

Evaluating the impact of data correction on precipitation forecasting in semi-arid basins using machine learning models

Ali Dalir Gabrabad^{1*}, Mahdi Aiyoubi², Jafar Soleimani³

¹M.Sc. Student, Department of Water Resources Management, Faculty of Agriculture and Natural Resources, University of Tehran, Tehran, Iran

²M.Sc. Student, Department of Desert Management and Control, Faculty of Agriculture and Natural Resources, University of Tehran, Tehran, Iran

³M.Sc. Student, Department of Watershed Science and Engineering, Faculty of Agriculture and Natural Resources, University of Tehran, Tehran, Iran

Corresponding Author's E-mail: alidalir@ut.ac.ir

Received: 10 June 2026 Accepted: 30 Aug. 2026
Revised: 08 Aug. 2026 Published: 21 Dec. 2024



Homepage: ijwer.uoz.ac.ir

Abstract: Precipitation forecasting in arid and semi-arid regions is challenging due to temporal variability, spatial heterogeneity, nonlinearity, and data uncertainty. This study evaluates the effects of outlier detection, missing value imputation, and machine learning models on precipitation forecasting across three semi-arid Iranian basins: Gavkhuni, Siahkooh, and Ghareh Ghom. A 24-year (2000–2024) ERA5 precipitation time series was extracted using Google Earth Engine, with 2000–2022 for training and 2023–2024 for testing. Median Absolute Deviation (MAD) and Isolation Forest were combined with k-Nearest Neighbors (kNN), Singular Spectrum Analysis (SSA), Random Forest (RF), and XGBoost. Performance was assessed using Kling–Gupta Efficiency (KGE). MAD (mean KGE ≈ 0.525) outperformed Isolation Forest (≈ 0.467), kNN (≈ 0.503) slightly exceeded SSA ($\approx 0/49$), and RF (≈ 0.510) was more robust than XGBoost (≈ 0.482). MAD, SSA, and XGBoost reached KGE ≈ 0.74 in Gavkhuni, while preprocessing substantially improved performance in Siahkooh. No configuration was universally optimal, highlighting basin-specific preprocessing and model selection.

Keyword: Mashine Learning, ERA5, Preprocess Data, Outlie

Introduction: Precipitation plays a fundamental role in sustaining human settlements in arid and semi-arid regions. However, rainfall exhibits temporal variability, spatial heterogeneity, and uncertainty, leaving limited tolerance for forecasting errors. Reliable forecasting supports flood preparedness, drought mitigation, irrigation planning, and agricultural sustainability (Kanani et al., 2023; Kundu et al., 2023).

Precipitation forecasting remains challenging because of atmospheric complexity and the influence of topographic, hydrological, and climatic factors. Dryland rainfall processes are highly nonlinear, while hydrological systems are non-stationary under climate change and anthropogenic pressures. Limited access to long-term, high-quality observations, particularly in developing countries, further constrains model reliability. Traditional approaches, including linear regression, ARIMA, and the Mann–Kendall test, often struggle to capture nonlinear dynamics and long-term dependencies associated with climate oscillations such as ENSO and IOD (Jabari et al., 2026; Patel et al., 2023; Sesethu & Piotr, 2022).

Machine learning (ML) provides a data-driven alternative for modeling complex nonlinear relationships. Algorithms such as Random Forest (RF), XGBoost, Support Vector Regression, and Multi-Layer Perceptron have shown

promising predictive performance. RF provides robustness against noise and overfitting, whereas XGBoost is effective for high-dimensional nonlinear datasets (Ghosh et al., 2023; Kanani et al., 2023).

Previous studies have focused mainly on humid regions and limited gauge networks, while systematic comparisons of RF and XGBoost under different preprocessing scenarios in highly variable semi-arid environments remain scarce (Adam et al., 2025; Jumadi et al., 2025). Therefore, this study evaluates precipitation prediction using remote sensing-based data and advanced ML models in arid and semi-arid regions.

Methodology: Precipitation data were obtained from the ECMWF ERA5 reanalysis acquired using Google Earth Engine. Data were spatially aggregated at the basin scale and organized as continuous time series. Outliers were identified using Median Absolute Deviation (MAD) and Isolation Forest, while an unprocessed dataset (Non) was retained as a baseline. MAD provides robust deviation-based detection, whereas Isolation Forest identifies rare observations through recursive partitioning.

