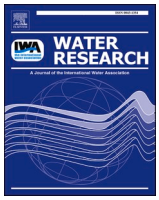
Water Research 255 (2024) 121499

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres



Data-driven evolution of water quality models: An in-depth investigation of innovative outlier detection approaches-A case study of Irish Water Quality Index (IEWQI) model

Md Galal Uddin [a,b,c,d,*], Azizur Rahman [e,f], Firouzeh Rosa Taghikhah [g], Agnieszka I. Olbert [a,b,c,d]

[a] School of Engineering, University of Galway, Ireland
[b] Ryan Institute, University of Galway, Ireland
[c] MaREI Research Centre, University of Galway, Ireland
[d] Eco-HydroInformatics Research Group (EHIRG), Civil Engineering, National University of Ireland Galway, Ireland
[e] School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga, Australia
[f] The Gulbali Institute of Agriculture, Water and Environment, Charles Sturt University, Wagga, Australia
[g] Business School, University of Sydney, Camperdown 2050, NSW, Australia

ARTICLE INFO

Keywords:
Data-driven models
Water quality models
IEWQI model
Data outliers
Machine learning

ABSTRACT

Recently, there has been a significant advancement in the water quality index (WQI) models utilizing data-driven approaches, especially those integrating machine learning and artificial intelligence (ML/AI) technology. Although, several recent studies have revealed that the data-driven model has produced inconsistent results due to the data outliers, which significantly impact model reliability and accuracy. The present study was carried out to assess the impact of data outliers on a recently developed Irish Water Quality Index (IEWQI) model, which relies on data-driven techniques. To the author's best knowledge, there has been no systematic framework for evaluating the influence of data outliers on such models. For the purposes of assessing the outlier impact of the data outliers on the water quality (WQ) model, this was the first initiative in research to introduce a comprehensive approach that combines machine learning with advanced statistical techniques. The proposed framework was implemented in Cork Harbour, Ireland, to evaluate the IEWQI model's sensitivity to outliers in input indicators to assess the water quality. In order to detect the data outlier, the study utilized two widely used ML techniques, including Isolation Forest (IF) and Kernel Density Estimation (KDE) within the dataset, for predicting WQ with and without these outliers. For validating the ML results, the study used five commonly used statistical measures.

The performance metric (R^2) indicates that the model performance improved slightly (R^2 increased from 0.92 to 0.95) in predicting WQ after removing the data outlier from the input. But the IEWQI scores revealed that there were no statistically significant differences among the actual values, predictions with outliers, and predictions without outliers, with a 95 % confidence interval at $p < 0.05$. The results of model uncertainty also revealed that the model contributed <1 % uncertainty to the final assessment results for using both datasets (with and without outliers). In addition, all statistical measures indicated that the ML techniques provided reliable results that can be utilized for detecting outliers and their impacts on the IEWQI model. The findings of the research reveal that although the data outliers had no significant impact on the IEWQI model architecture, they had moderate impacts on the rating schemes' of the model. This finding indicated that detecting the data outliers could improve the accuracy of the IEWQI model in rating WQ as well as be helpful in mitigating the model eclipsing problem. In addition, the results of the research provide evidence of how the data outliers influenced the data-driven model in predicting WQ and reliability, particularly since the study confirmed that the IEWQI model's could be effective for accurately rating WQ despite the presence of the data outliers in the input. It could occur due to the spatio-temporal variability inherent in WQ indicators.

However, the research assesses the influence of data input outliers on the IEWQI model and underscores important areas for future investigation. These areas include expanding temporal analysis using multi-year data, examining spatial outlier patterns, and evaluating detection methods. Moreover, it is essential to explore the real-

* Corresponding author at: Civil Engineering, College of Science and Engineering, University of Galway, Ireland.
E-mail address: mdgalal.uddin@universityofgalway.ie (M.G. Uddin).

https://doi.org/10.1016/j.watres.2024.121499
Received 27 September 2023; Received in revised form 9 February 2024; Accepted 19 March 2024
Available online 20 March 2024

world impacts of revised rating categories, involve stakeholders in outlier management, and fine-tune model parameters. Analysing model performance across varying temporal and spatial resolutions and incorporating additional environmental data can significantly enhance the accuracy of WQ assessment. Consequently, this study offers valuable insights to strengthen the IEWQI model's robustness and provides avenues for enhancing its utility in broader WQ assessment applications.

Moreover, the study successfully adopted the framework for evaluating how data input outliers affect the data-driven model, such as the IEWQI model. The current study has been carried out in Cork Harbour for only a single year of WQ data. The framework should be tested across various domains for evaluating the response of the IEWQI model's in terms of the spatio-temporal resolution of the domain. Nevertheless, the study recommended that future research should be conducted to adjust or revise the IEWQI model's rating schemes and investigate the practical effects of data outliers on updated rating categories. However, the study provides potential recommendations for enhancing the IEWQI model's adaptability and reveals its effectiveness in expanding its applicability in more general WQ assessment scenarios.

1. Introduction

In the field of environmental science and water resource management, comprehending water quality dynamics is imperative due to the intricate interplay between anthropogenic activities and the natural environment (Parween et al., 2022; Uddin et al., 2022a, 2020, 2018, 2017, 2023c, 2021). This understanding is facilitated by WQ models designed to simulate and predict various parameters (Burigato Costa et al., 2019; Chidiac et al., 2023; Ding et al., 2023; Uddin et al., 2023c; 2023b, 2021). Nevertheless, the increasing complexity and data intensity of these models underscore the challenge of identifying outliers, which significantly deviate from the norm (Jeong and Park, 2019; Lee, 2017; Najafabadi et al., 2015; Sivarajah et al., 2017; Smiti, 2020; Wu et al., 2021; Zhang and Thorburn, 2022). These outliers can skew model predictions, impair accuracy, and obstruct decision-making processes (Kang et al., 2017; Lee, 2017; Sivarajah et al., 2017). Recent studies have underscored the substantial impact of outliers on model outputs, thereby introducing significant uncertainty in final assessments (Jin et al., 2019; Lee, 2017; Liang et al., 2022; Orouji et al., 2013; Talagala et al., 2019). As consequence, the model(s) produces a considerable uncertainty to the final assessment (Angiulli and Fassetti, 2021; Harrington et al., 2021; Lee, 2017; Liang et al., 2022; Sharma and Seal, 2021; Wu et al., 2021; Zhang and Thorburn, 2022).

The Water Quality Index (WQI) serves as a widely adopted technique for assessing and monitoring WQ (Burić et al., 2023; Ding et al., 2023; Georgescu et al., 2023; Gupta and Gupta, 2021; Manna and Biswas, 2023; Mogane et al., 2023; Uddin, 2023; Uddin et al., 2020b; Uddin et al., 2023h; 2023b, 2023a, 2023c). This model adeptly converts specific WQ indicator data into dimensionless numbers while preserving essential information from raw measurements (Parween et al., 2022; Uddin et al., 2022a; 2022b; 2023d; 2021). The WQI model comprises five key components: (i) The indicator selection process is designed for selecting crucial WQ indicators, prioritizing them based on their relative significance; (ii) sub-index functions are used to transform diverse WQ data into a uniform, dimensionless form, ensuring there is no loss of essential information about each indicator; (iii) a weight generation technique is implemented for assigning weights to indicators based on their influence on overall water quality; (iv) an aggregation function is utilized to combine these weighted sub-index values into a single, comprehensive WQI score; and (v) a classification scheme interprets this score, categorizing WQ into levels such as "good," "fair," or "marginal," providing a clear and interpretable assessment of water conditions (Uddin et al., 2022a, 2023g, 2023e, 2023f, 2023d). Comprehensive information on the model's components is detailed in the study by Uddin et al. (2023d).

The development of numerous WQI models globally has facilitated WQ assessment in diverse aquatic environments including rivers, lakes, and groundwater (Gani et al., 2023; Georgescu et al., 2023; Mogane et al., 2023; Uddin et al., 2023d, 2023b, 2023c, 2021). An overview of these approaches, their model architectures, applications, and limitations, can be found in Uddin et al. (2021). These models' simplicity and accessibility have led to their widespread adoption (Burić et al., 2023; Ding et al., 2023; Uddin et al., 2023b, 2023h). However, recent research has brought to light uncertainties stemming from various aspects of these models, including indicator selection, weighting methods, sub-indexing functions, aggregation functions, and model interpretation schemes, as discussed in Uddin et al. (2022a, 2023c, 2023h, 2023b, 2023d, 2023f, 2023e, 2023g, 2021).

Addressing these challenges, an innovative approach, the "Irish Water Quality Index (IEWQI) model," has been introduced, focusing particularly on marine waters. Details of the model development process are presented in Uddin et al. (2022c, 2023d), while the IEWQI model's architecture and applications are outlined in Uddin et al. (2023d). Its effectiveness in assessing and monitoring surface WQ in different settings demonstrates its utility. To the best of the authors' knowledge, the IEWQI model represents the first instance of incorporating state-of-the-art machine learning (ML) and artificial intelligence (AI) techniques into WQI development. Since its development, the Irish Water Quality Index (IEWQI) model, it has garnered considerable national and international attention and adoption. In Ireland, it has been effectively used for assessing river, lake, and marine waters, as detailed in Uddin et al. (2023b, 2023d). The UK, particularly Northern Ireland, has utilized this model to compare with existing approaches like the "One-out, all-out" method (Uddin et al., 2023c). In Romania, the model's effectiveness and reliability were favorably compared to other WQI approaches (Georgescu et al., 2023), while China reported successful applications in sea water assessment and highlighted its bias-free assessment capabilities (Ding et al., 2023). Montenegro's application of the model in lake water assessments was noted in Burić et al. (2023), and South Africa's use of the model for lotic and lentic ecosystems was reported in Mogane et al. (2023). In India, a study emphasized the model's effectiveness in groundwater quality assessment (Manna and Biswas, 2023), and in Bangladesh, its efficacy for sub-tropical marine waters was outlined in Uddin et al. (2023h). Additionally, two recent studies, Gani et al. (2023) and Sajib et al. (2023, 2024), demonstrated its effectiveness in river water and real-time groundwater quality assessment, respectively.

Moreover, recent studies, including those by Uddin et al. (2021, 2022a, 2023) and supported by other researchers (Gupta and Gupta, 2021; Sutadian et al., 2016, 2018), have emphasized the need to address "eclipsing" in Water Quality Index (WQI) models, acknowledging its considerable impact on model precision. Eclipsing in WQI models is a phenomenon where scores overestimate water quality, often due to inappropriate sub-indexing aggregated rules and parameters weight values (Sutadian et al., 2016; Uddin et al., 2021). This can result in misleading representations of water quality, concealing actual instances of poor quality, as further detailed by Uddin et al. (2022c, 2023d). A few studies have revealed that the aggregation function(s) are the major source of the eclipsing problem in WQI computation. Several studies have reported that the actual scenarios of WQ do not reflect the computed WQI scores due to inaccurate WQI scores. It may occur if individual indicators exceed critical thresholds or are outliers. In that

case, the model produced higher WQI scores, which indicates the WQ was acceptable, despite some WQ indicators breaching the guideline values. Consequences result in a significant discrepancy between the actual state of WQ and the model's output, which can lead to inconsistent and unreliable assessments. Therefore, to avoid the model eclipsing problem, removing data outliers from the model's input could be one of the most effective approaches, particularly those at the lower end/higher tail (extreme higher) of the dataset (lower/higher outlier points). For example, the current study focuses on DOX levels in Cork Harbour, typically ranging between 70.8 (lowest concentration) and 115.75 (highest concentration), whereas the guideline values ranged from 72 % to 128 % of saturation (details in Table S4). Based on the DOX data points outliers (as shown in Fig. 4), in total four data points were found as outliers, whereas three were on the lower end (70.8, 78.68, and 81.5) and one on the higher end (115.75) of the DOX dataset. According to the methodology of the IEWQI model, these lower data points (outliers) can lead to an overestimation of WQ with breached indicator(s). Uddin et al. (2022c) provide details of the eclipsing issues in the IEWQI model, its nature, and how to determine this issue. Moreover, Figure S1 presents the relationship between data point(s) outliers, and eclipsing and their collective impact on the reliability of the IEWQI model. Therefore, as a data-driven model, it is essential to investigate the model sensitivity to high-dimensional WQ data (outliers).

For the purposes of the accurate assessment/modelling/simulating/forecasting, data outlier's detection and treatment are essential across various fields, including cyber security, healthcare, environmental monitoring, finance, etc. (Balamurali and Melkumyan, 2018; Berendrecht et al., 2022; Choi et al., 2021; Garces and Sbarbaro, 2009; Gui et al., 2017; Ha et al., 2014; Misra et al., 2020; Shah et al., 2023; Smiti, 2020). Outliers, often representing rare and unusual occurrences, can profoundly influence the quality and reliability of data-driven model(s) outcomes (Choi et al., 2023a, 2023b; El Alaoui et al., 2018; Gui et al., 2017; Harrington et al., 2021; Kim et al., 2022; Rahman and Harding, 2016; Yuan et al., 2018). To address this difficulty, the data scientists/analysts/researchers used a range of tools and techniques, from sophisticated ML and AI algorithms to traditional statistical approaches (Budhlakoti et al., 2020; Domański, 2020a; Duraj and Szczepaniak, 2021; Hansen et al., 2023; Kwak and Kim, 2017; Ojo et al., 2022; Shimizu, 2022; Smiti, 2020). Over the years, many tools and techniques have been developed to address this challenge, including typical statistical/mathematical approaches like Mahalanobis Distance (Cabana et al., 2021; Dashdondov and Kim, 2023; Etherington, 2021; Gyebnár et al., 2019; Leys et al., 2018; Todeschini et al., 2013), Robust Z-Score (Aggarwal et al., 2019; Berendrecht et al., 2022; Green, 2021; Haj--Hassan et al., 2020; Rousseeuw and Hubert, 2011; Templ et al., 2020; Yuen and Ortiz, 2017), and Histogram-Based Outlier Detection (HBOS) (Aguilera-Martos et al., 2023a, 2023b; Fahim et al., 2022; Fernández et al., 2022; Kalaycı and Ercan, 2018; Pei et al., 2021; Samariya and Ma, 2022; Smiti, 2020), Local Outlier Factor (LOF) (Alsini et al., 2021; Auskalnis et al., 2018; Chiu and Fu, 2003; Johannesen et al., 2022; Meenakshi and M, 2022; Mishra et al., 2019; Peng et al., 2021; Petkovski and Shehu, 2023a; Qiu et al., 2022; Wang et al., 2023; Xu et al., 2022, 2019) and Statistical Process Control (SPC) (Al Suwaidi et al., 2023; Baroudi et al., 2023; Raveendran et al., 2023; Tan et al., 2023; Yeganeh and Shongwe, 2023; Zeng et al., 2023) to cutting-edge Machine Learning (ML) and Artificial Intelligence (AI) approaches (Adeoye et al., 2023; Albahra et al., 2023; Ali et al., 2023; Chander and Kumaravelan, 2022; Chang et al., 2022; Hassan et al., 2022; Jamshidi et al., 2022a; Kokatnoor et al., 2022; Luley et al., 2023; Mentis et al., 2023; Milić et al., 2023; Nasir et al., 2022; Prasad et al., 2022; Sejr and Schneider-Kamp, 2021; Sikder and Batarseh, 2023; Varadharajan et al., 2022; Wei et al., 2023) like Isolation Forest, One-Class SVM, boosting algorithms, and Deep Learning (Albahra et al., 2023; Hassan et al., 2022; Ragab et al., 2022; Wei et al., 2023).

Recent several studies have revealed that the IF (AbuAlghanam et al., 2023; Carletti et al., 2023; Chen et al., 2023, 2022; Feng et al., 2022; Kabir et al., 2023; Liu et al., 2008; Mensi et al., 2023; Petkovski and Shehu, 2023b; Wang et al., 2023; Xu et al., 2023; Yin et al., 2023) and KDE (Choi et al., 2022; Hewitt et al., 2022; Lei et al., 2023; Matioli et al., 2018; Modak, 2023;Rosenberger et al., 2022; Zeng et al., 2023) outperforms other machine learning techniques when it comes to detecting data anomalies, outliers, or system defaults in various fields, including water research (Liu et al., 2020; Shi et al., 2023; Yin et al., 2023). Moreover, a number of research have successfully applied these techniques for detecting data outlier in various field including decision-making and risk management with extensive sensor data using IF technique (Tan et al., 2022); improved cybersecurity by detecting cyber anomalies by adopting IF (Ripan et al., 2021); credit card fraud detection utilizing IF (Krishna et al., 2023); anomalies detection in oil producing by applying IF (Fernandes et al., 2023) in order to optimize the impact of the data outlier in data-driven mode. Mensi et al. (2023) utilized IF approaches for identification outliers based on pairwise distances between two objects, while Xu et al. (2023) used the IF for detecting and mitigate the data anomalies in various data types. Buschjäger et al. (2022) adopted the IF for comparative analysis of detecting data outliers in fourteen different datasets successfully. Moreover, several recent studies have utilized the IF in water research, for examples- Yin et al. (2023) utilized this approaches for detecting the anomalies in simulation data within water level measurements a borehole groups, while Liu et al. (2020) and Wang et al. (2023) used the IF for detecting WQ anomaly and early warning. However, a few recent studies have revealed that the IF could be effective to detect the WQ anomaly or detect the data outliers in big WQ monitoring datasets in order to develop the data-driven model (Liu et al., 2020; Wang et al., 2023).

In functional data outlier detection, many studies have utilized the KDE technique successfully in various domain. Hernández et al. (2023) successfully used this technique for detecting data outliers in electrocardiogram signals and mortality curves datasets. Latecki et al. (2007) demonstrated its versatility in fields like network intrusion detection and video analysis. Tang and He (2017) used the KDE techniques for detecting data outliers in various real life datasets. Zhang et al. (2022) successfully demonstrated that the KDE algorithm effectiveness for improving predictive outcomes in datasets related to public safety, commodity trade, and network security. Several studies have utilized the KDE approaches for detecting WQ anomalies and early warning system (Rosenberger et al., 2022; Liu et al., 2020). However, the use of IF and KDE has not been investigated widely for detecting outliers or anomalies in WQ data-driven model(s) whereas most studies have focused on detecting the anomalies or outliers at particular WQ indicator(s). Therefore, the application of anomaly detection or outlier techniques in various fields has prompted the emergence of innovative methods in order to improve the reliability of WQ models through data-driven approaches (Aliashrafi et al., 2021; Jeong and Park, 2019; Jin et al., 2019; Kang et al., 2017; Orouji et al., 2013; Petkovski and Shehu, 2023a; Sarker, 2021).As discussed earlier, the research carried out a comparative investigation of between typical statistical approaches, and ML/AI approaches for detecting and removing data outlier from the data-driven IEWQI model in order to determine the impact of data-outlier on the model performance. For the purposes of detecting data outlies in input (WQ indicators) of the IEWQI model, the research also utilized five traditional statistical/mathematical approaches like Mahalanobis Distance, Robust Z-Score, Local Outlier Factor (LOF), Histogram-Based Outlier Detection (HBOS), and Statistical Process Control (SPC), alongside two novel ML/AI algorithms: the Isolation Forest, and Kernel Density Functions. By contrasting these diverse approaches, is to provide a comprehensive understanding of their strengths, limitations, and applicability for detecting input outliers of the IEWQI model and that can be helpful for investigating the anomalies in data-driven model across various domains particularly water research and management.

Therefore, the research aim was to investigate the impact of input

(various WQ indicators) outliers on newly developed advanced data-driven IEWQI model performance in predicting WQI scores. The study also proposed a comprehensive framework for handling data outliers in WQ models, especially WQI approaches. Fig. 1 presents the comprehensive conceptual framework of the research. In order to obtain the research aim, the study carried out by following objectives:

- To detect data outliers in input of the IEWQI model utilizing two identical including IF and KDE ML algorithms.
- To develop IEWQI model outcomes scenarios with outliers and removing them from the input.
- To evaluate the IEWQI model's performance under two distinct scenarios (with and without data outliers).
- To validate the outlier results by comparing them with the outcomes of five advanced statistical/mathematical techniques.
- To provide practical recommendations for improving/revising the IEWQI model's architecture and its effectiveness

As was pointed in previous section, the ultimate goal of the research was to investigate impact of data outliers on the IEWQI model by analysing the results of various techniques. For the purposes of the understanding the outlier's impacts on the IEWQI model, the research assumed hypothesis as follows:

Null Hypothesis (H0): Input data outliers treatment are important into WQ modelling, particularly in the case of the Irish WQ Index (IEWQI)

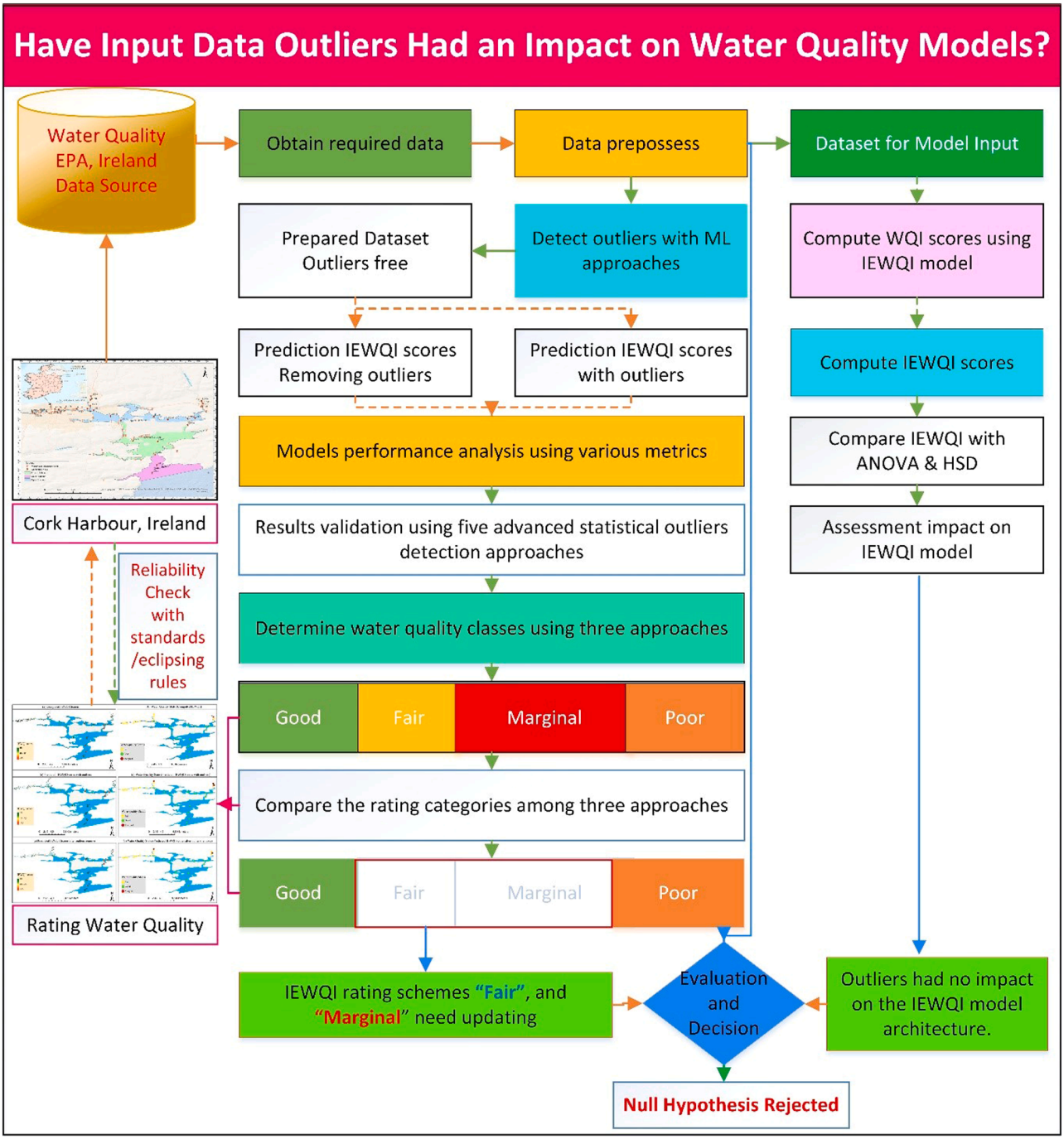


Fig. 1. A comprehensive conceptual framework for detecting data outliers in WQI approaches.

Model, will have a significant impact on model accuracy and reliability for providing accurate information of water quality.
Alternative Hypothesis (Ha): Input data outlier will not lead to a significant impact on model accuracy and reliability in terms of rating WQ accurately.

However, the research have been carried out to validate the alternative hypothesis (Ha) by utilizing the two robust algorithms and comparing a range of statistical/mathematical measures that could be improved the IEWQI model accuracy and reliability for assessing and predicting WQ accurately. The findings of the research could be helpful for understanding the data outlier's effects on WQ model like IEWQI, and also may offer valuable insights into the benefits of utilizing state-of-the-art outlier detection methods in WQ modelling, highlighting their significance in water resources management practices.

