Gorgan University of Agricultural Sciences and Natural Resources

Journal of Water and Soil Conservation

Print ISSN: 2322-2069
Online ISSN: 2322-2794

(OPEN ACCESS)

Forecasting Water Quality Parameter Using a Novel Kernel-Based Method with Feature Selection and Multivariate Decomposition

Masoud Dorfeshan[1], Iman Ahmadianfar[*2], Arvin Samadi Koucheksaraee[3]

1. Dept. of Mechanical Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran. E-mail: m.dorfeshan@gmail.com
2. Corresponding Author, Dept. of Civil Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran. E-mail: i.ahmadianfar@bkatu.ac.ir
3. Dept. of Civil, Construction and Environmental Engineering (Dept 2470), North Dakota State University, PO Box 6050, Fargo, ND, 58108-6050, USA. E-mail: arvin.samadi.k@ndsu.edu

Article Info

Article type:
Research Full Paper

Article history:
Received: 12.02.2024
Revised: 12.31.2024
Accepted: 03.10.2025

Keywords:
Decomposition,
Feature selection,
Forecasting,
Improved kernel ridge,
WQ

ABSTRACT

Background and Objectives: Precise forecasting of water quality (WQ) parameters, specifically PS (potential salinity), is critical for sustainable water utilization. In water-stressed regions like the Karun River in Iran, effective monitoring and prediction of the PS is not only important but also critical because of anthropogenic activities, climate change, and reduced inflows of freshwater. Therefore, effective machine learning (ML) models and appropriate input data is very important for monitoring and predicting WQ parameters. However, the influencing factors exhibit complex and non-linear relationships, and multicollinearity in the datasets makes it challenging for traditional ML models to address the problem. Limitations, thus, can result in inaccurate predictions, which obstruct the establishment of sustainable water management strategies. As mentioned above, accurate forecasting of PS is essential for water and soil conservation, because PS helps mitigate salinity-related degradation of agricultural lands and ensure the sustainability of vital ecosystems. This study supports the development of effective conservation strategies to maintain soil productivity and WQ in vulnerable regions by providing reliable predictions. To address these issues, the present study introduces a new hybrid model, IKRidge-GRM, which inherits the advantages of improved kernel ridge regression (IKRidge) and generalized ridge regression (GRM). The hybrid model integrates IKRidge's improved capacity to identify non-linearity with GRM's resilience against multicollinearity problems to improve the predictive performance of the PS prediction. This unique framework offers improved stability and interpretability of results, as well as increases forecast accuracy, making it a helpful tool for environmental monitoring and decision-making. The proposed strategy could aid policymakers and water resource managers in designing reasonable strategies to alleviate salinity issues, protect aquatic ecosystems, and ensure the long-term survival of vital water sources like the Karun River.

Materials and Methods: This study introduces a novel hybrid ML model based on two regression techniques, namely: generalized ridge regression (GRM) and improved kernel ridge regression (IKRidge), called IKRidge-GRM. The GRM effectively addresses multicollinearity and overfitting

issues using the iteratively reweighted least squares (IRLS) process. On the other hand, IKRidge incorporates a wavelet kernel function, optimized through the INFO algorithm, and the regularized locally weighted (RLW) approach, enabling it to capture complex, non-linear patterns in the data with high precision. This combination of techniques allows the hybrid model to overcome the limitations of traditional ML methods, making it particularly suitable for handling the intricate relationships inherent in WQ datasets. To further enhance the model's predictive accuracy, the IKRidge-GRM framework integrates a light gradient boosting machine (LGBM) for feature selection. It reduces dimensionality by identifying the most relevant input variables while eliminating redundant or irrelevant features. Additionally, the model employs multivariate variational mode decomposition (MVMD) to decompose the input data into high- and low-frequency components, allowing it to capture both short-term fluctuations and long-term trends in WQ parameters. The study utilized an extensive dataset comprising 48 years of monthly WQ data collected from the Farisat station on the Karun River. Nine keys WQ parameters, including magnesium (Mg), sulfate (SO_4^{2-}), calcium (Ca), discharge (Q), sodium (Na), bicarbonate (HCO3), chloride (Cl), electrical conductivity (EC), total dissolved solids (TDS) and pH, were used as inputs to forecast the PS three months ahead.

Results: The proposed IKRidge-GRM model accurately predicted PS values at the Farisat station, significantly outperforming baseline models (Ridge, DELM, and LSSVM) and their MVMD-enhanced versions. By leveraging its hybrid architecture and advanced feature extraction techniques, the MVMD-IKRidge-GRM model achieved remarkable results during the testing phase, with the highest correlation coefficient (R=0.977), the lowest RMSE (0.956), and the lowest MAPE (4.521). These metrics indicate the model's superior predictive accuracy and reliability in handling complex, non-linear relationships. The model also achieved high IA (0.988) and KGE (0.948) scores, underscoring its robustness and effectiveness in capturing the intricate dynamics of the PS variations. These results highlight the model's ability to uncover hidden patterns in the data and provide highly accurate predictions, even in challenging scenarios involving multicollinearity and non-linear dependencies. The model's exceptional performance was further confirmed by visual evaluations such as scatter plots, relative error plots, and Taylor diagrams. Scatter plots demonstrated that the MVMD-IKRidge-GRM model's predictions closely aligned with measured values, with minimal prediction intervals and narrow error distributions, reflecting its precision and consistency. Relative error plots revealed that the model exhibited the most compact and symmetric error distribution, with minimal bias and variability. Relative error plots also indicated the models' ability to generalize well across different data points. Taylor diagrams provided evidence of the model's strong agreement with reference data, showcasing its ability to balance accuracy, variability representation, and error minimization effectively. Residual analysis further confirmed the model's precision and reliability. Among all the models tested, the MVMD-IKRidge-GRM model achieved the smallest mean residual (-0.0073) and the lowest standard deviation (0.0613), demonstrating its ability to minimize prediction errors consistently. This level of precision is critical for practical applications, as it ensures that the model can provide reliable forecasts for decision-making in water resource management. The model's ability to integrate advanced regression techniques, feature selection, and frequency decomposition enhances its predictive capabilities. The ability also establishes the proposed model as a robust framework for addressing complex environmental challenges. These

findings emphasized the potential of the MVMD-IKRidge-GRM model as a powerful tool for sustainable water resource management, particularly in regions like the Karun River basin, where accurate and reliable predictions are essential for mitigating environmental degradation and ensuring long-term ecological balance.

Conclusion: The IKRidge-GRM model predicted PS values at the Farisat station on the Karun River. The findings demonstrated high accuracy and reliability across all evaluation metrics. The IKRidge-GRM model has the ability to uncover hidden patterns in complex, non-linear datasets. Its capacity to deliver precise predictions also highlights its potential as a valuable tool for environmental monitoring and management. By integrating advanced regression techniques, such as improved kernel ridge regression (IKRidge) and generalized ridge regression (GRM), with innovative feature selection and decomposition methods like light gradient boosting machine (LGBM) and multivariate variational mode decomposition (MVMD), the model effectively addresses challenges such as multicollinearity, overfitting, and non-linear relationships. This comprehensive framework ensures that the IKRidge-GRM model achieves superior predictive performance and maintains robustness and adaptability across diverse environmental conditions. This study emphasizes the importance of combining advanced ML techniques with effective preprocessing methods to develop reliable models for analyzing and forecasting complex environmental data. Integrating feature selection and frequency decomposition enhances the model's ability to extract meaningful information from high-dimensional datasets. This integration also enable the models to capture both short-term fluctuations and long-term trends in WQ parameters better. Such capabilities are essential for addressing the multifaceted challenges posed by environmental degradation, particularly in regions like the Karun River basin, where water resources are under significant stress due to anthropogenic activities and climate change.

Cite this article: Dorfeshan, Masoud, Ahmadianfar, Iman, Samadi Koucheksaraee, Arvin. 2025. Forecasting Water Quality Parameter Using a Novel Kernel-Based Method with Feature Selection and Multivariate Decomposition. *Journal of Water and Soil Conservation*, 32 (1), 105-127.

DOI: 10.22069/jwsc.2025.23310.3790
Publisher: Gorgan University of Agricultural Sciences and Natural Resources

نشریه پژوهش‌های حفاظت آب و خاک

شاپا چاپی: ۲۰۶۹-۲۳۲۲
شاپا آنلاین: ۲۷۹۴-۲۳۲۲

پیش‌بینی پارامتر کیفیت آب با استفاده از یک روش نوین کرنل‌محور همراه با انتخاب ویژگی و تجزیه چندمتغیره

مسعود درفشان[1]، ایمان احمدیان‌فر[2*]، آروین صمدی‌کوچکسرایی[3]

۱. گروه مکانیک، دانشکده فنی- مهندسی، دانشگاه صنعتی خاتم‌الانبیاء بهبهان، بهبهان، ایران. رایانامه: m.dorfeshan@gmail.com

۲. نویسنده مسئول، گروه عمران، دانشکده فنی - مهندسی، دانشگاه صنعتی خاتم‌الانبیاء بهبهان، بهبهان، ایران. رایانامه: i.ahmadianfar@bkatu.ac.ir

۳. دانشکده مهندسی عمران، ساخت و ساز و محیط زیست (دپارتمان ۲٤۷۰)، دانشگاه ایالتی داکوتای شمالی. رایانامه: arvin.samadi.k@ndsu.edu