Removed values were reconstructed using k-Nearest Neighbors (kNN) and Singular Spectrum Analysis (SSA). kNN estimates missing values from neighboring observations, while SSA decomposes time series into trend, oscillatory components, and noise for reconstruction. Combining outlier detection and imputation methods generated multiple corrected datasets.

Both corrected and Non datasets were modeled using Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Training used 1999–2022 data, with 2023–2024 reserved for independent testing. Hyperparameters were optimized using Grid Search (Table 1). All procedures were implemented in Python using Visual Studio Code. Model performance was evaluated using Kling–Gupta Efficiency (KGE), which integrates correlation, bias, and variability (Eq. (1)).

The study covered three semi-arid Iranian basins: Gavkhuni and Siahkooh in central Iran and Ghareh Ghom in eastern Iran. These basins exhibit low precipitation, irregular rainfall distribution, and high temporal variability. Gavkhuni experiences substantial surface and groundwater depletion; Siahkooh has complex topography and heterogeneous rainfall; and Ghareh Ghom is characterized by severe water scarcity and pronounced interannual variability. Their climatic fragility and heterogeneous rainfall regimes provide suitable testbeds for evaluating preprocessing and ML-based precipitation modeling using long-term remotely sensed datasets.

Results and Discussion: This study evaluated the combined effects of outlier detection, missing value imputation, and machine learning models on precipitation simulation performance using the Kling–Gupta Efficiency (KGE) metric. Results demonstrate that model performance was highly sensitive to preprocessing quality, and inappropriate outlier removal or imputation could substantially reduce predictive accuracy.

Across all basins and scenarios, MAD outperformed Isolation Forest, achieving a mean KGE of 0.525 versus 0.467, indicating greater robustness for hydrological time series. For missing-value reconstruction, kNN slightly exceeded SSA, with mean KGE values of 0.503 and 0.490, respectively. Among ML models, Random Forest (RF) achieved a higher overall mean KGE (0.510) than XGBoost (0.482), suggesting greater stability under heterogeneous and uncertain conditions.

Basin-specific results showed substantial differences. Gavkhuni exhibited the highest overall performance. For RF, combinations of MAD or Isolation Forest with kNN or SSA produced KGE values of 0.63–0.68, while the Non scenario performed considerably worse. The best result in the study was obtained by XGBoost + MAD + SSA, reaching a KGE of approximately 0.741, demonstrating the benefits of combined outlier removal and imputation.

Ghareh Ghom showed lower performance and greater preprocessing sensitivity. The Non scenario yielded the lowest KGE values for both models, whereas MAD combined with kNN or SSA increased KGE to approximately 0.49–0.51. For XGBoost, MAD + SSA achieved the highest KGE (0.513), while Isolation Forest combinations were weaker.

Siahkooh showed the strongest dependence on preprocessing. Under the Non scenario, RF achieved only 0.06, whereas XGBoost produced a negative KGE (−0.22). MAD + SSA increased performance to approximately 0.43–0.44. For RF, MAD-based combinations yielded 0.416–0.434, while XGBoost achieved its best result with MAD + kNN (0.441).

The weakest result was obtained using Isolation Forest + SSA (0.116).

Overall, interaction effects confirmed that no single preprocessing–model combination was universally optimal. The results indicate that robust outlier detection, appropriate missing-value reconstruction, and model selection should be adapted to basin-specific characteristics rather than applying a uniform strategy across heterogeneous semi-arid environments.

Conclusions: This study demonstrates that data preprocessing, particularly outlier detection and missing value imputation, is critical for improving hydrological simulation accuracy and, in heterogeneous basins with lower data quality, may be essential for achieving acceptable performance. MAD provided more stable and reliable outlier detection than Isolation Forest. Both kNN and SSA improved model performance, although kNN showed a slight overall advantage, while the SSA–MAD combination produced particularly strong results. Random Forest demonstrated greater overall robustness, whereas XGBoost, despite achieving the highest KGE in specific combinations, was more sensitive to preprocessing choices. No single combination of outlier detection, imputation, and machine learning model was optimal across all basins. Therefore, preprocessing and model selection should consider basin-specific characteristics, including climate, flow regime, and data quality. Adapting these strategies to local conditions can improve the reliability and applicability of machine learning models for hydrological simulation and water resources management.

