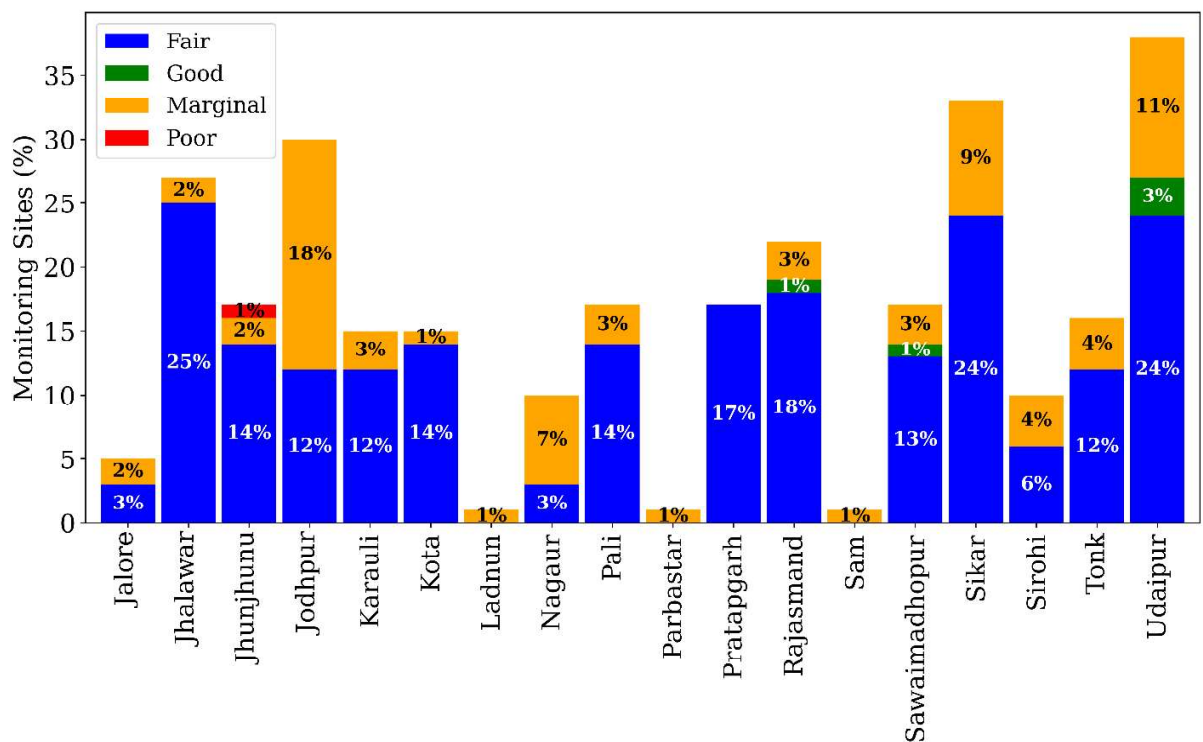


Fig. 14. Statistical summary of WQ status in different locations in the study area (cont.).

689

690 6. Discussion

691 This study presents an advanced approach for WQ assessment, utilizing key
692 physicochemical parameters (Table 3) to evaluate drinking water suitability in the arid and semi-
693 arid regions of the Rajasthan, India. The GW quality analysis highlights significant variations in
694 key physicochemical parameters, influencing its suitability for drinking purposes. For instance,
695 TH, EC, and TDS exhibited substantial variability, with 79%, 85%, and 84% of samples surpassing
696 the desirable limits (BIS, 2012), respectively. These findings suggest the influence of mineral
697 dissolution, high evaporation rates, and anthropogenic activities such as agriculture and industrial
698 discharge (Rahman et al., 2021). Among cations, Na^+ and K^+ concentrations exceeded desirable
699 limits in 51% and 22% of samples, respectively, likely due to mineral weathering, irrigation return
700 flow, and fertilizer application (Kumar et al., 2023). High Cl^- , SO_4^{2-} and NO_3^- levels were
701 detected in 44%, 36%, and 36% of samples, respectively, suggesting potential contamination from
702 agriculture, industrial effluents, and sewage infiltration. Notably, F^- concentrations exceeded
703 permissible limits in 17% of samples, posing a risk of fluorosis, a common issue in Rajasthan due
704 to the dissolution of fluoride-bearing minerals (Singh et al. 2015; Nizam et al., 2022c). Elevated
705 alkalinity (ALK) is a critical concern with 81% of samples exceeding desirable limits, reflecting
706 the dominance of CO_3^- and HCO_3^- ions in the aquifers of Rajasthan. Additionally, HCO_3^-
707 concentrations were high, exceeding limits in 89% of samples, indicate groundwater-rock
708 interactions as a major contributing factor (Rahman et al., 2021).

709 The spatial distribution of the GW quality showed a notable link with regional
710 physiography, climate, natural (weathering) or anthropogenic (mining, fertilizers, sewage, and
711 industrial) sources (Jiang et al., 2021; Rahman et al., 2021; Naik et al., 2022). The predominance

1
2
3
4 712 of appropriate WQ in southeastern Rajasthan (Jhalawar, Pratapgarh, Kota) is due to the unique
5
6
7 713 geology (Deccan and Vindhyan), offering better aquifer recharge and less mineral concentration,
8
9 714 subsequently diluted by higher precipitation in these regions (CGWB, 2022; 2023). In contrast,
10
11 715 the Jalore region within this area witnessed localized differences in GW quality due to varied
12
13
14 716 geology (Singh and Kumar, 2014; Singh and Mukherjee, 2015). Although agriculture exists, its
15
16 717 influence is negligible when compared to arid regions, though localized contamination from runoff
17
18
19 718 is possible (Singh et al. 2014). In contrast, western Rajasthan, particularly Jodhpur and Nagaur
20
21 719 suffers from high salinity and F⁻ due to dry arid conditions and mineral-rich geology (Chaudhuri
22
23
24 720 et al. 2024, Singh and Mukherjee, 2015). Similar problem documented in Sikar and Udaipur cities
25
26 721 driven by Aravali's lithology (Narisetty et al. 2023, Chaudhuri et al. 2024). Scanty rainfall
27
28
29 722 combined with high evaporation concentrate salts across these regions. Additionally, over-
30
31 723 extraction for household and agriculture use, coupled with fertilizer/pesticide runoff and industrial
32
33
34 724 waste, further degrades GW quality sources (Jiang et al., 2021; Naik et al., 2022; Singh and Kumar,
35
36 725 2014). The seemingly fair GW quality aquifers of Jhunjhunu region is deceptive, as localized
37
38 726 severe pollution from Industrial effluents, particularly from mining contaminates specific areas.
39
40
41 727 This highlights WQ based specific strategies to be implemented in safeguarding the inhabitants
42
43 728 and ecosystem of the affected areas.