2. Materials and method

2.1. Model application domain

This study focuses on Cork Harbour, located on the southwestern coast of Ireland, and represents a significant focal point within the field of geography and environmental research. Cork Harbour is characterized by its distinction as the largest natural Harbour in Ireland, and it exhibits the characteristics of a macro-tidal estuary (Comer et al., 2017; Hartnett and Nash, 2015; Uddin et al., 2020, 2023d). Its coastal environment is marked by a typical spring tide range of 4.2 m at the Harbour's entrance, resulting in relatively shallow water depths within the Harbour, particularly during spring tides (Hartnett et al., 2012; Uddin et al., 2022a). This phenomenon leads to the exposure of extensive mudflats and sandflats during low tide, significantly influencing the Harbour's ecological dynamics (Hartnett et al., 2012, 2011a; Uddin et al., 2023b). As one moves further towards the Harbour's mouth, the main channel deepens significantly, reaching depths of approximately 30 m. The Harbour is nourished by several rivers, with the River Lee being the most prominent contributor, accounting for approximately 75 % of the freshwater inflow into the estuary (Uddin et al., 2022a, 2023b). These freshwater inputs play a pivotal role in shaping the Harbour's environmental dynamics.

In addition, Cork Harbour is also notable for its substantial population and industrialization. Cork City, situated at the confluence of the River Lee and the Harbour, is home to approximately 125,000 residents. This urban center serves as a critical industrial hub for the south-western region of Ireland (Uddin et al., 2023e, 2023d). The surrounding hinterlands are characterized by intensive agricultural activities, exerting a direct influence on the WQ within the region. Moreover, Cork Harbour contends with WQ challenges stemming from effluent discharges originating from seven wastewater treatment plants (WWTPs) located within its catchment area. These WWTPs contribute to the intricate array of factors affecting WQ in the Harbour.

For the purpose of comprehensive regional assessment and analytical purposes, Cork Harbour has been partitioned into three distinct zones (see Fig. 2): (1) Upper Harbour, encompassing regions such as the River Lee, River Glashaboy, North Channel, and River Owenacurra; (2) Lower Harbour, comprising localities such as Passage West and Passage East; and (3) Outer Harbour, encompassing the River Owenboy and the

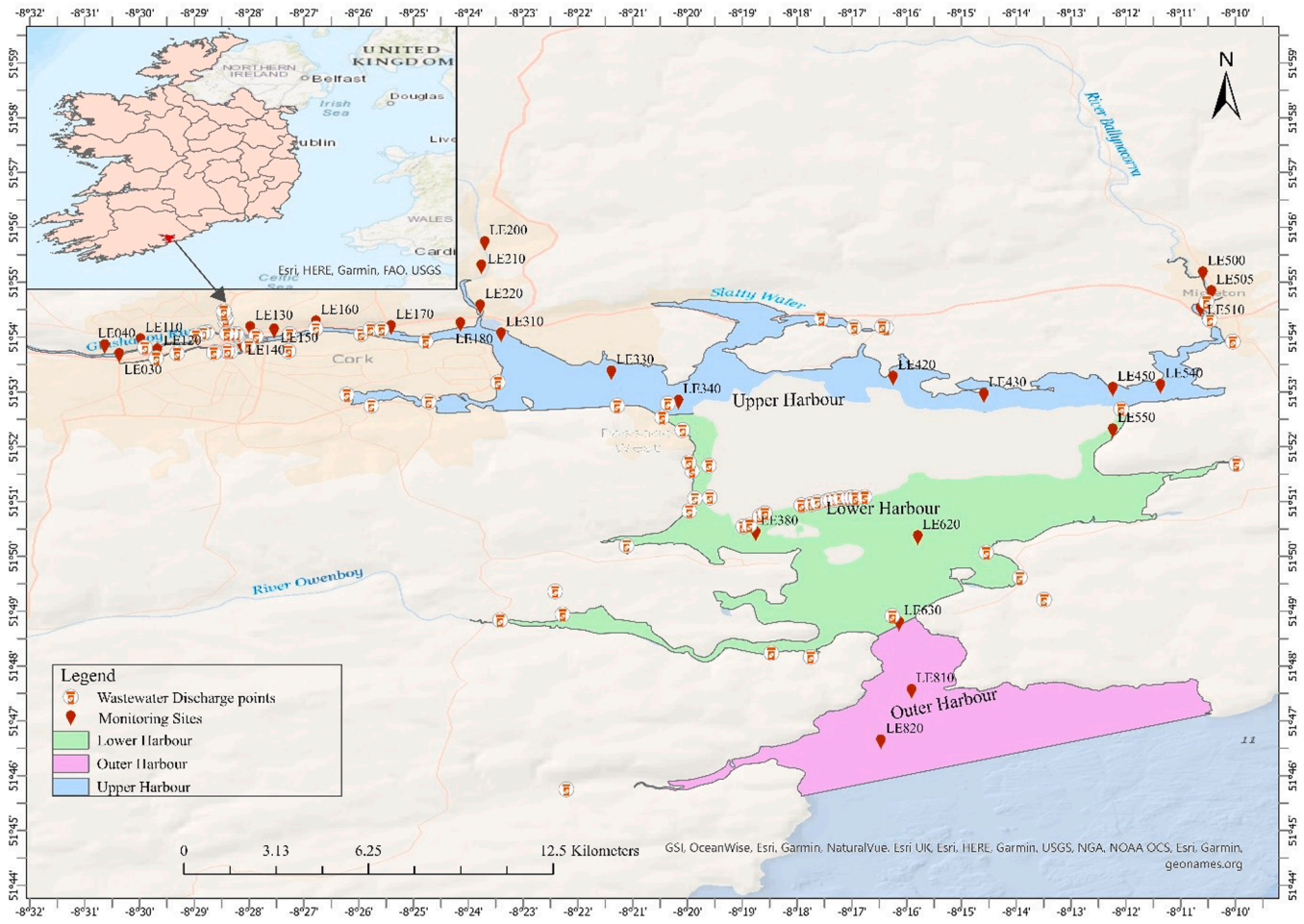


Fig. 2. Study area- monitoring sites and wastewater discharge points in Cork Harbour.

primary channel of the Harbour.

This study employs a proposed framework to evaluate coastal WQ within Cork Harbour, utilizing the Irish Water Quality Index (IEWQI) model. It also examines the outliers in data pertaining to various WQ indicators. The selection of Cork Harbour as the research site was motivated by its previous utilization in an IEWQI model study conducted by the authors, ensuring the availability of pertinent data for validation purposes. Furthermore, the designation of Cork Harbour as a Special Protection Area (SPA) underscores its significance in the realm of environmental conservation (Hartnett et al., 2011b; Olbert et al., 2017; Uddin et al., 2022a). The Harbour's substantial economic potential, coupled with the challenges posed by population density, industrialization, agriculture, and wastewater discharges, accentuates the complexity of managing its WQ and ecological well-being (Uddin et al., 2022a, 2023d). Ongoing efforts to monitor and mitigate these challenges are imperative for the sustainable stewardship of this invaluable coastal ecosystem.

2.2. Data description

To compute IEWQI scores, the research considered into account nine key WQ indicators, including pH, dissolved oxygen (DOX), salinity (SAL), biological oxygen demand (BOD5), water temperature (TEMP), transparency (TRAN), total organic nitrogen (TON), molybdate-reactive phosphorus (MRP), and dissolved inorganic nitrogen (DIN) by adopting the methodology proposed by (Uddin et al., 2023d). Table 1 provides a comprehensive list of these selected WQ indicators, along with their units and guideline values. The guideline values for several WQ indicators, excluding DOX, MRP, and DIN, were sourced from a variety of national and international guidelines. Conversely, the guideline values for DOX, MRP, and DIN were established using the methodology outlined in Uddin et al. (2022c, 2023d). Details of the methodology can be found in Uddin et al. (2023d).

The data for these WQ indicators was collected in the year 2022 from 29 out of 37 monitoring sites within the EPA database. This data is publicly accessible at https://www.catchments.ie/data/. The EPA upholds data reliability and accuracy through a stringent quality control system (refer to EPA 2021). The selection of monitoring sites was based on data availability and the coverage of geographical attributes within the domain, aligning with the requirements of the IEWQI model input. Fig. 2 presents the locations of these sites and wastewater discharge points. Detailed information regarding each WQ indicator across different monitoring sites in Cork Harbour can be found in Table S2.

To ensure data consistency, the study considered on the annual means of indicator measurements for the year 2022. The depth-averaged concentrations of these WQ indicators were determined by calculating the annual means for each respective indicator. It's important to note that SAL was exclusively used to establish guideline values for the moving thresholds of nutrient enrichment indicators (DOX, MRP, and DIN). It is noted that, the study primarily focused on analysing data from a single year (2022) obtained from the EPA database. It also utilized annual averages, a method that could potentially overlook short-term fluctuations in water quality.

Table 1
Guidelines values of various WQ indicators for marine waters.

WQ indicators	Unit	Standard threshold	
		Lower	Upper
TEMP[a]	°C	–	25
pH[b]	–	5	9
TRAN[c]	m/depth	>1	–
DOX[d]	% sat	72	128
BOD_5[a]	mg/l	0.0	7
DIN[d]	mg/l	0.0	1.14
TON[e]	mg/l as N	0.0	2
MRP[d]	mg/l as P	0.0	0.5

[a] EPA-Ireland (2001) recommended values for surface water.
[b] Estuary Monitoring Manual for pH, EPA, USA.
[c] EPA Bathing Water Quality Regulations 2008 (Ref. No. 79/2008).
[d] ATSEBI guide values, standard values were obtained based on median value of SAL. In this study, SAL median value was found 22 psu, 25 psu, 20 psu, 20 psu, 24 psu and 21 ppt during 2017, 2018, 2019, 2020, 2021 and 2022 respectively.
[e] The European Communities Regulations for quality of surface water intended for the abstraction of drinking water, 1989 (S.I. No. 294/1989).

2.3. IEWQI scores computation

To calculate WQI scores, various WQI approaches are currently in use. However, recent studies have revealed that these existing approaches introduce a significant level of uncertainty into the final assessments (Ding et al., 2023; Uddin, 2023; Uddin et al., 2022a, 2023d, 2023f, 2023h, 2023b, 2023c, 2021). To the best of the authors' knowledge (Uddin et al., 2023d), have developed an improved approach known as the "Irish Water Quality Index (IEWQI) model" especially focusing the marine waters. This model has proven effective in assessing marine waters, reducing uncertainty, and enhancing the reliability of assessment results (Ding et al., 2023; Uddin et al., 2023c, 2023h, 2023b).

Furthermore, several recent research studies have reported that the IEWQI model can significantly reduce uncertainty when assessing marine (Uddin et al., 2023c, 2023h), transitional, and coastal waters (Uddin et al., 2022a, 2023b, 2023f, 2023g, 2023e), as well as river waters (Ding et al., 2023; Gani et al., 2023; Georgescu et al., 2023; Mogane et al., 2023), compared to other existing approaches, with an uncertainty level of less than 1 % (Burić et al., 2023; Ding et al., 2023; Georgescu et al., 2023; Uddin et al., 2023h, 2023b, 2023c, 2023f). Numerous recent studies have also indicated that the IEWQI model is efficient in providing bias-free results, minimizing eclipsing and ambiguity issues compared to other methods (Ding et al., 2023; Gani et al., 2023; Mogane et al., 2023; Uddin et al., 2023b, 2023h, 2023c).

As the aim of the research, the research utilized the IEWQI model for computing WQI scores, following the methodology outlined in Uddin et al. (2023d). Detailed of the model's attributes and functions can be found in Uddin et al. (2022a, 2023c, 2023d, 2023e). Like other WQI approaches, the IEWQI model also consists of five crucial elements, and Fig. 3 presents the foundational architecture of the IEWQI model, as shown below:

Here, we presents the IEWQI model's attributes and functions of various components as below:

2.3.1. Input selection technique

The initial step in the WQI model involves the selection of essential WQ indicators (input) (Parween et al., 2022; Uddin et al., 2022a, 2022e, 2021). This process commonly uses various tools and techniques such as Principal Component Analysis, correlation analysis, based on literature suggestions, Delphi technique, expert judgment, and analytical hierarchical process (Gupta and Gupta, 2021; Uddin et al., 2021). Recent studies highlight the impact of inappropriate indicator selection on introducing model uncertainty, leading to the adoption of machine learning algorithms like gradient boosting and random forest to enhance the selection process (Uddin et al., 2022a, 2022d). In the IEWQI model, the random forest technique was recommended to select the crucial input indicators, the model recommended eight indicators including DOX, MRP, DIN, TON, BOD5, pH, TEMP, and TRAN, for computing IEWQI scores (for indicator details, refer to Section 2.2), with comprehensive procedures outlined in Uddin et al. (2023d).

2.3.2. Sub-index function

Sub-index functions are another crucial components that are widely used in standardizing WQ indicators to a uniform scale (Uddin et al., 2021, Uddin et al., 2022a). Many approaches, including linear interpolation and rating curve functions, have been widely used for this

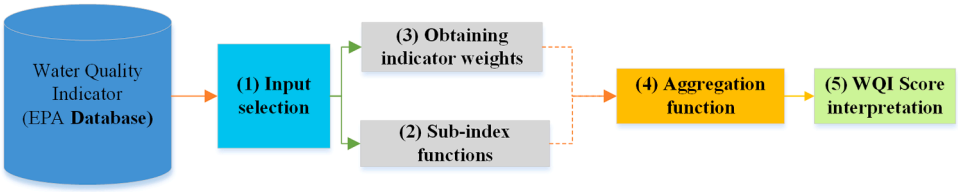


Fig. 3. Architecture of the IEWQI model' according to Uddin et al. (2023d).

purpose. However, recent several research revealed that these methods have been produces a considerable amount of uncertainty due to the ambiguity, and eclipsing issues (Gupta and Gupta, 2021; Sutadian et al., 2016; Uddin et al., 2021). To mitigate this problem, the authors developed a set of new linear interpolation rescaling functions, ensuring precision and removing ambiguity from the model (Uddin et al., 2022c).

In the IEWQI model, three newly developed linear interpolation functions, accompanied by specific binary rules (see Table S1), were utilized to convert the various WQ indicators into the Sub-index scores (unit less). These functions are as follows:

$$SI = (SI_u - SI_l) - \frac{(SI_u \times WQ_m)}{(STD_u - STD_l)} \quad (1)$$

$$SI = \frac{(WQ_m - STD_l)}{(STD_u - STD_l)} \times SI_u \quad (2)$$

$$SI_i = (SI_u - SI_l) - \frac{(WQ_m - STD_l)}{(STD_u - STD_l)} \quad (3)$$

where, SI represents the sub-index value of the indicator (s), SI_u denotes the upper threshold value (100 for "excellent") for the respective water class, SI_l is threshold lower value (0 for "poor") of the respective water class, STD_u indicates the upper limit threshold for the standard concerning the indicator, STD_l represents the lower limit threshold for the standard concerning the indicator, and WQ_m corresponds to the measured or actual concentration of the WQ indicator. Details of threshold ranges are provides in Table 1.

2.3.3. Obtaining indicator weight values

The weighting of WQ indicators is another important components, and existing WQI approaches utilized a range of tools and techniques for generating indicators weight values based on their relative importance, including subjective methods and objective mathematical functions (Gupta and Gupta, 2021; Sutadian et al., 2016; Uddin et al., 2021, 2022c, 2022d). Recently, several studies have reported that ambiguous weight values produced due to the in appropriate weighting approaches (Uddin et al., 2022c, 2022b). To mitigate this situation, the authors developed a novel approach combining ML and objective-based mathematical functions, resulting in more accurate weight values (Uddin et al., 2022c), whereas the random forest was used to rank of the indicators, with rank sum mathematical functions effectively reducing model uncertainty (Uddin et al., 2022d). Weight computation can be defined as follows:

$$w_i = \frac{n+1-i}{\sum_{j=1}^{n} j} = \frac{2(n+1-i)}{n(n+1)}, \; i = 1, 2, 3 \ldots\ldots, n \quad (4)$$

where n represents the total number of ranked indicators; i denotes the rank of the i^{th} indicator rank; j signifies the summation of ranks, and w represents the associated weight value. It is noted that using the indicators weight values are provided in Table S3.

2.3.4. Aggregation function

The final component of the IEWQI model is aggregation function that is used to combine SI and weight values into a single WQI score. Various aggregation functions, including weighted and unweighted methods, have been used in existing WQI approaches. A number of research have reported that the ambiguity issues in existing aggregation functions could be generated the misclassification of WQ in final rating system due to the over estimation or under estimation problems of the aggregation process (Uddin et al., 2021, 2022c, 2023d, 2023f). For the purposes of avoiding the this problem, the IEWQ model introduced the weighted quadratic mean (WQM) aggregation function, which outperformed other approaches in order to minimize ambiguity and uncertainty (Ding et al., 2023; Georgescu et al., 2023; Manna and Biswas, 2023; Sajib et al., 2023, 2024; Uddin et al., 2022a, 2023f). The IEWQI aggregation function can be defined as follows:

$$IEWQI = \sqrt{\sum_{i}^{n} w_i s_i^2} \quad (5)$$

where s_i is the sub-index value for indicator i; w_i is weight value of respective WQ variables, and n is the number of indicators.

2.3.5. IEWQI score translation

The ultimate goal of the Water Quality Index (WQI) approach is to assess and rate the quality of water (Georgescu et al., 2023; Parween et al., 2022; Uddin et al., 2022a, 2022b; 2023b, 2023d, 2023f) lighted a significant issue known as the "metaphoring problem" resulting from inappropriate classification schemes (Burić et al., 2023; Ding et al., 2023; Mogane et al., 2023; Uddin et al.,2023g, 2023d). Details of the "metaphoring problem" can be found in Uddin et al. (2023g). To address this issue, the authors have introduced a novel classification scheme for evaluating marine waters. The process of developing this classification scheme is detailed in Uddin et al. (2022c), and the validation results are presented in Uddin et al. (2023g). For the purpose of assessing WQ using IEWQI scores, the research has adopted the classification schemes outlined in Uddin et al. (2023e). Table S4 provides the details classification schemes.

2.4. Outlier detection techniques

Outlier detection is a vital issue for exploring or identifying the extreme data points in a dataset(s) (Garces and Sbarbaro, 2009; Luo and Paal, 2023; Rousseeuw and Hubert, 2011;Rahman, 2019; Smiti, 2020). Gradually, this topic has gained much more attention for detecting data anomalies/ outliers in any datasets across various fields including health care systems, finance, natural resources management, cybersecurity and it anomalies detection, fault detection etc. In recent years, there has been an increasing interest in detecting outliers for developing the data-driven models, at particularly in water resources research and its modelling with big monitoring datasets. A range of tools and techniques have utilized to detect data anomalies/outliers like as Z-Score, Percentile-based, Isolation Forest, and Local Outlier Factor use statistical methods (Domański, 2020b; Ha et al., 2015; Obikee et al., 2014; Smiti, 2020; Yuen and Ortiz, 2017). Recently several studies have widely utilized various ML techniques for detecting data outliers or anomalies across various fields, the IF and KDE one of them outperformed algorithms compared to others techniques (Chen et al., 2022; Kabir et al., 2023; Mensi et al., 2023, 2021; Tokovarov and Karczmarek, 2022; Hewitt et al., 2022; Rosenberger et al., 2022; Zeng et al., 2023). In the context of modelling WQ, data outliers are crucial issue for monitoring

and management water resources more accurately, especially large simulation model(s) or data-driven approaches (Garces and Sbarbaro, 2009; Ragab et al., 2022; Shah et al., 2023). Commonly, the research have used some basics statistical approaches such as Z-Score and Interquartile Range for detecting the anomalies in WQ indicators like dissolved oxygen, pH, and water temperature (Aggarwal et al., 2019; Pei et al., 2021). Recently, a few research have utilized ML algorithms such as IF, Local Outlier Factor, and KDE to identify the changing the concentration level of conductivity and it contaminates level (Abdulghafoor and Mohamed, 2022; Mensi et al., 2023; Tokovarov and Karczmarek, 2022). Many studies have revealed that these techniques could be effective compared to others algorithms to detect the data outliers in terms of enhancing model accuracy as well as accurate assessment of WQ (Chen et al., 2022; Hewitt et al., 2022; Jin et al., 2019b; Kabir et al., 2023; Mensi et al., 2023, 2021; Najman and Zieliński, 2021; Ripan et al., 2021; Tokovarov and Karczmarek, 2022; Tan et al., 2022).

However, as discussed in earlier section, in the domain of data-driven modelling, particularly for the IEWQI, the implementation of IF and KDE techniques presents a sophisticated approach to managing complex datasets. The IF algorithm is particularly use at handling high-dimensional data, a common characteristic of various indicators in the WQ model(s) (Chen et al., 2022; Kabir et al., 2023; Mensi et al., 2023, 2021; Najman and Zieliński, 2021; Ripan et al., 2021; Tan et al., 2022); that could be potential for efficiently managing the multi-dimensional nature of environmental data (Mensi et al., 2021; Najman and Zieliński, 2021; Toufigh and Ranjbar, 2023; Yin et al., 2023). Moreover, a few studies have revealed that the IF may an integrated approaches for its computational efficiency (due to its sensitivity to anomalies) in order to detect the data outliers/anomalies in multi-dimensional data such as large datasets typically associated with environmental monitoring (Liu et al., 2020; Mensi et al., 2021; Tan et al., 2022). A few studies have reported that the IF is particularly sensitive anomalies even in the detection of small deviations, this approaches could be effective to develop a strong early warning systems for the management of environment by reducing data anomalies in datasets. (Toufigh and Ranjbar, 2023; Xu et al., 2023).

Contrary, the KDE offers a non-parametric approach to data analysis, a significant advantage when dealing with environmental data that may not adhere to standard distributions (Cao et al., 2023; Latecki et al., 2007; Villa and Lozano, 2020; Zhao et al., 2022). This flexibility allows KDE to model the probability density function of various WQ parameters comprehensively (Cao et al., 2023; Gallego et al., 2022; Han et al., 2019; Hernández et al., 2023). Furthermore, KDE's flexibility in responding to variances in data is especially effective in dynamic datasets such as environmental monitoring database, where seasonal and environmental variables can cause fluctuations in WQ indicators (Han et al., 2019; Hernández et al., 2023; Latecki et al., 2007).

Therefore, the research utilized IF and KDE techniques for detecting the data outliers and their impacts on the IEWQI model, because a number of recent studies revealed that these techniques outperformed in detecting water quality. Details of the conceptual framework for detecting the data outliers are provided in Fig. 1. The following sections briefly discussed the IF and KDE below:

2.4.1. Isolation forest (IF) algorithm

IF is an innovative and powerful approaches to detect the rare and abnormal data points in multidimensional dataset. (Kabir et al., 2023; Petkovski and Shehu, 2023b; Wang et al., 2023). Unlike traditional methods relying on complex metrics, it leverages the uniqueness and rarity of outliers in WQ data (Chen et al., 2023; Petkovski and Shehu, 2023b; Wang et al., 2023; Yin et al., 2023). Several recent studies have reveals that this approach efficiently detects anomalies by recursively partitioning the data, outperforming established methods (Feng et al., 2022; Liu et al., 2008; Oliveira et al., 2022; Tokovarov and Karczmarek, 2022). A number of recent studies have revealed that the IF algorithms outperformed compared to other techniques for detecting data outliers/anomalies across various domain (Misra et al., 2020; Shah et al., 2023; Wang et al., 2023; Yin et al., 2023). Therefore, the research utilized this algorithm to detect/identify outliers in various WQ indicators and model them following Yin et al. (2023). The IF algorithm mathematical can be defined as follows:

Given a dataset X (herein is the WQ indicators dataset, as mentioned in Section 2.2) with n data points (current research used 29 points) of each WQ indicators, where each data point is represented as x_i in d-dimensional space, the IF algorithm can be described using the following steps:

- Random Partitioning: Select a random feature f from the d features (various WQ indicators). Choose a random split value v between the minimum and maximum values of f.
- The dataset is partitioned into two subsets: one comprising data points with feature values less than v, and the other containing values greater than or equal to v. This partitioning process is then recursively applied to each subset until a predetermined depth is attained or the subset comprises only a single data point.
- Path Length Calculation: For each data point of x_i, calculate the path length $h(x_i)$ from the root node to the terminal node containing x_i.
- Scoring: once the calculate the average path length for each data point to each WQ indicator: $E(h(x_i)) = c(n)$, where $c(n)$ is a normalization factor based on the average path length for balanced binary trees.