اطلاعات مقاله

نوع مقاله:
مقاله کامل علمی- پژوهشی

تاریخ دریافت: ۰۳/۰۹/۱۲
تاریخ ویرایش: ۰۳/۱۰/۱۱
تاریخ پذیرش: ۰۳/۱۲/۲۰

واژه‌های کلیدی:
انتخاب ویژگی،
پیش‌بینی،
تجزیه،
کرنل ریج بهبودیافته،
کیفیت آب

چکیده

سابقه و هدف: پیش‌بینی دقیق پارامترهای کیفیت آب، به‌ویژه پتانسیل شوری (PS)، برای استفاده پایدار از منابع آبی بسیار مهم است. در مناطق تحت فشار آبی مانند رودخانه کارون در ایران، نظارت و پیش‌بینی مؤثر PS نه تنها اهمیت دارد بلکه به‌دلیل فعالیت‌های انسانی، تغییرات آب و هوا و کاهش ورودی‌های آب شیرین، حیاتی است. بنابراین، مدل‌های یادگیری ماشین (ML) مؤثر و داده‌های ورودی مناسب برای نظارت و پیش‌بینی پارامترهای کیفیت آب بسیار مهم هستند. با این‌حال، عوامل تأثیرگذار روابط پیچیده و غیرخطی را نمایش می‌دهند و چندخطی بودن در داده‌ها، باعث می‌شود که مدل‌های سنتی ML نتوانند به‌طور مؤثر به این مشکل پرداخته و پیش‌بینی‌های درستی ارائه دهند که مانع ایجاد استراتژی‌های پایدار مدیریت آب می‌شود. همان‌طور که ذکر شد، پیش‌بینی دقیق PS برای حفظ آب و خاک ضروری است زیرا PS به کاهش تخریب مرتبط با شوری زمین‌های کشاورزی کمک می‌کند و بر پایداری اکوسیستم‌های حیاتی تأثیر می‌گذارد. این مطالعه از توسعه استراتژی‌های مؤثر حفاظت برای حفظ بهره‌وری خاک و کیفیت آب در مناطق آسیب‌پذیر با ارائه پیش‌بینی‌های قابل اعتماد پشتیبانی می‌کند. برای پرداختن به این مسائل، پژوهش حاضر یک مدل هیبریدی جدید به نام IKRidge-GRM را معرفی می‌کند که مزایای رگرسیون کرنل ریج بهبودیافته (IKRidge) و رگرسیون ریج تعمیم‌یافته (GRM) را به دربر می‌گیرد. این مدل هیبریدی ظرفیت بهبود یافته IKRidge برای شناسایی عدم‌خطی بودن را با روش GRM در برابر مسائل چندخطی ترکیب می‌کند تا عملکرد پیش‌بینی PS را بهبود بخشد. این چارچوب منحصر به فرد، ثبات و تفسیرپذیری بهتری از نتایج را ارائه می‌دهد و دقت پیش‌بینی را افزایش می‌دهد، بنابراین ابزاری مفید برای نظارت بر محیط زیست و تصمیم‌گیری به حساب می‌آید. استراتژی پیشنهادی می‌تواند به سیاست‌گذاران و مدیران منابع آبی در طراحی استراتژی‌های معقول برای کاهش مسائل شوری، حفاظت از اکوسیستم‌های آبی و تضمین بقای پایدار منابع آبی حیاتی مانند رودخانه کارون کمک کند.

مواد و روش‌ها: این مطالعه یک مدل جدید هیبریدی یادگیری ماشین براساس دو تکنیک رگرسیون، یعنی رگرسیون ریج تعمیم‌یافته (GRM) و رگرسیون کرنل ریج بهبودیافته (IKRidge) را معرفی می‌کند که IKRidge-GRM نامیده می‌شود. GRM به‌طور مؤثر مشکلات چندخطی بودن و بیش‌برازش را با استفاده از فرآیند حداقل مربعات وزنی تکراری (IRLS) برطرف می‌کند. از سوی دیگر، IKRidge یک تابع کرنل موجک را که از طریق الگوریتم INFO بهینه‌سازی شده است و هم‌چنین رویکرد وزنی محلی تنظیم شده را شامل می‌شود که به آن اجازه می‌دهد الگوهای پیچیده و غیرخطی را در داده‌ها با دقت بالا شناخته و شناسایی کند. این ترکیب از تکنیک‌ها به مدل هیبریدی اجازه می‌دهد تا محدودیت‌های روش‌های سنتی ML را برطرف کند و آن را به‌طور ویژه مناسب برای رسیدگی به روابط پیچیده در داده‌های کسفیت آب بسازد. برای افزایش بیش‌تر دقت پیش‌بینی مدل، چارچوب IKRidge-GRM یک ماشین تقویت گرادیان سبک (LGBM) را برای انتخاب ویژگی‌ها ادغام می‌کند. این فرآیند با شناسایی مرتبط‌ترین متغیرهای ورودی و حذف ویژگی‌های زائد یا نامربوط، ابعاد داده‌ها را کاهش می‌دهد. علاوه بر این، مدل از تجزیه مؤلفه‌های مدور چندمتغیره (MVMD) استفاده می‌کند تا داده‌های ورودی را به مؤلفه‌های با فرکانس بالا و پایین تجزیه کند و به این ترتیب به آن امکان می‌دهد که هم نوسانات کوتاه‌مدت و هم روندهای بلندمدت در پارامترهای کیفیت آب را شناسایی کند. این مطالعه از یک مجموعه داده گسترده متشکل از ٤٨ سال داده‌های ماهانه کیفیت آب جمع‌آوری‌شده از ایستگاه فاریسات در رودخانه کارون استفاده کرد. نه پارامتر کلیدی، شامل منیزیم (Mg)، سولفات (SO_4^{-2})، کلسیم (Ca)، دبی(Q)، سدیم (Na)، بیکربنات (HCO_3)، کلرید (Cl)، هدایت الکتریکی(EC)، مواد جامد حل‌شده کل (TDS) و pH، به‌عنوان ورودی‌ها برای پیش‌بینی PS سه ماه آینده مورد استفاده قرار گرفتند.

نتایج: مدل پیشنهادی IKRidge-GRM پیش‌بینی‌های دقیقی از مقادیر PS در ایستگاه فاریسات با عملکردی به‌مراتب بهتر از مدل‌های پایه (Ridge، DELM و LSSVM)، و نسخه‌های تقویت‌شده MVMD آن‌ها ارائه کرد. با استفاده از معماری هیبریدی و تکنیک‌های پیشرفته استخراج ویژگی، مدل MVMD-IKRidge-GRM در مرحله آزمایش نتایج چشمگیری کسب کرده است، با بالاترین ضریب همبستگی (R=۰/۹۷۷)، کم‌ترین RMSE (۰/۹۵٦) و کم‌ترین MAPE (٤/٥٢١). این معیارها نشان‌دهنده دقت پیش‌بینی و قابلیت اطمینان بالای مدل در رسیدگی به روابط پیچیده و غیرخطی هستند. این مدل همچنین رتبه بالایی در IA (۰/۹۸۸) و KGE (۰/۹٤۸) کسب کرده است که قدرت و کارایی آن را در شناسایی دینامیک‌های پیچیده تغییرات PS نشان می‌دهد. این نتایج به توانایی مدل در کشف الگوهای پنهان در داده‌ها و ارائه پیش‌بینی‌های بسیار دقیق حتی در سناریوهای چالش‌برانگیز که شامل چندخطی بودن و وابستگی‌های غیرخطی است، تأکید می‌کند. عملکرد استثنائی مدل هم‌چنین با ارزیابی‌های بصری مانند نمودارهای پراکندگی، نمودارهای نسبت خطا و دیاگرام‌های تیلور تأیید شد. نمودارهای پراکندگی نشان‌دهنده این بود که پیش‌بینی‌های مدل MVMD-IKRidge-GRM به‌خوبی با مقادیر اندازه‌گیری‌شده هم‌راستا بودند، با حداقل فواصل پیش‌بینی و توزیع‌های خطای باریک، که دقت و ثبات آن را منعکس می‌کند. نمودارهای نسبت خطا نشان دادند که مدل توزیع خطای متراکم و متقارن‌تری را با کم‌ترین بایاس و تغییرپذیری به نمایش گذاشت. هم‌چنین، نمودارهای نسبت خطا توانایی مدل‌ها را در تعمیم‌پذیری خوب در سراسر نقاط داده

مختلف نشان دادند. دیاگرام‌های تیلور شواهدی از توافق قوی مدل با داده‌های مرجع ارائه داد و نشان داد که مدل به‌خوبی می‌تواند دقت، نمایندگی تغییرپذیری و حداقل خطا را متعادل سازد. تحلیل باقی‌مانده هم‌چنین دقت و قابلیت اطمینان مدل را تأیید کرد. در میان تمامی مدل‌های آزمایش شده، مدل MVMD-IKRidge-GRM کوچک‌ترین میانگین باقی‌مانده (۰/۰۰۷۳-) و کم‌ترین انحراف معیار (۰/۰۶۱۳) را به‌دست آورد، که نشان‌دهنده توانایی آن در به حداقل رساندن خطاهای پیش‌بینی به‌صورت پیوسته است. این سطح از دقت برای کاربردهای عملی بسیار حیاتی است زیرا اطمینان حاصل می‌کند که مدل می‌تواند پیش‌بینی‌های قابل اعتمادی برای تصمیم‌گیری در مدیریت منابع آب ارائه دهد. توانایی مدل در ادغام تکنیک‌های پیشرفته رگرسیون، انتخاب ویژگی و تجزیه فرکانس، قابلیت‌های پیش‌بینی آن را افزایش می‌دهد. این توانایی همچنین مدل پیشنهادی را به‌عنوان یک چارچوب توانمند برای پرداختن به چالش‌های پیچیده محیط زیست تثبیت می‌کند. این یافته‌ها بر قدرت مدل MVMD-IKRidge-GRM به‌عنوان ابزاری قوی برای مدیریت پایدار منابع آب تأکید کردند، به‌ویژه در مناطقی مانند حوضه رودخانه کارون که پیش‌بینی‌های دقیق و قابل‌اعتماد برای کاهش تخریب محیطی و تضمین تعادل اکولوژیکی بلندمدت ضروری است.