44
45
46 729 Analysis of modeled results shows that the RMS-WQI model effectively predicted WQI
47
48 730 scores, with high consistency between predicted and actual WQI ($R^2 = 0.99$), substantiating the
49
50
51 731 effective application of the RMS-WQI model in accurate WQ assessment. This high R^2 value is
52
53 732 notably higher than those reported in previous studies using similar WQI modeling approaches.
54
55
56 733 For instance, significantly lower R^2 (below 0.85) have been reported in arid aquifers studies in
57
58 734 Saudi Arabia and Iran (Masoud et al., 2022; Jalali et al., 2024). This highlights the enhanced

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

735 predictive capability of the RMS-WQI model. The significantly lower values of metrics such as
736 MSE, RMSE, PERI, and high NSE (few exceptional sites), further support the high accuracy and
737 reliability of the model's predicted WQI. These findings align with, and in some cases, exceeded
738 the accuracy observed in ML and statistical models based WQI prediction in different regions
739 Africa and India, suggesting that the RMS-WQI model offers a more robust and precise prediction
740 (El Bilali et al., 2020; Aoulmi et al., 2021; Ibrahim et al., 2023; Uddin 2023a). Overall, the model
741 performance deployed for different aquifers spatially located tens to thousands of km apart in this
742 study was excellent which further demonstrate robustness of RMS-WQI model in precise WQI
743 prediction with least uncertainty. Given the relative uniform climatic condition in the study region,
744 extending model performance validation to aquifers in diverse climatic settings is crucial.
745 Furthermore, translating the model's numerical ranking into categorical WQ classes would provide
746 valuable insights. As demonstrated by the data, a significant portion of the samples, both actual
747 and predicted, fall within both fair and marginal categories with a mean WQI of less than 60. This
748 indicates a moderate to high potential health risk for humans consuming water in the study region.

749 In contrast, uncertainty and sensitivity revealed significant eclipsing and ambiguity issues.
750 These issues may stem from inconsistencies in data collection or standardization, which affect the
751 precise calculation of a single score (Dawood et al., 2021; Wan et al., 2022; Gani et al., 2023;
752 Uddin et al., 2023c; Tripathy and Mishra, 2024). This issue of inconsistencies is not unique to this
753 study. Cho et al. (2020) recently highlighted potential sources of noise and bias in modeled WQ,
754 primarily due to variations in sampling protocols, analytical methods, and data reporting formats
755 across monitoring stations or time periods. Studies have also that the selection of weighting factors
756 and sub-index aggregation methods in WQI calculations can result in overlapping influences,
757 where certain parameters are masked, and ambiguity, where different parameter combinations

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

758 yield similar WQI scores (Akhtar et al 2021; Uddin 2022; Pham, 2020). Despite these limitations,
759 the models captured observed range of variation in WQ indicators. Notably, the overall
760 performance of our model aligns with the findings from previous studies, supporting the reliability
761 of our results. However, inherent model limitations, data quality, and the statistical approach of
762 data selection require further investigation to enhance model performance.

763 The combination of RMS-WQI model along with the state-of-the-art XGB model enhanced
764 the prediction of GW quality in the state of Rajasthan and promoted the use of ML/AI models for
765 future WQ prediction. The results also indicate that RMS-WQI model mitigates the model
766 uncertainty, eclipsing, and ambiguity problems, making it a reliable tool for assessing WQ status
767 in the study area. This improvement is likely due to RMS-WQI's ability to handle complex, non-
768 linear relationships within the data, a feature that has been validated by other studies using similar
769 robust statistical methods. AI-based ground WQ forecasting has demonstrated that these robust
770 methods can significantly reduce the impact of outliers and data inconsistencies, resulting in more
771 reliable model predictions (Weerts et al., 2011; Rahmati et al., 2019; Esen et al., 2024; Wang et
772 al., 2025). Additionally, the integration of the ML/AI model demonstrates its superior capability
773 over conventional methods in predicting WQ in the study area, making it suitable for WQ
774 monitoring and management. This aligns with the growing body of research supporting ML
775 applications in WQ forecasting. Ahmadi et al (2022) in their analysis of numerous studies,
776 concluded that ML models consistently outperform traditional statistical methods in terms of
777 prediction accuracy and ability to handle complex datasets. The effectiveness of the XGB model,
778 along with enhanced by the Optuna optimization technique, was also validated using advanced
779 statistical methods, inducing PDF, CDF, Turkey pair-wise ANOVA analysis, NSE, MEF, and
780 PREI test. These rigorous validation techniques are critical for ensuring applicability ML models

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

781 in the environmental field. Our findings offer valuable insights for WQ monitoring and
782 management initiatives in the water-scarce regions of western India. The model results can be used
783 to delineate potential GW health risk zones and inform targeted policy-making strategies for
784 suitable water supply and management practices. In summary, this research provides a valuable
785 understanding of the ML-based GW quality prediction in arid and semi-arid parts of India using
786 available GW datasets. These outcomes have crucial implications in resource and environmental
787 management, as they provide a foundation for data-driven decision-making and policy formulation
788 for sustainable resource distribution.

789 Given the increasing stress on groundwater resources in Rajasthan due to arid climatic
790 conditions and anthropogenic influences, a comprehensive policy framework is essential to ensure
791 sustainable water management. This study highlights the need for state-wide groundwater quality
792 monitoring using advanced machine learning-based RMS-WQI models to improve assessment
793 accuracy. Integrating AI-driven WQI models into government monitoring programs can aid real-
794 time decision-making and policy formulation. Additionally, strict regulations on excessive
795 groundwater extraction and contamination sources, such as unregulated irrigation return flow and
796 industrial discharge, should be enforced through mandatory WQ audits aligned with BIS standards.
797 Given the elevated Na^+ , Cl^- , SO_4^{2-} , and NO_3^- levels in certain districts, developing district-specific
798 mitigation strategies, including controlled fertilizer use and improved wastewater treatment, is
799 crucial. Finally, public awareness campaigns and community-driven groundwater monitoring
800 initiatives should be promoted to enhance stakeholder engagement, ensuring long-term
801 groundwater sustainability in Rajasthan while advancing India's commitment to SDG-6.

1
2
3
4 **802 7. Conclusions**

5
6
7 **803** This study examined groundwater (GW) quality in arid and semi-arid regions of Rajasthan
8
9 **804** (India) using the RMS-WQI model combined with ML algorithms to understand WQ patterns
10
11 **805** better. Major conclusions are given below.