Following a rule of thumb, data points displaying notably shorter average path lengths than the normalized average are indicative of potential outliers. Usually, the IF algorithm is very efficient and scalable to detect data outliers compared to other techniques for managing high-dimensional datasets without relying on the predetermined distance metrics between the root node(s) and terminal node(s) (Carletti et al., 2023; Chen et al., 2023, 2022; Feng et al., 2022; Kabir et al., 2023). It should be noted that several recent studies have successfully applied IF techniques for detecting data outliers or anomalies across various fields. This is the first initiative, and the research utilized this approach for detecting outliers and the impact of them on the WQ model, particularly the IEWQI model. Therefore, its applicability might vary across various datasets, necessitating meticulous parameter tuning for optimal performance.

2.4.2. Kernel density estimation (KDE)

KDE is an effective statistical method that is widely used for estimating the probability density function for continuous random variable based on given dataset(s). Currently, this technique is not used only for estimating the probability density function, but also recently this approach widely utilized for detecting the data outliers/anomalies in various high-dimension dataset(s) by addressing the region(s) of low data density (Humbert et al., 2022; Rosenberger et al., 2022). Commonly, it measures data point concentration in feature space regions to detect lower-density areas where outliers tend to occur (Hewitt et al., 2022; Matioli et al., 2018). Being non-parametric, KDE flexibly adapts various data distribution function(s), making it valuable for complex and multi-modal datasets (Lei et al., 2023; Liu et al., 2020b; Zeng et al., 2023a). Recently several studies utilized this approach for detecting the data outliers across various field such as streaming data, various real-field sensor data, network fault analysis data etc. (Liu et al., 2020; Rosenberger et al., 2022; Wahid et al., 2018; Xu et al., 2016; Zheng et al., 2017). while a very limited studies can be found in WQ or its indicators (Jiang et al., 2022; Panjei et al., 2022; Piñeiro Di Blasi et al., 2015; Talagala et al., 2019; Tang and He, 2017). Recent a few studies have revealed that it can be used to detect outliers in a WQ model by highlighting areas of low data density (Oliveira et al., 2022; Rosenberger et al., 2022; Zeng et al., 2023). Consequently, the study performed this approach for detecting the data outliers in WQ dataset by adopting the methodology of Zeng et al. (2023). Details of the

methodological outlined are presented in Zeng et al. (2023). The mathematical expression of the KDE's function as follows:

Given a dataset of WQ measurements $x = \{x_1, x_2, \ldots, x_n\}$, the kernel density estimates at a specific point x is given by:

$$KDE(x) = \left(\frac{1}{n \times h}\right) \times \sum K\left(\frac{x - x_i}{h}\right) \quad (6)$$

where n is the number of data points (the research considered 29 data points) of various WQ indicators in the dataset, h is the bandwidth, a smoothing parameter that determines the width of the kernel function, x_i represents individual data points of different WQ indicators from the dataset, and K is the kernel function, typically a symmetric probability density function (e.g., Gaussian or Epanechnikov kernel). It is noted that the kernel function K is chosen to be a positive function that integrates to 1 and has its maximum at 0. It "spreads" the influence of each data point x_i to nearby points, allowing the KDE to provide a smooth estimate of the underlying data distribution.

2.5. Advanced statistical approaches

For the purposes of the validation of the outlier detection results of IF and KDE, the research utilized five advanced statistical tools and techniques including (i) Mahalanobis distance, (ii) Robust Z score, (iii) Local Outlier Factor (LOF), (iv) Histogram-Based Outlier Detection scores (HBOS), and (iv) Statistical Process Control (SPC); these have widely used for detecting data outliers in various field including water research domain (Aggarwal et al., 2019; IAguilera-Martos et al., 2023b; Alsini et al., 2021; Etherington, 2021; Fahim et al., 2022; Johannesen et al., 2022; Leys et al., 2018; Pei et al., 2021; Zeng et al., 2023). These techniques effective to identify observations that deviate significantly from the normal pattern(s) (Cabana et al., 2021; Pei et al., 2021). While traditional methods like mean and standard deviation-based approaches are commonly used, advanced statistical techniques offer more accurate and effective solutions for detecting data outliers/anomalies in a dataset (Panjei et al., 2022; Wang et al., 2023; Meenakshi and M, 2022). These methods consider various factors, such as correlation among variables within a dataset, local data density, and resistance to extreme values (Panjei et al., 2022). Usually, the Mahalanobis distance measures multivariate distances, considering variable correlations but assuming multivariate normality of various variables (Cabana et al., 2021; Leys et al., 2018), whereas the robust Z score is adaptable to outliers and non-normal distributions, offering a standardized deviation measure from the median, making it suitable for skewed or heavy-tailed data (Pei et al., 2021; Yin and Liu, 2022). On the other hand, the LOF is effective to identify the local anomalies by assessing or comparing the local density deviations, adapting to diverse data attributes, although this approach effectively utilized for detecting clustered outliers in high-dimensional datasets (Wang et al., 2023; Meenakshi and M, 2022). It is noted that principally the LOF requires parameter selection for dealing high-dimensional spaces (Alsini et al., 2021). In general, the HBOS relies on thresholding histogram bins, providing a straightforward approach for univariate or low-dimensional data with skewed distributions, contingent on the selection of appropriate bin width and thresholds (Fahim et al., 2022; Pei et al., 2021). Most commonly this technique used for detecting data anomalies/outlier using the histograms nature and patterns in univariate data (Fahim et al., 2022). Moreover, the study used the SPC technique for detecting the outlies in IEWQI model. Commonly, this method is used in process monitoring, utilizing control charts and historical data to detect systematic changes or any sudden anomalies at any data point, which could be effective in pinpointing specific sources of data outliers within the system(s) or model(s) (Gessa et al., 2022; Minne et al., 2012). These methods collectively contribute to enhancing the reliability and integrity of research findings in terms of detecting the data outliers.

However, utilizing these advanced techniques could be enhanced the accuracy of the analyses and assure the reliability of the results. Several recent studies have revealed that these methods are useful for validating the detection of outliers and ensuring the stability of the data over time (Aggarwal et al., 2019; Aguilera-Martos et al., 2023b; Aliashrafi et al., 2021; Alsini et al., 2021; Balamurali and Melkumyan, 2018; Berendrecht et al., 2022; Choi et al., 2023b; Etherington, 2021; Fahim et al., 2022; Johannesen et al., 2022; Lee, 2017; Leys et al., 2018; Ottosen and Kumar, 2019; Parra-Plazas et al., 2023; Pei et al., 2021; Rangeti et al., 2015; Shah et al., 2023; Sivarajah et al., 2017; Yin and Fang, 2021; Yuan et al., 2018; Zeng et al., 2023). Therefore, based on the literature, the study utilized these techniques for validating the results of incorporating into ML outlier detection approaches. Details of the each technique are presented in below:

(i) Mahalanobis Distance

Mahalanobis distance is a versatile metric that considers the correlations among variables. It measures the number of standard deviations an observation is away from the mean (Etherington, 2021). Mostly this approach used for dealing with complex inter-variable relationships datasets (Cabana et al., 2021). Data outliers are determined based on high Mahalanobis distances that indicate their distinctness from the rest of the data points (Etherington, 2021; Leys et al., 2018). The research implemented this technique according to the approaches of Leys et al. (2018). The Mahalanobis distance can be defined as follow:

$$D_{M(x)} = \sqrt{\{(x - \mu)^T \Sigma^{\{-1\}(x - \mu)}\}} \quad (7)$$

where x is the data point of WQ indicators being considered, μ is the mean vector of the indicator's dataset, and Σ is the covariance matrix of the input dataset.

Commonly, high Mahalanobis distances indicate that a data point is far from the mean, considering the correlations among variables, while data outliers are identified by setting a threshold on the Mahalanobis distance (Cabana et al., 2021; Etherington, 2021; Leys et al., 2018; Todeschini et al., 2013).

(i) Robust Z-Score

Usually, Z-scores are used to normalize data and assume that data distribution is normal, while robust Z-scores are sensitive to extreme values in a dataset that could be effective in addressing the data outliers or abnormally distributed data points in datasets (Aggarwal et al., 2019; Pei et al., 2021). Instead of using the mean and standard deviation, it employs the median and the median absolute deviation (MAD) (Leys et al., 2013; Owolabi et al., 2021; Yin and Liu, 2022). The MAD provides a robust measure of variability, making the Z-score calculation more resistant to the influence of outliers or non-normally distributed data (Jamshidi et al., 2022; Leys et al., 2013; Prabhakar et al., 2022; Singh and Kundu, 2022). Several recent studies have revealed that robust Z-score approach could be more effective to detect the data outliers compared the typical Z-score (Aggarwal et al., 2019; Leys et al., 2013; Prabhakar et al., 2022). Therefore, the study utilized the robust Z-score approach for further validation of the ML outcomes by adopting the framework of Jamshidi et al. (2022) for computing the robust Z score. It can be mathematically defined as follow:

$$Z_{R(x)} = \frac{|x - Median(X)|}{MAD(X) \times 1.4826} \quad (8)$$

where x is the data point (each WQ indicators) being considered, median (X) is the median value of the dataset, and MAD(X) is the median absolute deviation of various WQ indicators in the given dataset.

(i) Local Outlier Factor (LOF)

LOF assesses the local density deviation of a data point with respect to its neighbours (Alsini et al., 2021; Fredianto and Putri, 2023) while data points with significantly lower density compared to their neighbours are considered outliers or anomalies (Lee et al., 2011; Qiu et al., 2022). Recently many studies have revealed that the LOF is particularly effective for identifying outliers in data that exhibit varying densities, clusters, or subclusters (Auskalnis et al., 2018; Lee et al., 2011; Meenakshi and M, 2022; Wang et al., 2023). . The research used the methodology for estimating the LOF scores in approaching of Alsini et al. (2021). It can be defined as follows:

$$LOF_{(x)} = \frac{\overline{\mu}}{\mu_x} \tag{9}$$

where $\overline{\mu}$ is the average local density of neighbors' indicators, and μ_x is the local density of x^{th} WQ indicator.

(i) Histogram-Based Outlier Detection (HBOS)

To detect the data outliers, recently several studies have utilized the HBOS leverages histogram information method (I. Aguilera-Martos et al., 2023b; Fahim et al., 2022; Pei et al., 2021). Compared to the typical approaches, it constructs histograms for each feature (WQ indicator) and calculates a score based on the combined density of features (WQ indicators) (Berendrecht et al., 2022; Fahim et al., 2022; Kwak and Kim, 2017). Recent a number of research has reported that this method could be efficient and suitable for detecting the data outliers in high-dimensional data (Aguilera-Martos et al., 2023a; Berendrecht et al., 2022; Smiti, 2020). Consequently, the research utilized this approach according to the methodology of Aguilera-Martos et al. (2023a) for obtaining the HBOS scores. It calculates the product of the densities of each feature that can be presented as:

$$HBOS\ Score(x) = \prod from\ i = 1\ to\ d\ [\ P(x_i)] \tag{10}$$

where x_i represents each WQ indicator of the data point, and $P(x_i)$ is the density of the indicator.

(i) Statistical Process Control (SPC)

The SPC is widely used to identify process problems or anomalies in a system, although this technique is not primarily designed for outlier detection. In recent years, a few studies successfully have utilized this approach for detecting outliers or anomalies across various field in long-term monitoring datasets or any systematic process (Knoth and Schmid, 2004; Minne et al., 2012; Qiu, 2020, 2019; Seim et al., 2006; Tegegne et al., 2022). In addition, in recent studies, this technique has been widely adopted to improve data accuracy, identify abnormal attributes in data, maintain process stability, and detect anomalies that may signify outliers or deviations that could potentially enhance understanding and be helpful for making decisions across various domains (Gessa et al., 2022; Minne et al., 2012; Pérez-Benítez et al., 2023; von Rosing et al., 2015; Zhang and Liu, 2019). Usually, the SPC process includes a range of tools and techniques including control charts, run charts, Pareto charts, histograms, box plots, process capability analysis, exponential smoothing, time series analysis, multivariate control charts, capability analysis, process behavior Charts, CUSUM (Cumulative Sum) control charts, and EWMA (Exponentially Weighted Moving Average) control charts (Gorsky, 2020; Jin et al., 2019; Knoth and Schmid, 2004; Minne et al., 2012; Seim et al., 2006). Most studies have revealed that particularly (i) Shewhart Control Charts (X-bar average), (ii) CUSUM Control Charts, and (iii) EWMA Control Charts are effective for detecting the data outlier or any abnormal pattern in both (long-term and short term) monitoring dataset(s) (Boaventura et al., 2022; Gessa et al., 2022; Gorsky, 2020; Jin et al., 2019; Qiu, 2020, 2019; Seim et al., 2006; Zhang and Liu, 2019). Therefore, the research utilized these approaches in order to validate outlier results from IF and KDE algorithms (Baseman, 2020; Boaventura et al., 2022; Gessa et al., 2022; Gorsky, 2020; Jin et al., 2019; Minne et al., 2012; Zhang and Liu, 2019). The research adopted these techniques following the methodology of Minne et al. (2012). Details of the approaches can be found in Minne et al. (2012).

2.6. Comparison the impact of outliers on IEWQI model

To evaluate the impact of data outliers on the model, the study compared IEWQ scores between datasets with outliers and datasets where outliers were eliminated. To substantiate these comparative results, hypotheses were formulated, as disscussed in Section 1. For the purposes of the statistical validation of the comparison results and hypothesis, the research utilized the F-test, incorporating Tukey's Honestly Significant Difference (HSD), because several recent water research studies have used this approach (Festus Biosengazeh et al., 2020; Gessa et al., 2022; Uddin et al., 2022a, 2022b; 2023f, 2023e, 2023h, 2023b, 2023c, 2023d) . The F-test, also termed analysis of variance (ANOVA), evaluates group mean differences among multiple groups, this technique mostly used for the comparison the variance or variability between groups or among groups across various field (Dobie and Wilson, 1996; Mayer et al., 1994; Sureiman and Mangera, 2020; Wilcox, 2003). Contrary, Tukey's HSD test, a prevalent post hoc method, identifies significantly distinct group means when the F-test yields significance. Commonly, the Turkey HSD test utilize for the comparison among datasets/tests/methods/models' outputs (Midway et al., 2020; Nanda et al., 2021; Rouder et al., 2016). In this research follows the approach outlined in Uddin et al. (2023f). Details of the methodology can be found in Uddin et al. (2023f).

It is noted that, for visualizing spatiotemporal data, ArcGIS Pro 3.1.1 was used in this research. All statistical and ML/AI analyses were conducted using the Python programming language within the Google Colab framework, which offers advantages including cloud-based accessibility, pre-installed libraries, powerful hardware acceleration, and seamless integration with Google services.

2.7. Sensitivity analysis

In the domain of WQ modelling, understanding how input parameters influence model outcomes is essential for effective water resources management and policy development (He et al., 2015; Singh and Rashmi, 2014). Sensitivity analysis plays a crucial role in assessing the robustness and reliability of data-driven WQ models, particularly in understanding the intricate relationships between model inputs and outputs. A range of tools and techniques are used for assessing model sensitivity, and the coefficient of determination (R^2) is one of the most widely adopted and effective methods (Chen et al., 2020; Chicco et al., 2021; Hamby, 1995, 1994; He et al., 2015; Suvarna et al., 2022; Zhang et al., 2022). Utilizing R^2 in WQ modelling provides a comprehensive understanding of how input parameter changes affect model outcomes, enhancing model reliability for addressing critical issues in WQ management (Chicco et al., 2021; Hamby, 1995). Commonly, R^2 score ranges from 0 to 1, higher R^2 values indicate a better fit between the model and the data, demonstrating how well the model captures data variation (Chicco et al., 2021; Hamby, 1995, 1994; He et al., 2015; Suvarna et al., 2022; Zhang et al., 2022). In recent several water research studies have widely utilized the R^2 incorporating the ML approaches for assessing the model sensitivity (Ding et al., 2023; Ibrahim et al., 2023; Uddin et al., 2022a; 2022b, 2023b, 2023c, 2023h, 2023e, 2023f, 2023d; Zhang et al., 2022; Zhang et al., 2022). Therefore, the current research follows the established methodology outlined by Uddin et al. (2022a) for leveraging R^2 in analysing IEWQI model's sensitivity. Detailed insights into the methodology can be explored in the comprehensive work of Uddin et al. (2022c).

2.8. Uncertainty analysis

Uncertainty analysis is a critical component of WQ modelling, much like other modelling approaches, especially when utilizing water quality index (WQI) methodologies. Understanding and quantifying the level of uncertainty within a WQI model is an essential step in improving the reliability of WQ models in order to rate the WQ accurately. Several recent studies have revealed that existing WQI models produced a significant amount of uncertainty due to their model architectures (Burić et al., 2023; Ding et al., 2023; Georgescu et al., 2023; Mogane et al., 2023; Parween et al., 2022; Uddin et al., 2023f, 2023h, 2023b, 2023c, 2021). As results, the existing approaches are contributed uncertainty to the final assessment results (Sutadian et al., 2016; Uddin et al., 2021, 2023d, 2023f). To address the estimation of IEWQI model uncertainty, this study used the approaches outlined by Uddin et al. (2023f). To the best of the authors' knowledge, Uddin et al. (2023f) provides the first comprehensive approach to systematically and mathematically compute WQI model uncertainty at each step. Additionally, numerous recent studies in the field of water research have reported the effectiveness of this approach in computing model uncertainty, especially within the context of WQI models (Burić et al., 2023; Ding et al., 2023; Georgescu et al., 2023; Mogane et al., 2023; Parween et al., 2022; Sajib et al., 2023; Uddin et al., 2023f, 2023h, 2023b, 2023c, 2024; 2021. Detailed information about this framework can be found in Uddin et al. (2023f).

3. Results

3.1. Overview of the various WQ indicators in CORK harbour

The research employed boxplot analysis according to the approaches of Kwak and Kim (2017) to assess the distribution and identify outliers within a dataset encompassing various WQ indicators, specifically DOX, MRP, DIN, SAL, BOD, pH, TEMP, TON, and TRAN. This approach has been increasingly adopted in recent studies for the detection and comparison of WQ attributes against recommended guideline values. Recently, several studies have revealed that this could be helpful for identifying the extreme concentrations in the measured values (Dovoedo and Chakraborti, 2015; Li et al., 2016; van Zoest et al., 2018; Zhao and Yang, 2019).

Additionally, the presented boxplots in Fig. 4 provided a visually informative depiction of the central tendencies and the spread of these indicators. This visualization facilitated the precise identification of potential outliers by highlighting data points that fell outside the interquartile range represented by the boxplot whiskers. Notably, with the exception of MRP, SAL, and pH (Fig. 3), most WQ indicators exhibited potential outliers. These outliers signify significant deviations in the concentrations of these indicators, either surpassing or falling below the expected range for Cork Harbour waters. Furthermore, the statistical summary in Fig. 4, coupled with a comprehensive examination of WQ attributes, revealed that the majority of these indicators adhered to permissible limits with the exception for TRAN, DIN, and TON. Similar findings have been reported in numerous recent studies in the literature (EPA, 2022, 2021; Uddin et al., 2023g, 2023d, 2023b). However, the results of the boxplot offered a robust means of identifying and subsequently investigating values that deviated significantly from the expected ranges for each indicator, contributing valuable insights into the WQ dynamics of the studied environment.

Correlation analysis is widely used as an effective approach that could be helpful for understanding the relationship between or among variables, determining patterns and trends that are useful to improve the data, and developing any model(s). Many recent water research studies have utilized this technique for pre-possessing the data in order to develop various WQ models (Haas et al., 2018; Jayaweera and Aziz, 2018; Qian et al., 2024). In addition, this method has been increasingly used in recent studies to assess the influence of various input attributes on model outputs in water research. Therefore, for the purposes of the dependency analysis of the IEWQI score, this research utilized the Pearson correlation technique to investigate the relationship between WQ indicators and the IEWQI score. Fig. 5 presents the correlation results of various WQ indicators between IEWQI scores in Cork Harbour over the study period. From the Fig. 5, the correlation results indicate that most WQ indicators have a significant impact on IEWQI scores, with the exceptions being TEMP and DOX. In comparison to the results, pH, SAL, and TRAN exhibit higher positive influences on IEWQI scores, while the remaining indicators reveal more substantial negative effects. The result of the correlation reveals that water pH, Sal and TRAN have a significant positive impact on maintaining overall "good" water quality. Contrary to this, TON, DIN, BOD5, and MRP show a significant negative impact on overall WQ (IEWQI scores), which indicates these indicators should be monitored regularly. Negative relationships between

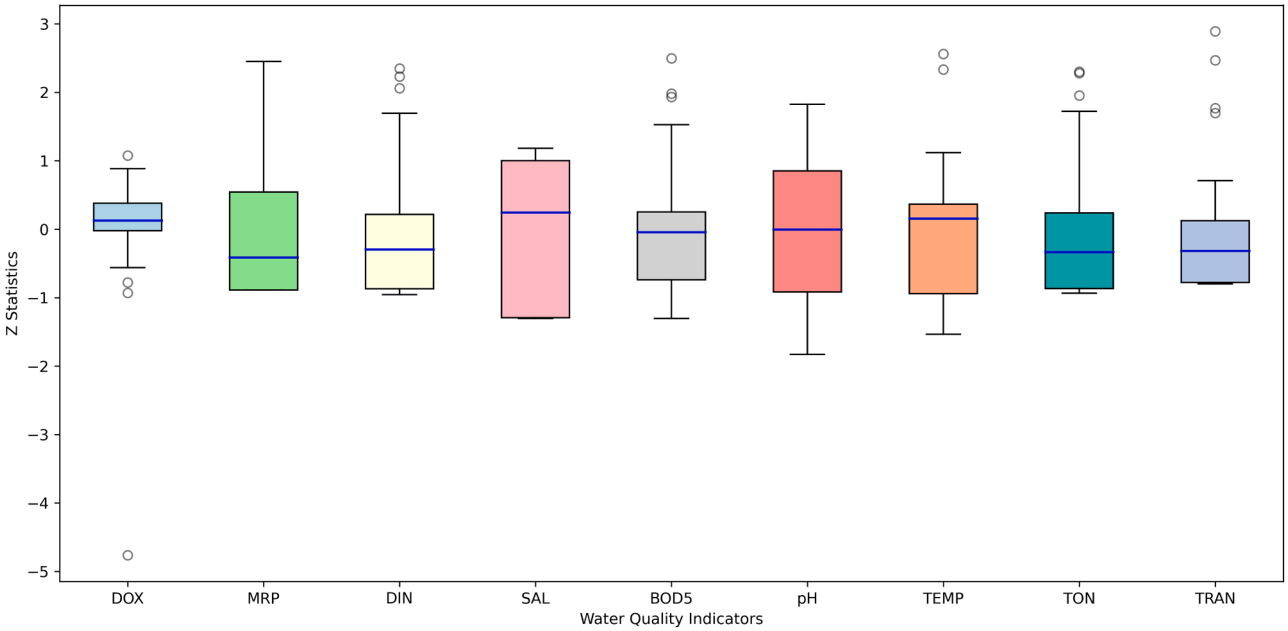


Fig. 4. Statistical summary with data outliers of various WQ indicators in Cork Harbour.

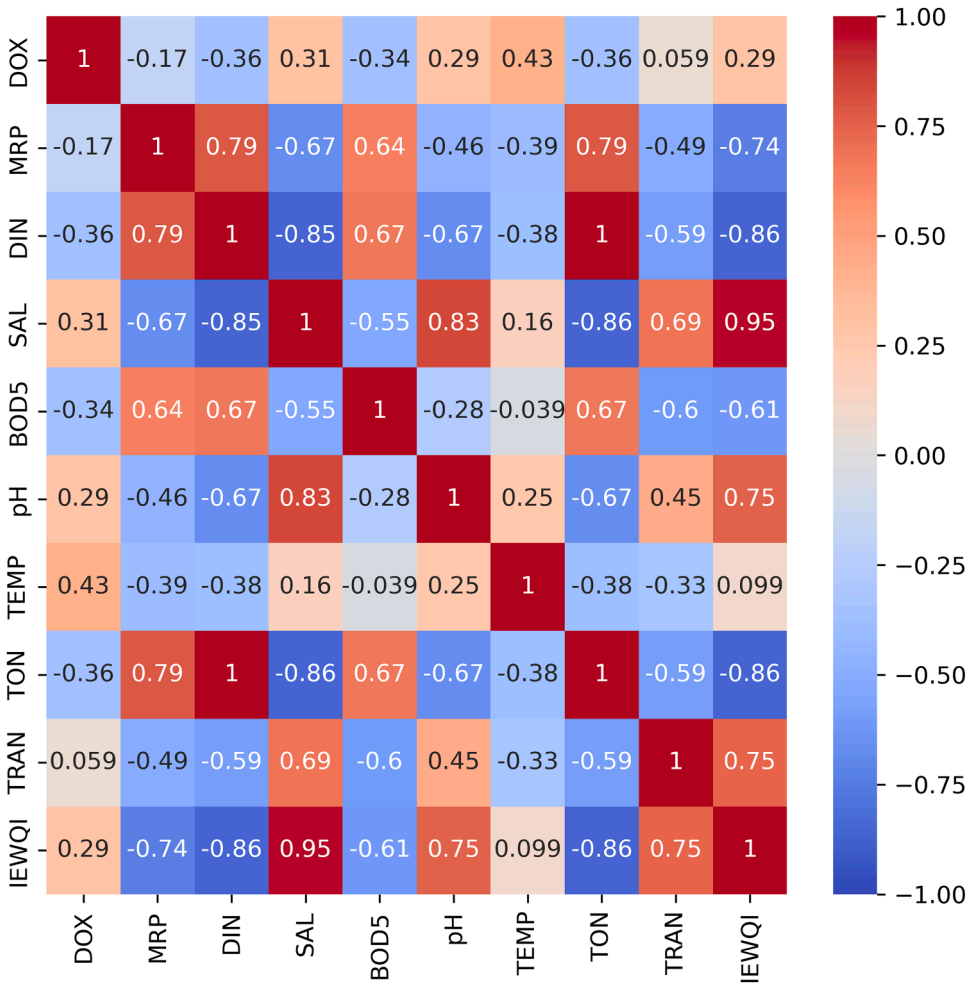


Fig. 5. Pearson correlation of among WQ indicators and between IEWQI scores over the year.

indicators and IEWQI scores suggest that these initiators can be lead overall WQ in Harbour. However, based on the correlation results, it affirms the IEWQI scores' computation using reliable dependencies in indicators.