© 2024 University of Zabol, Zabol, Iran.



This is an open access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0 license) (<http://creativecommons.org/licenses/by/4.0/>)

ارزیابی اثر اصلاح داده بر پیش‌بینی بارش در حوضه‌های نیمه‌خشک با استفاده از مدل‌های یادگیری ماشین

علی دلیر گبرآباد^{۱*}، مهدی ایوبی^۲، محمدجعفر سلیمانی^۳

^۱ دانشجوی کارشناسی ارشد، گروه مدیریت منابع آب، دانشکده کشاورزی و منابع طبیعی، دانشگاه تهران، تهران، ایران
^۲ دانشجوی کارشناسی ارشد، گروه مدیریت و کنترل بیابان، دانشکده کشاورزی و منابع طبیعی، دانشگاه تهران، تهران، ایران
^۳ دانشجوی کارشناسی ارشد، گروه علوم و مهندسی آبخیز، دانشکده کشاورزی و منابع طبیعی، دانشگاه تهران، تهران، ایران

پست الکترونیکی نگارنده مسئول: alidalir@ut.ac.ir

وبگاه نشریه: ijwer.uoz.ac.ir



تاریخ دریافت: ۱۴۰۵/۰۳/۲۰ تاریخ پذیرش: ۱۴۰۵/۰۶/۰۸
تاریخ ویرایش: ۱۴۰۵/۰۵/۱۷ تاریخ چاپ: ۱۴۰۳/۱۰/۰۱

چکیده: پیش‌بینی بارش در مناطق خشک و نیمه‌خشک به دلیل نوسانات زمانی، ناهمگنی مکانی، غیرخطی بودن و عدم قطعیت داده‌ها چالش برانگیز است. این مطالعه اثر ترکیبی شناسایی داده‌های پرت، پرکردن داده‌های گمشده و مدل‌های یادگیری ماشین را در سه حوضه نیمه‌خشک ایران، گاوخونی، سیاه‌کوه و قره‌قوم، بررسی کرد. سری زمانی ۲۴ ساله بارش ERA5 (2000–2024) با Google Earth Engine استخراج شد؛ داده‌های ۲۰۲۰–۲۰۲۲ برای آموزش و ۲۰۲۳–۲۰۲۴ برای آزمون استفاده شدند. روش‌های MAD و Isolation Forest برای شناسایی پرت و kNN و SSA برای بازسازی داده‌ها، همراه با RF و XGBoost ارزیابی شدند. عملکرد با شاخص KGE سنجیده شد. MAD (0.525) عملکرد بهتری از Isolation Forest (0.467) و kNN اندکی بهتر از SSA (0.490) بود. RF نیز میانگین KGE بالاتری (۰/۵۱۰) نسبت به XGBoost (0.482) داشت. بیشترین KGE در گاوخونی با ترکیب MAD+SSA+XGBoost برابر ۰/۷۴۱ به دست آمد. نتایج نشان داد هیچ پیکربندی واحدی برای همه حوضه‌ها بهینه نیست و انتخاب روش پیش‌پردازش و مدل باید متناسب با ویژگی‌های هر حوضه انجام شود.

کلیدواژه‌ها: یادگیری ماشین، ERA5، پیش‌پردازش داده، داده پرت.

۱- مقدمه

با این حال، بارش در این پهنه‌های اقلیمی با تغییرپذیری شدید زمانی و مکانی، عدم قطعیت بالا و الگوهای نامنظم همراه است. در چنین شرایطی، به دلیل محدودیت منابع و شکنندگی اقلیمی، حاشیه خطای پیش‌بینی بسیار اندک است. پیش‌بینی

بارش در مناطق خشک و نیمه‌خشک نقشی حیاتی در توسعه و حفظ جوامع انسانی دارد، زیرا در این مناطق یکی از منابع اصلی تامین آب برای مصارف مختلف محسوب می‌شود.

نادرست بارش می‌تواند خسارات مالی، تخریب زیرساخت‌ها و حتی جراحات و مرگ‌ومیر به دنبال داشته باشد، درحالی‌که پیش‌بینی‌های قابل‌اعتماد امکان توسعه سیستم‌های هشدار سریع برای سیلاب، برنامه‌ریزی پیشگیرانه آبیاری در دوره‌های کم‌بارش و کاهش خسارات اقتصادی را فراهم می‌کنند (Kanani et al., 2023; Kundu et al., 2023; Sesethu & Piotr, 2022).