- 12
13
14 **806** • GW exhibited a mean pH of 7.9, indicating slightly alkaline water suitable for multiple
15
16 **807** uses. However, 18% of samples exceeded permissible TH limits. This, coupled with
17
18 **808** elevated EC and TDS levels, poses challenges for health and industrial applications.
19
20
21 **809** Elevated Na^+ , Cl^- , and NO_3^- levels were linked to natural processes and anthropogenic
22
23 **810** activities like agriculture and poor waste disposal, impacting water salinity and human
24
25 **811** health.
- 26
27
28 **812** • The RMS-WQI model, validated through spatial analysis, showed marginal to poor WQ
29
30 **813** across the study area, with improved scores in districts like Banswara and Chittorgarh. The
31
32 **814** model's predictive capabilities were robust, demonstrated by low RMSE, MSE, and MAE
33
34 **815** values, and reduced PABE. Hyperparameter tuning using Optuna optimization further
35
36 **816** enhanced model performance, validated by HSD test.
- 37
38
39 **817** • Sensitivity analysis confirmed the model's robustness, demonstrating the highest goodness-
40
41 **818** of-fit. Uncertainty analysis revealed minimal discrepancies between real, predicted, and
42
43 **819** simulated data, with overall model uncertainty remaining below 2%. However, model
44
45 **820** eclipsing and ambiguity were observed in 25% and 54% of sampling sites, respectively,
46
47 **821** suggesting areas for improvement.
- 48
49 **822** • Efficiency analysis using NSE and MEF indicated a generally bias-free model with good
50
51 **823** performance, though some sites exhibited higher PREI scores, highlighting the need for
52
53 **824** multiple evaluation metrics.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

825 Future research should refine the RMS-WQI model to address model eclipsing and
826 ambiguity issues, integrate more WQ indicators, and use advanced ML techniques. Expanding
827 spatial analysis and incorporating real-time monitoring data will enhance model precision. Further
828 consideration should be given to the seasonal GW data to retain the model's performance in
829 capturing seasonal variation of GW quality. Notwithstanding these limitations, the configured
830 model could provide insightful information on improving the status of GW resources for arid and
831 semi-arid regions. Essentially, waste management, sustainable agricultural practices, and
832 policymaking are crucial for mitigating GW contamination and ensuring long-term water resource
833 sustainability.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

834 **Acknowledgement**

835 The authors gratefully acknowledge the editor's and anonymous reviewers' contributions to the
836 improvement of this paper. Additionally, the authors extend their gratitude to the Eco-
837 HydroInformatics Research Group (EHIRG) at the School of Engineering, College of Science and
838 Engineering, University of Galway, Ireland, for providing computational laboratory facilities
839 essential for completing this research.

840 **Data availability:** The analyzed data supporting the findings of this study are provided in the
841 manuscript and can also be obtained from the corresponding author upon request.

842 **Ethics approval:** Not applicable

843 **Consent to participate:** All authors duly participated.

844 **Consent for publication:** All authors hereby consent to publish this manuscript.

845 **Competing interests/Conflict of interest:** The authors declare no competing interests.

846 **Declaration of Competing Interest**

847 The authors declare that they have no known competing financial interests or personal
848 relationships that could have appeared to influence the work reported in this paper.