نتیجه‌گیری: مدل IKRidge-GRM مقادیر PS را در ایستگاه فاریسات در رودخانه کارون پیش‌بینی کرد. یافته‌ها دقت و قابلیت اطمینان بالایی را در تمام معیارهای ارزیابی نشان دادند. مدل IKRidge-GRM توانایی کشف الگوهای پنهان در داده‌های پیچیده و غیرخطی را دارد. ظرفیت آن برای ارائه پیش‌بینی‌های دقیق، هم‌چنین پتانسیل آن را به‌عنوان ابزاری ارزشمند برای نظارت و مدیریت محیط زیست نشان می‌دهد. با ادغام تکنیک‌های رگرسیون پیشرفته، هم‌چون رگرسیون کرنل ریج بهبود یافته (IKRidge) و رگرسیون ریج تعمیم‌یافته (GRM) با روش‌های نوآورانه انتخاب ویژگی و تجزیه، هم‌چون ماشین تقویت گرادیان سبک (LGBM) و تجزیه مؤلفه‌های چندمتغیره (MVMD)، مدل به‌طور مؤثری چالش‌هایی مانند چندخطی بودن، بیش‌برازش و روابط غیرخطی را برطرف می‌کند. این چارچوب جامع اطمینان می‌دهد که مدل IKRidge-GRM عملکرد پیش‌بینی فوق‌العاده‌ای را حاصل کرده و در شرایط محیطی متنوع، مقاوم و انعطاف‌پذیر باقی می‌ماند. این مطالعه بر اهمیت ترکیب تکنیک‌های پیشرفته ML با روش‌های مؤثر پیش‌پردازش تأکید می‌کند تا مدل‌های قابل اعتمادی را برای تحلیل و پیش‌بینی داده‌های پیچیده محیط زیست توسعه دهند. ادغام انتخاب ویژگی و تجزیه فرکانس، توانایی مدل را در استخراج اطلاعات معنادار از مجموعه داده‌های با ابعاد بالا افزایش می‌دهد. این یکپارچگی هم‌چنین به مدل‌ها اجازه می‌دهد نوسانات کوتاه‌مدت و روندهای بلندمدت در پارامترهای کیفیت آب را بهتر شناسایی کنند. چنین قابلیت‌هایی برای پرداختن به چالش‌های چندوجهی ناشی از تخریب محیط زیست، به‌ویژه در مناطقی مانند حوضه رودخانه کارون که منابع آبی تحت فشار قابل‌توجهی به‌دلیل فعالیت‌های انسانی و تغییرات اقلیمی هستند، ضروری است.

استناد: درفشان، مسعود، احمدیان‌فر، ایمان، صمدی‌کوچکسرایی، آروین (۱۴۰۴). پیش‌بینی پارامتر کیفیت آب با استفاده از یک روش نوین کرنل‌محور همراه با انتخاب ویژگی و تجزیه چندمتغیره. *پژوهش‌های حفاظت آب و خاک*، ۳۲ (۱)، ۱۲۷-۱۰۵.
DOI: 10.22069/jwsc.2025.23310.3790

ناشر: دانشگاه علوم کشاورزی و منابع طبیعی گرگان

Introduction

Water, as a vital resource for life, agriculture, industry, and the preservation of biodiversity, plays a fundamental role in the development of societies (Bui et al., 2020; Chang et al., 2015). With increasing demand and urbanization, water consumption has risen significantly (Salarijazi et al., 2024; Zhou et al., 2024). Simultaneously, Water pollution in major rivers worldwide, such as the Karun River in Iran, is primarily caused by industrial, agricultural, and urban activities. This pollution has escalated into a severe crisis, threatening public health, economic growth, and sustainable development (Ahmadianfar, Shirvani-Hosseini, He et al., 2022; Asadollah et al., 2021). WQ prediction is an effective tool for managing this crisis because it provides valuable information for water-dependent industries and resource managers (Chatterjee et al., 2017; Chen et al., 2024). These predictions contribute to better planning, pollution reduction, and water usage optimization, thereby positively impacting the economy and public health (Gharemahmudli and Seyed Hamidreza Sadeghi, 2024; Zahiri et al., 2024).

Many studies have been conducted to develop the WQ prediction models (Deng et al., 2015; Huang et al., 2018; Jamei et al., 2021). These models mainly use two approaches: physics-based (PB) or ML methods (Qiu et al., 2020). Physics-based models, designed based on hydrodynamic laws, require deep knowledge of physical and chemical processes and detailed information about pollution sources and tributaries (Han et al., 2021). In contrast, ML methods rely only on historical data to establish mathematical relationships between parameters without considering complex theories or model calibration (Wu et al., 2021). One more advantage for ML models is their ability to transfer and apply to different locations easily. The mentioned reasons have promoted many researchers to widely explore and use ML methods.

Artificial intelligence (AI) advancements have increased attention on ML models for data-driven modeling (Ahmadianfar, Shirvani- Hosseini, Samadi-Koucheksaraee et al., 2022; Ahmed et al., 2019). For example, in the study by Barzegar et al. (2016), the accuracy of four different models in predicting the salinity of the Ajichay River was evaluated (Barzegar et al., 2016). The results showed that the adaptive neuro-fuzzy inference system (ANFIS) model performed better than the artificial neural network (ANN) model. Additionally, the hybrid wavelet-ANFIS and wavelet-ANN models, using the db4 wavelet transform, had higher prediction accuracy than the ANFIS and ANN models. These findings indicated that wavelet-based hybrid methods could improve the performance of WQ prediction models. In another study, Haddad et al. (2017) compared the performance of genetic programming (GP) and the least squares support vector regression (LSSVR) model in predicting the WQ of the Sefidrud River (Bozorg-Haddad et al., 2017). They used principal component analysis (PCA) to select effective inputs and applied the genetic algorithm (GA) to optimize the parameters of the LSSVR model. The results showed that the GA-LSSVR model had higher accuracy than the GP model, confirming the importance of hybrid and optimization methods in WQ modeling.

In another study, Ahmadianfar et al. (2020) investigated the W-LWLR method, a combination of wavelet transforms and locally weighted linear regression, for predicting the electrical conductivity (EC) of the Sefidrud River (Ahmadianfar, Jamei, et al., 2020). A comparison of this method with models such as SVR (support vector regression (SVR)), W-SVR (wavelet SVR), ARIMA (Autoregressive Integrated Moving Average (ARIMA)) and W-ARIMA showed that W-LWLR had higher accuracy. This study highlighted that combining local methods with wavelet analysis was able to improve the accuracy of WQ parameter predictions. In an additional analysis, Ahmadianfar et al. (2022) combined the ANFIS model with an adaptive hybrid optimization method, including particle swarm optimization and differential evolution (ANFIS-DEPSO), to predict the electrical conductivity (EC) of the Maroon River in Iran (Ahmadianfar, Shirvani-Hosseini, He, et al., 2022). Due to wavelet analysis in the proposed hybrid model, the A-DEPSO-ANFIS model showed better performance than other tested models.

In the last few years, Wai et al. (2024) used GRU (gated recurrent unit) and LSTM (long short-term memory) models to predict WQ indices (Wai, et al., 2024). The VMD-LSTM (variational mode decomposition with LSTM) model, after decomposing input signals using EMD (empirical mode decomposition) and VMD methods, achieved better performance with a MAPE (mean absolute percentage error) of 1.9237% and a KGE (Kling-Gupta efficiency (KGE)) of 0.6761 compared to other models. Additionally, Jamei et al. (2024) used gaussian process regression (GPR) to predict the monthly sodium adsorption ratio (SAR) of the Zayandehrud River (Jamei et al., 2024). Applying the Boruta-SHAP method for feature selection, and TVF-EMD (time-varying filter-based EMD) and VMD for the decomposition of the input variables improved the accuracy of the GPR model. In addition, through the integration of climatological and geospatial data, Satish et al. (2024) enhanced WQ forecasts for the Godavari River Basin in India (Satish, et al., 2024). Their work distinguished nitrate levels as being associated with climate and land use factors. A stacked ANN meta-model, augmented with XGB, RF, and Extra Trees, exhibits enhanced predictive performance. In another study, Kandasamy et al. (2025) presented a hybrid structure that integrates remote sensing with ML methods to estimate chlorophyll-a values in rivers (Kandasamy et al., 2025). To accurately estimate Chl-a concentrations, they employed ensemble CatBoost and NBeats models. According to the results, the CatBoost model was able to make more accurate predictions.

Traditional ML models such as LSSVR, ANFIS, and GP have succeeded in the WQ prediction. However, they rely on complex architecture settings and optimization processes, which require high computational resources and make real-time applications difficult. Additionally, these models face limitations in understanding the complex and nonlinear dynamics of WQ data, which are influenced by external factors such as climate change and human activities. These limitations reduce the accuracy and stability of predictions. Developing a model that can integrate these complex methods into a unified and efficient framework remains challenging. The model should identify complex patterns in the data while still working efficiently. The model must combine the best features of different methods to be effective while minimizing their weaknesses. This goal can be achieved using innovative hybrid strategies or algorithms that leverage recent advancements in ML and data processing. This approach addresses current and future needs in WQ management and supports sustainable environmental development.