Furthermore, the research employed SPC techniques, including three well-established Shewhart Control Charts for central tendency monitoring, CUSUM Control Charts for detecting subtle mean deviations, and EWMA Control Charts for effectively tracking trends and shifts over time in various WQ indicators at diverse monitoring sites within Cork Harbour. Through the continuous monitoring of critical parameters such as pH, turbidity, and chemical levels of various indicators such as DIN, TON, MRP etc., SPC aids in the early detection of variations (deviations from standards), maintenance of safety standards, and prompt issue resolution, that could be helpful for maintain "good" WQ status. Figure S2 illustrates the SPC results for various WQ indicators in Cork Harbour. The SPC results indicate a general improvement trend for most WQ indicators, with the exception of SAL and pH, both of which exhibit no discernible trend (Fig. S5; Fig. S7). For example, the correlation results indicate that increasing DIN, MRP, TON, and BOD5 concentrations can lead to higher pollution levels in Cork Harbour. It could be controlled by continuous monitoring these indicators Integration of these control charts (Fig. S2- Fig. S11) establishes a robust and comprehensive monitoring framework for these WQ indicators that could be effective in detecting anomalies in monitoring data and furnishing timely information for proactive WQ management and enhancement in Cork Harbour.

3.2. *Initial screening of data outliers in input dataset*

To facilitate a comprehensive investigation of the relationships between WQ indicators, we employed the Isolation Forest algorithm. This method has gained prominence in recent studies for its effectiveness in detecting data outliers across various fields. In Fig. 6, we present the results of pairwise comparisons of outliers among different WQ indicators in Cork Harbour.

Notably, in Fig. 6(1), when examining the association between DOX and MRP, the study identified a total of 2 outliers (whereas the red dot(s) indicates the data outlier). Remarkably, these outliers exhibited consistency across all pairs of indicators, suggesting a shared source or an underlying correlation among these anomalous data points. These findings offer valuable insights into the interconnections among outliers across diverse WQ parameters. This understanding can play a pivotal role in diagnosing potential issues related to inputs for the IEWQI model, ultimately enhancing its accuracy and reliability, which holds significant academic and practical importance.

3.3. *ML models performance*

In the domain of ML/AI model evaluation, the comprehensive assessment of predictive models stands as a critical endeavour, essential for their effectiveness in real-world applications. A range of meticulously devised evaluation techniques and metrics is available to facilitate a thorough scrutiny of model performance. For the purposes of the detecting outliers, the research utilized two widely used algorithms: (i)

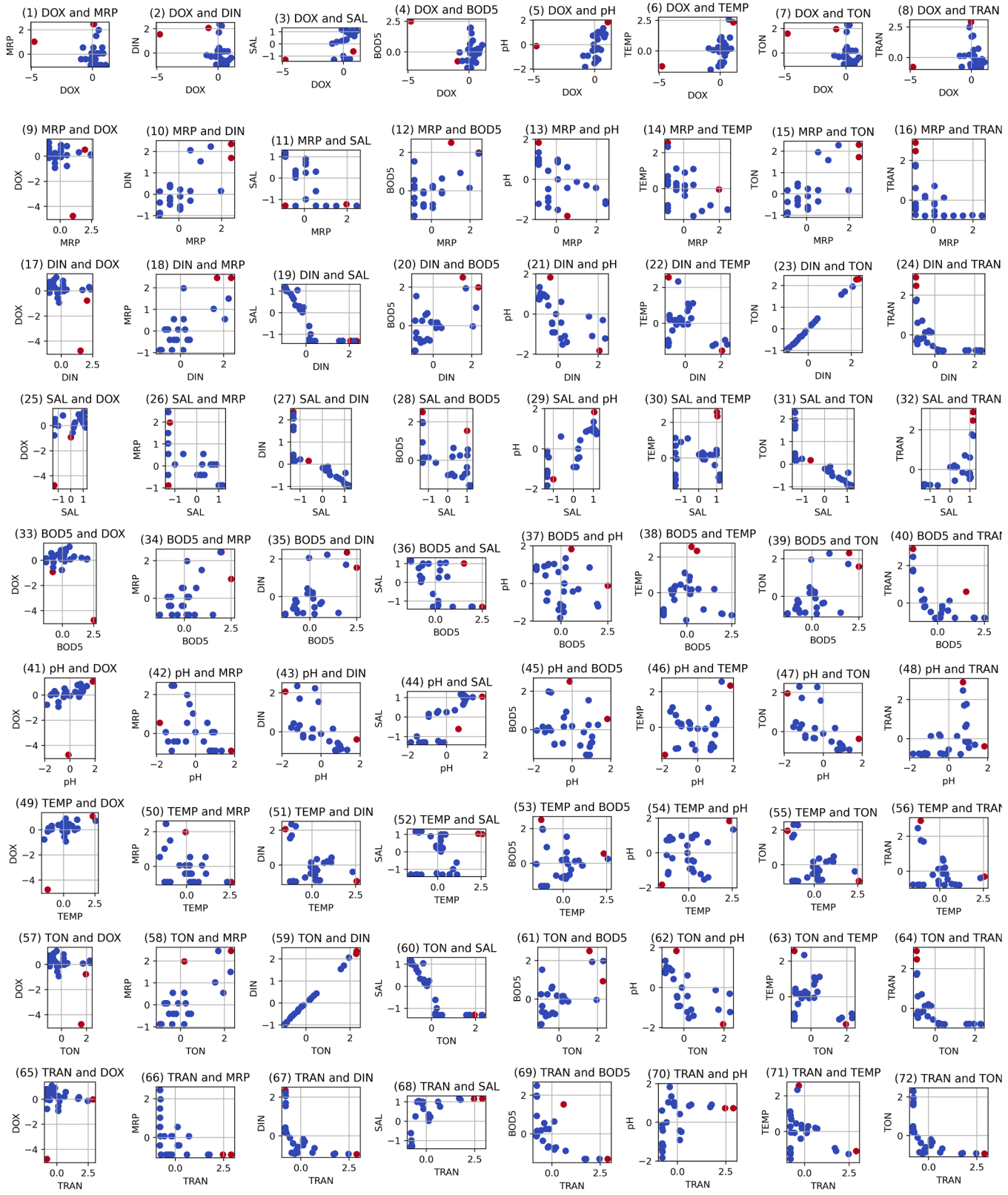


Fig. 6. Pair-wise comparison of outlier in various WQ indicators in Cork harbour.

IF, and (ii) KDE techniques in order to predict IEWQI scores included outliers and after outlier's removal. To evaluate the model performance in predicting IEWQI scores under both (with outliers and without outliers), the research used three widely recognized evaluation metrics: mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). These metrics were selected due to their extensive use in recent water research, specifically in the context of evaluating predictive models while accounting for the influence of outliers—a crucial facet of data pre-processing. Fig. 7 presents the comparative results of the various performance metrics for the both algorithms.

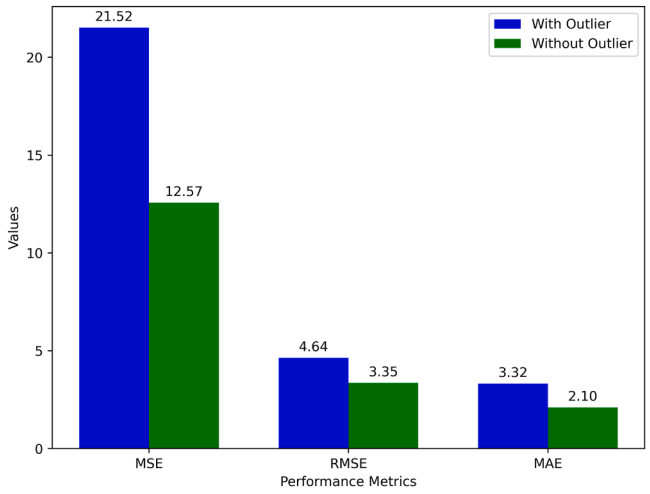


Fig. 7. Comparison of models performance with outliers and after removing outliers.

In the presence of outliers, it can be seen from the Fig. 7, the IEWQI model exhibited performance metrics that included an MSE of 21.52, indicating a notable dispersion of prediction errors, accompanied by a RMSE of 4.64 and a MAE of 3.32. These metrics offer valuable insights into the model's predictive capabilities under such conditions (Ding et al., 2023; Georgescu et al., 2023; Uddin et al., 2022b). However, the removal of outliers led to a significant transformation in its performance profile (Fig. 7). Notably, the MSE showed substantial improvement, registering at 12.57, reflecting heightened accuracy with fewer instances of scattered errors (Fig. 8). Concurrently, the RMSE saw a marked reduction to 3.355, signifying an elevation in prediction precision, while the MAE contracted to 2.1, underscoring consistent accuracy in predictions (Fig. 8).

This comparative analysis effectively underscores the pivotal role that the detection and removal of outliers play in elevating the model's accuracy, thus reinforcing the centrality of these processes in data preprocessing. The results of the evaluation metrics hold particular significance in scenarios where precision in predictions and informed decision-making is of paramount importance, a common requirement in domains such as WQ modelling and related fields.

3.4. Model sensitivity analysis

To assess the model sensitivity, the research utilized the coefficient of determination (R^2), because recently a few studies have used this technique to evaluate the sensitivity of the water model to variations in input data. Commonly, this metric quantifies the proportion of the variance in a model's predictions that can be explained by its independent variables. In this study, we utilize this approaches to evaluate the sensitivity of two distinct outlier detection techniques: the IF Algorithm and KDE function. Fig. 8, and Figs. 9 presents the R^2 results for the IF, and KDE, respectively. It can be seen from the Fig. 8, when the IF algorithm was applied with outliers present, it yielded an R^2 of 0.92 that indicates 92 % of the variability in the model's predictions can be attributed to the independent variables, while the remaining 8 % is either unaccounted (unexplainable) for or attributed to other factors, potentially including the presence of outliers (Fig. 8a). However, upon the removal of outliers, the R^2 increased to 0.95, signifying a substantial enhancement in the model's sensitivity (Fig. 8b).

Similarly, when employing KDE, the R^2 value with outliers present was 0.92, aligning with the IF Algorithm's initial result (Fig. 9a). Yet, after the removal of outliers, the R^2 improved to 0.95, echoing the sensitivity enhancement observed with the IF Algorithm (Fig. 9b). This results of R^2 indicates that the presence of outliers significantly impacted the model's performance, and their removal led to more accurate and sensitive predictions. However, these results underscore the substantial influence of outliers on model sensitivity and highlight the importance of robust outlier detection techniques in refining model performance, particularly when precision in predictions is crucial.

However, the increase in R^2 using both techniques (IF and KDE) revealed that the model's reliability and accuracy improved after outliers were removed from the input. In addition, it indicates that the selected input of IEWQI (indicators) can explain a larger proportion of the overall WQ (IEWQI score) attributes. Moreover, the results of the sensitivity analysis reveal that the performance of the IEWQI model could be enhanced after removing the input outliers from the model in order to rate the accurate water quality.

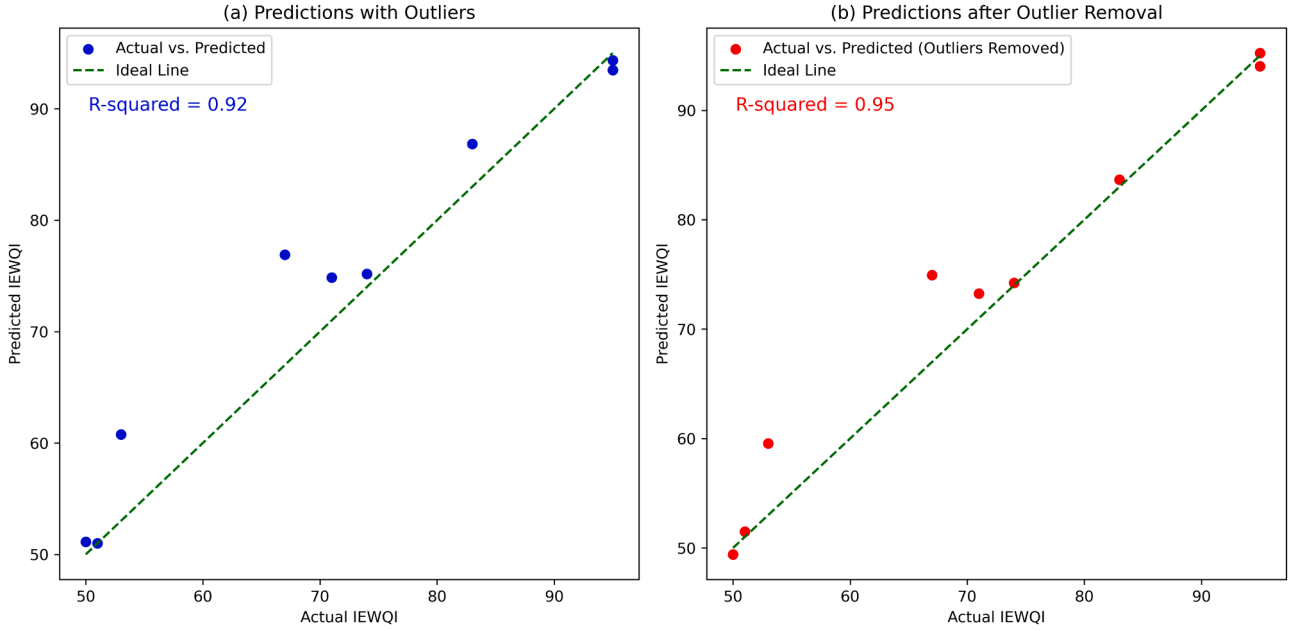


Fig. 8. Performance of the IF algorithms for predicting IEWQI score.

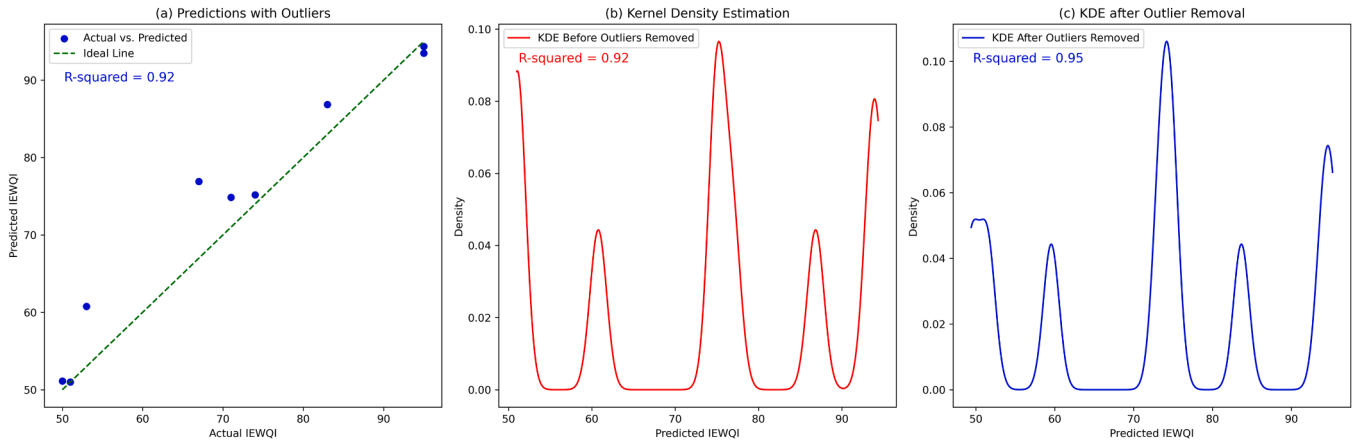


Fig. 9. Performance of the KDF algorithms for predicting IEWQI score.

3.5. Comparison of IEWQI prediction results

The advanced IEWQI model was harnessed in this research to compute WQI scores, with comprehensive details about the IEWQI methodology available in Uddin et al. (2023d). For the prediction of IEWQI scores, two distinct approaches, as discussed in Section 2.4, were employed. Fig. 10 provides a comparative insight into the IEWQI scores resulting from both techniques (IF and KDE), considering datasets with and without outliers at various monitoring sites within Cork Harbour. The findings in Fig. 10 reveal minimal disparities between the actual (computed) IEWQI scores and those with outliers removed across most monitoring sites, except for LE030, LE170, LE450, LE810, and LE820. The negligible changes in IEWQI scores between actual, predicted (with outliers), and predicted (outliers removed) datasets indicate that, in terms of model robustness, the IEWQI model is effective in handling the input data outliers without significantly affecting the model's accuracy. It is noted that the real-world WQ data is varied and can be subject to outlier patterns due to various factors such as measurement errors, extreme events, and existing pressures like agricultural, domestic, etc. However, the IEWQI score had minimal discrepancies between both datasets, indicating that the model is effective for overall WQ assessment and can be utilized as a global tool for monitoring water quality.

3.6. Validation of ML outlier's results

The research employed a range of statistical techniques to investigate the influence of outliers on IEWQI scores across nine WQ indicators: DOX, MRP, DIN, SAL, BOD5, pH, TEMP, TON, and TRAN. These methods are fundamental tools for detecting and assessing outliers and are commonly used in various scientific and data-driven fields. To validate and cross-check the results obtained from machine learning approaches, in this study, we applied nine statistical methods (see details in Section 2.5).

Fig. 11a presents the boxplot analysis, providing a visual summary of computed, predicted (with outliers), and predicted (remove outliers) IEWQI scores, highlighting significant deviations in raw data. Fig. 11b illustrates the histogram analysis, revealing unusual spikes or gaps in score distributions, indicating atypical WQ conditions, whereas Fig. 11c shows the scatter plots, helping to identify monitoring sites where scores did not align with expectations based on contextual information. Fig. 11d displays Quantile-Quantile (Q-Q) plots, exposing subtle deviations from expected statistical patterns, while Fig. 11e visualizes

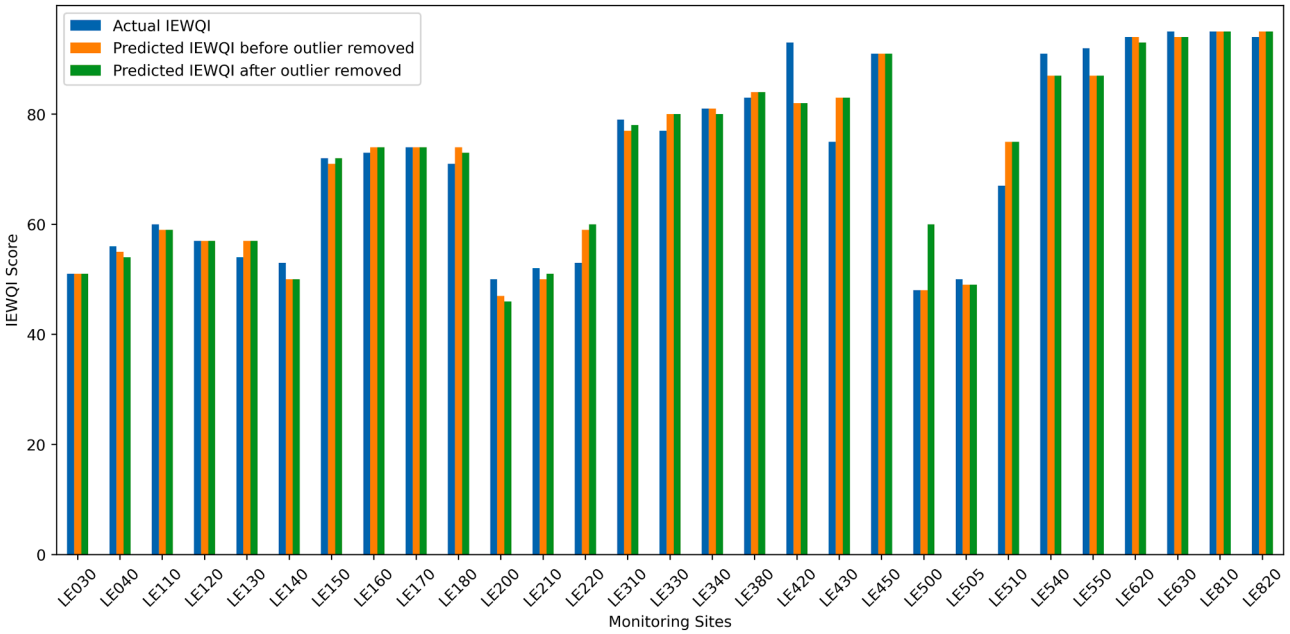


Fig. 10. Point-based comparison between actual, predicting with outliers, and after removing outliers of IEWQI score across various monitoring sites in Cork harbour.

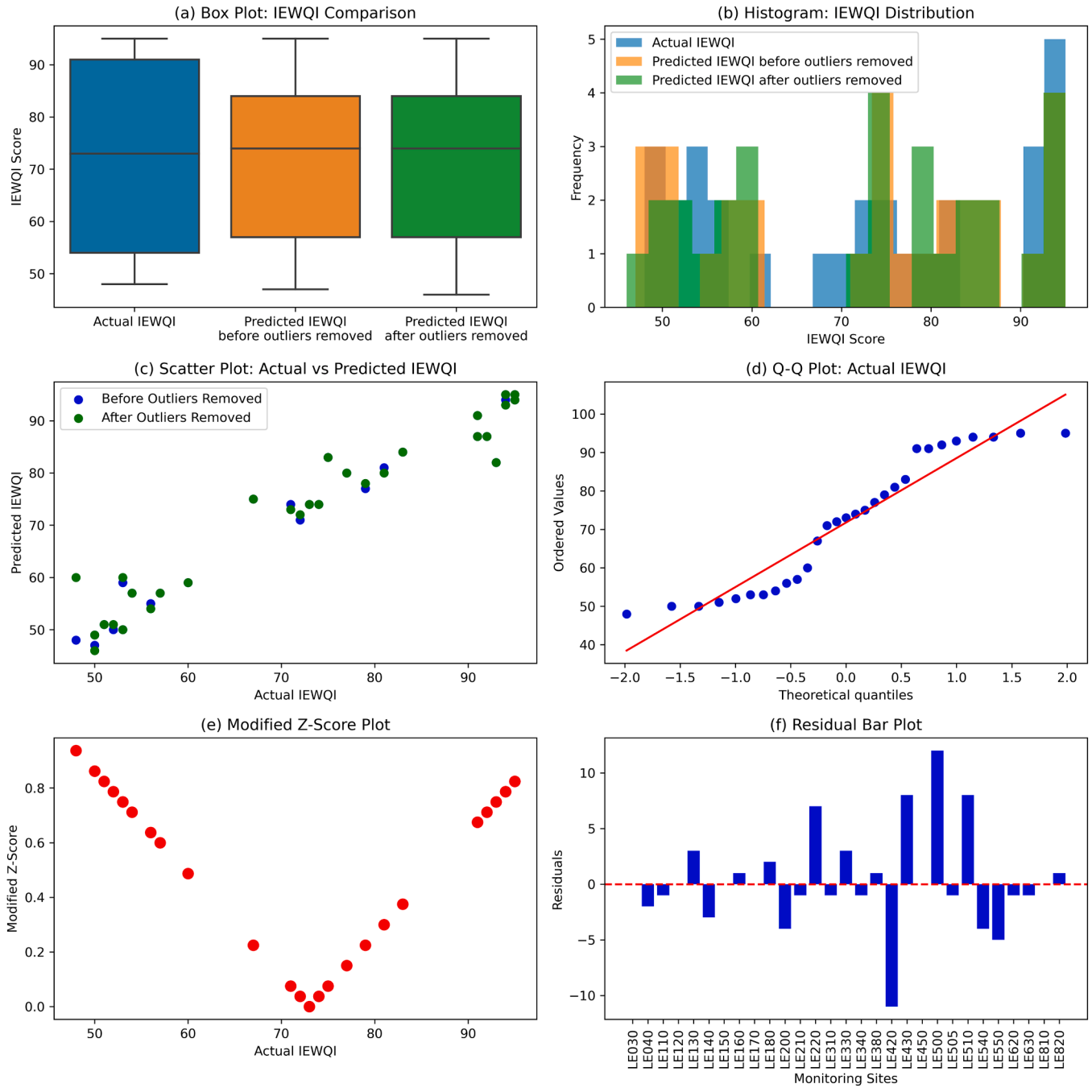


Fig. 11. Outlier's results visualize using various advanced statistical techniques.

modified Z-scores, pinpointing scores significantly differing from the dataset's median. Fig. 11f presents the prediction errors of IEWQI scores at each monitoring site using a residual error bar plot. These analyses collectively offered a comprehensive perspective on outliers, facilitating informed decisions regarding their significance and impact.