1
2
3
4 **849** **References**

- 5
6
7 **850** Ahmad, I., Khan, S.A., Shoeb, M., Islam, S., Khan, M.A., 2025. Hydrochemical facies distribution,
8 **851** controlling mechanisms and natural background concentrations of major pollutants in Ganga-
9 **852** Yamuna interfluvial aquifer, India. *Environmental Pollution* 368, 125694.
10 **853** doi:10.1016/j.envpol.2025.125694
11
12 **854** Ahmadi, A., Olyaei, M., Heydari, Z., Emami, M., Zeynolabedin, A., Ghomlaghi, A., Daccache, A., Fogg,
13 **855** G.E. and Sadegh, M., 2022. Groundwater level modeling with machine learning: a systematic
14 **856** review and meta-analysis. *Water*, 14(6), p.949.
15
16 **857** Akhtar, N., Ishak, M.I.S., Ahmad, M.I., Umar, K., Md Yusuff, M.S., Anees, M.T., Qadir, A. and Ali
17 **858** Almanasir, Y.K., 2021. Modification of the water quality index (WQI) process for simple
18 **859** calculation using the multi-criteria decision-making (MCDM) method: a review. *Water*, 13(7),
20 **860** p.905.
21
22 **861** Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter
23 **862** Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on*
24 **863** *Knowledge Discovery and Data Mining* 2623–2631. doi:10.1145/3292500.3330701
25
26 **864** Amaranto, A., Pianosi, F., Solomatine, D., Corzo, G., Muñoz-Arriola, F., 2020. Sensitivity analysis of data-
27 **865** driven groundwater forecasts to hydroclimatic controls in irrigated croplands. *Journal of Hydrology*
28 **866** 587, 124957. doi:10.1016/j.jhydrol.2020.124957
29
30 **867** Amiri, V., Rezaei, M., Sohrabi, N., 2014. Groundwater quality assessment using entropy weighted water
31 **868** quality index (EWQI) in Lenjanat, Iran. *Environmental Earth Sciences* 72, 3479–3490.
32 **869** doi:10.1007/s12665-014-3255-0
33
34 **870** Aoulmi, Y., Marouf, N., Amireche, M., Kisi, O., Shubair, R.M. and Keshtegar, B., 2021. Highly accurate
35 **871** prediction model for daily runoff in semi-arid basin exploiting metaheuristic learning algorithms.
36 **872** *Ieee Access*, 9, pp.92500-92515.
37
38 **873** APHA. (2017). *Standard methods for the examination of water and wastewater* (23rd ed.). American Public
39 **874** Health Association.
40
41 **875** Bahrami, A., Bahrami, M., Haghani, E., 2024. Groundwater quality assessment for potable using WQI and
42 **876** GIS technology in the south of Iran. *Sustainable Water Resources Management* 10, 177.
43 **877** doi:10.1007/s40899-024-01155-7
44
45 **878** Bamal, A., Uddin, M.G., Olbert, A.I., 2024. Harnessing Machine Learning for Assessing Climate Change
46 **879** Influences on Groundwater Resources: A Comprehensive Review. *Heliyon* 10, e37073.
47 **880** <https://doi.org/10.1016/j.heliyon.2024.e37073>
48
49 **881** BIS, 2012. *Indian Standards Specification for Drinking Water*. IS: 10500: 2012. 2nd Rev. New Delhi.
50 **882** Bureau of Indian Standard.
51
52 **883** Boo, K.B.W., El-Shafie, A., Othman, F., Khan, M.M.H., Birima, A.H., Ahmed, A.N., 2024. Groundwater
53 **884** level forecasting with machine learning models: A review. *Water Research* 252, 121249.
54 **885** doi:10.1016/j.watres.2024.121249
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4 886 Budholiya, K., Shrivastava, S.K., Sharma, V., 2022. An optimized XGBoost based diagnostic system for
5 887 effective prediction of heart disease. *Journal of King Saud University - Computer and Information*
6 888 *Sciences* 34, 4514–4523. doi:<https://doi.org/10.1016/j.jksuci.2020.10.013>
- 8 889 Carvalho, A., Costa, R., Neves, S., Oliveira, C.M., Bettencourt da Silva, R.J.N., 2021. Determination of
9 890 dissolved oxygen in water by the Winkler method: Performance modelling and optimisation for
10 891 environmental analysis. *Microchemical Journal* 165, 106129. doi:10.1016/j.microc.2021.106129
- 12 892 Census, 2011. Houselisting and housing census schedule, Government of India.
- 14 893 CGWB, 2022. Central Ground Water Board, Ground Water Year Book Rajasthan 2021 – 2022. Department
15 894 of Water Resources, River Development and Ganga Rejuvenation Ministry of Jal Shakti.
- 17 895 CGWB, 2023. Central Ground Water Board Report on Dynamic Ground Water Resources of Rajasthan
18 896 (As on 31st March, 2022)
- 20 897 Chaudhuri, R., Sahoo, S., Debsarkar, A. and Hazra, S., 2024. Fluoride contamination in groundwater—a
21 898 review. *Geospatial Practices in Natural Resources Management*, pp.331-354.
- 23 899 Che Nordin, N.F., Mohd, N.S., Koting, S., Ismail, Z., Sherif, M., El-Shafie, A., 2021. Groundwater quality
24 900 forecasting modelling using artificial intelligence: A review. *Groundwater for Sustainable*
25 901 *Development* 14, 100643. doi:10.1016/j.gsd.2021.100643
- 27 902 Chidiac, S., Najjar, P. El, Ouaini, N., Rayess, Y. El, Azzi, D. El, 2023. A comprehensive review of water
28 903 quality indices (WQIs): history, models, attempts and perspectives. *Reviews in Environmental*
29 904 *Science and Bio/Technology* 22, 349–395. doi:10.1007/s11157-023-09650-7
- 31 905 Cho, K.H., Pachepsky, Y., Ligaray, M., Kwon, Y. and Kim, K.H., 2020. Data assimilation in surface water
32 906 quality modeling: A review. *Water Research*, 186, p.116307.
- 34 907 Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G., Gharari, S., Freer,
35 908 J.E., Whitfield, P.H., Shook, K.R., Papalexiou, S.M., 2021. The Abuse of Popular Performance
36 909 Metrics in Hydrologic Modeling. *Water Resources Research* 57, 1–16.
37 910 doi:10.1029/2020WR029001
- 39 911 Dai, Z., Samper, J., 2004. Inverse problem of multicomponent reactive chemical transport in porous media:
40 912 Formulation and applications. *Water Resour. Res.* 40, 1–18.
41 913 <https://doi.org/10.1029/2004WR003248>
- 43 914 Dai, Z., Wolfsberg, A., Lu, Z., Ritzi, R., 2007. Representing aquifer architecture in macrodispersivity
44 915 models with an analytical solution of the transition probability matrix. *Geophys. Res. Lett.* 34, 1–
45 916 6. <https://doi.org/10.1029/2007GL031608>
- 47 917 Dawood, T., Elwakil, E., Novoa, H.M., Gárate Delgado, J.F., 2021. Toward urban sustainability and clean
48 918 potable water: Prediction of water quality via artificial neural networks. *Journal of Cleaner*
49 919 *Production* 291, 125266. doi:10.1016/j.jclepro.2020.125266
- 51 920 Dimple, Singh, P.K., Rajput, J., Kumar, D., Gaddikeri, V., Elbeltagi, A., 2023. Combination of
52 921 discretization regression with data-driven algorithms for modeling irrigation water quality indices.
53 922 *Ecological Informatics* 75, 102093. doi:10.1016/j.ecoinf.2023.102093
- 54
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4 923 Ding, F., Hao, S., Zhang, W., Jiang, M., Chen, L., Yuan, H., Wang, N., Li, W., Xie, X., 2025. Using multiple
5 924 machine learning algorithms to optimize the water quality index model and their applicability. *Ecol.*
6 925 *Indic.* 172, 113299. <https://doi.org/10.1016/j.ecolind.2025.113299>
- 8
9 926 El Bilali, A. and Taleb, A., 2020. Prediction of irrigation water quality parameters using machine learning
10 927 models in a semi-arid environment. *Journal of the Saudi Society of Agricultural Sciences*, 19(7),
11 928 pp.439-451.
- 13 929 El Bilali, A., Abdeslam, T., Ayoub, N., Lamane, H., Ezzaouini, M.A., Elbeltagi, A., 2023. An interpretable
14 930 machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting
15 931 daily pan evaporation. *J. Environ. Manage.* 327, 116890.
16 932 <https://doi.org/https://doi.org/10.1016/j.jenvman.2022.116890>
- 18 933 Elzain, H.E., Chung, S.Y., Venkatramanan, S., Selvam, S., Ahemd, H.A., Seo, Y.K., Bhuyan, M.S., Yassin,
19 934 M.A., 2023. Novel machine learning algorithms to predict the groundwater vulnerability index to
20 935 nitrate pollution at two levels of modeling. *Chemosphere* 314.
21 936 doi:10.1016/j.chemosphere.2022.137671
- 24 937 Esen, Ö., Yıldırım, D.Ç. and Yıldırım, S., 2024. A quantile regression approach to assess the impact of
25 938 water-related environmental innovations on water stress. *Technological Forecasting and Social*
26 939 *Change*, 203, p.123343.
- 28 940 Esmailbeiki, F., Nikpour, M.R., Singh, V.K., Kisi, O., Sihag, P., Sanikhani, H., 2020. Exploring the
29 941 application of soft computing techniques for spatial evaluation of groundwater quality variables.
30 942 *Journal of Cleaner Production* 276, 124206. doi:10.1016/j.jclepro.2020.124206
- 32 943 Faruq, O., Malak, M.A., Hossain, N.J., Sami, M.S., Sajib, A.M., 2025. Investigating the relationship
34 944 between land use and water quality in urban water bodies. *Clean. Water* 3, 100070.
35 945 <https://doi.org/10.1016/j.clwat.2025.100070>
- 37 946 Gani, M.A., Sajib, A.M., Siddik, M.A., Md Moniruzzaman, 2023. Assessing the impact of land use and
38 947 land cover on river water quality using water quality index and remote sensing techniques.
39 948 *Environmental Monitoring and Assessment* 195, 449. doi:10.1007/s10661-023-10989-1
- 41 949 Haggerty, R., Sun, J., Yu, H., Li, Y., 2023. Application of machine learning in groundwater quality
42 950 modeling - A comprehensive review. *Water Research* 233, 119745.
43 951 doi:10.1016/j.watres.2023.119745
- 45 952 Haidery, A., Umar, R., Khan, I., 2024. Seasonal variation and spatial distribution of heavy metal (loid)s
46 953 concentration in groundwater and surface water from hard-rock terrain, Ranchi, India,
47 954 *Environment, Development and Sustainability*. Springer Netherlands. doi:10.1007/s10668-024-
48 955 04658-7
- 50
51 956 Horton, R.K., 1965. An index number system for rating water quality. *Journal of Water Pollution Control*
52 957 *Federation* 37, 300–306.
- 54 958 Ibrahim Ahmed Osman, A., Najah Ahmed, A., Chow, M.F., Feng Huang, Y., El-Shafie, A., 2021. Extreme
55 959 gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain*
56 960 *Shams Engineering Journal* 12, 1545–1556. doi:10.1016/j.asej.2020.11.011
- 58 961 Ibrahim, H., Yaseen, Z.M., Scholz, M., Ali, M., Gad, M., Elsayed, S., Khadr, M., Hussein, H., Ibrahim,
59 962 H.H., Eid, M.H. and Kovács, A., 2023. Evaluation and prediction of groundwater quality for

1
2
3
4 963 irrigation using an integrated water quality indices, machine learning models and GIS approaches:
5 964 A representative case study. *Water*, 15(4), p.694.