This paper developed a novel machine learning (ML) model named GKRidge, which builds upon the principles of generalized ridge regression and kernel ridge regression while incorporating the concept of regularized locally weighted regression. The model is further enhanced by integrating a feature-selection mechanism based on a Light Gradient Boosting Machine (LGBM) model, which is optimized using the Weighted Mean of Vectors (INFO) strategy (Ahmadianfar, Heidari, et al., 2022). This innovative combination enables the optimal selection of input variables to ensure robust predictive capabilities. Additionally, the Multivariate Variational Mode Decomposition (MVMD) technique is applied to the input variables to decompose their components effectively, thereby addressing noise and improving the predictive accuracy of the model.

The proposed approach is designed to enhance both computational performance and the stability of existing predictive modeling frameworks. By addressing limitations such as overfitting, instability, and inefficiency associated with traditional models, this method delivers a more accurate and reliable solution for water quality (WQ) prediction. It not only ensures greater predictive accuracy through rigorous feature selection and input variable decomposition but also introduces improvements in model computational efficiency, scalability, and consistency. As a result, GKRidge signifies a significant step forward in the development of advanced ML-based approaches for tackling complex WQ prediction challenges, providing a versatile and effective tool for researchers and practitioners in environmental modeling and data analysis.

Material and method

The proposed ML model

Generalized ridge regression

The Generalized Ridge Regression Method (GRM) is introduced to address issues such as multicollinearity and overfitting. This method combines Ridge regression and generalized linear model (GLM) (Nelder, and Wedderburn, 1972) to develop a powerful and flexible model. In the GRM, the model coefficients are continuously updated through the iteratively reweighted least squares (IRLS) process. The GRM selects an appropriate link function, its derivative, and a variance function for different distributions (such as normal, binomial, gamma, or Poisson). The link function is used at each iteration to calculate the mean response and linear prediction. Then the weight matrix and pseudo-response variable are generated. The basic formula of GRM is defined as follows,

$$\beta = (X^T \times \omega \times X)^{-1} \times (X^T \times \omega \times z) \quad (1)$$

in which

$$\omega = diag\left(\frac{h'(\eta))}{Var(Y)}\right) \quad (2)$$

$$z = \mu + \frac{y-\eta}{h'(\eta))} \quad (3)$$

where, μ represents the linear predictor for the observed values in a GLM. η denotes the mean of the response variables. $h'(\eta))$ represents the derivative of the link function h with respect to η. $Var(Y)$ indicates the variance function associated with the distribution of the response variable Y. Here, to improve the performance of the GLM, a regularization coefficient ρ_1 is used.

Using the main formula of GLM (Eq. (3)), the GRM is derived by adding ridge regularization. To achieve this, a regularization term (ρ_1) is included in the below equation (Eq. (4)). As a result, the coefficient for the GRM method is expressed as follows,

$$\beta = (X^T \times \omega \times X + \rho_1 \times UM)^{-1} \times (X^T \times \omega \times z) \quad (4)$$

where ρ_1 denotes the regularization factor. UM expresses the unit matrix. Eq. (5) is used to determine the predicted value ($\hat{y}_{GRM}$) generated by the GRM model.

$$\hat{y}_{GRM} = X\beta \quad (5)$$

Improved kernel ridge regression

Kernel ridge regression (KRidge) (Vovk, 2013) improves upon ridge regression by using kernel methods to handle non-linear relationships in data. While ridge regression addresses linear models and reduces overfitting with a penalty term, KRidge extends these by capturing complex, non-linear patterns. This makes KRidge more effective for modeling intricate datasets, offering greater accuracy and adaptability in scenarios where relationships are not purely linear. The predicted value ($\hat{y}_{KRidge}$) is calculated using Eq. (6).

$$\hat{y}_{KRidge} = X\alpha \tag{6}$$

In which

$$\alpha = (K + \rho_2 UM)^{-1} X^T \mathrm{y} \tag{6-1}$$

Where α is the regression factor, and ρ_2 and K denote the regularization coefficient and the kernel function, respectively. This research used the wavelet kernel function, which is defined as follows:

$$K_{jl} = \cos\left(a_1 \times \frac{-(x_j - x_l')}{a_2}\right) \times exp\left(\frac{-||x_j - x_l'||^2}{4 \times a_3}\right) \tag{7}$$

where a_1, a_2, and a_3 are the kernel function coefficients. In addition, the INFO optimization approach was employed to identify the best possible values for these factors.

To improve the forecasting accuracy of KRidge, this research proposed new input variable coefficients derived by the regularized locally weighted (RLW) approach. The proposed model is called improved kernel ridge regression (IKRidge). The core equation of the RLW method is expressed as follows,

$$\phi = (X^T \times \omega \times X + \rho_3 \times UM)^{-1} \times (X^T \times \omega \times y) \tag{8}$$

where ω is the wavelet kernel function. ρ_3 is the regularization coefficient. The ϕ is applied to generate a new kernel function according to Eq. (9),

$$X_{new} = \phi X \tag{9}$$

$$K_{new} = K(X_{new}, X_{new,l}) \tag{10}$$

where K_{new} is an improved version of K, obtained based on X_{new}.

Therefore, Eq. (6-1) is reformulated as,

$$\alpha_{new} = (K_{new} + \rho_2 UM)^{-1} X^T \mathrm{y} \tag{11}$$

and

$$\hat{y}_{IKRidge} = X\alpha_{new} \tag{12}$$

where α_{new} indicates a new coefficient for KRidge, achieved based on K_{new}.

Hybrid of IKRidge and GRM models

This study proposed an innovative hybrid regression model for predicting the irrigation water quality indexes (IWQIs). The foundation of the proposed model lies in the integration of two powerful regression techniques: the previously discussed IKRidge and the GRM, collectively referred to as the IKRidge-GRM model. To combine these models, the following relationship was established:

$$\hat{y}_{IKRidge-GRM} = c \times \hat{y}_{IKRidge} + (1 - c) \times \hat{y}_{GRM} \quad (13)$$

where $\hat{y}_{IKRidge-GRM}$ is the forecasted value obtained by using the $\hat{y}_{IKRidge}$ and $\hat{y}_{GRM}$. c is a positive number within the range of [0, 1] that calculated by the INFO algorithm. Figure 1 depicts the structure of proposed IKRidge-GRM method.

Feature selection method

The performance of ML models could deteriorate when an excessive number of parameters are included. Instead of relying on traditional input selection methods that primarily focus on linear relationships, this study adopted the LGBM (light gradient boosting machine) (Ke et al., 2017), a data filtering technique and a nonlinear method, to enhance model accuracy. LGBM employs a histogram-based approach for decision tree learning, which simplifies data by discretizing continuous features into bins. This process not only accelerates training but also minimizes memory usage while preserving high accuracy. LGBM is known for handling large datasets effectively and providing fast and accurate predictions. In this study, LGBM was used to simplify the forecasting process by focusing on data with higher gradients and using an automatic feature selection method. This helped reduce the number of input variables and identify the most important ones.

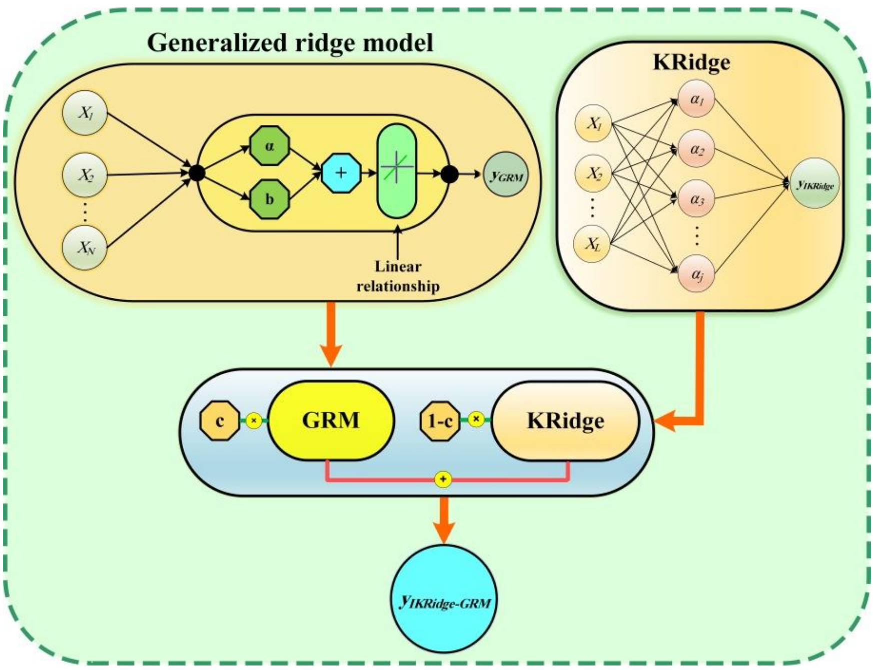


Figure 1. Schematic of IKRidg-GRM model.

Decomposition method

Decomposition methods play a vital role in simplifying complex datasets by breaking them into smaller, more manageable parts. Decomposition methods also make the data easier to interpret and analyze. These techniques also uncover both high and low-frequency components, which are essential for enhancing the accuracy and efficiency of ML models. One prominent approach for multivariate data decomposition is the multivariate variational mode decomposition (MVMD) (ur Rehman and Aftab, 2019). This method depends on two key parameters, namely: the total number of decompositions (ND) and the quadratic penalty term (ψ). The former denotes the number of intrinsic mode functions (IMFs) extracted from the data. It is notable that setting ND too high can lead to mode aliasing, where modes overlap. While a low value for ND results in incomplete decomposition and insufficient feature extraction. Meanwhile, ψ influences the bandwidth of the IMFs, directly affecting the quality of the decomposition process. In order to get trustworthy results, it is essential to choose the correct values for ND and ψ. In this research, a trial-and-error approach was used to identify the optimal values for these parameters, ensuring effective decomposition and improved performance of the model.