Data outliers affect the model(s) in various ways. In most cases, outliers induce bias in parameter estimation, particularly for linear regression models, whereas the extreme data point(s) can lead the model to bias estimation. Consequently, the developed model(s)/tool(s) may not accurately capture the underlying relationships (coefficients) in the data. Notably, the boxplot analysis in Fig. 11a presents statistical summaries for actual (computed), predicted (with outlier), and predicted (remove outlier) IEWQI scores, revealing significant deviations from computed IEWQI scores, indicative of potential outliers in the raw data. Histograms in Fig. 11b illustrated score distributions, showing irregular spikes or gaps that signalled unusual WQ conditions. Scatter plots in Fig. 11c compared computed IEWQI scores with predicted scores (with and without outliers), pinpointing a few monitoring sites with score inconsistencies based on their context. Quantile-Quantile (Q-Q) plots in Fig. 11d displayed slight deviations from expected statistical patterns, while modified Z-scores in Fig. 10e identified scores significantly different from the dataset's median.

To calculate modified Z-scores, the research employed the median and Median of Absolute Deviations (MAD), a robust alternative to the standard deviation, particularly suitable for datasets containing outliers or non-normal distributions. In the computed IEWQI scores, the median score was 73, with a MAD of 1. In contrast, the predicted IEWQI scores (with and without outliers) exhibited a median score of 74 and a MAD of

2 (Fig. 11e), indicating that the typical deviation of data points from the median was 2 units. In comparison, the MAD for actual IEWQI scores was 1.

Finally, residual bar plots in Fig. 11f assessed the performance of IEWQI predictive models, highlighting sites with unusually large prediction errors. Noteworthy, higher prediction errors were observed at monitoring sites LE220, LE420, LE430, LE500, and LE510. The residual results indicate that it is expected that there are higher prediction errors at the highest outlier's data point monitoring sites (Table 2; Fig. 11f). Most statistical models assume that the provided data follows a normal distribution pattern. Data outliers can breach this assumption, as results from other statistical measures like validity tests, confidence intervals, etc. can also be disrupted. It is noted that data outliers also can disrupt or destroy the patterns and trends of the data. These types of attributes of data can reduce the model's prediction capabilities as well as prediction accuracy because models struggle to generalize the model due to hidden patterns or the presence of outliers in the dataset.

Furthermore, the study utilized the power of three advanced statistical techniques, Mahalanobis Distance, Robust Z-Score, and LOF, to determine the presence of outliers at various monitoring sites in Cork Harbour. The outcomes of this outlier detection endeavour are meticulously detailed in Table 2. Notably, the LOF algorithm computed negative LOF scores (−1), with the exception of the LE200 site, suggesting the absence of outliers in its dataset. In contrast, both the Mahalanobis Distance and Robust Z-Score methods detected potential outliers in specific instances. A comparison of these techniques reveals slight variations in their results. In a collective sense, these findings underscore the presence of data outliers in several WQ indicators across most monitoring sites. However, it's important to acknowledge that a handful of sites deviate from this overarching trend, as highlighted by the bold entries in Table 2.

However, the findings from the array of statistical measures underscore the significant influence of extreme values in WQ indicators at various monitoring sites on IEWQI scores. When these techniques are applied within the context of IEWQI scores, they collectively assume a pivotal role in outlier detection. This identification process is instrumental in enabling precise management and the enhancement of WQ in specific areas, thereby providing invaluable insights for targeted interventions. Furthermore, it's worth noting that the results obtained through these advanced statistical approaches align with those from the machine learning outcomes, enhancing the effectiveness of the research conclusions.

Table 2
Outlier's detection results across various monitoring sites in Cork Harbour.

Sites	Mahalanobis distance	Robust Z-Score	LOF Score	Total outliers
LE030	4.264430036	3.012725392	−1	2
LE040	2.722021896	1.217878456	−1	1
LE110	3.206944796	1.209253254	−1	1
LE120	2.613896288	1.217878456	−1	1
LE130	2.655958883	1.124151266	−1	1
LE140	3.892577676	3.372453797	−1	2
LE150	2.481435526	3.519942443	−1	1
LE160	**2.298308742**	**2.293268582**	**−1**	**0**
LE170	**1.663939993**	**0.965870768**	**−1**	**0**
LE180	2.704684441	1.348981519	−1	1
LE200	3.377469814	4.046944557	1	2
LE210	3.711430268	3.276097975	−1	2
LE220	3.29670053	2.233913395	−1	1
LE310	**1.822379878**	**0.789647718**	**−1**	**0**
LE330	2.375760935	2.518098835	−1	1
LE340	**2.38813437**	**1.487847264**	**−1**	**0**
LE380	2.655402387	0.832771258	−1	1
LE420	4.268769366	2.908050324	−1	2
LE430	4.115057178	3.147623544	−1	2
LE450	3.955184662	1.595746431	−1	1
LE500	5.138108549	16.26871712	1	2
LE505	3.227999724	4.046944557	−1	2
LE510	2.870098421	0.674490759	−1	1
LE540	2.805895524	1.093574351	−1	1
LE550	**2.491468557**	**0.764422861**	**−1**	**0**
LE620	2.086780424	3.027273232	−1	1
LE630	2.101291728	2.924115822	−1	1
LE810	3.199304842	4.65398624	−1	2
LE820	2.61630713	4.039009371	−1	2

3.7. Results of uncertainty in assessing impact of outliers on IEWQI scores

In our pursuit of estimating the uncertainty associated with our predictive models, this study utilized the inferential error bar techniques with a 95 % confidence interval according to the approach of Uddin et al. (2023f). Fig. 9 shows the results of the 95 % confidence interval error bars at each monitoring site in Cork Harbour, providing valuable insights into the inherent uncertainty of the predictive models concerning IEWQI scores. Table 3 presents a comparative analysis of the statistical summary that was conducted to assess model uncertainty using a 95 % confidence interval, with the significance level (alpha) set at $p < 0.000$, and the degrees of freedom for this analysis determined to be 28. Through a meticulous process, the study computed confidence intervals for the expected IEWQI scores following the precise inclusion of outliers and their removal from the dataset.

Fig. 12 presents that there were no significant differences in IEWQI scores between the actual, predicted (with outliers), and predicted (outliers removed) data across various monitoring sites in Cork Harbour. Moreover, the length of the error bars mostly appears shorter, indicating minimal variation. These results suggest that there is no significant uncertainty associated with IEWQI scores in both models, with outliers and after their removal from the input data. Furthermore, Table 3 reveals that there were no significant deviations in IEWQI scores from the mean of the actual (71.76 ± 16.86), predicted with outliers (71.72 ± 16.48), and predicted with outliers removed (72.10 ± 15.92). Notably, the standard deviation of IEWQI scores slightly decreased after removing outliers from the input data, suggesting that the IEWQI model could introduce less than 1 % uncertainty in assessing and predicting water quality.

To validate the error bar results, t-statistics were utilized. Table 4 presents the *t*-test results of IEWQI models for the comparison of different datasets. These results offer valuable insights into the comparison between 'Actual IEWQI' values and two sets of predicted values: 'Predicted IEWQI before outlier removal' and 'Predicted IEWQI after outlier removal.' For the first comparison, the *t*-test yielded a t-statistic of approximately 0.050 and a corresponding *p*-value of about 0.961. These values indicate that there is minimal difference between the means of the 'Actual IEWQI' and 'Predicted IEWQI before outlier removal' datasets, and this difference is not statistically significant. Similarly, for the second comparison, the *t*-test produced a t-statistic of approximately −0.423 and a *p*-value of around 0.676, once again signifying negligible differences between the 'Actual IEWQI' and 'Predicted IEWQI after outlier removal' datasets, without statistical significance. In both cases, the *p*-values are considerably greater than the conventional significance level of 0.05, suggesting that any disparities

Table 3
Comparative analysis of statistical summary for model uncertainty with 95 % confidence interval at $p < 0.000$ whereas degree of freedom was 28.

Statistical attributes	IEWQI Scores		
	Actual	Predicted	
		with outliers	remove outliers
mean	71.76	71.72	72.10
Std.	16.86	16.48	15.92
min	48.00	47.00	46.00
25 %	54.00	57.00	57.00
50 %	73.00	74.00	74.00
75 %	91.00	84.00	84.00
max	95.00	95.00	95.00

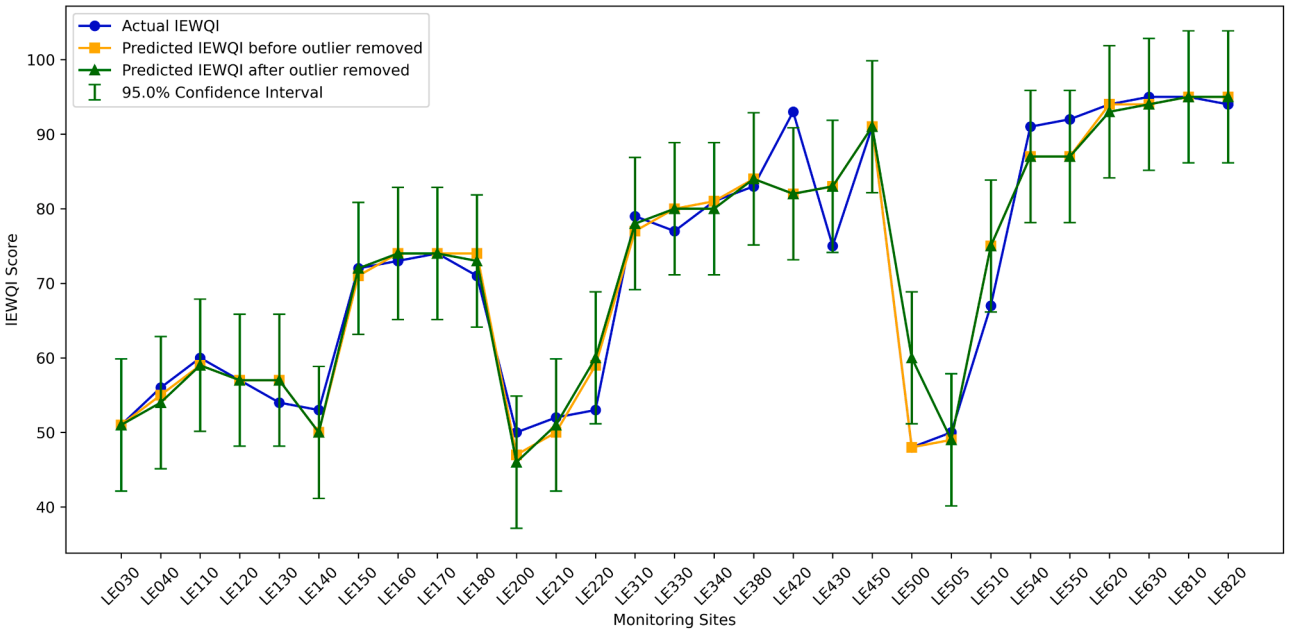


Fig. 12. Comparison of IEWQI scores (computed with predicted including outliers and without outliers) means with 95 % CI bars when n is 29, $p < 0.000$.

Table 4
t-test results for comparing the predicting accuracy of IEWQI model.

Input scenarios	t-statistic	*p*-value
Actual vs. Predicted (with outliers)	0.04995	0.9605
Actual vs. Predicted (remove outliers)	−0.42263	0.6757

observed are likely due to random variation rather than meaningful distinctions in the data.

However, considering the results of model uncertainty under different scenarios, it is evident that if the analysis were to be conducted repeatedly on numerous samples drawn from the same population, we could expect the actual mean IEWQI score to fall within this interval for approximately 95 % of those samples. In terms of reliability, the results indicate that the IEWQI model could be highly effective in predicting and assessing marine waters with minimal bias, contributing to less than 1 % of uncertainty in the assessment.

For the purposes of testing the hypothesis, Tukey's Honestly Significant Difference (HSD) comparison analysis was utilized to compare the different datasets obtained from the IF and KDE approaches (predicted IEWQI scores with and without outliers), which is an effective analysis to identify the differences among datasets. Tukey's HSD test is widely used advanced statistical approach for multiple comparisons (Midway

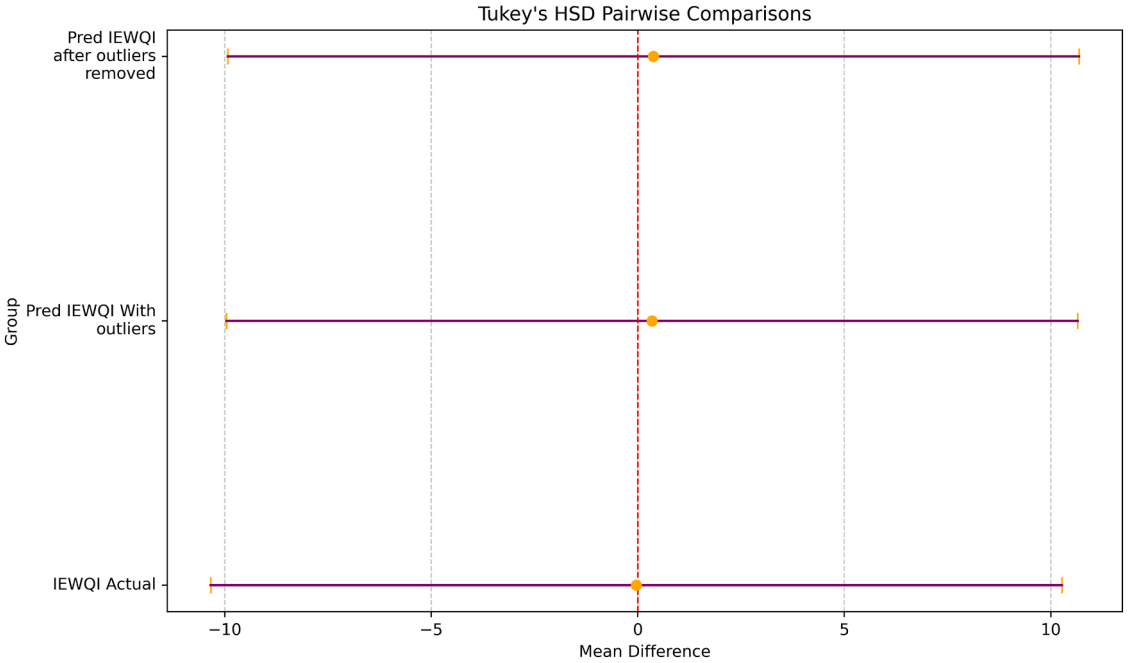


Fig. 13. Comparison of outliers impacts on IEWQI scores among three approaches with 95 % confidence interval from Turkey HSD analysis (the vertical dashed line indicates the point where the difference between the means is equal to zero or similarity of both approaches statistical attributes, the refers to the means are equal of both techniques.

et al., 2020; Nanda et al., 2021). It could be effective to identify the significant differences between or among models, techniques/methods/datasets (Midway et al., 2020; Nanda et al., 2021; Rouder et al., al.,2016; Uddin et al., 2023). This technique is widely adopted in various scientific fields for the comparison of various test results (Esnaola et al., 2018; Kim, 2015; Lee and Lee, 2018; Midway et al., 2020; Nanda et al., 2021; Rouder et al., 2016). Recently, several water research studies have utilized this approach for the comparison of various WQ models (Uddin et al., 2023f, 2022a, 2022b, 2023h). Fig. 13 presents the comparison results among three approaches. It can be observed from the figure that there were no statistically significant differences ($p < 0.05$) among the three techniques at a 95 % confidence level, with $F = 0.0047$ and df = 2. The mean IEWQI scores showed variation ranging from +0.3793 to −0.0345 (Fig. 13). Based on the results of the ANOVA analysis, the null hypothesis was rejected, indicating that input data outliers do not significantly impact the IEWQI scores.

3.8. Comparison of WQ status

For the purpose of assessing WQ status, a novel classification scheme (Table S4) utilized in this study. Fig. 14 provides a statistical summary of the WQ status, while Fig. 15 presents spatial distribution of IEWQI scores and point-based representation of WQ status across various monitoring sites in Cork Harbour. Both approaches classified WQ into "good," "fair," and "marginal" categories. Recent research has consistently reported similar WQ patterns across various monitoring sites in Cork Harbour (EPA, 2022; Uddin et al., 2023b, 2023d). When comparing these quality states among computed IEWQI, predicted (with outliers), and predicted (remove outliers), slight variations were observed. In terms of reliable assessment of water quality, this variability of WQ classes has a practical impact on the correct rating of the water quality. Due to the inaccurate assessment of water quality, it may influence the decision-makers to take proper initiative to manage water resources. In the case of actual scores, 34 % (10) of sites fell into the "good" class, 62 % (18) into the "fair" class, and 3 % (1) into the "marginal" class, respectively. Comparatively, minor differences were found in the results of the other approaches.

Table S5 provides a comprehensive comparison of the number and

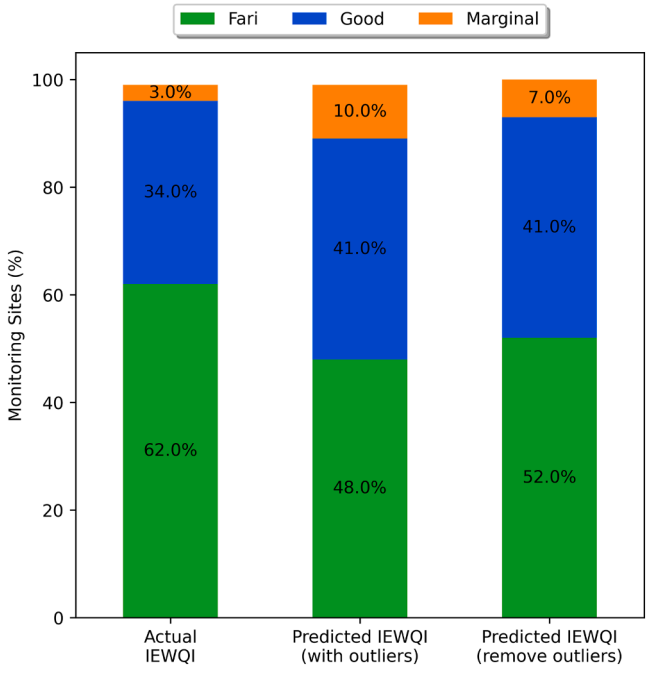


Fig. 14. Comparison of WQ rating between actual (computed), with and without Outliers.

percentage of sites categorized as "Good," "Fair," and "Marginal" across different models. The comparison highlights the distribution of sites across various status categories (Good, Fair, and Marginal) for the Actual IEWQI, Predicted IEWQI (with outlier), and Predicted IEWQI (remove outliers) models. It's evident that each model predicts a different number of sites within each status category, resulting in variations in the percentage distribution. For example, the "Fair" category comprises 18 sites (62.07 %) in the Actual IEWQI model, 14 sites (48.28 %) in the Predicted (with outlier) model, and 15 sites (51.72 %) in the Predicted IEWQI (remove outliers) model, indicating that input data outliers had a slight impact on the model-predicted WQ classifications. As depicted in Fig. 15, the spatial variation of IEWQI scores and the influence of outliers on WQ classifications are evident. In general, most monitoring sites exhibited similar WQ states compared to the computed ratings, with the exception of LE500 and LE505 (Table S5).

Fig. 15 highlights that in the case of computed classes, only "marginal" WQ was assigned to monitoring site LE500 (User ID 21). This classification remained consistent when using data with outliers. However, a notable difference (its shows "fair") emerged after removing outliers from the input data. Additionally, LE505 also received a "marginal" rating for both datasets, while the computed rating was "fair" (Fig. 15d, f). A comparison of the newly assigned ratings with and without outliers clearly indicates that these sites exhibited severe eclipsing problems, following the approach proposed by Uddin et al. (2022a). By comparing the determined rating of water quality, and the indicators guidelines values, it can be seen from Table 1, in the cases of LE500 and LE505 respectively four indicators (TRAN, DIN, DOX, and TON), and (TRAN, DIN, TON, and MRP) has breached for both sites. According to the methodology of Uddin et al. (2022a), both sites should be ranked "poor" categories, but models ranked it "fair" to "marginal" classes. It seems that both models (actual, predicted with and without outliers) have significantly suffered overestimation problems. According to the classification schemes, as provided in Table S4, if the "Fair (IEWQI scores = 50−79)", and "marginal (IEWQI scores = 30 −49)" schemes score intervals revised according to the finding of this research, the eclipsing problem may be resolved from the IEWQI approach that could be effective for assessing or rating the WQ accurately and final assessment would be more reliable in terms of their actual observation. However, the suggested revision of the classification schemes would be updated, and the IEWQI model could be utilized for assessing the WQ and monitoring more accurately in terms of the presence of data outliers. This approach could be globally acceptable for rating WQ as well as a potential tool for sustainably managing water resources.

However, a comprehensive analysis of WQ status from three different aspects suggests that the classification of "fair" and "marginal" should be revisited and updated to address this issue, incorporating outlier detection techniques. Furthermore, the results reveal that data input outliers had no significant impact on the IEWQI model architecture, indicating that the model remains effective in rating marine waters while optimizing data input anomalies.

4. Discussion

The research was conducted to investigate the data outliers' impact on the recently developed data-driven Irish Water Quality Index (IEWQ) model. Recently several studies have revealed that data outliers or anomalies have a significant impact on model performance. Consequently, the data outlier's treatments/solutions gradually increased in data-driven modelling approaches (Lee, 2017; Liang et al., 2022; Orouji et al., 2013). To date, a range of statistical and mathematical tools and techniques developed for detecting data outliers/anomalies in datasets (Choi et al., 2021; Garces and Sbarbaro, 2009; Gui et al., 2017; Ha et al., 2014; Misra et al., 2020; Shah et al., 2023). Recently, state-of-the-art ML/AI technology is widely used to detect the data anomalies/outliers in high-dimensional datasets across various fields in order to improve the data accuracy for the decision-making process (Duraj and

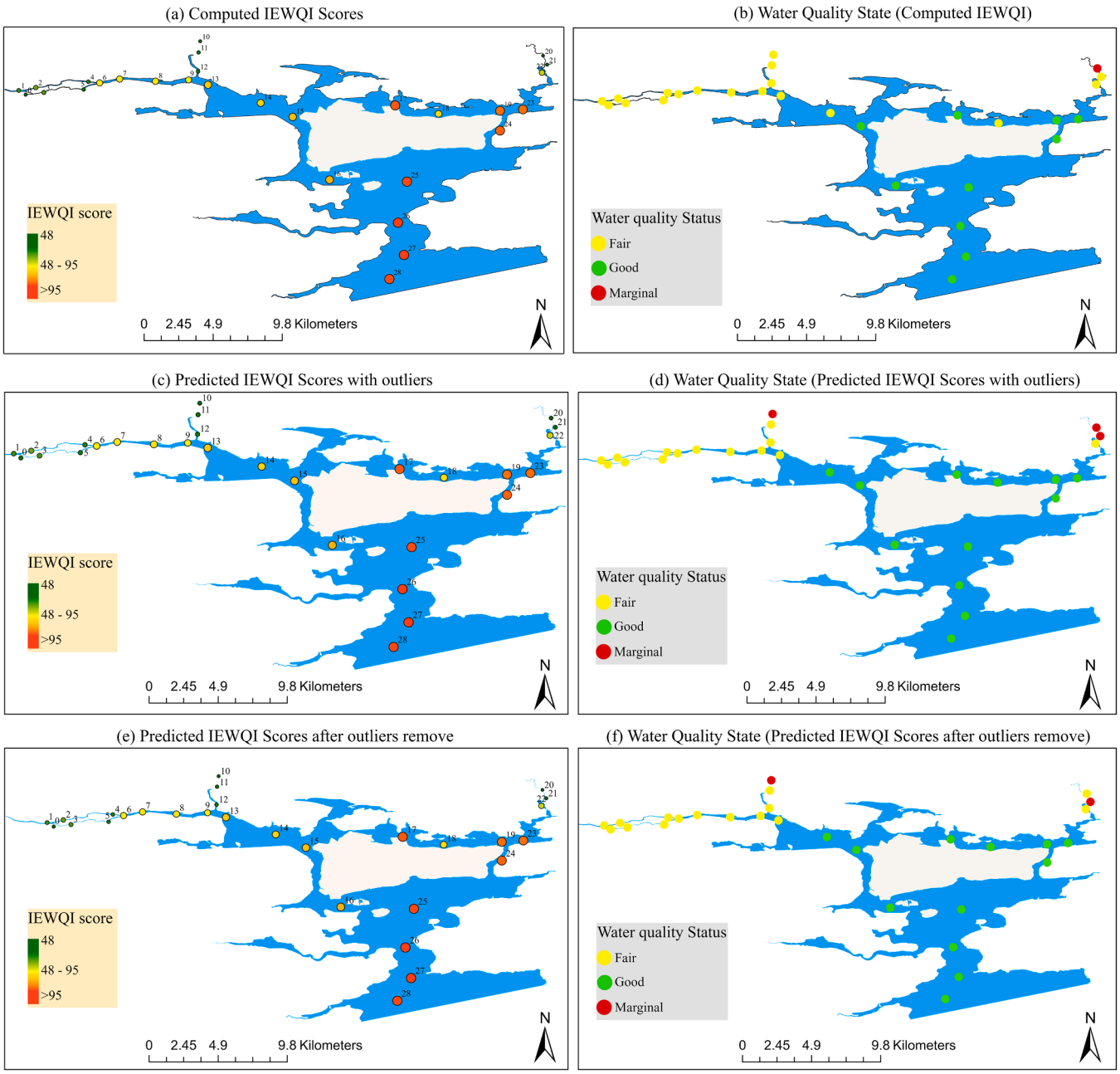


Fig. 15. Comparison of spatial distribution of IEWQI scores and WQ status among actual IEWQI, Predicted (with outliers), and Predicted (remove outliers) across different monitoring sites in Cork Harbour.