7 965 Jahan, A., Rai, N. and Khan, M.U., 2025. Groundwater Contamination Characterization and Source
8 966 Apportionment of Heavy Metals and Associated Source-Specific Health Risk Appraisal Using
9 967 Monte Carlo Simulation Coupled with PCA-MLR and PMF Models in the Middle Gangetic Basin,
10 968 India. *Exposure and Health*, pp.1-32. <https://doi.org/10.1007/s12403-025-00693-5>

13 969 Jalali, M., Jalali, M. and Morrison, L., 2024. Groundwater hydrogeochemical processes, water quality
14 970 index, and probabilistic health risk assessment in an arid and semi-arid environment (Hamedan,
15 971 Iran). *Groundwater for Sustainable Development*, 26, p.101255.

17 972 Jalan, S., Chouhan, D.S., Chaure, S., Vyas, A., 2023. Ground Water Quality and Its Impact on Human
18 973 Health in Dungarpur District of Rajasthan, India. *Int. Arch. Photogramm. Remote Sens. Spat. Inf.*
19 974 *Sci. - ISPRS Arch.* 48, 421–428. [https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-421-](https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-421-2023)
21 975 [2023](https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-421-2023)

23 976 Jiang, C., Zhao, Q., Zheng, L., Chen, X., Li, C., Ren, M., 2021. Distribution, source and health risk
24 977 assessment based on the Monte Carlo method of heavy metals in shallow groundwater in an area
25 978 affected by mining activities, China. *Ecotoxicology and Environmental Safety* 224.
26 979 doi:10.1016/j.ecoenv.2021.112679

28 980 Khan, I., Ayaz, M., 2024. Sensitivity analysis-driven machine learning approach for groundwater quality
29 981 prediction: Insights from integrating ENTROPY and CRITIC methods. *Groundwater for*
30 982 *Sustainable Development* 26, 101309. doi:10.1016/j.gsd.2024.101309

32 983 Khan, I., Umar, R., 2019. Environmental risk assessment of coal fly ash on soil and groundwater quality,
33 984 Aligarh, India. *Groundwater for Sustainable Development* 8, 346–357.
34 985 doi:10.1016/j.gsd.2018.12.002

36 986 Khan, I., Umar, R., 2024a. Improving evaluation of groundwater heavy metal(loid)s pollution efficiencies:
37 987 Insights from novel Shannon entropy-weight and one-way ANOVA analysis. *Groundwater for*
38 988 *Sustainable Development* 24, 101052. doi:10.1016/j.gsd.2023.101052

41 989 Khan, I., Umar, R., 2024b. Machine Learning-driven Optimization of Water Quality Index: A Synergistic
42 990 ENTROPY-CRITIC Approach Using Spatio-Temporal Data. *Earth Systems and Environment* 8,
43 991 1453–1475. doi:10.1007/s41748-024-00500-2

45 992 Khosravi, K., Rezaie, F., Cooper, J.R., Kalantari, Z., Abolfathi, S., Hatamiafkoueih, J., 2023. Soil water
46 993 erosion susceptibility assessment using deep learning algorithms. *Journal of Hydrology* 618,
47 994 129229. doi:10.1016/j.jhydrol.2023.129229

49 995 Khouri, L., Bashar Al-Moufti, M., 2022. Selection of suitable aggregation function for estimation of water
50 996 quality index for the Orontes River. *Ecological Indicators* 142, 109290.
51 997 doi:10.1016/j.ecolind.2022.109290

53 998 Krishnaraj, A., Honnasiddaiah, R., 2022. Remote sensing and machine learning based framework for the
54 999 assessment of spatio-temporal water quality in the Middle Ganga Basin. *Environmental Science*
55 1000 *and Pollution Research* 64939–64958. doi:10.1007/s11356-022-20386-9