Metric performance

The present study uses seven error metrics to evaluate the ML methods. These metrics are root mean square error (RMSE), mean absolute percentage error (MAPE), correlation coefficient (R), Vicis symmetric distance (VSD), index of agreement (I_A), Kling-Gupta Efficiency (KGE), and median absolute error (MdAE), which are defined as,

$$R=\frac{\sum_{i=1}^{N}\left(PS_{M,i}-\overline{PS}_M\right)\times\left(PS_{F,i}-\overline{PS}_F\right)}{\sqrt{\sum_{i=1}^{N}(PS_{M,i}-\overline{PS}_M)^2\times\sum_{i=1}^{M}(PS_{F,i}-\overline{PS}_F)^2}} \tag{14}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{PS_{M,i}-PS_{F,i}}{PS_{M,i}}\right|\times 100 \tag{15}$$

$$MdAE=median_{i=1,\ldots,N}\left|PS_{M,i}-PS_{F,i}\right| \tag{16}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(PS_{M,i}-PS_{F,i})^2} \tag{17}$$

$$KGE=1-\sqrt{(R-1)^2+\left(SD(PS_{M,i})/SD(PS_{F,i})-1\right)^2+(\overline{PS}_M/\overline{PS}_F-1)^2} \tag{18}$$

$$IA = 1 - \frac{\sum_{i=1}^{N}\left(PS_{F,i} - PS_{M,i}\right)^2}{\sum_{i=1}^{N}\left(\left|\left(PS_{F,i} - \overline{PS_F}\right)\right| + \left|\left(PS_{M,i} - \overline{PS_M}\right)\right|\right)^2}, 0 < IA \leq 1 \quad (19)$$

$$VSD = \sum_{i=1}^{M} \frac{(PS_{M,i} - PS_{F,i})^2}{\min(PS_{M,i}, PS_{F,i})} \quad (20)$$

Where $PS_{M,i}$ and $PS_{F,i}$ are the PS amounts of measured and forecasted, respectively. SD is the standard deviation. $\overline{PS_M}$ and $\overline{PS_F}$ are the average amounts of PS for measured and forecasted values.

Case study

The Karun River is the longest and most important river in Iran, running 950 kilometers through the southwest of the Khuzestan Plain. Over the past ten years, its WQ has gotten worse because of factories, too much water being taken, farming, and inadequate sewage systems for both industrial and domestic uses. To take care of the river and its ecosystem, it is important to predict WQ accurately. Figure 2 illustrates the locations of stations used to monitor WQ throughout the Khuzestan Plain. The study utilized 48 years (1968-2015) of monthly WQ data collected from the Farisat station. A total of nine WQ parameters were analyzed as input variables, namely: magnesium (Mg), sulfate (SO_4^{-2}), calcium (Ca), discharge (Q), sodium (Na), bicarbonate (HCO3), chloride (Cl), electrical conductivity (EC), total dissolved solids (TDS) and pH. Potential salinity (PS) was selected as the target variable. A time-series graph of PS is shown in Figure 3, while Table 1 summarizes the statistical properties of the data, such as maximum, mean, minimum, and standard deviation. The PS was calculated using Eq. (21).

$$PS = Cl + \left(\frac{SO4}{2}\right) \quad (21)$$

Table 1. Statistical analysis of all WQ parameters at the Farisat station.

Parameter	Max	Min	Mean	SD
Na (mg/L)	25.04	1.90	9.29	4.21
Mg (mg/L)	9.61	0.03	3.25	1.35
Ca (mg/L)	17.40	1.75	5.44	2.38
Cl (mg/L)	26.80	2.16	9.06	4.08
HCO_3 (mg/L)	5.26	0.47	2.91	0.68
SO_4 (mg/L)	20.48	0.52	5.97	3.26
PH	9.00	6.01	7.94	0.35
Q (m^3/s)	3016.00	3.71	529.45	498.86
EC ($\mu S/cm$)	3980.00	623.00	1766.96	618.25
TDS (mg/L)	2473.00	21.38	1129.08	400.42
PS	31.40	2.93	12.05	5.11

Model development

To predict potential salinity (PS) at the Farisat station, advanced and carefully designed models were employed. The framework incorporated four cutting-edge ML models: IKRidge-GRM, Ridge, Deep ELM (DELM) (Fayaz, and Kim, 2018), and LSSVM. Additionally, the methodology utilized the LGBM feature selection technique alongside the MVMD decomposition method. The primary objective of the study was to predict PS values three months into the future (t + 3). The proposed framework to forecast the PS parameter is displayed in Figure 2. The model development process was structured into three key stages, which are described as follows,

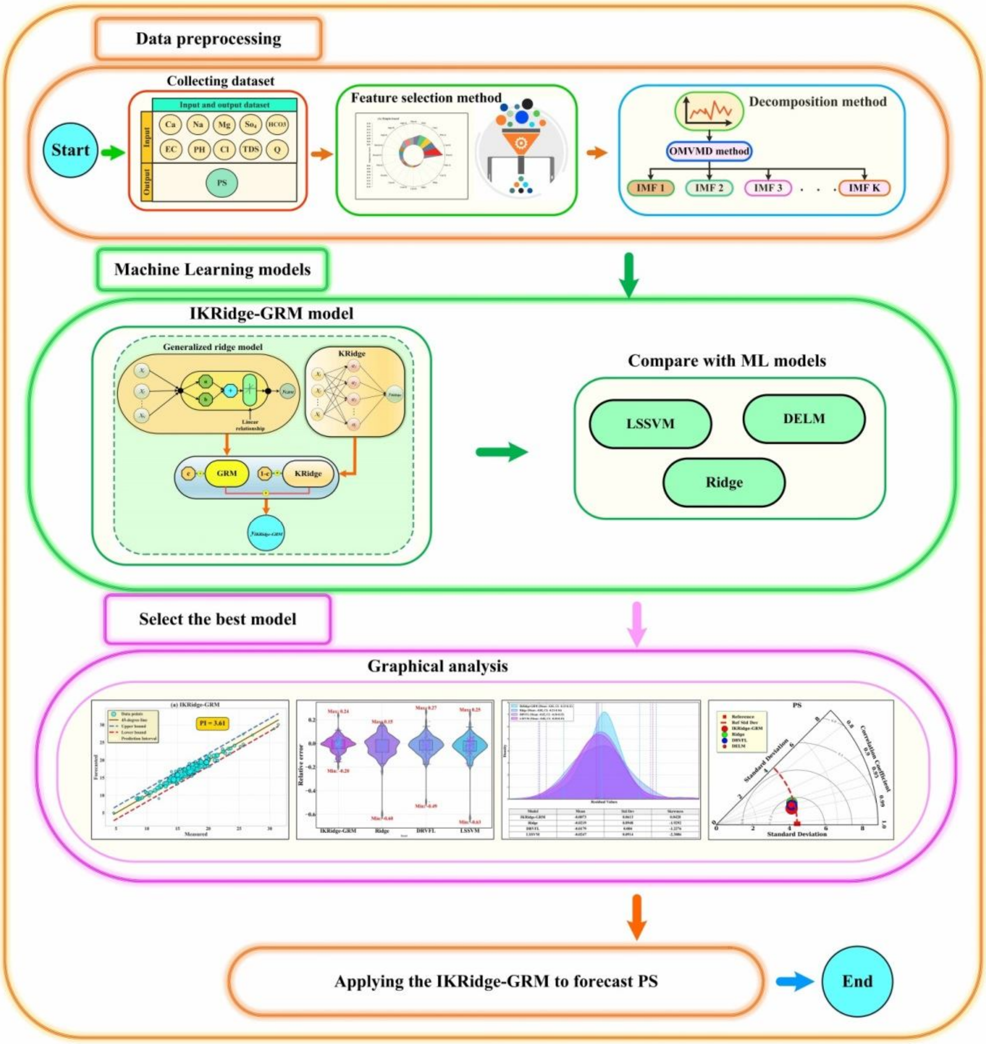


Figure 2. Proposed framework to forecast the PS.

Determination of input variables using feature selection

In the present research, the optimal input variables were identified using the LGBM model for feature selection. This method determines the most critical time delays, with each input variable incorporating a 10-time lag. The selected features for the Farisat station along with their importance scores are depicted in Figure 3 (A) and (B). For example, Table 2 lists the most significant features for forecasting PS(t+3). At the Farisat station, a total of 20 PS-related features were identified as the most relevant.

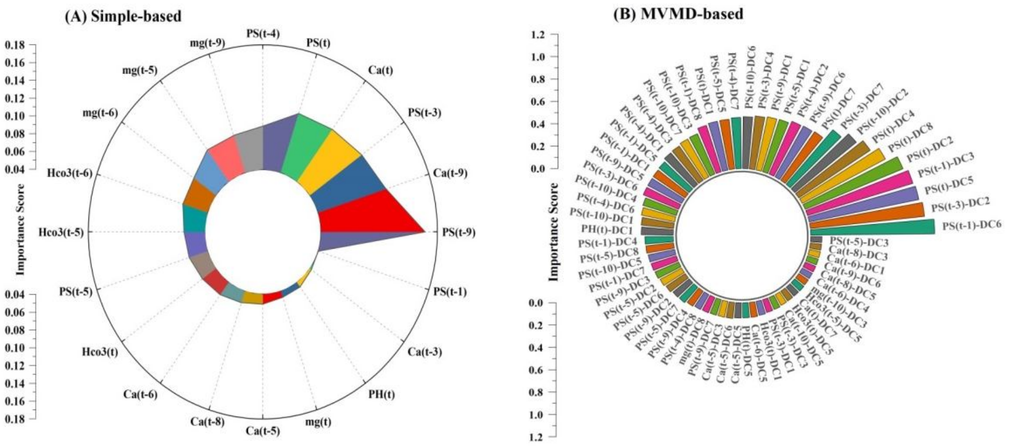


Figure 3. Selected features for (A) simple and (B) MVMD-based models.