Szczepaniak, 2021; Hansen et al., 2023). Specially, long-term monitoring database should require investigating the data patterns, hide structure of data and trend of data. The presence of the data outliers or extreme values or anomalies disrupted the data patterns and trend of the historical datasets (Ojo et al., 2022; Smiti, 2020). Therefore, it is essential to detect the data outliers or anomalies. For the purposes of detecting data outliers in input of the IEWQI model, the study was utilized the IF and KDE algorithms, because several recent research especially focusing water research studies have reported that these techniques outperformed compared to others approaches in terms of dealing monitoring water quality datasets (Liu et al., 2020; Yin et al., 2023; Wang et al., 2023). Therefore, the research adopted these approaches for detecting data outliers/anomalies in water quality monitoring dataset in Cork Harbour, Ireland – as a case study.

According to the IEWQI model architecture, eight water quality indicators (see Section 2.2) used as model inputs. By comparing the WQ indicators' measured concentration with regarding guideline values (see Table 1), most indicators found within the guideline values except for the TRAN, DIN, and TON (Table S4). For the purposes of initial investigation for determining the individual WQ indicators, the study utilized the boxplot, correlation and SPC analysis techniques, because recently several research used these approaches to assess the data distribution and detect outliers in various WQ indicators. The boxplot analysis results revealed potential outliers in most indicators, while significant data deviation was found from the expected values (Fig. 4). The Pearson correlation results also highlighted the influences of these indicators on the overall WQ (IEWQI scores), emphasizing their role in assessing WQ in Cork Harbour (Fig. 5). Additionally, statistical process control (SPC) techniques were employed to monitor trends and shifts in water quality indicators over time. Most indicators exhibited improvement trends, with a few exceptions (Fig. S2 – Fig. S11). The SPC results revealed that particularly DIN, MRP, TON and BOD5 should be monitored frequently

for maintain the “good” water quality status in Cork Harbour. The results of various WQ indicators are in line with previous studies in Cork Harbour (Uddin et al., 2022a, 2022b, 2023d, 2023g).

For further investigation of the data outliers in water quality datasets, the IF and KDE ML algorithms utilized to predict IEWQI scores with and without outliers. The removal of outliers significantly improved model performance, reducing MSE (21.52, and 12.57), MAE (3.32, and 2.10), and RMSE (4.64, and 3.35) respectively for prediction model with data outliers and without outliers (Fig. 7), demonstrating improved accuracy and sensitivity (R^2 from 0.92 to 0.95) after outlier removal (Figs. 8 and 9). Sensitivity results revealed that both (IF and KDE) techniques could be effective to remove data outliers significantly from the IEWQI model input in order to enhance model performance for rating water quality accurately. It’s also reported that the IEWQI model could be explained the more 95 % of variability of the input features (Fig. 8; Fig. 9). In literature, recent several studies have also reported similar results for detecting data outliers in WQ datasets (Jiang et al., 2022; Panjei et al., 2022; Piñeiro Di Blasi et al., 2015; Talagala et al., 2019; Tang and He, 2017)

Comparison of the IEWQI scores among models outputs (actual, predicted with and removal outliers), the IEWQI model exhibited minimal discrepancies among them, indicating its effectiveness in handling input data outliers (Figs. 10 and 12). In addition, the study also assessed the impact of outliers on water quality assessment using a range of statistical techniques as discussed in Section 2.5; revealing their influence on IEWQI scores. All statistical tools revealed that a slight improvement (nearly 3 % of explainable features increased) of the model after removal outliers (Figs. 8 and 9). The research also assessed the uncertainty of the model under various scenarios (actual, predicted with and without outliers), the uncertainty results showed minimal uncertainty associated with IEWQI scores (<1 %) for both models, even in the presence of outliers, reinforcing the reliability of the assessment (Fig. 12). These results are in line with those of previous studies on IEWQI model in adopting various domain in the world (Ding et al., 2023; Manna and Biswas, 2023; Sajib et al., 2023; Uddin et al., 2023d;2023f; 2023h).

However, the ultimate goal of the IEWQI model is to rate WQ. By comparing, both models were classified WQ into "good," "fair," and "marginal" categories. Minor differences were observed across various monitoring sites in Cork Harbour through these categories when considering input data outliers except for the LE500 and LE505 (Table S4). The results of rating WQ indicates the challenges in accurately rating water quality with data outliers. The results also revealed that the model had suffered the severe eclipsing problem (over estimation) at particular sampling sites. Based on the analysis, the study suggested revisiting and updating the classification schemes, especially for the "fair" and "marginal" categories, to address the issue of outliers and improve the accuracy of water quality assessment using the IEWQI approach.

Furthermore, the research compared the results obtained from machine learning and statistical methods to validate the outlier detection process and their impact on the IEWQI model. The findings from both approaches aligned, and revealed that effectiveness of IEWQI model for assessing the WQ considering the input data outliers. Therefore, the results and findings of the research are concluded that removal outliers from the model input can enhance the model accuracy. Although, the study considered only short term WQ dataset for investigating the data outliers, in future research should be focused using long-term and spatial variability of WQ datasets. However, this comprehensive analysis of WQ indicators, outlier detection, model performance, and classification schemes highlights the importance of outlier removal for accurate WQ assessment. The study’s findings contribute to improve WQ monitoring and management practices, emphasizing the need for revisiting and updating classification schemes to address outlier-related challenges in IEWQI model for generalized application across global aspects.

5. Conclusion

Data-driven model(s) are significantly impacted by data outliers. In recent years, treatment and remedies for it have drawn a lot of attention. As a result, numerous methods and instruments, encompassing statistical, mathematical, and empirical techniques, have been created thus far to identify data outliers or abnormalities within datasets. Recently, the state-of-the-art technology of ML and AI has been widely used to detect data outliers across various fields. Several recent studies across various fields, including water research, have reported that data outliers have a substantial impact on the performance of the model. Therefore, it is essential to investigate the data outliers/anomalies in any data-driven model(s) that can be helpful for identifying hidden information, data patterns, and trends, especially in long-term monitoring datasets, in order to improve the data quality as well as increase the model performance. As part of the advancement of the world’s first data-driven “Irish Water Quality Index (IEWQI) model,” which is widely used for assessing and rating water quality, particularly that which is designed for marine waters. Hence, the aim of this research was to investigate the impact of data input outliers on the output of the data-driven IEWQI model. Being a data-driven WQI approach and the first systematic mathematical tool for assessing and monitoring marine waters, it is crucial to analyse its response to various outliers in WQ indicators. A number of research objectives were considered in order to obtain the research goal (see Section 1). For the purposes of the research aim, a case study was conducted in Cork Harbour, Ireland. This involved incorporating advanced two ML algorithms (IF and KDE) to detect data outliers and predict WQ within Cork Harbour. For the validation purposes of the ML outcomes, the research also utilized a series of advanced statistical and mathematical tools and techniques. From the research results, several key findings are outlined below:

- The analysis demonstrated a significant improvement in the coefficient of determination (R^2), increasing from 0.92 to 0.95 when data outliers were removed from the model input. This improvement suggests that outliers have a substantial impact on the predictive performance of the IEWQI model. The results of the model’s sensitivity highlight the importance of effectively identifying and treating oultiers when assessing water quality using data-driven approaches. The findings also place emphasis on the data pre-processing in developing the prediction model(s) for water quality. Therefore, the results can be supported to the environmental managers that could be utilized, enhancing the model’s accuracy to assess water quality accurately, leading to more informed decisions regarding environmental management strategies.
- A comparative analysis of IEWQI scores revealed that there were no significant differences in IEWQI scores among the three techniques (actual, predicted with and withour outliers), despite the presence of data outliers in model’s input, suggests that the IEWQI model architecture remains robust and resilient to variations in input data. This findings indicates that although outliers can affect specific data points, the overall assessment of WQ that the IEWQI model provides is not significantly alter by them. This findings affirms that the IEWQI model could be effective reliably assess WQ even in the presence of outliers in input attributes, providing consistent and accurate results that may contribute to sustainable water resources management. End-users/stakeholders/environmental managers can be utilized the IEWQI model for assessing WQ in any geographical extent with confidence in the model’s ability to generate reliable assessments regardless of data outliers/anomalies. It should be noted that the WQ indicators attributes and their anomalies can differ in terms of geospatial resulotion of the domains.The interpretation of scores (translated into the rating WQ) within the three approaches showed a slight difference, indicating that outliers indeed play a role in influencing the rating of WQ. This variation in categorization highlights the sensitivity of WQ ratings to the presence of outliers in

the model data. The findings of the rating variation underscore the importance of considering outliers in the assessment process to ensure consistency and accuracy in categorizing WQ. This finding can be helpful for environmental managers/policymakers to handale data outliers with extra care in rating WQ more accurately.

- Additionally, the results of newly assigned ratings, with and without outliers, clearly indicated that a few monitoring sites exhibited severe eclipsing problems due to the impact of outliers, have practical implications for WQ assessment as well as aquatic environment management. Therefore, the study suggests that the "fair" and "marginal" categories of the rating schemes for the IEWQI model should be revised in terms of data reliability and updated to ensure bias free assessments. In addition, by updating rating schemes to reflect more accurate assessments of WQ, environmental managers/policymakers can mitigate the risk of misinterpretation (bias free) and ensure more informed decision-making regarding water resource management and its pollution control measures. This findings supports to understanding of the challenges associated with WQ assessment with data outliers, emphasizing the require for adaptive rating framework to further advancement of the IEWQI model reliability with evolving data dynamics attributes.

Future research in WQ assessment should prioritize integrating spatio-temporal data and exploring the model's sensitivity to using various factors including hydrodynamics attributes of domains. It is important to note that although the research utilized short-term WQ data (only one year of water quality WQ data considering average of each indicators), future research should consider a range of spatio-temporal data (long-term monitoring data) to investigate the model's sensitivity to input outliers in terms of spatio-temporal resolution of water bodies. The research also recommends, by incorporating comprehensive data from additional monitoring sites and employing advanced modelling techniques that could be effective to understand of how WQ varies across different geo-spatial resolution and over time. In addition, it could be potentially benefited for further improving model should investigating the model's sensitivity to other WQ indicators like biological WQ indicators like faecal coliform, algae information etc., and anthropogenic activities such as different pressures - agricultural, domestic, industrial etc., that can help identify key factors of WQ dynamics and improve the model's predictive accuracy. Furthermore, exploring hybrid modelling approaches that combine IEWQI model with hydrodynamics models can lead to more holistic analyses and facilitate more effective environmental management strategies. Overall, using the research findings and considering these highlighted areas in future research endeavours would be contributed to the advancement of more accurate and reliable WQ modelling tools for monitoring and managing water resources.

However, the study highlighted the critical impact of outliers on water quality assessment, emphasizing the necessity for robust outlier detection methods in environmental modelling including WQ models like IEWQI. By revealing the significant influence of outliers on the IEWQI model's performance, it recommends for revised and updating existing rating schemes to enhance reliability in order to assess accurate WQ. The research findings and results clearly demonstrated how data outliers affect the model performance that could be utilized widely in any environmental modelling approaches including predicting models to further advancement of the sustainable environment management along water resources. Moreover, the findings of the research may support to identifying the outlier-related challenges in predictive modelling to obtain reliable information using data-driven approaches.

In conclusion, this study underscores the efficiency of predictive models in WQ evaluation while acknowledging the need for ongoing research to address limitations and unravel the complexities of Cork Harbour's aquatic environment. The research findings have profound implications for sustainable WQ governance, contributing to the preservation of this vital ecosystem for future generations. These insights also have broader relevance in the field of environmental science, bridging the gap between predictive modelling and effective ecological conservation. Moreover, the findings of the research highlighted the importance of ongiong research in advancing the existing WQ models to affirm their effectivnees in sustainable environment management with focusing water resources. In addition, the research also revelas that the addressing the impact of data outliers on environmental modelling such as IEWQI model, the proposed framework could be an effective approaches for furtherther improvement of environmental moddeling approaches in terms of increasing accuarcy and reliability. However, these findings offer significant support for environmental managers/policymakers to practice the sustainable environmental resources management by advancing their exiting approaches considering data outliers role in modelling approaches.

CRediT authorship contribution statement

Md Galal Uddin: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Azizur Rahman:** Methodology, Validation, Writing – review & editing. **Firouzeh Rosa Taghikhah:** Writing – review & editing. **Agnieszka I. Olbert:** Data curation, Funding acquisition, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors gratefully acknowledge the editor's and anonymous reviewers' contributions to the improvement of this paper. The authors also sincerely acknowledge the Eco HdroInformatics Research Group (EHIRG), School of Engineering, College of Science and Engineering, University of Galway, Ireland for providing computational laboratory facilities to complete this research. We also would like to thanks to the Environmental Protection Agency of Ireland, for providing essential water quality data. The research was also supported by the Environmental Protection Agency, Ireland for the AquaCop project [Grant Ref No: 2022-NE-1128].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.121499.

References

Abdulghafoor, S.A., Mohamed, L.A., 2022. A local density-based outlier detection method for high dimension data. Int. J. Nonlinear Anal. Appl. 13, 1683–1699. https://doi.org/10.22075/ijnaa.2022.5784.

AbuAlghanam, O., Alazzam, H., Alhenawi, E., Qatawneh, M., Adwan, O., 2023. Fusion-based anomaly detection system using modified isolation forest for internet of things. J. Ambient. Intell. Humaniz. Comput. 14, 131–145. https://doi.org/10.1007/s12652-022-04393-9.

Adeoye, J., Hui, L., Su, Y.-X., 2023. Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer. J. Big. Data 10, 28. https://doi.org/10.1186/s40537-023-00703-w.

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., Sharma, N., 2019. Detection of spatial outlier by using improved Z-score test. In: 2019 3rd International Conference on

Trends in Electronics and Informatics (ICOEI), pp. 788–790. https://doi.org/10.1109/ICOEI.2019.8862582.
Aguilera-Martos, I., García-Barzana, M., García-Gil, D., Carrasco, J., López, D., Luengo, J., Herrera, F., 2023a. Multi-step histogram based outlier scores for unsupervised anomaly detection: ArcelorMittal engineering dataset case of study. Neurocomputing. 544, 126228 https://doi.org/10.1016/j.neucom.2023.126228.
Aguilera-Martos, I., Luengo, J., Herrera, F., et al., 2023b. Revisiting histogram based outlier scores: strengths and weaknesses. In: García Bringas, P., Pérez García, H., Martínez de Pisón, F.J., Martínez Álvarez, F., Troncoso Lora, A., Herrero, Á., et al. (Eds.), Hybrid Artificial Intelligent Systems. Springer Nature Switzerland, Cham, pp. 39–48.
Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S.V.S., Ockunzzi, S., Lallo, D., Hu, B., Rashidi, H.H., 2023. Artificial intelligence and machine learning overview in pathology & laboratory medicine: a general review of data preprocessing and basic supervised concepts. Semin. Diagn. Pathol. 40, 71–87. https://doi.org/10.1053/j.semdp.2023.02.002.
Aliashrafi, A., Zhang, Y., Groenewegen, H., Peleato, N.M., 2021. A review of data-driven modelling in drinking water treatment. Rev. Environ. Sci. Biotechnol. 20, 985–1009. https://doi.org/10.1007/s11157-021-09592-y.
Ali, M.S., Islam, M.K., Das, A.A., Duranta, D.U.S., Haque, Mst.F., Rahman, M.H., 2023. A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: machine learning insights. Biomed. Res. Int. 2023, 8583210 https://doi.org/10.1155/2023/8583210.
Alsini, R., Almakrab, A., Ibrahim, A., Ma, X., 2021. Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor. Constr. Build. Mater. 270, 121396 https://doi.org/10.1016/j.conbuildmat.2020.121396.
Al Suwaidi, D., Haridy, S., Al Zaylaie, M., Shamsuzzaman, M., Bashir, H., Maged, A., Arab, M.G., 2023. Early detection of adverse conditions in deep excavations using statistical process control. Innov. Infrastruct. Sol. 8, 93. https://doi.org/10.1007/s41062-023-01054-4.
Angiulli, F., Fassetti, F., 2021. Uncertain distance-based outlier detection with arbitrarily shaped data objects. J. Intell. Inf. Syst. 57, 1–24. https://doi.org/10.1007/s10844-020-00624-7.
Auskalnis, J., Paulauskas, N., Baskys, A., 2018. Application of local outlier factor algorithm to detect anomalies in computer network. Elektronika ir Elektrotechnika 24, 96–99. https://doi.org/10.5755/j01.eie.24.3.20972.
Balamurali, M., Melkumyan, A., 2018. Detection of outliers in geochemical data using ensembles of subsets of variables. Math. Geosci. 50, 369–380. https://doi.org/10.1007/s11004-017-9716-8.
Baroudi, H., Huy Minh Nguyen, C.I., Maroongroge, S., Smith, B.D., Niedzielski, J.S., Shaitelman, S.F., Melancon, A., Shete, S., Whitaker, T.J., Mitchell, M.P., Yvonne Arzu, I., Duryea, J., Hernandez, S., El Basha, D., Mumme, R., Netherton, T., Hoffman, K., Court, L., 2023. Automated contouring and statistical process control for plan quality in a breast clinical trial. Phys. ImAging Radiat. Oncol. 28, 100486 https://doi.org/10.1016/j.phro.2023.100486.
Baseman, H.S., 2020. Chapter 1 - Process validation: design and planning. In: Gorsky, I., Baseman, H.S. (Eds.), Principles of Parenteral Solution Validation. Academic Press, pp. 9–31. https://doi.org/10.1016/B978-0-12-809412-9.00001-0.
Berendrecht, W., van Vliet, M., Griffioen, J., 2022. Combining statistical methods for detecting potential outliers in groundwater quality time series. Environ. Monit. Assess. 195, 85. https://doi.org/10.1007/s10661-022-10661-0.
Boaventura, L.L., Ferreira, P.H., Fiaccone, R.L., 2022. On flexible statistical process control with artificial intelligence: classification control charts. Expert. Syst. Appl. 194, 116492 https://doi.org/10.1016/j.eswa.2021.116492.
Budhlakoti, N., Rai, A., Mishra, D.C., 2020. Statistical approach for improving genomic prediction accuracy through efficient diagnostic measure of influential observation. Sci. Rep. 10, 8408. https://doi.org/10.1038/s41598-020-65323-3.
Burić, D., Mijanović, I., Doderović, M., Mihajlović, J., Trbić, G., 2023. Assessment of the environmental quality of Lake Skadar and its ur-ban surroundings in Montenegro. Eur. J. Geogr. 14, 76–87. https://doi.org/10.48088/ejg.d.bur.14.2.076.087.
Burigato Costa, C.M., da, S., da Silva Marques, L., Almeida, A.K., Leite, I.R., de Almeida, I.K., 2019. Applicability of water quality models around the world—A review. Environ. Sci. Pollut. Res. 26, 36141–36162. https://doi.org/10.1007/s11356-019-06637-2.
Buschjäger, S., Honysz, P.-J., Morik, K., 2022. Randomized outlier detection with trees. Int J Data Sci Anal 13 (2), 91–104. https://doi.org/10.1007/s41060-020-00238-w.
Cabana, E., Lillo, R.E., Laniado, H., 2021. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. Statistical Papers 62, 1583–1609. https://doi.org/10.1007/s00362-019-01148-1.
Cao, X.-Y., Feng, D.-C., Beer, M., 2023. A KDE-based non-parametric cloud approach for efficient seismic fragility estimation of structures under non-stationary excitation. Mech. Syst. Signal. Process. 205, 110873 https://doi.org/10.1016/j.ymssp.2023.110873.
Carletti, M., Terzi, M., Susto, G.A., 2023. Interpretable Anomaly Detection with DIFFI: depth-based feature importance of Isolation Forest. Eng. Appl. Artif. Intell. 119, 105730 https://doi.org/10.1016/j.engappai.2022.105730.
Chander, B., Kumaravelan, G., 2022. Outlier detection strategies for WSNs: a survey. J. King Saud Univ. 34, 5684–5707. https://doi.org/10.1016/j.jksuci.2021.02.012.
Chang, V., Bhavani, V.R., Xu, A.Q., Hossain, M.A., 2022. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthc. Anal. 2, 100016 https://doi.org/10.1016/j.health.2022.100016.
Chen, G., Zhang, S., Zhang, C., Qian, X., 2023. A study on the prediction model of dam seepage volume based on isolated forest-multiple stepwise linear regression. In: 2023 4th International Conference on Computer Engineering and Application (ICCEA), pp. 465–468. https://doi.org/10.1109/ICCEA58433.2023.10135319.
Chen, Y., Zhao, Z., Wu, H., Chen, X., Xiao, Q., Yu, Y., 2022. Fault anomaly detection of synchronous machine winding based on isolation forest and impulse frequency response analysis. Measurement 188, 110531. https://doi.org/10.1016/j.measurement.2021.110531.
Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., El Azzi, D., 2023. A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. Rev. Environ. Sci. Biotechnol. 22, 349–395. https://doi.org/10.1007/s11157-023-09650-7.
Chiu, A.L.M., Fu, A.W., 2003. Enhancements on local outlier detection. In: Seventh International Database Engineering and Applications Symposium, 2003. Proceedings, pp. 298–307. https://doi.org/10.1109/IDEAS.2003.1214939.
Chen, S., Ren, Y., Friedrich, D., Yu, Z., Yu, J., 2020. Sensitivity analysis to reduce duplicated features in ANN training for district heat demand prediction. Energy AI 2, 100028. https://doi.org/10.1016/j.egyai.2020.100028.
Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Comput. Sci. 7, e623. https://doi.org/10.7717/peerj-cs.623.
Choi, J., Kim, B., Im, S., Yoo, G., 2022. Supervised multivariate kernel density estimation for enhanced plasma etching endpoint detection. IEEE Access 10, 25580–25590. https://doi.org/10.1109/ACCESS.2022.3155513.
Choi, Y., Bhadriaju, B., Cho, H., Lim, J., Han, I.-S., Moon, I., Kwon, J.S.-I., Kim, J., 2023a. Data-driven modeling of multimode chemical process: validation with a real-world distillation column. Chem. Eng. J. 457, 141025 https://doi.org/10.1016/j.cej.2022.141025.
Choi, Y., Bhadriaju, B., Cho, H., Lim, J., Han, I.-S., Moon, I., Kwon, J.S.-I., Kim, J., 2023b. Data-driven modeling of multimode chemical process: validation with a real-world distillation column. Chem. Eng. J. 457, 141025 https://doi.org/10.1016/j.cej.2022.141025.
Choi, Y., Park, H., Roh, J., Lim, J., Han, I., Kim, D.-K., Jeon, S., Nam, H.-G., Moon, I., Cho, H., Kim, J., 2021. A data-based predictive model for distillation column of bio-based 2,3-butanediol. In: Türkay, M., Gani, R. (Eds.), Computer Aided Chemical Engineering. Elsevier, pp. 1005–1011. https://doi.org/10.1016/B978-0-323-88506-5.50155-8.
Comer, J., Indiana Olbert, A., Nash, S., Hartnett, M., 2017. Development of high-resolution multi-scale modelling system for simulation of coastal-fluvial urban flooding. Nat. Hazards Earth Syst. Sci. 17, 205–224. https://doi.org/10.5194/NHESS-17-205-2017.
Dashdondov, K., Kim, M.-H., 2023. Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction. Neural Process. Lett. 55, 265–277. https://doi.org/10.1007/s11063-021-10663-y.
Ding, F., Zhang, W., Cao, S., Hao, S., Chen, L., Xie, X., Li, W., Jiang, M., 2023. Optimization of water quality index models using machine learning approaches. Water. Res. 243, 120337 https://doi.org/10.1016/j.watres.2023.120337.
Dobie, R.A., Wilson, M.J., 1996. A comparison of t-test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids. J. Acoust. Soc. Am. 100, 2236–2246. https://doi.org/10.1121/1.417933.
Domański, P.D., 2020a. Study on Statistical Outlier Detection and Labelling. International Journal of Automation and Computing 17, 788–811. https://doi.org/10.1007/s11633-020-1243-2.
Domański, P.D., 2020b. Study on statistical outlier detection and labelling. Int. J. Autom. Comput. 17, 788–811. https://doi.org/10.1007/s11633-020-1243-2.
Dovoedo, Y.H., Chakraborti, S., 2015. Boxplot-based outlier detection for the location-scale family. Commun. Stat. Simul. Comput. 44, 1492–1513. https://doi.org/10.1080/03610918.2013.813037.
Duraj, A., Szczepaniak, P.S., 2021. Outlier detection in data streams — a comparative study of selected methods. Procedia Comput. Sci. 192, 2769–2778. https://doi.org/10.1016/j.procs.2021.09.047.
El Alaoui, I., Gahi, Y., Messoussi, R., Todoskoff, A., Kobi, A., 2018. Big Data analytics: a comparison of tools and applications. In: Ben Ahmed, M., Boudhir, A.A. (Eds.), Innovations in Smart Cities and Applications. Springer International Publishing, Cham, pp. 587–601.
EPA, 2022. Water Quality in 2022: An Indicators Report. Wexford.
EPA, 2021. Urban Waste Water Treatment in 2021.
Esnaola, A., Arrizabalaga-Escudero, A., González-Esteban, J., Elosegi, A., Aihartza, J., 2018. Determining diet from faeces: Selection of metabarcoding primers for the insectivore Pyrenean desman (Galemys pyrenaicus). PLoS ONE 13. https://doi.org/10.1371/journal.pone.0208986.
Etherington, T.R., 2021. Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterising a multivariate location and scatter method. PeerJ. 9, e11436. https://doi.org/10.7717/peerj.11436.
Fahim, P., Vaezi, N., Shahraki, A., Khoshnevisan, M., 2022. An integration of genetic feature selector, histogram-based outlier score, and deep learning for wind turbine power prediction. Energy Sources Part A 44, 9342–9365. https://doi.org/10.1080/15567036.2022.2129876.
Feng, Y., Cai, W., Yue, H., Xu, J., Lin, Y., Chen, J., Hu, Z., 2022. An improved X-means and isolation forest based methodology for network traffic anomaly detection. PLoS ONE 17, e0263423–.
Fernandes, W., Komati, K.S., Assis de Souza Gazolli, K., 2023. Anomaly detection in oil-producing wells: a comparative study of one-class classifiers in a multivariate time series dataset. J. pet. explor. Prod. Technol. https://doi.org/10.1007/s13202-023-01710-6.
Fernández, Á., Bella, J., Dorronsoro, J.R., 2022. Supervised outlier detection for classification and regression. Neurocomputing. 486, 77–92. https://doi.org/10.1016/j.neucom.2022.02.047.