- 1
2
3
4 1001 Kumar, M., Mishra, G. V., 2024. Causes and Impacts of Water Pollution on Various water bodies in the
5 1002 State of Rajasthan, India: A Review. *Environ. Ecol.* 42, 645–654.
6 1003 <https://doi.org/10.60151/envec/qiyj5706>
7
8 1004 Kumar, M., Panday, D.P., Bhagat, C., Herbha, N., Agarwal, V., 2023. Demystifying the decadal shift in the
9 1005 extent of groundwater in the coastal aquifers of Gujarat, India: A case of reduced extent but
10 1006 increased magnitude of seawater intrusion. *Science of the Total Environment* 898.
11 1007 doi:10.1016/j.scitotenv.2023.165451
12
13 1008 Lap, B.Q., Phan, T.-T.-H., Nguyen, H. Du, Quang, L.X., Hang, P.T., Phi, N.Q., Hoang, V.T., Linh, P.G.,
14 1009 Hang, B.T.T., 2023. Predicting Water Quality Index (WQI) by feature selection and machine
15 1010 learning: A case study of An Kim Hai irrigation system. *Ecological Informatics* 74, 101991.
16 1011 doi:<https://doi.org/10.1016/j.ecoinf.2023.101991>
17
18 1012 Liu, Q., Gui, D., Zhang, L., Niu, J., Dai, H., Wei, G., Hu, B.X., 2022. Simulation of regional groundwater
19 1013 levels in arid regions using interpretable machine learning models. *Sci. Total Environ.* 831.
20 1014 <https://doi.org/10.1016/j.scitotenv.2022.154902>
21
22 1015 Liu, R., Li, G., Wei, L., Xu, Y., Gou, X., Luo, S., Yang, X., 2022. Spatial prediction of groundwater
23 1016 potentiality using machine learning methods with Grey Wolf and Sparrow Search Algorithms.
24 1017 *Journal of Hydrology* 610, 127977. doi:10.1016/j.jhydrol.2022.127977
25
26 1018 Manzar, M.S., Benaafi, M., Costache, R., Alagha, O., Mu'azu, N.D., Zubair, M., Abdullahi, J., Abba, S.I.,
27 1019 2022. New generation neurocomputing learning coupled with a hybrid neuro-fuzzy model for
28 1020 quantifying water quality index variable: A case study from Saudi Arabia. *Ecological Informatics*
29 1021 70, 101696. doi:10.1016/j.ecoinf.2022.101696
30
31 1022 Masoud, M., El Osta, M., Alqarawy, A., Elsayed, S. and Gad, M., 2022. Evaluation of groundwater quality
32 1023 for agricultural under different conditions using water quality indices, partial least squares
33 1024 regression models, and GIS approaches. *Appl Water Sci* 12: 244 [online]
34
35 1025 McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. *Journal*
36 1026 *of Hydrologic Engineering* 11, 597–602. doi:10.1061/(ASCE)1084-0699(2006)11:6(597)
37
38 1027 Merchán, D., Sanz, L., Alfaro, A., Pérez, I., Goñi, M., Solsona, F., Hernández-García, I., Pérez, C., Casali,
39 1028 J., 2020. Irrigation implementation promotes increases in salinity and nitrate concentration in the
40 1029 lower reaches of the Cidacos River (Navarre, Spain). *Science of The Total Environment* 706,
41 1030 135701. doi:10.1016/j.scitotenv.2019.135701
42
43 1031 Mohammadpour, A., Gharehchahi, E., Gharaghani, M.A., Shahsavani, E., Golaki, M., Berndtsson, R.,
44 1032 Khaneghah, A.M., Hashemi, H., Abolfathi, S., 2024. Assessment of drinking water quality and
45 1033 identifying pollution sources in a chromite mining region. *J. Hazard. Mater.* 480, 136050.
46 1034 <https://doi.org/10.1016/j.jhazmat.2024.136050>
47
48 1035 Mukate, S., Wagh, V., Panaskar, D., Jacobs, J.A., Sawant, A., 2019. Development of new integrated water
49 1036 quality index (IWQI) model to evaluate the drinking suitability of water. *Ecological Indicators* 101,
50 1037 348–354. doi:10.1016/j.ecolind.2019.01.034
51
52 1038 Muniz, D.H.F., Malaquias, J. V., Lima, J.E.F.W., Oliveira-Filho, E.C., 2020. Proposal of an irrigation water
53 1039 quality index (IWQI) for regional use in the Federal District, Brazil. *Environmental Monitoring*
54 1040 *and Assessment* 192. doi:10.1007/s10661-020-08573-y
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 1041 Naik, M.R., Mahanty, B., Sahoo, S.K., Jha, V.N., Sahoo, N.K., 2022. Assessment of groundwater
5 1042 geochemistry using multivariate water quality index and potential health risk in industrial belt of
6 1043 central Odisha, India. *Environmental Pollution* 303. doi:10.1016/j.envpol.2022.119161
7
8 1044 Nallakaruppan, M.K., Gangadevi, E., Shri, M.L., Balusamy, B., Bhattacharya, S., Selvarajan, S., 2024.
9 1045 Reliable water quality prediction and parametric analysis using explainable AI models. *Sci. Rep.*
10 1046 14, 1–24. <https://doi.org/10.1038/s41598-024-56775-y>
11
12 1047 Narisetty, N.G., Tripathi, G., Kanga, S., Singh, S.K., Meraj, G., Kumar, P., Đurin, B. and Matijević, H.,
13 1048 2023. Integrated multi-model approach for assessing groundwater vulnerability in Rajasthan’s
14 1049 semi-arid zone: incorporating DRASTIC and SINTACS variants. *Hydrology*, 10(12), p.231.
15
16 1050 Nath, S., Mathew, A., Khandelwal, S., Shekar, P.R., 2023. Rainfall and temperature dynamics in four Indian
17 1051 states: A comprehensive spatial and temporal trend analysis. *HydroResearch* 6, 247–254.
18 1052 <https://doi.org/https://doi.org/10.1016/j.hydres.2023.09.001>
19
20 1053 Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., Righetti, M., 2024.
21 1054 Applications of XGBoost in water resources engineering: A systematic literature review (Dec
22 1055 2018–May 2023). *Environmental Modelling and Software* 174, 105971.
23 1056 doi:10.1016/j.envsoft.2024.105971
24
25 1057 Nisa, F.U., Umar, R., 2023. Evaluation of physicochemical and microbiological parameters, and their
26 1058 correlation in Himalayan Spring Water Systems: A case study of District Kulgam of Kashmir
27 1059 Valley, India, Western Himalaya. *Environmental Monitoring and Assessment* 195, 441.
28 1060 doi:10.1007/s10661-023-11025-y
29
30 1061 Nizam, S., Acharya, T., Dutta, S., Sen, I.S., 2022c. Occurrence, sources, and spatial distribution of fluoride
31 1062 in the Ganga alluvial aquifer, India. *Environmental Geochemistry and Health* 1975–1989.
32 1063 doi:10.1007/s10653-022-01319-4
33
34 1064 Nizam, S., Dutta, S., Sen, I.S., 2022a. Geogenic controls on the high levels of uranium in alluvial aquifers
35 1065 of the Ganga Basin. *Applied Geochemistry* 143, 105374. doi:10.1016/j.apgeochem.2022.105374
36
37 1066 Nizam, S., Virk, H.S., Sen, I.S., 2022b. High levels of fluoride in groundwater from Northern parts of Indo-
38 1067 Gangetic plains reveals detrimental fluorosis health risks. *Environmental Advances* 8, 100200.
39 1068 doi:10.1016/j.envadv.2022.100200
40
41 1069 Nong, X., He, F., Chen, L., Wei, J., 2025. Integrated machine learning-based optimization framework for
42 1070 surface water quality index comparing coastal and non-coastal cases of Guangxi, China. *Mar.*
43 1071 *Pollut. Bull.* 213, 117564. <https://doi.org/10.1016/j.marpolbul.2025.117564>
44
45 1072 Palar, P.S., Zuhail, L.R., Shimoyama, K., 2023. Enhancing the explainability of regression-based
46 1073 polynomial chaos expansion by Shapley additive explanations. *Reliab. Eng. Syst. Saf.* 232, 109045.
47 1074 <https://doi.org/10.1016/j.ress.2022.109045>
48
49 1075 Pandey, D.K., Bahadur, T., 2009. A review of the stratigraphy of Marwar Supergroup of west-central
50 1076 Rajasthan. *Journal of the Geological Society of India* 73, 747–758. doi:10.1007/s12594-009-0060-
51 1077 6
52
53 1078 Pandey, S., Mohapatra, G., Arora, R., 2023. Groundwater quality, human health risks and major driving
54 1079 factors in arid and semi-arid regions of Rajasthan, India. *Journal of Cleaner Production* 427.
55 1080 doi:10.1016/j.jclepro.2023.139149
56
57
58
59
60
61
62
63
64
65