Table 2. Selected features for simple-based models.

Target	Selected input features
PS (t+3)	PS(t-9), PS(t-3), Ca(t-9), PS(t), Ca(t), PS(t-4), Mg(t-5), Mg(t-6), Mg(t-9), Ca(t-8), HCO3(t-6), Hco3(t-5), PS(t-5), PH(t), Mg(t), PS(t-1), HCO3(t), Ca(t-3), Ca(t-5), Ca(t-6), Ca(t-3)

Decomposition of input variables

This study utilized the MVMD method to decompose input features. The MVMD method simplified the signals before feeding them into the ML models for hybrid implementation. The key adjustable parameters for the MVMD method were the mode decomposition factor (ND) and the penalty variable (ψ). These parameters were determined through a trial-and-error approach, with the optimal values identified as ND = 8 and ψ = 420. A total of 160 input variables were decomposed using MVMD (8 IMFs × 20). To further refine the data, the LGBM model was applied to select the most significant features, reducing the dimensionality by retaining only 35% of the total variables. This process resulted in 56 selected features, as illustrated in Figure 2 (B), which were used for PS forecasting.

Adjustment of ML models

Tuning the hyperparameters of ML algorithms is a critical aspect of model development. Relying on solutions derived from local optima can result in less accurate models and biased evaluations of forecasting methods. Therefore, employing advanced and robust optimization methods is essential for effectively addressing complex optimization problems (Abdollahi and Ahmadianfar, 2021; Ahmadianfar, Bozorg-Haddad et al., 2020; Ahmadianfar et al., 2021). This study utilizes the weIghted meaN oF VectOrs (INFO) optimizer (Ahmadianfar, Heidari, et al., 2022). The INFO is an advanced algorithm that enhances ML models by balancing exploration and exploitation. The INFO updates vector positions through three key processes, namely: an updating rule for generating new vectors, vector combination for refining solutions, and a local search for avoiding suboptimal results. These processes are designed to improve convergence, accuracy, and to find optimal solutions efficiently. Therefore, the INFO method was employed in this study to optimize the key hyperparameters of the IKRidge-GRM model. Additionally, other ML models, such as LSSVM, Ridge, and LSSVM, also utilized the parameter adjustments provided by the INFO approach. As a result, the optimal parameter values for both the simple ML models and the OMVMD-based ML models are presented in Table 3.

Table 3. Optimal parameter values determined for all ML models.

WQI	Methods	Values of parameters
Simple	IKRidge-GRM	$a_1 = 1.62E+09$, $a_2 = 4.13E+08$, $a_3 = 3.01E+08$, $\rho_1 = 2.44E+09$ $\rho_2 = 1.38E+04$, $\rho_3 = 1.42E+04$
	LSSVM	$\gamma = 2.13E+01, \sigma = 4.12E+03$
	DELM	NoNr = [300, 300], aFc = selu, RegF = 1.21E-03
	Ridge	Ridge cefficient = 183
MVMD	IKRidge-GRM	$a_1 = 4.92E+04$, $a_2 = 2.00E+06$, $a_3 = 8.55E+05$, $\rho_1 = 9.71E-01$ $\rho_2 = 1.98E+04$, $\rho_3 = 1.83E+04$
	LSSVM	$\gamma = 3.12E+03, \sigma = 2.11E+03$
	DELM	NoNr = [5000, 5000], aFc = selu, RegF = 4.32E-04
	Ridge	Ridge cefficient = 0.15

NoNr * = Number of neurons, aFc* = Activation Function, Neuron number, RegF * = regularization factor

Result and discussion

Assessment of ML models using statistical metrics

Table 4 compares the performance of various models based on several metrics, including R, RMSE, MAPE, IA, MdAE, and KGE. Among the models, the MVMD-IKRidge-GRM consistently demonstrated the best performance across both training and testing datasets. The MVMD-IKRidge-GRM achieved the highest R values with 0.982 for training and 0.977 for testing. The remarkable R values denote a robust connection between anticipated and observed values, indicating that the model proficiently captures the fundamental patterns in the data. Additionally, the MVMD-IKRidge-GRM had the lowest RMSE (0.737 for training and 0.956 for testing) and MAPE (6.580 for training and 4.521 for testing), reflecting its comparable accuracy and predictive reliability. The I_A values (0.991 for training and 0.988 for testing) and KGE scores (0.953 for training and 0.948 for testing) further confirmed its robustness and overall effectiveness. These metrics collectively highlight the MVMD-IKRidge-GRM as the most accurate and reliable model.

In contrast, the baseline models (e.g., the IKRidge-GRM, LSSVM, DRVFL, and Ridge) performed significantly worse, with much lower R values (e.g., 0.392 for the IKRidge-GRM and 0.312 for Ridge in testing) and higher RMSE and MAPE values. The models' inadequate ability to capture data dependencies is shown by these lower correlation coefficients, leading to inferior predicted accuracy. These baseline models' much higher RMSE and MAPE values suggest that their forecasts are unreliable. As an example, the RMSE and MAPE of the IKRidge-GRM were 4.310 and 20.923, respectively, which were much higher than those of the MVMD-IKRidge-GRM. These results highlighted the enormous potential for improvement in the conventional methods. The MVMD-enhanced versions of these models (e.g., the MVMD-LSSVM, MVMD-DRVFL, and MVMD-Ridge) showed notable improvements over their non-MVMD counterparts, with higher R values and lower errors. The MVMD-LSSVM achieved an R value of 0.967 and an RMSE of 1.165 in testing, which demonstrated a substantial enhancement. However, their performance was still inferior to that of the MVMD-IKRidge-GRM. Overall, the MVMD-IKRidge-GRM was the best-performing model, offering the most accurate and reliable predictions across all metrics.

Assessment of ML models using scatter plot

Figure 4 compares the performance of four models (the IKRidge-GRM, Ridge, DRVFL, and LSSVM) based on their prediction intervals (PI) and the alignment of forecasted versus measured values (scatter plot). The prediction interval (PI) quantifies the uncertainty in predictions, with lower PI values indicating higher confidence and precision. The PI value for the IKRidge-GRM was 3.61, suggesting the proposed model had the most precise predictions with minimal uncertainty. The data points for the IKRidge-GRM were closely aligned with the 45-degree line, indicating strong agreement between forecasted and measured values. Additionally, the upper and lower bounds of the prediction interval were narrower compared to the other models, further emphasizing its superior predictive accuracy and reliability.

In contrast, the other models (the Ridge, DRVFL, and LSSVM) exhibited higher PI values of 4.85, 4.32, and 4.33, respectively. The PI values for the mentioned models indicated greater uncertainty in their predictions. Ridge performed the worst, with the largest PI and a wider spread of data points around the 45-degree line, reflecting lower accuracy. The DRVFL and LSSVM performed slightly better than Ridge but still fell short of the IKRidge-GRM in terms of precision and alignment with the measured values. Consequently, the IKRidge-GRM was the best-performing model in this comparison, offering the most accurate and reliable predictions with the least uncertainty.

Table 4. Statistical results of simple- and MVMD-based methods.

Model		R	RMSE	MAPE	IA	MdAE	KGE
MVMD-IKRidge-GRM	train	0.982	0.737	6.580	0.991	0.500	0.953
	test	0.977	0.956	4.521	0.988	0.535	0.948
IKRidge-GRM	train	0.581	3.296	33.295	0.673	2.194	0.369
	test	0.392	4.310	20.923	0.590	2.335	0.286
MVMD-LSSVM	train	0.981	0.758	6.420	0.990	0.483	0.948
	test	0.967	1.165	6.157	0.982	0.759	0.946
LSSVM	train	0.592	3.143	28.604	0.696	1.869	0.401
	test	0.374	4.989	22.208	0.538	3.144	0.207
MVMD-DRVFL	train	0.983	0.721	6.125	0.991	0.501	0.953
	test	0.967	1.139	5.878	0.983	0.731	0.952
DRVFL	train	0.510	3.359	30.534	0.636	2.039	0.322
	test	0.358	4.505	21.404	0.529	2.279	0.207
MVMD-Ridge	train	0.983	0.736	6.393	0.991	0.522	0.946
	test	0.959	1.277	6.887	0.978	0.853	0.949
Ridge	train	0.569	3.576	33.227	0.333	2.278	0.058
	test	0.312	6.996	32.666	0.440	5.433	0.014