Festus Biosengazeh, N., Estella Buleng Tamungang, N., Nelson Alakeh, M., Antoine david, M., 2020. Analysis and water quality control of alternative sources in Bangolan, Northwest Cameroon. J. Chem. 2020, 5480762 https://doi.org/10.1155/2020/5480762.

Fredianto, Putri, D.A.P., 2023. Comparison of the interquartile range algorithm and local outlier factor on Australian weather data sets. AIP. Conf. Proc. 2727, 040010 https://doi.org/10.1063/5.0141897.

Gallego, J.A., Osorio, J.F., Gonzalez, F.A., 2022. Fast Kernel density estimation with density matrices and random Fourier features. In: Bicharra Garcia, A.C., Ferro, M., Rodríguez Ribón, J.C. (Eds.), Advances in Artificial Intelligence – IBERAMIA 2022. Springer International Publishing, Cham, pp. 160–172.

Gani, M.A., Sajib, A.M., Siddik, M.A., Moniruzzaman, M., 2023. Assessing the impact of land use and land cover on river water quality using water quality index and remote sensing techniques. Environ. Monit. Assess. 195, 449. https://doi.org/10.1007/s10661-023-10989-1.

Garces, H., Sbarbaro, D., 2009. Outliers detection in environmental monitoring data. IFAC Proc.s 42, 330–335. https://doi.org/10.3182/20091014-3-CL-4011.00060.

Georgescu, P.-L., Moldovanu, S., Iticescu, C., Calmuc, M., Calmuc, V., Topa, C., Moraru, L., 2023. Assessing and forecasting water quality in the Danube River by using neural network approaches. Sci. Total Environ. 879, 162998 https://doi.org/10.1016/j.scitotenv.2023.162998.

Gessa, A., Marin, E., Sancha, P., 2022. A practical application of statistical process control to evaluate the performance rate of academic programmes: implications and suggestions. Quality Assurance in Education 30, 571–588. https://doi.org/10.1108/QAE-03-2022-0065.

Gorsky, I., 2020. Chapter 6 - Use of statistics in process validation. In: Gorsky, I., Baseman, H.S. (Eds.), Principles of Parenteral Solution Validation. Academic Press, pp. 115–136. https://doi.org/10.1016/B978-0-12-809412-9.00005-8.

Green, J.A., 2021. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. Health Psychol. Behav. Med. 9, 436–455. https://doi.org/10.1080/21642850.2021.1920416.

Gui, G., Pan, H., Lin, Z., Li, Y., Yuan, Z., 2017. Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection. KSCE J. Civil Eng. 21, 523–534. https://doi.org/10.1007/s12205-017-1518-5.

Gupta, S., Gupta, S.K., 2021. A critical review on water quality index tool: genesis, evolution and future directions. Ecol. Inform. 63 https://doi.org/10.1016/j.ecoinf.2021.101299.

Gyebnár, G., Klimaj, Z., Entz, L., Fabó, D., Rudas, G., Barsi, P., Kozák, L.R., 2019. Personalized microstructural evaluation using a Mahalanobis-distance based outlier detection strategy on epilepsy patients' DTI data – Theory, simulations and example cases. PLoS ONE 14, e0222720–.

Haj-Hassan, A., Habib, C., Nassar, J., 2020. Real-time spatio-temporal based outlier detection framework for wireless body sensor networks. In: 2020 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1–6. https://doi.org/10.1109/ANTS50601.2020.9342827.

Han, Q., Ma, S., Wang, T., Chu, F., 2019. Kernel density estimation model for wind speed probability distribution with applicability to wind energy assessment in China. Renew. Sustain. Energy Rev. 115, 109387 https://doi.org/10.1016/j.rser.2019.109387.

Ha, J., Seok, S., Lee, J.-S., 2015. A precise ranking method for outlier detection. Inf. Sci. 324, 88–107. https://doi.org/10.1016/j.ins.2015.06.030.

Ha, J., Seok, S., Lee, J.-S., 2014. Robust outlier detection using the instability factor. Knowl. Based. Syst. 63, 15–23. https://doi.org/10.1016/j.knosys.2014.03.001.

Haas, J.C., Switanek, M., Birk, S., 2018. Analysis of hydrological data with correlation matrices: technical implementation and possible applications. Environ. Earth. Sci. 77, 310. https://doi.org/10.1007/s12665-018-7469-4.

He, X., Gou, W., Liu, Y., Gao, Z., 2015. A practical method of nonprobabilistic reliability and parameter sensitivity analysis based on space-filling design. Math. Probl. Eng. 2015, 1–12. https://doi.org/10.1155/2015/561202.

Hernández, N., Muñoz, A., Martos, G., 2023. Density kernel depth for outlier detection in functional data. Int. J. Data Sci. Anal. 16, 481–488. https://doi.org/10.1007/s41060-023-00420-w.

Hamby, D.M., 1995. A comparison of sensitivity analysis techniques. Health Phys. 68, 195–204. https://doi.org/10.1097/00004032-199502000-00005.

Hamby, D.M., 1994. A review of techniques for parameter sensitivity. Environ. Monit. Assess. 32, 135–154.

Hansen, J., Ahern, S., Earnest, A., 2023. Evaluations of statistical methods for outlier detection when benchmarking in clinical registries: a systematic review. BMJ Open. 13, e069130 https://doi.org/10.1136/bmjopen-2022-069130.

Harrington, L.J., Schleussner, C.-F., Otto, F.E.L., 2021. Quantifying uncertainty in aggregated climate change risk assessments. Nat. Commun. 12, 7140. https://doi.org/10.1038/s41467-021-27491-2.

Hartnett, M., Dabrowski, T., Olbert, A.I., 2011a. A new formula to calculate residence times of tidal waterbodies. Proc. Inst. Civil Eng. 164, 243–256. https://doi.org/10.1680/wama.2011.164.5.243.

Hartnett, M., Nash, S., 2015. An integrated measurement and modeling methodology for estuarine water quality management. Water Sci. Eng. 8, 9–19. https://doi.org/10.1016/j.wse.2014.10.001.

Hartnett, M., Nash, S., Olbert, I., 2012. An integrated approach to trophic assessment of coastal waters incorporating measurement, modelling and water quality classification. Estuar. Coast. Shelf. Sci. https://doi.org/10.1016/j.ecss.2011.08.012.

Hartnett, M., Wilson, J.G., Nash, S., 2011b. Irish estuaries: water quality status and monitoring implications under the water framework directive. Mar. Policy. 35, 810–818. https://doi.org/10.1016/j.marpol.2011.01.010.

Hassan, A.F., Barakat, S., Rezk, A., 2022. Towards a deep learning-based outlier detection approach in the context of streaming data. J. Big Data 9, 120. https://doi.org/10.1186/s40537-022-00670-8.

Hewitt, J., Gelfand, A.E., Quick, N.J., Cioffi, W.R., Southall, B.L., DeRuiter, S.L., Schick, R.S., 2022. Kernel density estimation of conditional distributions to detect responses in satellite tag data. Anim. Biotelem. 10, 28. https://doi.org/10.1186/s40317-022-00299-7.

Humbert, P., Bars, B.Le, Minvielle, L., 2022. Robust kernel density estimation with median-of-means principle. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, pp. 9444–9465.

Ibrahim, A., Ismail, A., Juahir, H., Iliyasu, A.B., Wailare, B.T., Mukhtar, M., Aminu, H., 2023. Water quality modelling using principal component analysis and artificial neural network. Mar. Pollut. Bull. 187, 114493 https://doi.org/10.1016/j.marpolbul.2022.114493.

Jamshidi, E.J., Yusup, Y., Kayode, J.S., Kamaruddin, M.A., 2022. Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: a case study on surface water temperature. Ecol. Inform. 69, 101672 https://doi.org/10.1016/j.ecoinf.2022.101672.

Jayaweera, C.D., Aziz, N., 2018. Reliability of principal component analysis and pearson correlation coefficient, for application in artificial neural network model development, for water treatment plants. IOP. Conf. Ser. Mater. Sci. Eng. 458, 012076 https://doi.org/10.1088/1757-899X/458/1/012076.

Jeong, J., Park, E., 2019. Comparative applications of data-driven models representing water table fluctuations. J. Hydrol. 572, 261–273. https://doi.org/10.1016/j.jhydrol.2019.02.051.

Jiang, Z., Chen, C., Li, N., Wang, H., Wang, P., Zhang, C., Ma, F., Zhang, Z., Huang, Y., Qi, J., Chen, W.-Q., 2022. Advancing UN comtrade for physical trade flow analysis: addressing the issue of outliers. Resour. Conserv. Recycl. 186, 106524 https://doi.org/10.1016/j.resconrec.2022.106524.

Jin, J., Vandenplas, C., Loosveldt, G., 2019a. The Evaluation of Statistical Process Control Methods to Monitor Interview Duration During Survey Data Collection, 9. Sage Open. https://doi.org/10.1177/2158244019854652, 2158244019854652.

Jin, T., Cai, S., Jiang, D., Liu, J., 2019b. A data-driven model for real-time water quality prediction and early warning by an integration method. Environ. Sci. Pollut. Res. 26, 30374–30385. https://doi.org/10.1007/s11356-019-06049-2.

Johannesen, N.J., Kolhe, M.L., Goodwin, M., 2022. Evaluating anomaly detection algorithms through different grid scenarios using k-nearest neighbor, iforest and local outlier factor. In: 2022 7th International Conference on Smart and Sustainable Technologies (SpliTech), pp. 1–6. https://doi.org/10.23919/SpliTech55088.2022.9854355.

Kabir, S., Shufian, A., Zishan, M.S.R., 2023. Isolation forest based anomaly detection and fault localization for solar PV system. In: 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 341–345. https://doi.org/10.1109/ICREST57604.2023.10070033.

Kalaycı, İ., Ercan, T., 2018. Anomaly detection in wireless sensor networks data by using histogram based outlier score method. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6. https://doi.org/10.1109/ISMSIT.2018.8567262.

Kang, G., Gao, J.Z., Xie, G., 2017. Data-driven water quality analysis and prediction: a survey. In: 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), pp. 224–232. https://doi.org/10.1109/BigDataService.2017.40.

Kim, H.-Y., 2015. Statistical notes for clinical researchers: post-hoc multiple comparisons. Restor Dent Endod 40 (2), 172–176. https://doi.org/10.5395/rde.2015.40.2.172.

Kim, M., Cho, S., Jang, K., Hong, S., Na, J., Moon, I., 2022. Data-driven robust optimization for minimum nitrogen oxide emission under process uncertainty. Chem. Eng. J. 428, 130971 https://doi.org/10.1016/j.cej.2021.130971.

Knoth, S., Schmid, W., 2004. Control charts for time series: a review. In: Lenz, H.-J., Wilrich, P.-T. (Eds.), Frontiers in Statistical Quality Control 7. Physica-Verlag HD. Heidelberg, pp. 210–236.

Kokatnoor, S.A., Reddy, V., Balachandran, K., 2022. Deducing Water Quality Index (WQI) by comparative supervised machine learning regression techniques for India region. In: Saraswat, M., Sharma, H., Balachandran, K., Kim, J.H., Bansal, J.C. (Eds.), Congress on Intelligent Systems. Singapore. Springer Nature Singapore, pp. 727–742.

Krishna, M.H., K, N., Charmitha, G., Vignesh, T., Ch, V., Kuchibhotla, S.. Studies on Anomaly Detection Techniques. https://doi.org/10.1109/ICCMC56507.2023.10083885.

Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: management of missing values and outliers. Korean J. Anesthesiol. 70, 407–411. https://doi.org/10.4097/kjae.2017.70.4.407.

Latecki, L.J., Lazarevic, A., Pokrajac, D., 2007. Outlier detection with kernel density functions. In: Perner, P. (Ed.), Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, pp. 61–75.

Lee, I., 2017. Big data: dimensions, evolution, impacts, and challenges. Bus. Horiz. 60, 293–303. https://doi.org/10.1016/j.bushor.2017.01.004.

Lee, J., Kang, B., Kang, S.-H., 2011. Integrating independent component analysis and local outlier factor for plant-wide process monitoring. J. Process. Control 21, 1011–1021. https://doi.org/10.1016/j.jprocont.2011.06.004.

Lee, S., Lee, D.K., 2018. What is the proper way to apply the multiple comparison test? Korean J. Anesthesiol. 71 (5), 353–360. https://doi.org/10.4097/kja.d.18.00242.

Lei, X., Xia, Y., Wang, A., Jian, X., Zhong, H., Sun, L., 2023. Mutual information based anomaly detection of monitoring data with attention mechanism and residual

learning. Mech. Syst. Signal. Process. 182, 109607 https://doi.org/10.1016/j.ymssp.2022.109607.
Leys, C., Klein, O., Dominicy, Y., Ley, C., 2018. Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. J. Exp. Soc. Psychol. 74, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011.
Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49, 764–766. https://doi.org/10.1016/j.jesp.2013.03.013.
Liang, W., Tadesse, G.A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., Zou, J., 2022. Advances, challenges and opportunities in creating data for trustworthy AI. Nat. Mach. Intell. 4, 669–677. https://doi.org/10.1038/s42256-022-00516-1.
Li, A., Feng, M., Li, Y., Liu, Z., 2016. Application of outlier mining in insider identification based on boxplot method. Procedia Comput. Sci. 91, 245–251. https://doi.org/10.1016/j.procs.2016.07.069.
Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17.
Liu, J., Wang, P., Jiang, D., Nan, J., Zhu, W., 2020a. An integrated data-driven framework for surface water quality anomaly detection and early warning. J. Clean. Prod. 251, 119145 https://doi.org/10.1016/j.jclepro.2019.119145.
Liu, F., Yu, Y., Song, P., Fan, Y., Tong, X., 2020b. Scalable KDE-based top-n local outlier detection over large-scale data streams. Knowl. Based. Syst. 204, 106186 https://doi.org/10.1016/j.knosys.2020.106186.
Luley, P.-P., Deriu, J.M., Yan, P., Schatte, G.A., Stadelmann, T., 2023. From concept to implementation: the data-centric development process for AI in industry. In: 2023 10th IEEE Swiss Conference on Data Science (SDS), pp. 73–76. https://doi.org/10.1109/SDS57534.2023.00017.
Luo, H., Paal, S.G., 2023. A novel outlier-insensitive local support vector machine for robust data-driven forecasting in engineering. Eng. Comput. https://doi.org/10.1007/s00366-022-01781-9.
Manna, A., Biswas, D., 2023. Assessment of drinking water quality using water quality index: a review. Water Conserv. Sci. Eng. 8, 6. https://doi.org/10.1007/s41101-023-00185-0.
Matioli, L.C., Santos, S.R., Kleina, M., Leite, E.A., 2018. A new algorithm for clustering based on kernel density estimation. J. Appl. Stat. 45, 347–366. https://doi.org/10.1080/02664763.2016.1277191.
Mayer, D.G., Stuart, M.A., Swain, A.J., 1994. Regression of real-world data on model output: an appropriate overall test of validity. Agric Syst 45, 93–104. https://doi.org/10.1016/S0308-521X(94)90282-8.
Meenakshi, S., M, N.D., 2022. Performance enhancement of unsupervised hardware trojan detection algorithm using clustering-based local outlier factor technique for design security. In: 2022 IEEE International Test Conference India (ITC India), pp. 1–8. https://doi.org/10.1109/ITCIndia202255192.2022.9854569.
Mensi, A., Tax, D.M.J., Bicego, M., 2023. Detecting outliers from pairwise proximities: proximity isolation forests. Pattern. Recognit. 138, 109334 https://doi.org/10.1016/j.patcog.2023.109334.
Mensi, A., Franzoni, A., Tax, D.M.J., Bicego, M., Torsello, A., Rossi, L., Pelillo, M., Biggio, B., 2021. An alternative exploitation of isolation forests for outlier detection. In: Robles-Kelly, A. (Ed.), Structural, Syntactic, and Statistical Pattern Recognition. Springer International Publishing, Cham, pp. 34–44.
Mentis, A.-F.A., Lee, D., Roussos, P., 2023. Applications of artificial intelligence−machine learning for detection of stress: a critical overview. Mol. Psychiatry. https://doi.org/10.1038/s41380-023-02047-6.
Midway, S., Robertson, M., Flinn, S., Kaller, M., 2020. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. PeerJ. 8, 1–26. https://doi.org/10.7717/peerj.10387.
Milić, S.D., Đurović, Ž., Stojanović, M.D., 2023. Data science and machine learning in the IIoT concepts of power plants. Int. J. Electric. Power Energy Syst. 145, 108711 https://doi.org/10.1016/j.ijepes.2022.108711.
Minne, L., Eslami, S., de Keizer, N., de Jonge, E., de Rooij, S.E., Abu-Hanna, A., 2012. Statistical process control for validating a classification tree model for predicting mortality – a novel approach towards temporal validation. J. Biomed. Inform. 45, 37–44. https://doi.org/10.1016/j.jbi.2011.08.015.
Mishra, S., Chawla, M., Abraham, A., Dutta, P., Mandal, J.K., Bhattacharya, A., 2019. A comparative study of local outlier factor algorithms for outliers detection in data streams. In: Dutta, S. (Ed.), Emerging Technologies in Data Mining and Information Security. Springer Singapore, Singapore, pp. 347–356.
Misra, S., Osogba, O., Powers, M., 2020. Chapter 1 - Unsupervised outlier detection techniques for well logs and geophysical data. In: Misra, S., Li, H., He, J. (Eds.), Machine Learning for Subsurface Characterization. Gulf Professional Publishing, pp. 1–37. https://doi.org/10.1016/B978-0-12-817736-5.00001-6.
Modak, S., 2023 A new interpoint distance-based clustering algorithm using kernel density estimation. Commun. Stat. Simul. Comput. 1–19. https://doi.org/10.1080/03610918.2023.2179071.
Mogane, L.K., Masebe, T., Msagati, T.A.M., Ncube, E., 2023. A comprehensive review of water quality indices for lotic and lentic ecosystems. Environ. Monit. Assess. 195, 926. https://doi.org/10.1007/s10661-023-11512-2.
Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. J. Big. Data 2, 1. https://doi.org/10.1186/s40537-014-0007-7.
Najman, K., Zieliński, K., 2021. Outlier detection with the use of isolation forests. In: Jajuga, K., Najman, K., Walesiak, M. (Eds.), Data Analysis and Classification. Springer International Publishing, Cham, pp. 65–79.
Nanda, A., Mohapatra, Dr.B.B., Mahapatra, Abikesh Prasada Kumar, Mahapatra, Abiresh Prasad Kumar, Mahapatra, Abinash Prasad Kumar, 2021. Multiple comparison test by Tukey's honestly significant difference (HSD): do the confident level control type I error. Int. J. Stat. Appl. Math. 6, 59–65. https://doi.org/10.22271/maths.2021.v6.i1a.636.
Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. J. Water. Process. Eng. 48, 102920 https://doi.org/10.1016/j.jwpe.2022.102920.
Obikee, A.C., Ebuh, G.U., Obiora-Ilouno, H.O., 2014. Comparison of outlier techniques based on simulated data. Open. J. Stat. 04, 536–561. https://doi.org/10.4236/ojs.2014.47051.
Ojo, O.T., Fernández Anta, A., Lillo, R.E., Sguera, C., 2022. Detecting and classifying outliers in big functional data. Adv. Data Anal. Classif. 16, 725–760. https://doi.org/10.1007/s11634-021-00460-9.
Olbert, A.I., Comer, J., Nash, S., Hartnett, M., 2017. High-resolution multi-scale modelling of coastal flooding due to tides, storm surges and rivers inflows. A Cork City example. Coast. Eng. https://doi.org/10.1016/j.coastaleng.2016.12.006.
Oliveira, P., Duarte, M.S., Novais, P., 2022. Applying anomaly detection models in wastewater management: a case study of nitrates concentration in the effluent. In: Bicharra Garcia, A.C., Ferro, M., Rodríguez Ribón, J.C. (Eds.), Advances in Artificial Intelligence – IBERAMIA 2022. Springer International Publishing, Cham, pp. 65–76.
Orouji, H., O, B.H., Fallah-Mehdipour, E., Mariño, M.A., 2013. Modeling of water quality parameters using data-driven models. J. Environ. Eng. 139, 947–957. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000706.
Ottosen, T.-B., Kumar, P., 2019. Outlier detection and gap filling methodologies for low-cost air quality measurements. Environ. Sci. Process. Impacts. 21, 701–713. https://doi.org/10.1039/C8EM00593A.
Owolabi, O., Okoh, D., Rabiu, B., Obafaye, A., Dauda, K., 2021. A median absolute deviation-neural network (MAD-NN) method for atmospheric temperature data cleaning. MethodsX. 8, 101533 https://doi.org/10.1016/j.mex.2021.101533.
Panjei, E., Gruenwald, L., Leal, E., Nguyen, C., Silvia, S., 2022. A survey on outlier explanations. The VLDB Journal 31, 977–1008. https://doi.org/10.1007/s00778-021-00721-1.
Parra-Plazas, J., Gaona-Garcia, P., Plazas-Nossa, L., 2023. Time series outlier removal and imputing methods based on Colombian weather stations data. Environ. Sci. Pollut. Res. 30, 72319–72335. https://doi.org/10.1007/s11356-023-27176-x.
Parween, S., Siddique, N.A., Mahammad Diganta, M.T., Olbert, A.I., Uddin, M.G., 2022. Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. Environ. Sustain. Indicat. 16 https://doi.org/10.1016/j.indic.2022.100202.
Pei, F., Miao, Z., Wang, J., 2021. Dynamic SLAM system using histogram-based outlier score to improve anomaly detection. In: 2021 China Automation Congress (CAC), pp. 4909–4913. https://doi.org/10.1109/CAC53003.2021.9728124.
Peng, Y., Yang, Y., Xu, Y., Xue, Y., Song, R., Kang, J., Zhao, H., 2021. Electricity theft detection in AMI based on clustering and local outlier factor. IEEE Access 9, 107250–107259. https://doi.org/10.1109/ACCESS.2021.3100980.
Pérez-Benítez, B.E., Tercero-Gómez, V.G., Khakifirooz, M., 2023. A review on statistical process control in healthcare: data-driven monitoring schemes. IEEe Access 11, 56248–56272. https://doi.org/10.1109/ACCESS.2023.3282569.
Petkovski, A., Shehu, V., 2023. Anomaly detection on univariate sensing time series data for smart aquaculture using K-means, isolation forest, and local outlier factor. In: 2023 12th Mediterranean Conference on Embedded Computing (MECO), pp. 1–5. https://doi.org/10.1109/MECO58584.2023.10154991.
Piñeiro Di Blasi, J.I., Martínez Torres, J., García Nieto, P.J., Alonso Fernández, J.R., Díaz Muñiz, C., Taboada, J., 2015. Analysis and detection of functional outliers in water quality parameters from different automated monitoring stations in the Nalón River Basin (Northern Spain). Environ. Sci. Pollut. Res. 22, 387–396. https://doi.org/10.1007/s11356-014-3318-5.
Prabhakar, P., Arora, S., Khosla, A., Beniwal, R.K., Arthur, M.N., Arias-Gonzáles, J.L., Areche, F.O., 2022. Cyber security of smart metering infrastructure using median absolute deviation methodology. Secur. Commun. Netw. 2022, 6200121 https://doi.org/10.1155/2022/6200121.
Prasad, D.V.V., Venkataramana, L.Y., Kumar, P.S., Prasannamedha, G., Harshana, S., Srividya, S.J., Harrinei, K., Indraganti, S., 2022. Analysis and prediction of water quality using deep learning and auto deep learning techniques. Sci. Total Environ. 821, 153311 https://doi.org/10.1016/j.scitotenv.2022.153311.
Qian, Q., He, M., Sun, F., Liu, X., 2024. Monitoring and evaluation of the water quality of the Lower Neches River, Texas, USA. Water Sci. Eng. 17, 21–32. https://doi.org/10.1016/j.wse.2023.10.002.
Qiu, P., 2020. Big Data? Statistical process control can help! Am. Stat. 74, 329–344. https://doi.org/10.1080/00031305.2019.1700163.
Qiu, P., 2019. Some recent studies in statistical process control. In: Lio, Y., Ng, H.K.T., Tsai, T.-R., Chen, D.-G. (Eds.), Statistical Quality Technologies: Theory and Practice. Springer International Publishing, Cham, pp. 3–19. https://doi.org/10.1007/978-3-030-20709-0_1.
Qiu, Y., Dong, T., Lin, D., Zhao, B., Cao, W., Jiang, F., 2022. Fault diagnosis for lithium-ion battery energy storage systems based on local outlier factor. J. Energy Storage 55, 105470. https://doi.org/10.1016/j.est.2022.105470.
Ragab, M., Farouk, S., Sabir, M., 2022. Outlier detection with optimal hybrid deep learning enabled intrusion detection system for ubiquitous and smart environment. Sustain. Energy Technol. Assessm. 52, 102311 https://doi.org/10.1016/j.seta.2022.102311.
Rahman, A., 2019. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. Int. J. Artif. Intell. 17, 44–65.
Rahman, A., Harding, A., 2016. Small area estimation and microsimulation modeling, Small Area Estimation and Microsimulation Modeling. CRC Press. https://doi.org/10.1201/9781315372143.