1
2
3
4 1081 Parween, S., Siddique, N.A., Mahammad Diganta, M.T., Olbert, A.I., Uddin, M.G., 2022. Assessment of
5 1082 urban river water quality using modified NSF water quality index model at Siliguri city, West
6 1083 Bengal, India. *Environmental and Sustainability Indicators* 16, 100202.
7 1084 doi:10.1016/j.indic.2022.100202
8
9
10 1085 Patel, P.S., Pandya, D.M., Shah, M., 2022. A review on various mathematical techniques for groundwater
11 1086 quality analysis and assessment. *Materials Today: Proceedings* 8–11.
12 1087 doi:10.1016/j.matpr.2022.08.456
13
14 1088 Pham, H.N., 2020. Relative water quality index (ReWQI)—a new method for aggregate water quality
15 1089 assessment. *Water and Environment Journal*, 34, pp.873-883.
16
17 1090 Rahman, A., Mondal, N.C., Tiwari, K.K., 2021. Anthropogenic nitrate in groundwater and its health risks
18 1091 in the view of background concentration in a semi arid area of Rajasthan, India. *Scientific Reports*
19 1092 11, 1–13. doi:10.1038/s41598-021-88600-1
20
21 1093 Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher,
22 1094 J., Cipullo, S., Ahmad, B.B. and Bui, D.T., 2019. Predicting uncertainty of machine learning
23 1095 models for modelling nitrate pollution of groundwater using quantile regression and UNEEC
24 1096 methods. *Science of the Total Environment*, 688, pp.855-866.
25
26
27 1097 Rajkumar, H., Naik, P.K., Rishi, M.S., 2020. A new indexing approach for evaluating heavy metal
28 1098 contamination in groundwater. *Chemosphere* 245. doi:10.1016/j.chemosphere.2019.125598
29
30 1099 Rajkumar, H., Naik, P.K., Rishi, M.S., 2022. A comprehensive water quality index based on analytical
31 1100 hierarchy process. *Ecological Indicators* 145. doi:10.1016/j.ecolind.2022.109582
32
33 1101 Rasool, U., Yin, X., Xu, Z., Rasool, M.A., Senapathi, V., Hussain, M., Siddique, J., Trabucco, J.C., 2022.
34 1102 Mapping of groundwater productivity potential with machine learning algorithms: A case study in
35 1103 the provincial capital of Baluchistan, Pakistan. *Chemosphere* 303, 135265.
36 1104 doi:10.1016/j.chemosphere.2022.135265
37
38
39 1105 Roy, A.B., Jakhar, S.R., 2002. *Geology of Rajasthan (Northwest India) precambrian to recent*. Scientific
40 1106 Publishers.
41
42 1107 Sadeghi-Lari, A., Bahrami, M., Dastandaz, T., 2024. Temporal and spatial variations of groundwater
43 1108 quantity and quality for drinking and irrigation purposes in the arid and hot weather of Southern
44 1109 Iran. *Physics and Chemistry of the Earth, Parts A/B/C* 134, 103582. doi:10.1016/j.pce.2024.103582
45
46 1110 Sahour, H., Gholami, V., Vazifedan, M., 2020. A comparative analysis of statistical and machine learning
47 1111 techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *Journal*
48 1112 *of Hydrology* 591, 125321. doi:https://doi.org/10.1016/j.jhydrol.2020.125321
49
50
51 1113 Sajib, A.M., Bamal, A., Diganta, M.T.M., S.M. Ashekuzzaman, Rahman, A., Olbert, A.I., Uddin, M.G.,
52 1114 2025. Novel groundwater quality index (GWQI) model : A reliable approach for the assessment of
53 1115 groundwater. *Results Eng.* 25, 104265. https://doi.org/10.1016/j.rineng.2025.104265
54
55 1116 Sajib, A.M., Diganta, M.T.M., Moniruzzaman, M., Rahman, A., Dabrowski, T., Uddin, M.G., Olbert, A.I.,
56 1117 2024. Assessing water quality of an ecologically critical urban canal incorporating machine
57 1118 learning approaches. *Ecological Informatics* 80, 102514. doi:10.1016/j.ecoinf.2024.102514
58
59
60
61
62
63
64
65