پیش‌بینی بارش به دلیل پیچیدگی سامانه‌های جوی و تاثیر هم‌زمان عوامل توپوگرافی، هیدرولوژیکی و اقلیمی، فرایندی دشوار است. با این حال، پیش‌بینی دقیق آن برای مدیریت منابع آب سطحی و زیرزمینی، برنامه‌ریزی آبیاری، تولید برق‌آبی و سازگاری با تغییر اقلیم اهمیت زیادی دارد. پیش‌بینی‌های عملیاتی و فصلی می‌توانند در برنامه‌ریزی بلندمدت مقابله با سیلاب و خشکسالی، تخصیص آب برای مصارف کشاورزی، خانگی و صنعتی، پیش‌بینی تولید انرژی برق‌آبی و پیش محیط‌زیست مورد استفاده قرار گیرند. همچنین، ارزش اقتصادی پیش‌بینی بارش از طریق کاهش پیامدهای خشکسالی و بهبود مدیریت منابع در سال‌های عادی افزایش می‌یابد (Jabari et al., 2026; Patel et al., 2023; Sesethu & Piotr, 2022).

مدل‌سازی بارش با چالش‌های متعددی همراه است. فرآیندهای بارشی، به‌ویژه در مناطق خشک، ماهیتی غیرخطی دارند و سامانه‌های هیدرولوژیکی نیز ذاتاً ناپایستا هستند؛ به‌گونه‌ای که ویژگی‌های آماری آن‌ها در طول زمان تحت تاثیر تغییر اقلیم و فعالیت‌های انسانی تغییر می‌کند. علاوه بر این، دسترسی محدود به داده‌های باکیفیت و بلندمدت، به‌ویژه در مناطق خشک و کشورهای در حال توسعه، قابلیت اعتماد مدل‌ها را محدود می‌کند. روش‌های آماری سنتی مانند رگرسیون خطی، ARIMA و آزمون من-کندال، اگرچه کاربرد گسترده‌ای دارند، در نمایش روابط غیرخطی و پردازش حجم بالای داده‌های هواشناسی با محدودیت مواجه‌اند (Adam et al., 2025; Omary et al., 2012; Tuysuzoglu et al., 2023).

یادگیری ماشین (ML) به‌عنوان رویکردی داده‌محور، امکان یادگیری روابط پیچیده و غیرخطی را بدون نیاز به فرض‌های توزیعی قوی فراهم می‌کند. برخلاف روش‌های سنتی، الگوریتم‌های یادگیری ماشین می‌توانند الگوهای غیرخطی و وابستگی‌های بلندمدت را که در حوضه‌های تحت تاثیر محرک‌های اقلیمی بزرگ‌مقیاس مانند نوسان ال‌نینو-جنوبی (ENSO) و دوگانگی اقیانوس هند (IOD) اهمیت دارند، شناسایی کنند. الگوریتم‌هایی مانند جنگل تصادفی (RF)، تقویت گرادیان حدی (XGBoost)، رگرسیون بردار پشتیبان و پرسپترون چندلایه در پیش‌بینی بارش عملکرد امیدوارکننده‌ای نشان داده‌اند (Hajirad et al., 2026; Dalir & Gabrabad et al., 2026). در میان این روش‌ها، مدل‌های یادگیری جمعی، به‌ویژه RF، به دلیل مقاومت نسبی در برابر نویز و توانایی در مدل‌سازی برهم‌کنش‌های پیچیده، مورد توجه قرار گرفته‌اند (Ghosh et al., 2023; Kanani et al., 2023; Amrehn et al., 2018).

با وجود گسترش کاربرد ML، بخش قابل‌توجهی از مطالعات پیش‌بینی بارش بر مناطق مرطوب و تعداد محدودی ایستگاه باران‌سنجی متمرکز بوده است. در مقابل، مناطق خشک و نیمه‌خشک با تغییرپذیری شدید، روابط غیرخطی و نویز بالای داده‌ها کمتر مورد توجه قرار گرفته‌اند. اگرچه RF و XGBoost از الگوریتم‌های قدرتمند یادگیری ماشین محسوب می‌شوند، مقایسه نظام‌مند عملکرد آن‌ها تحت سناریوهای مختلف پیش‌پردازش در محیط‌های خشک و بسیار متغیر همچنان محدود است (Adam et al., 2025; Beck et al., 2019; Jumadi et al., 2025).