Rangeti, I., Dzwairo, B., Barratt, G.J., Otieno, F.A.O., 2015. Validity and errors in water quality data — a review. In: Lee, T.S. (Ed.), Research and Practices in Water Quality. IntechOpen, Rijeka. https://doi.org/10.5772/59059 p. Ch. 4.
P.T., A.C.P., R. Raveendran, V., R, G.R., Bhasi, S., Kinhikar, R.A., 2023. Moving towards process-based radiotherapy quality assurance using statistical process control Physica Medica 112, 102651 https://doi.org/10.1016/j.ejmp.2023.102651.
Ripan, R.C., Sarker, I.H., Anwar, M.M., Furhad, Md.H., Rahat, F., Hoque, M.M., Sarfraz, M., 2021. An isolation forest learning based outlier detection approach for effectively classifying cyber anomalies. In: Abraham, A., Hanne, T., Castillo, O., Gandhi, N., Nogueira Rios, T., Hong, T.-P. (Eds.), Hybrid Intelligent Systems. Springer International Publishing, Cham, pp. 270–279.
Rosenberger, J., Müller, K., Selig, A., Bühren, M., Schramm, D., 2022. Extended kernel density estimation for anomaly detection in streaming data. Procedia CIRP. 112, 156–161. https://doi.org/10.1016/j.procir.2022.09.065.
Rouder, J.N., Engelhardt, C.R., McCabe, S., Morey, R.D., 2016. Model comparison in ANOVA. Psychon. Bull. Rev. 23, 1779–1786. https://doi.org/10.3758/s13423-016-1026-5.
Rousseeuw, P.J., Hubert, M., 2011. Robust statistics for outlier detection. WIREs Data Min. Knowl. Discov. 1, 73–79. https://doi.org/10.1002/widm.2.
Sajib, A.M., Diganta, M.T.M., Moniruzzaman, M., Rahman, A., Dabrowski, T., Uddin, M. G., Olbert, A.I., 2024. Assessing water quality of an ecologically critical urban canal incorporating machine learning approaches. Ecol. Inform., 102514 https://doi.org/10.1016/j.ecoinf.2024.102514.
Sajib, A.M., Diganta, M.T.M., Rahman, A., Dabrowski, T., Olbert, A.I., Uddin, M.G., 2023. Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Groundw. Sustain. Dev. 23, 101049 https://doi.org/10.1016/j.gsd.2023.101049.
Samariya, D., Ma, J., 2022. Anomaly detection on health data. In: Traina, A., Wang, H., Zhang, Y., Siuly, S., Zhou, R., Chen, L. (Eds.), Health Information Science. Springer Nature Switzerland, Cham, pp. 34–41.
Sarker, I.H., 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN. Comput. Sci. 2, 377. https://doi.org/10.1007/s42979-021-00765-8.
Seim, A., Andersen, B., Sandberg, W.S., 2006. Statistical process control as a tool for monitoring nonoperative time. Anesthesiology 105, 370–380. https://doi.org/10.1097/00000542-200608000-00021.
Sejr, J.H., Schneider-Kamp, A., 2021. Explainable outlier detection: what, for whom and why? Mach. Learn. Appl. 6, 100172 https://doi.org/10.1016/j.mlwa.2021.100172.
Shah, A., Ali, B., Wahab, F., Ullah, I., Amesho, K.T.T., Shafiq, M., 2023. Entropy-based grid approach for handling outliers: a case study to environmental monitoring data. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-023-26780-1.
Sharma, K.K., Seal, A., 2021. Outlier-robust multi-view clustering for uncertain data. Knowl. Based. Syst. 211, 106567 https://doi.org/10.1016/j.knosys.2020.106567.
Shi, H., Guo, J., Deng, Y., Qin, Z., 2023. Machine learning-based anomaly detection of groundwater microdynamics: case study of Chengdu, China. Sci. Rep. 13 (1), 14718. https://doi.org/10.1038/s41598-023-38447-5.
Shimizu, Y., 2022. Multiple desirable methods in outlier detection of univariate data With R source codes. Front. Psychol. 12.
Sikder, M.N.K., Batarseh, F.A., 2023. 7 - Outlier detection using AI: a survey. In: Batarseh, F.A., Freeman, L.J. (Eds.), AI Assurance. Academic Press, pp. 231–291. https://doi.org/10.1016/B978-0-32-391919-7.00020-2.
Singh, G., Kundu, S., 2022. Outlier and trend detection using approximate median and median absolute deviation. In: 2022 5th International Conference on Computational Intelligence and Networks (CINE), pp. 1–6. https://doi.org/10.1109/CINE56307.2022.10037489.
Singh, K., Rashmi, P., 2014. Water quality management using statistical analysis and time-series prediction model 425–434. https://doi.org/10.1007/s13201-014-0159-9.
Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001.
Smiti, A., 2020. A critical overview of outlier detection methods. Comput. Sci. Rev. 38, 100306 https://doi.org/10.1016/j.cosrev.2020.100306.
Sureiman, O., Mangera, C.M., 2020. F-Test of overall significance in regression analysis simplified. J. Pract. Cardiovasc. Sci. 6.
Sutadian, A.D., Muttil, N., Yilmaz, A.G., Perera, B.J.C., 2016. Development of river water quality indices—a review. Environ. Monit. Assess. https://doi.org/10.1007/s10661-015-5050-0.
Suvarna, M., Araújo, T.P., Pérez-Ramírez, J., 2022. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO2 hydrogenation. Appl. Catal. B 315. https://doi.org/10.1016/j.apcatb.2022.121530.
Talagala, P.D., Hyndman, R.J., Leigh, C., Mengersen, K., Smith-Miles, K., 2019. A feature-based procedure for detecting technical outliers in water-quality data from in situ sensors. Water. Resour. Res. 55, 8547–8568. https://doi.org/10.1029/2019WR024906.
Tan, X., Yang, J., Rahardja, S., 2022. Sparse random projection isolation forest for outlier detection. Pattern. Recognit. Lett. 163, 65–73. https://doi.org/10.1016/j.patrec.2022.09.015.
Tang, B., He, H., 2017. A local density-based approach for outlier detection. Neurocomputing. 241, 171–180. https://doi.org/10.1016/j.neucom.2017.02.039.
Tan, H.Q., Lew, K.S., Wong, Y.M., Chong, W.C., Koh, C.W.Y., Chua, C.G.A., Yeap, P.L., Ang, K.W., Lee, J.C.L., Park, S.Y., 2023. Detecting outliers beyond tolerance limits derived from statistical process control in patient-specific quality assurance. J. Appl. Clin. Med. Phys. e14154. https://doi.org/10.1002/acm2.14154 n/a.
Tegegne, D.A., Kitaw, D., Berhan, E., 2022. Advances in statistical quality control chart techniques and their limitations to cement industry. Cogent. Eng. 9, 2088463 https://doi.org/10.1080/23311916.2022.2088463.
Templ, M., Gussenbauer, J., Filzmoser, P., 2020. Evaluation of robust outlier detection methods for zero-inflated complex data. J. Appl. Stat. 47, 1144–1167. https://doi.org/10.1080/02664763.2019.1671961.
Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., Filzmoser, P., 2013. Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection. Anal. Chim. Acta 787, 1–9. https://doi.org/10.1016/j.aca.2013.04.034.
Toufigh, V., Ranjbar, I., 2023. Unsupervised deep learning framework for ultrasonic-based distributed damage detection in concrete: integration of a deep auto-encoder and Isolation Forest for anomaly detection. Struct. Health Monit. https://doi.org/10.1177/14759217231183143, 14759217231183144.
Tokovarov, M., Karczmarek, P., 2022. A probabilistic generalization of isolation forest. Inf. Sci. 584, 433–449. https://doi.org/10.1016/j.ins.2021.10.075.
Uddin, G., 2023. Development of a Novel Water Quality Index Model Using Data Science Approaches. University of Galway, Ireland. http://hdl.handle.net/1020 379/17786.
Uddin, G., Nash, S., Olbert, A.I., 2022e. Optimization of Parameters in a Water Quality Index Model Using Principal Component Analysis. Research Publishing Services, pp. 5739–5744. https://doi.org/10.3850/iahr-39wc2521711920221326.
Uddin, Md.G., Moniruzzaman, Md., Khan, M., 2017. Evaluation of groundwater quality using CCME water quality index in the rooppur nuclear power plant area, Ishwardi, Pabna, Bangladesh. Am. J. Environ. Protect. 5, 33–43. https://doi.org/10.12691/env-5-2-2.
Uddin, Md.G., Moniruzzaman, Md., Quader, M.A., Hasan, Md.A., 2018. Spatial variability in the distribution of trace metals in groundwater around the Rooppur nuclear power plant in Ishwardi, Bangladesh. Groundw Sustain Dev. https://doi.org/10.1016/J.GSD.2018.06.002.
Uddin, Md.G., Nash, S., Diganta, M.T.M., Rahman, A., Olbert, A.I., 2022a. A comparison of geocomputational models for validating geospatial distribution of water quality index. In: Priyanka, H., Rahman, A., Basant agarwal, Binita Tiwari (Eds.), Computational Statistical Methodologies and Modeling for Artificial Intelligence. CRC Press, Taylor & Francis Publisher, USA.
Uddin, Md.G., Nash, S., Olbert, A.I., 2020a. Assessment of water quality using water quality index (WQI) models and advanced geostatistical technique. In: Ruane, K., Jaksic, V. (Eds.), Civil Engineering Research in Ireland 2020 (CERI2020) Conference. Cork, Ireland. Civil Engineering Research Association of Ireland, pp. 582–587.
Uddin, Md.G., Nash, S., Talas, M., Diganta, M., Rahman, A., Olbert, A.I., 2022b. Robust machine learning algorithms for predicting coastal water quality index. J. Environ. Manage. https://doi.org/10.1016/j.jenvman.2022.115923.
Uddin, M.G., Diganta, M.T.M., Sajib, A.M., Hasan, Md.A., Moniruzzaman, Md., Rahman, A., Olbert, A.I., Moniruzzaman, M., 2023a. Assessment of hydrogeochemistry in groundwater using water quality index model and indices approaches. Heliyon 9, e19668. https://doi.org/10.1016/j.heliyon.2023.e19668.
Uddin, M.G., Diganta, M.T.M., Sajib, A.M., Rahman, A., Nash, S., Dabrowski, T., Ahmadian, R., Hartnett, M., Olbert, A.I., 2023b. Assessing the impact of COVID-19 lockdown on surface water quality in Ireland using advanced Irish water quality index (IEWQI) model. Environ. Pollut. 336, 122456 https://doi.org/10.1016/j.envpol.2023.122456.
Uddin, M.G., Jackson, A., Nash, S., Rahman, A., Olbert, A.I., 2023c. Comparison between the WFD approaches and newly developed water quality model for monitoring transitional and coastal water quality in Northern Ireland. Sci. Total Environ. 901, 165960 https://doi.org/10.1016/j.scitotenv.2023.165960.
Uddin, M.G., Nash, S., Olbert, A.I., 2021. A review of water quality index models and their use for assessing surface water quality. Ecol. Indic. 122, 107218 https://doi.org/10.1016/j.ecolind.2020.107218.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023d. A sophisticated model for rating water quality. Sci. Total Environ. 869, 161614 https://doi.org/10.1016/j.scitotenv.2023.161614.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023e. Assessing optimization techniques for improving water quality model. J. Clean. Prod. 385, 135671 https://doi.org/10.1016/j.jclepro.2022.135671.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023f. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. Water. Res. 229, 119422 https://doi.org/10.1016/j.watres.2022.119422.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023g. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Saf. Environ. Protect. 169, 808–828. https://doi.org/10.1016/j.psep.2022.11.073.
Uddin, M.G., Nash, S., Rahman, A., Dabrowski, T., Olbert, A.I., 2024. Data-driven modelling for assessing trophic status in marine ecosystems using machine learning approaches. Environ. Res. 242, 117755 https://doi.org/10.1016/j.envres.2023.117755.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022c. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. Water. Res. 219 https://doi.org/10.1016/j.watres.2022.118532.
Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022d. Development of a water quality index model - a comparative analysis of various weighting methods. In: Çiner, Prof. Dr.A. (Ed.), Mediterranean Geosciences Union Annual Meeting (MedGU-21). Istanbul. Spinger, pp. 1–6.
Uddin, M.G., Rahman, A., Nash, S., Diganta, M.T.M., Sajib, A.M., Moniruzzaman, M., Olbert, A.I., 2023h. Marine waters assessment using improved water quality model incorporating machine learning approaches. J. Environ. Manage 344, 118368. https://doi.org/10.1016/j.jenvman.2023.118368.

Uddin, M.G., Nash, Stephen, Olbert, A.I., 2020b. Application of water quality index models to an Irish Estuary. In: Kieran Runae, V.J. (Ed.), Civil and Environmental Research. Civil and Environmental Research, Cork, Ireland, pp. 576–581.
Uddin, M.G., Stephen, Nash, Olbert, A.I., 2023h. Development of an efficient water quality model using cutting-edge artificial intelligence techniques. In: College of Science and Engineering, Inaugural Research and Innovation Day 2023. University of Galway, Ireland.
van Zoest, V.M., Stein, A., Hoek, G., 2018. Outlier detection in urban air quality sensor networks. Water. Air. Soil. Pollut. 229, 111. https://doi.org/10.1007/s11270-018-3756-7.
Varadharajan, C., Appling, A.P., Arora, B., Christianson, D.S., Hendrix, V.C., Kumar, V., Lima, A.R., Müller, J., Oliver, S., Ombadi, M., Perciano, T., Sadler, J.M., Weierbach, H., Willard, J.D., Xu, Z., Zwart, J., 2022. Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? Hydrol. Process. 36, e14565. https://doi.org/10.1002/hyp.14565.
Villa, G., Lozano, S., 2020. Data envelopment analysis and non-parametric analysis. In: Charles, V., Aparicio, J., Zhu, J. (Eds.), Data Science and Productivity Analytics. Springer International Publishing, Cham, pp. 121–160. https://doi.org/10.1007/978-3-030-43384-0_5.
von Rosing, M., Scheer, A.-W., Zachman, J.A., Jones, D.T., Womack, J.P., von Scheel, H., 2015. Phase 3: process concept evolution. In: von Rosing, M., Scheer, A.-W., von Scheel, H. (Eds.), The Complete Business Process Handbook. Morgan Kaufmann, Boston, pp. 37–77. https://doi.org/10.1016/B978-0-12-799959-3.00003-3.
Wahid, A., Rao, A.C.S., Deb, K., 2018. A relative kernel-density based outlier detection algorithm. In: 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), pp. 1–7. https://doi.org/10.1109/SKIMA.2018.8631526.
Wang, H., Yang, S., Liu, Y., Li, Q., 2023a. A novel abnormal data detection method based on dynamic adaptive local outlier factor for the vibration signals of rotating parts. Meas. Sci. Technol. 34, 085118 https://doi.org/10.1088/1361-6501/accbda.
Wang, L., Dong, H., Cao, Y., Hou, D., Zhang, G., 2023b. Real-time water quality detection based on fluctuation feature analysis with the LSTM model. J. Hydroinform. 25, 140–149. https://doi.org/10.2166/hydro.2023.127.
Wei, L., Niraula, D., Gates, E.D.H., Fu, J., Luo, Y., Nyflot, M.J., Bowen, S.R., El Naqa, I. M., Cui, S., 2023. Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. Br. J. Radiol., 20230211 https://doi.org/10.1259/bjr.20230211.
Wilcox, R.R., 2003. 12 - Multiple comparisons. In: Wilcox, R.R. (Ed.), Applying Contemporary Statistical Techniques. Academic Press, Burlington, pp. 407–456. https://doi.org/10.1016/B978-012751541-0/50033-X.
Wu, Z.Y., Chew, A., Meng, X., Cai, J., Pok, J., Kalfarisi, R., Lai, K.C., Hew, S.F., Wong, J. J., 2021. Data-driven and model-based framework for smart water grid anomaly detection and localization. AQUA - Water Infrastruct. Ecosyst. Soc. 71, 31–41. https://doi.org/10.2166/aqua.2021.091.
Xu, H., Pang, G., Wang, Y., Wang, Y., 2023. Deep isolation forest for anomaly detection. IEEe Trans. Knowl. Data Eng. 1–14. https://doi.org/10.1109/TKDE.2023.3270293.
Xu, H., Zhang, L., Li, P., Zhu, F., 2022. Outlier detection algorithm based on k-nearest neighbors-local outlier factor. J. Algorithm. Comput. Technol. 16 https://doi.org/10.1177/17483026221078111, 17483026221078112.
Xu, Y., Xu, N., Feng, X., 2016. A new outlier detection algorithm based on kernel density estimation for ITS. In: 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 258–262. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.67.
Xu, Z., Kakde, D., Chaudhuri, A., 2019. Automatic hyperparameter tuning method for local outlier factor, with applications to anomaly detection. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 4201–4207. https://doi.org/10.1109/BigData47090.2019.9006151.
Yeganeh, A., Shongwe, S.C., 2023. A novel application of statistical process control charts in financial market surveillance with the idea of profile monitoring. PLoS. One 18, e0288627–.
Yin, H., Wu, Q., Yin, S., Dong, S., Dai, Z., Soltanian, M.R., 2023. Predicting mine water inrush accidents based on water level anomalies of borehole groups using long short-term memory and isolation forest. J. Hydrol. 616, 128813 https://doi.org/10.1016/j.jhydrol.2022.128813.
Yin, S., Liu, H., 2022. Wind power prediction based on outlier correction, ensemble reinforcement learning, and residual correction. Energy 250, 123857. https://doi.org/10.1016/j.energy.2022.123857.
Yin, Z., Fang, X., 2021. An Outlier-Robust Point and Interval Forecasting System for Daily PM2.5 Concentration. Front. Environ. Sci. 9.
Yuan, Z., Zhang, X., Feng, S., 2018. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures. Expert. Syst. Appl. 112, 243–257. https://doi.org/10.1016/j.eswa.2018.06.013.
Yuen, K.-V., Ortiz, G.A., 2017. Outlier detection and robust regression for correlated data. Comput. Methods Appl. Mech. Eng. 313, 632–646. https://doi.org/10.1016/j.cma.2016.10.004.
Zeng, X., Shen, S.-H., Shen, H., Luo, D.-Y., 2023a. Statistical process control for the analysis of quality control in urodynamics: a potential new approach for quality review of urodynamics. Neurourol. Urodyn. 42, 289–296. https://doi.org/10.1002/nau.25081.
Zeng, Z., Huang, R., Xiao, R., Lin, X., Zhang, S., 2023b. Anomaly detection for high-dimensional dynamic data stream using stacked habituation autoencoder and union kernel density estimator. Concurr. Comput. 35, e7718. https://doi.org/10.1002/cpe.7718.
Zhang, J., Fu, P., Meng, F., Yang, X., Xu, J., Cui, Y., 2022a. Estimation algorithm for chlorophyll-a concentrations in water from hyperspectral images based on feature derivation and ensemble learning. Ecol. Inform. 71 https://doi.org/10.1016/j.ecoinf.2022.101783.
Zhang, W., Huang, W., Tan, J., Guo, Q., Wu, B., 2022b. Heterogeneous catalysis mediated by light, electricity and enzyme via machine learning: paradigms, applications and prospects. Chemosphere. https://doi.org/10.1016/j.chemosphere.2022.136447.
Zhang, W., Liu, Y., 2019. Chapter 19 - Model validation of control systems with an application in abnormal driving state detection. In: Zhang, L., Zeigler, B.P., laili, Y. (Eds.), Model Engineering for Simulation. Academic Press, pp. 419–429. https://doi.org/10.1016/B978-0-12-813543-3.00019-6.
Zhang, Y., Thorburn, P.J., 2022. Handling missing data in near real-time environmental monitoring: a system and a review of selected methods. Fut. Gener. Comput. Syst. 128, 63–72. https://doi.org/10.1016/j.future.2021.09.033.
Zhao, H., Jiang, X., Wang, B., Cheng, X., Xu, S., 2022. Towards smart monitoring of systems: an integrated non-parametric Bayesian KDE and LSTM approach for anomaly detection of rotating machinery under uncertainties. Struct. Health Monit. 22, 1984–2001. https://doi.org/10.1177/14759217221117277.
Zhao, C., Yang, J., 2019. A robust skewed boxplot for detecting outliers in rainfall observations in real-time flood forecasting. Adv. Meteorol. 2019, 1795673 https://doi.org/10.1155/2019/1795673.
Zheng, Z., Jeong, H.-Y., Huang, T., Shu, J., 2017. KDE based outlier detection on distributed data streams in multimedia network. Multimed. Tools. Appl. 76, 18027–18045. https://doi.org/10.1007/s11042-016-3681-y.