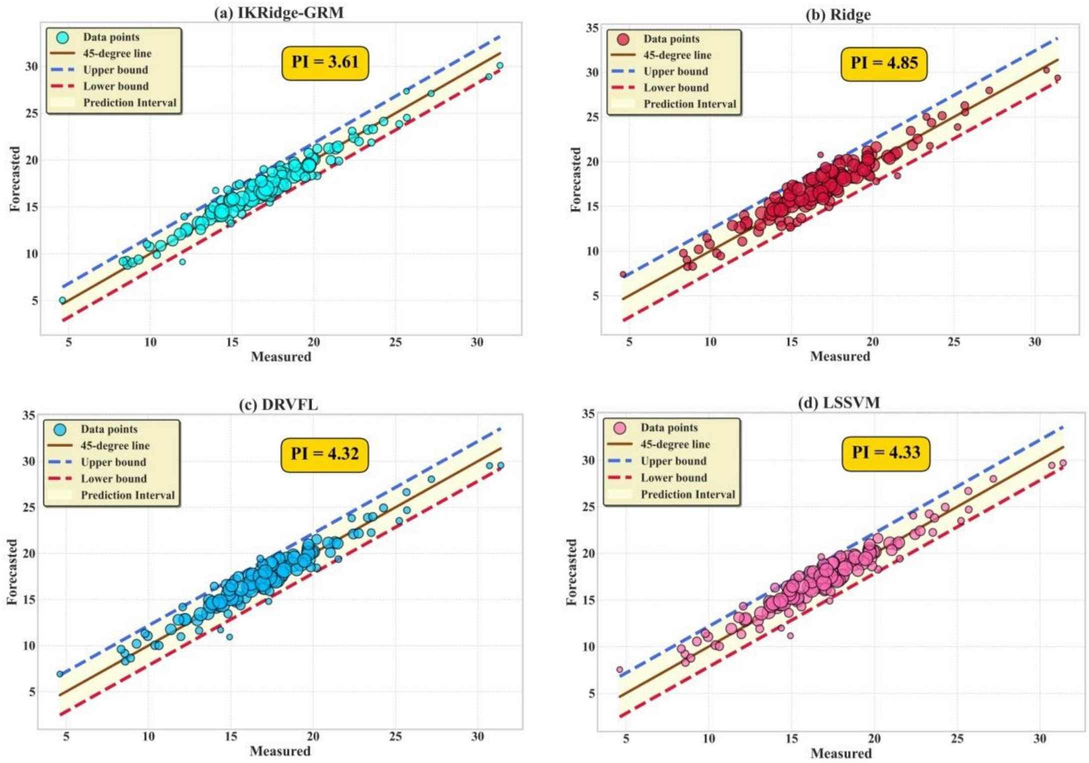


Figure 4. Scatter plot for all ML methods.

Assessment of ML models using relative error plot

Figure 5 presents violin plots comparing the relative error distributions of four models, namely: the IKRidge-GRM, Ridge, DRVFL, and LSSVM. Among the models, the IKRidge-GRM demonstrated the most compact and symmetric error distribution, with a minimum relative error of -0.20 and a maximum of 0.24. The figures indicated higher accuracy and consistency compared with other tested models. The Ridge model exhibited a wider spread, with a minimum error of -0.60 and a maximum of 0.15, reflecting greater variability and less reliability. Similarly, the DRVFL and LSSVM model showed larger error ranges, with the DRVFL model spanning from -0.49 to 0.27 and LSSVM from -0.63 to 0.25. The boxplots within the violins further highlighted that the IKRidge-GRM had the smallest interquartile range. Consequently, the IKRidge-GRM outperformed the other models by achieving the most precise and stable predictions.

Assessment of ML models using relative Taylor diagram

The Taylor diagram visually compares the performance of four models (the IKRidge-GRM, Ridge, DRVFL, and DELM) against a reference dataset based on three metrics, namely: standard deviation, correlation coefficient, and centered root mean square error (cRMSE). From Figure 6, the IKRidge-GRM was the best-performing method, as it was closest to the reference point (red square) in terms of both correlation coefficient and standard deviation. It achieved a high correlation coefficient (close to 1.0), indicating strong agreement with the reference data. The IKRidge-GRM's standard deviation closely also matched the reference value, reflecting accurate variability representation. In contrast, the Ridge, DRVFL, and DELM model were farther from the reference point, with slightly lower correlation coefficients and deviations from the reference standard deviation. Based on these results, the IKRidge-GRM demonstrated the best balance of accuracy, variability representation, and error minimization, making it the most reliable model in this comparison.

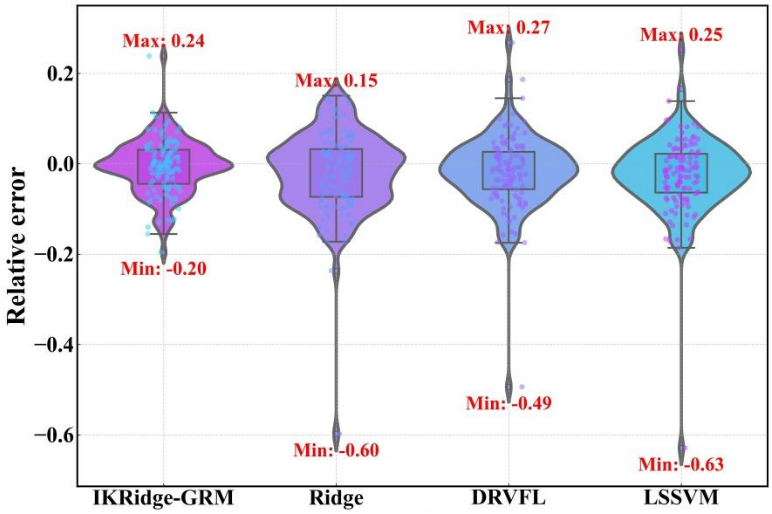


Figure 5. Violin plot of relative error for four ML methods.

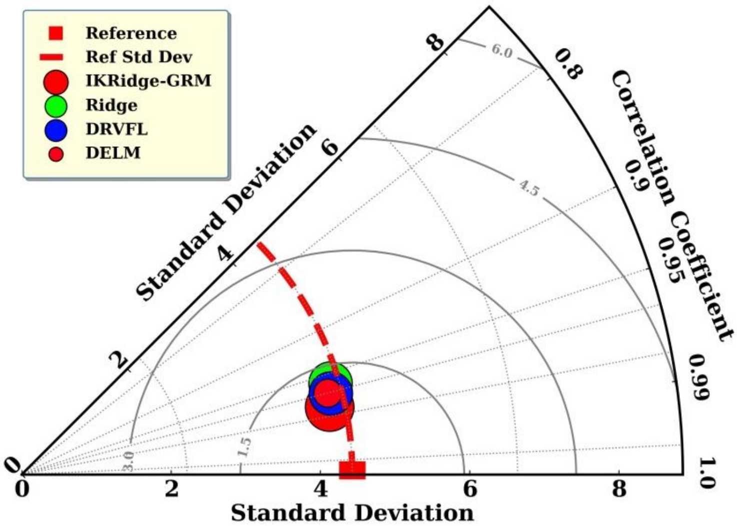
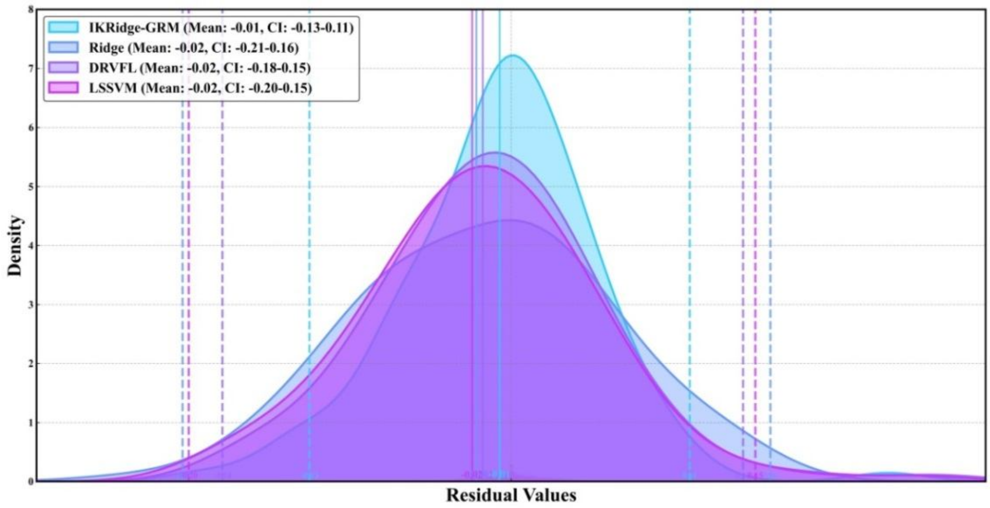


Figure 6. Taylor plot for four ML methods.

Assessment of ML models using residual distributions plot

Figure 7 compares the residual distributions of four models, namely: the IKRidge-GRM, Ridge, DRVFL, and LSSVM. The IKRidge-GRM model exhibited the smallest mean residual (-0.0073), indicating the least bias, and the lowest standard deviation (0.0613), suggesting the highest precision. Additionally, its skewness (0.0428) was close to zero, indicating a nearly symmetric residual distribution. In contrast, the other models (Ridge, DRVFL, and LSSVM) had higher standard deviations and more pronounced negative skewness, indicating less precision and asymmetry in their residuals. Based on these metrics, the IKRidge-GRM model was the best-performing model, as it demonstrates the most accurate and consistent residual distribution.



Model	Mean	Std Dev	Skewness
IKRidge-GRM	-0.0073	0.0613	0.0428
Ridge	-0.0219	0.0948	-1.9292
DRVFL	-0.0179	0.084	-1.2276
LSSVM	-0.0247	0.0914	-2.3086

Figure 7. Density distribution of residual values for four ML methods.