یکی از عوامل مهم در عملکرد مدل‌های یادگیری ماشین که گاهی کمتر مورد توجه قرار می‌گیرد، کیفیت داده‌های ورودی است. سری‌های زمانی هیدرولوژیکی در مناطق خشک ممکن است شامل داده‌های پرت، مقادیر گمشده و ناهماهنگی‌های اندازه‌گیری باشند. محصولات بازتحلیل مانند ERA5، با وجود پوشش مکانی گسترده، می‌توانند تحت تاثیر توپوگرافی پیچیده، فرایندهای جوی و عدم قطعیت ذاتی داده‌های

پیش‌پردازش و مدل‌سازی در مناطق کم‌داده و اقلیم‌های خشک و نیمه‌خشک ارائه کرده و به بهبود قابلیت اعتماد پیش‌بینی بارش و مدیریت پایدار منابع آب کمک کند.

۲- روش‌شناسی

۲-۱- گردآوری داده‌ها

در این پژوهش، داده‌های مورد استفاده از محصولات بازتحلیلی سنجنش از دور ERA5 استخراج شدند. داده‌ها شامل سری‌های زمانی پیوسته با طول ۲۴ سال در بازه زمانی ۲۰۰۰ میلادی تا ۲۰۲۴ میلادی است. فرآیند استخراج داده‌ها در محیط Google Earth Engine (GEE) انجام شد که امکان دسترسی، پردازش و استخراج کارآمد داده‌های بلندمدت اقلیمی را فراهم می‌کند. داده‌های استخراج‌شده برای هر حوضه آبریز به‌صورت سری زمانی تجمیع و برای مراحل بعدی تحلیل آماده شدند (Hersbach et al., 2020). داده‌های بارش مورد استفاده در این پژوهش از مجموعه‌داده بازتحلیلی ERA5-Land در محیط Google Earth Engine استخراج شدند. داده‌های اولیه دارای تفکیک زمانی ساعتی و تفکیک مکانی $0.1^\circ \times 0.1^\circ$ بودند. متغیر total precipitation برای محدوده هر حوضه استخراج شد. داده‌های ساعتی ابتدا به مقادیر روزانه تجمیع و سپس با جمع مقادیر روزانه به بارش ماهانه تبدیل شدند. واحد اولیه بارش متر (m) بود و با استفاده از ضریب تبدیل ۱۰۰۰ به میلی‌متر (mm) تبدیل شد. برای هر حوضه، میانگین مکانی مقادیر تمام پیکسل‌های واقع در محدوده حوضه محاسبه شد و مقدار حاصل به‌عنوان بارش ماهانه نماینده حوضه مورد استفاده قرار گرفت.

۲-۲- پیش‌پردازش داده‌ها و حذف داده‌های پرت

پس از استخراج داده‌ها، مرحله پیش‌پردازش با هدف بهبود کیفیت داده‌ها انجام شد. در این مرحله، داده‌های پرت موجود در سری‌های زمانی با استفاده از دو روش مختلف شناسایی و حذف شدند Median Absolute Deviation (MAD) و Isolation Forest (IF)

بازتحلیل قرار گیرند. داده‌های نویزی یا ناقص می‌توانند عملکرد حتی الگوریتم‌های پیچیده یادگیری ماشین را کاهش دهند. بنابراین، پیش‌پردازش داده نباید صرفاً مرحله‌ای مقدماتی تلقی شود، بلکه باید به‌عنوان بخشی اساسی از فرایند مدل‌سازی مورد ارزیابی قرار گیرد.