1
2
3
4 1119 Sajib, A.M., Diganta, M.T.M., Rahman, A., Dabrowski, T., Olbert, A.I., Uddin, M.G., 2023. Developing a
5 1120 novel tool for assessing the groundwater incorporating water quality index and machine learning
6 1121 approach. *Groundwater for Sustainable Development* 23, 101049. doi:10.1016/j.gsd.2023.101049
7
8 1122 Samani, S., 2021. Assessment of groundwater sustainability and management plan formulations through
9 1123 the integration of hydrogeological, environmental, social, economic and policy indices.
10 1124 *Groundwater for Sustainable Development* 15, 100681. doi:10.1016/j.gsd.2021.100681
11
12 1125 Seifi, A., Dehghani, M., Singh, V.P., 2020. Uncertainty analysis of water quality index (WQI) for
13 1126 groundwater quality evaluation: Application of Monte-Carlo method for weight allocation.
14 1127 *Ecological Indicators* 117, 106653. doi:https://doi.org/10.1016/j.ecolind.2020.106653
15
16 1128 Sharif, O., Hasan, M.Z., Rahman, A., 2022. Determining an effective short term COVID-19 prediction
17 1129 model in ASEAN countries. *Scientific Reports* 12, 5083. doi:10.1038/s41598-022-08486-5
18
19 1130 Singh, C.K., Mukherjee, S., 2015. Aqueous geochemistry of fluoride enriched groundwater in arid part of
20 1131 Western India. *Environ Sci Pollut Res* 22, 2668–2678 <https://doi.org/10.1007/s11356-014-3504-5>
21
22 1132 Singh, R.B. and Kumar, A., 2014. Vulnerability of Agriculture to Climate Change in Arid Regions: A Case
23 1133 Study of Western Rajasthan, India. *Vulnerability of Land systems in Asia*, pp.77-90.
24
25 1134 Singha, S., Pasupuleti, S., Singha, S.S., Singh, R., Kumar, S., 2021. Prediction of groundwater quality using
26 1135 efficient machine learning technique. *Chemosphere* 276, 130265.
27 1136 doi:10.1016/j.chemosphere.2021.130265
28
29 1137 Solanki, A., Agrawal, H., Khare, K., 2015. Predictive Analysis of Water Quality Parameters using Deep
30 1138 Learning. *International Journal of Computer Applications* 125, 29–34.
31 1139 doi:10.5120/ijca2015905874
32
33 1140 Srinivas, R., Bhakar, P., Singh, A.P., 2015. Groundwater Quality Assessment in Some Selected Area of
34 1141 Rajasthan, India Using Fuzzy Multi-criteria Decision Making Tool. *Aquatic Procedia* 4, 1023–
35 1142 1030. doi:10.1016/j.aqpro.2015.02.129
36
37 1143 Sunita, Ghosh, T., 2024. Groundwater hydro-geochemical inferences and eXplainable Artificial
38 1144 Intelligence augmented groundwater quality prediction in arid and semi-arid segment of Rajasthan,
39 1145 India. *Groundw. Sustain. Dev.* 26, 101272. <https://doi.org/10.1016/j.gsd.2024.101272>
40
41 1146 Talukdar, S., Shahfahad, Ahmed, S., Naikoo, M.W., Rahman, A., Mallik, S., Ningthoujam, S., Bera, S.,
42 1147 Ramana, G. V., 2023. Predicting lake water quality index with sensitivity-uncertainty analysis using
43 1148 deep learning algorithms. *Journal of Cleaner Production* 406, 136885.
44 1149 doi:https://doi.org/10.1016/j.jclepro.2023.136885
45
46 1150 Tripathy, K.P., Mishra, A.K., 2024. Deep learning in hydrology and water resources disciplines: concepts,
47 1151 methods, applications, and research directions. *Journal of Hydrology* 628, 130458.
48 1152 doi:10.1016/j.jhydrol.2023.130458
49
50 1153 Uddin, M.G., Imran, M.H., Sajib, A.M., Hasan, M.A., Diganta, M.T.M., Dabrowski, T., Olbert, A.I.,
51 1154 Moniruzzaman, M., 2024a. Assessment of human health risk from potentially toxic elements and
52 1155 predicting groundwater contamination using machine learning approaches. *Journal of Contaminant*
53 1156 *Hydrology* 261. doi:10.1016/j.jconhyd.2024.104307
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 1157 Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022. A comprehensive method for improvement of water
5 1158 quality index (WQI) models for coastal water quality assessment. *Water Research* 219, 118532.
6 1159 doi:10.1016/j.watres.2022.118532
7
8
9 1160 Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023a. A novel approach for estimating and predicting
10 1161 uncertainty in water quality index model using machine learning approaches. *Water Research* 229,
11 1162 119422. doi:10.1016/j.watres.2022.119422
12
13 1163 Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023b. Assessing optimization techniques for improving
14 1164 water quality model. *Journal of Cleaner Production* 385, 135671.
15 1165 doi:https://doi.org/10.1016/j.jclepro.2022.135671
16
17 1166 Uddin, M.G., Rana, M.M.S.P., Diganta, M.T.M., Bamal, A., Sajib, A.M., Abioui, M., Shaibur, M.R.,
18 1167 Ashekuzzaman, S.M., Nikoo, M.R., Rahman, A., Moniruzzaman, M., Olbert, A.I., 2024b.
19 1168 Enhancing groundwater quality assessment in coastal area: A hybrid modeling approach. *Heliyon*
20 1169 10, e33082. doi:10.1016/j.heliyon.2024.e33082
21
22
23 1170 Wan, H., Xu, R., Zhang, M., Cai, Y., Li, J., Shen, X., 2022. A novel model for water quality prediction
24 1171 caused by non-point sources pollution based on deep learning and feature extraction methods.
25 1172 *Journal of Hydrology* 612, 128081. doi:10.1016/j.jhydrol.2022.128081
26
27 1173 Wang, S., Cao, W., Hu, X., Zhong, H. and Sun, W., 2025. A Selective Overview of Quantile Regression
28 1174 for Large-Scale Data. *Mathematics*, 13(5), p.837.
29
30 1175 Wang, S., Peng, H., Hu, Q., Jiang, M., 2022. Analysis of runoff generation driving factors based on
31 1176 hydrological model and interpretable machine learning method. *J. Hydrol. Reg. Stud.* 42, 101139.
32 1177 <https://doi.org/10.1016/j.ejrh.2022.101139>
33
34 1178 Wang, X., Zheng, W., Tian, W., Gao, Y., Wang, X., Tian, Y., Li, J., Zhang, X., 2022. Groundwater
35 1179 hydrogeochemical characterization and quality assessment based on integrated weight matter-element
36 1180 extension analysis in Ningxia, upper Yellow River, northwest China. *Ecological Indicators* 135.
37 1181 doi:10.1016/j.ecolind.2021.108525
38
39
40 1182 Weerts, A.H., Winsemius, H.C. and Verkade, J.S., 2011. Estimation of predictive hydrological uncertainty
41 1183 using quantile regression: examples from the National Flood Forecasting System (England and
42 1184 Wales). *Hydrology and Earth System Sciences*, 15(1), pp.255-265.
43
44
45 1185 WHO, 2017. *Guidelines for Drinking-water Quality*, Fourth edi. ed. World Health Organization.
46 1186 doi:https://www.who.int/publications/i/item/9789241549950
47
48
49 1187 Wu, J., Zhang, Y., Zhou, H., 2020. Groundwater chemistry and groundwater quality index incorporating
50 1188 health risk weighting in Dingbian County, Ordos basin of northwest China. *Chemie der Erde* 80,
51 1189 125607. doi:10.1016/j.chemer.2020.125607
52
53 1190 Xiong, H., Guo, X., Wang, Y., Xiong, R., Gui, X., Hu, X., Li, Y., Qiu, Y., Tan, J., Ma, C., 2023. Spatial
54 1191 prediction of groundwater potential by various novel boosting-based ensemble learning models in
55 1192 mountainous areas. *Geocarto International* 38. doi:10.1080/10106049.2023.2274870
56
57 1193 Xiong, Y., Luo, J., Liu, X., Liu, Y., Xin, X., Wang, S., 2022. Machine learning-based optimal design of
58 1194 groundwater pollution monitoring network. *Environmental Research* 211, 113022.
59 1195 doi:10.1016/j.envres.2022.113022
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Zhang, K., Wang, X., Liu, T., Wei, W., Zhang, F., Huang, M., Liu, H., 2024. Enhancing water quality prediction with advanced machine learning techniques: An extreme gradient boosting model based on long short-term memory and autoencoder. *J. Hydrol.* 644, 132115. <https://doi.org/10.1016/j.jhydrol.2024.132115>

Zhang, Q., Xu, P., Qian, H., 2020. Groundwater Quality Assessment Using Improved Water Quality Index (WQI) and Human Health Risk (HHR) Evaluation in a Semi-arid Region of Northwest China. *Exposure and Health* 12, 487–500. doi:10.1007/s12403-020-00345-w

Zhang, X., Zhao, R., Wu, X., Mu, W., Wu, C., 2022. Delineating the controlling mechanisms of arsenic release into groundwater and its associated health risks in the Southern Loess Plateau, China. *Water Research* 219, 118530. doi:10.1016/j.watres.2022.118530

Zhu, Y., Dai, H., Yuan, S., 2023. The competition between heterotrophic denitrification and DNRA pathways in hyporheic zone and its impact on the fate of nitrate. *J. Hydrol.* 626, 130175. <https://doi.org/10.1016/j.jhydrol.2023.130175>