Conclusion

This study developed a novel ML model named IKRidge-GRM, which combines generalized ridge regression with kernel ridge regression, incorporating a regularized locally weighted approach. Indeed, the main novelty of this research is the development of a new ML model (IKRidge-GRM) for forecasting the PS parameter. The proposed method employs a set of regulated weights and the GRM model to improve prediction accuracy. The proposed framework model employed an LGBM-based optimization technique using the INFO algorithm to achieve optimal input variable selection. Furthermore, the MVMD method braked down input variables, enhancing prediction accuracy. Unlike existing methods, this hybrid model uniquely integrates feature selection, input variable decomposition, and an advanced ML framework. The primary objective was to improve computational efficiency and stability while delivering a more precise and dependable solution for the WQ prediction.

The IKRidge-GRM model was utilized to predict the PS parameter at the Farisat station in Iran, demonstrating superior performance compared to both standard models and those enhanced with MVMD. The integration of MVMD significantly boosted the model's effectiveness, achieving an impressive testing R value of 0.977 and an RMSE of 0.956. These results highlighted the model's capability to uncover complex data patterns and produce highly reliable predictions. The MVMD-IKRidge-GRM model has proven to be a powerful tool for precise environmental forecasting, offering a robust framework for addressing challenges in predicting environmental parameters. Its ability to integrate advanced decomposition techniques like MVMD with ML ensures improved accuracy and stability, making it a valuable approach for handling complex datasets. Furthermore, the model's consistent performance across various parameters underscores its adaptability and reliability, positioning it as a promising solution for environmental monitoring and decision-making processes. By combining innovative methodologies, the IKRidge-GRM model sets a new benchmark for predictive accuracy in environmental studies.

Future research could focus on integrating the IKRidge-GRM model with deep learning or hybrid approaches to better capture temporal and spatial dependencies in environmental data. Expanding its application to diverse environmental parameters, locations, and extreme conditions can validate its robustness.

Data availability

The data required for conducting this research were obtained from the Khuzestan Water and Power Organization.

Conflict of interest

There are no conflicts of interest in this article, and this has been confirmed by all the authors.

Credit authorship contribution statement

Masoud Dorfeshan: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Investigation.

Iman Ahmadianfar: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Software, Visualization, Investigation.

Arvin Samadi-Koucheksaraei: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Investigation.

Financial support

This research did not receive any financial support.

Acknowledgements

The authors would like to thank Behbahan Khatam Alanbia University of Technology for facilitating the completion of this research.

Ethics declarations

The authors have adhered to ethical principles in conducting and publishing this work, and this has been confirmed by all of them.

Reference

1. Abdollahi, A., & Ahmadianfar, I. (2021). Multi-mechanism ensemble interior search algorithm to derive optimal hedging rule curves in multi-reservoir systems. *Journal of Hydrology*. 598, 126211.
2. Ahmadianfar, I., Bozorg-Haddad, O., & Chu, X. (2020). Gradient-based optimizer: A new Metaheuristic optimization algorithm. *Information Sciences*. 540, 131-159.
3. Ahmadianfar, I., Heidari, A. A., Gandomi, A. H., Chu, X., & Chen, H. (2021). RUN beyond the metaphor: an efficient optimization algorithm based on Runge Kutta method. *Expert Systems with Applications*. 181, 115079.
4. Ahmadianfar, I., Heidari, A. A., Noshadian, S., Chen, H., & Gandomi, A. H. (2022). INFO: An Efficient Optimization Algorithm based on Weighted Mean of Vectors. *Expert Systems with Applications*. 116516.
5. Ahmadianfar, I., Jamei, M., & Chu, X. (2020). A novel hybrid wavelet-locally weighted linear regression (W-LWLR) model for electrical conductivity (EC) prediction in surface water. *Journal of Contaminant Hydrology*. 232, 103641.
6. Ahmadianfar, I., Shirvani-Hosseini, S., He, J., Samadi-Koucheksaraee, A., & Yaseen, Z. M. (2022). An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. *Scientific Reports*. 12 (1), 1-34.
7. Ahmadianfar, I., Shirvani-Hosseini, S., Samadi-Koucheksaraee, A., & Yaseen, Z. M. (2022). Surface water sodium (Na+) concentration prediction using hybrid weighted exponential regression model with gradient-based optimization. *Environmental Science and Pollution Research*. 1-26.
8. Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*. 578, 124084.

9.Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering.* 9 (1), 104599.

10.Barzegar, R., Adamowski, J., & Moghaddam, A. A. (2016). Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. *Stochastic Environmental Research and Risk Assessment.* 30 (7), 1797-1819.

11.Bozorg-Haddad, O., Soleimani, S., & Loáiciga, H. A. (2017). Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming. *Journal of Environmental Engineering*. 143 (7), 4017021.

12.Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment.* 721, 137612.

13.Chang, F. J., Tsai, Y. H., Chen, P. A., Coynel, A., & Vachaud, G. (2015). Modeling water quality in an urban river using hydrological factors–Data driven approaches. *Journal of Environmental Management.* 151, 87-96.

14.Chatterjee, S., Sarkar, S., Dey, N., Sen, S., Goto, T., & Debnath, N. C. (2017). Water quality prediction: Multi objective genetic algorithm coupled artificial neural network based approach. 2017 IEEE 15th International Conference on Industrial Informatics (INDIN). 963-968.

15.Chen, H., Ahmadianfar, I., Liang, G., & Heidari, A. A. (2024). Robust kernel extreme learning machines with weighted mean of vectors and variational mode decomposition for forecasting total dissolved solids. *Engineering Applications of Artificial Intelligence*. 133, 108587.

16.Deng, W., Wang, G., & Zhang, X. (2015). A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting. *Chemometrics and Intelligent Laboratory Systems*. 149, 39-49.

17.Fayaz, M., & Kim, D. (2018). A prediction methodology of energy consumption based on deep extreme learning machine and comparative analysis in residential buildings. *Electronics*. 7 (10), 222.

18.Gharemahmudli, S., & Seyed Hamidreza Sadeghi, V. S. S. (2024). Changeability of saline soil surface due to soil cyanobacteria inoculation using image processing. *Water and Soil Conservation.* 31 (2), 119-137.

19.Han, Y., Aziz, T. N., Del Giudice, D., Hall, N. S., & Obenour, D. R. (2021). Exploring nutrient and light limitation of algal production in a shallow turbid reservoir. *Environmental Pollution*. 269, 116210.

20.Huang, M., Tian, D., Liu, H., Zhang, C., Yi, X., Cai, J., Ruan, J., Zhang, T., Kong, S., & Ying, G. (2018). A hybrid fuzzy wavelet neural network model with self-adapted fuzzy-means clustering and genetic algorithm for water quality prediction in rivers. *Complexity*. 2018.

21.Jamei, M., Ahmadianfar, I., Karbasi, M., Jawad, A. H., Farooque, A. A., & Yaseen, Z. M. (2021). The assessment of emerging data-intelligence technologies for modeling Mg^{+2} and SO_4^{-2} surface water quality. *Journal of Environmental Management.* 300, 113774.

22.Jamei, M., Ali, M., Karbasi, M., Karimi, B., Jahannemaei, N., Farooque, A. A., & Yaseen, Z. M. (2024). Monthly sodium adsorption ratio forecasting in rivers using a dual interpretable glass-box complementary intelligent system: Hybridization of ensemble TVF-EMD-VMD, Boruta-SHAP, and eXplainable GPR. *Expert Systems with Applications*. 237, 121512.

23. Kandasamy, L., Mahendran, A., Sangaraju, S. H. V., Mathur, P., Faldu, S. V., & Mazzara, M. (2025). Enhanced remote sensing and deep learning aided water quality detection in the Ganges River, India supporting monitoring of aquatic environments. *Results in Engineering*. 25, 103604.
24. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 30.
25. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 135 (3), 370-384.
26. Qiu, R., Wang, Y., Wang, D., Qiu, W., Wu, J., & Tao, Y. (2020). Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River. *Science of The Total Environment*. 737, 139729.
27. Salarijazi, M., Ahmadianfar, I., & Yaseen, Z. M. (2024). Prediction enhancement for surface water sodium adsorption ratio using limited inputs: Implementation of hybridized stacked ensemble model with feature selection algorithm. Physics and Chemistry of the Earth, Parts a/b/C. 134, 103561.
28. Satish, N., Anmala, J., Rajitha, K., & Varma, M. R. R. (2024). A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. *Ecological Informatics*. 80, 102500.
29. ur Rehman, N., & Aftab, H. (2019). Multivariate variational mode decomposition. *IEEE Transactions on Signal Processing*. 67 (23), 6039-6052.
30. Vovk, V. (2013). Kernel ridge regression. In *Empirical inference: Festschrift in honor of vladimir n. vapnik* (pp. 105-116). Springer.
31. Wai, K. P., Koo, C. H., Huang, Y. F., & Chong, W. C. (2024). Decomposed intrinsic mode functions and deep learning algorithms for water quality index forecasting. *Neural Computing and Applications*. 1-20.
32. Wu, C., Zhang, X., Wang, W., Lu, C., Zhang, Y., Qin, W., Tick, G. R., Liu, B., & Shu, L. (2021). Groundwater level modeling framework by combining the wavelet transform with a long short-term memory data-driven model. *Science of The Total Environment*. 783, 146948.
33. Zahiri, J., Cheraghi, M., & Salarijazi, M. (2024). Simulating chlorophyll a in dam reservoirs using remote sensing and data-driven approaches. *Water and Soil Conservation*. 31 (3), 85-108.
34. Zhou, X., Leng, Y., Salarijazi, M., Ahmadianfar, I., & Farooque, A. A. (2024). Development of forecasting of monthly SAR time series in river systems: A multivariate data decomposition-based hybrid approach. *Process Safety and Environmental Protection*. 188, 1355-1375.