در راستای رفع این شکاف، پژوهش حاضر اثرات ترکیبی روش‌های شناسایی داده‌های پرت، تکنیک‌های بازسازی مقادیر گمشده و مدل‌های یادگیری ماشین را بر پیش‌بینی بارش در سه حوضه نیمه‌خشک ایران، شامل گاوخونی، سیاه‌کوه و قره‌قروم، بررسی می‌کند. در این پژوهش، داده‌های بارش ERA5 به‌کار گرفته شده و دو روش شناسایی داده‌های پرت شامل انحراف مطلق از میانه (MAD) و Isolation Forest، دو روش بازسازی داده‌های گمشده شامل kNN و تحلیل طیف منفرد (SSA)، و دو مدل یادگیری ماشین RF و XGBoost به‌صورت ترکیبی ارزیابی می‌شوند. برخلاف مطالعاتی که پیش‌پردازش و مدل‌سازی را به‌صورت مراحل جداگانه بررسی کرده‌اند، این پژوهش یک چارچوب یکپارچه برای تحلیل تعامل میان کیفیت داده، روش پیش‌پردازش و عملکرد مدل ارائه می‌دهد.

نوآوری اصلی پژوهش، ارزیابی نظام‌مند برهم‌کنش روش‌های پیش‌پردازش و مدل‌های یادگیری ماشین در شرایط متنوع هیدرو-اقلیمی است. استفاده از ERA5 امکان پوشش مکانی گسترده‌تر و دسترسی به سری زمانی پیوسته را فراهم می‌کند، درحالی‌که ترکیب MAD و Isolation Forest برای شناسایی داده‌های پرت و kNN و SSA برای بازسازی داده‌های گمشده، یک خط لوله جامع پیش‌پردازش ایجاد می‌کند. همچنین، با مقایسه مستقیم پیکربندی‌های مختلف پیش‌پردازش و مدل، پژوهش مشخص می‌کند که آیا انتخاب راهبرد پیش‌پردازش می‌تواند به اندازه یا حتی بیشتر از انتخاب مدل بر دقت پیش‌بینی اثرگذار باشد. ارزیابی این چارچوب در سه حوضه با ویژگی‌های هیدرو-اقلیمی متفاوت، امکان بررسی وابستگی عملکرد به شرایط محلی را فراهم می‌سازد. در نهایت، نتایج می‌تواند راهنمایی عملی برای انتخاب راهبردهای مناسب

مختلف برآورد شدند. روش‌های مورد استفاده شامل k-Singular Spectrum و Nearest Neighbors (kNN) Analysis (SSA) هستند.

روش kNN بر اساس شباهت بین مشاهدات مجاور، مقادیر گم‌شده را برآورد می‌کند، در حالی که روش SSA با تجزیه سری زمانی به مولفه‌های ساختاری، قادر به بازسازی مقادیر حذف‌شده است. بدین ترتیب، چندین سری زمانی اصلاح‌شده با ترکیب روش‌های حذف داده پرت و روش‌های برآورد تولید شد. همچنین، سری زمانی اولیه بدون حذف داده پرت و بدون برآورد مقادیر گم‌شده به‌عنوان مرجع (Non) در تحلیل‌ها لحاظ گردید (Cover & Hart, 1967; Golyandina et al., 2001). تنظیمات روش‌های kNN و SSA در [جدول ۱](#) نمایش داده شده‌اند.

به‌منظور افزایش شفافیت طراحی آزمایش‌ها، تمامی سناریوهای مورد بررسی در این پژوهش در [جدول ۳](#) نمایش داده شده است. [این جدول](#)، ترکیب هر روش تشخیص داده‌های پرت با روش‌های مختلف برآورد داده‌های مفقود را نشان می‌دهد و مشخص می‌کند که هر سناریو چگونه ایجاد شده است. همچنین سناریوی بدون پیش‌پردازش نیز به‌عنوان حالت مرجع در [جدول](#) معرفی شده است.

۲-۴- مدل‌سازی با یادگیری ماشین

در مرحله مدل‌سازی، سری‌های زمانی اصلاح‌شده و سری زمانی بدون اصلاح به‌صورت جداگانه به‌عنوان ورودی مدل‌های یادگیری ماشین مورد استفاده قرار گرفتند.

جدول ۱ پارامترهای روش‌های حذف داده پرت

Table 1 Parameters of the outlier detection methods

Method	Parameter	Value
MAD	Threshold	3
	constant	1.48
Isolation Forest	n_estimators	100
	max_samples	auto
	contamination	auto
	max_features	1
	bootstrap	False

روش MAD به‌عنوان یک روش آماری مقاوم، بر پایه انحراف مطلق از میانه عمل کرده و برای داده‌هایی با توزیع نامتقارن مناسب است. در مقابل Isolation Forest یک روش مبتنی بر یادگیری ماشین است که با ایزوله‌سازی مشاهدات نادر، داده‌های پرت را شناسایی می‌کند. به‌منظور مقایسه اثر حذف داده‌های پرت، سری‌های زمانی بدون هیچ‌گونه اصلاح (Non) نیز حفظ شدند (Leys et al., 2013; Liu et al., 2008).

به‌منظور جلوگیری از ایجاد نشت اطلاعات، کلیه مراحل مدل‌سازی و ارزیابی به‌گونه‌ای طراحی شدند که داده‌های آموزش و آزمون در فرآیند ارزیابی عملکرد مدل‌ها از یکدیگر مستقل باقی بمانند. از آنجا که هدف این پژوهش مقایسه اثر سناریوهای مختلف کیفیت داده بر عملکرد مدل‌های یادگیری ماشین بوده است، تمامی نتایج در قالب یک مطالعه مقایسه‌ای بر روی سناریوهای اصلاح داده تفسیر شده‌اند و این موضوع در نسخه اصلاح‌شده مقاله به‌طور شفاف بیان شده است.

پارامترهای اجرایی هر دو روش نیز در این بخش ارائه شده‌اند. در روش MAD، مقدار آستانه تشخیص داده‌های پرت، نحوه اعمال روش بر سری‌های زمانی و شیوه پردازش داده‌ها قبل از اجرای الگوریتم به‌صورت کامل تشریح شده است. همچنین مشخص شده است که تشخیص داده‌های پرت برای هر حوضه به‌صورت مستقل انجام شده و نحوه لحاظ کردن تغییرات فصلی در فرآیند تشخیص نیز توضیح داده شده است.

برای الگوریتم Isolation Forest نیز تمامی پارامترهای اجرایی شامل مقدار contamination، تعداد درخت‌ها (n_estimators)، مقدار random_state، نوع داده‌های ورودی (تک‌متغیره یا چندمتغیره) و ویژگی‌های مورد استفاده در فرآیند تشخیص داده‌های پرت در [جدول ۲](#) نمایش داده شده است تا امکان بازتولید کامل آزمایش‌ها فراهم شود.

۲-۳- برآورد مقادیر گم‌شده

در مرحله بعد، داده‌هایی که در اثر حذف داده‌های پرت از سری‌های زمانی حذف شده بودند، با استفاده از دو روش

ارزیابی مستقیم عملکرد مدل‌ها انجام شد. از آنجاکه استفاده از متغیرهای متعدد هواشناسی و اقلیمی می‌تواند موجب افزایش پیچیدگی مدل و تاثیرگذاری عوامل جانبی بر نتایج شود، در این مطالعه تلاش شد با محدود کردن ورودی‌ها به متغیرهای اصلی و پایه، اثر واقعی فرآیند اصلاح داده‌ها بر عملکرد مدل‌ها به‌صورت شفاف‌تر و قابل‌تفسیرتر مورد بررسی قرار گیرد. همچنین استفاده از شماره ماه به‌عنوان یک متغیر زمانی، امکان در نظر گرفتن الگوی فصلی بارش را برای مدل فراهم کرده است.

لازم به ذکر است که هدف اصلی این پژوهش، توسعه یک مدل عملیاتی برای پیش‌بینی بارش آینده نبوده، بلکه بررسی نظام‌مند اثر روش‌های مختلف پیش‌پردازش داده شامل حذف داده‌های پرت و برآورد مقادیر مفقودی بر عملکرد مدل‌های یادگیری ماشین بوده است. به همین منظور، تمامی سناریوهای پیش‌پردازش تحت یک چارچوب مدل‌سازی یکسان مورد ارزیابی قرار گرفتند تا اثر کیفیت داده‌ها به‌صورت مستقل بررسی شود. در این پژوهش، تمرکز اصلی بر مقایسه عملکرد سناریوهای مختلف اصلاح داده بوده و نه ارائه یک مدل پیش‌بینی عملیاتی برای ماه‌های آینده. بنابراین، ساختار مدل به‌گونه‌ای طراحی شد که تنها امکان ارزیابی تاثیر فرآیندهای پیش‌پردازش بر عملکرد مدل‌های یادگیری ماشین فراهم شود و نتایج حاصل صرفاً در همین چارچوب تفسیر شده‌اند.

۲-۵- پیاده‌سازی و ارزیابی

کلیه مراحل پردازش، تحلیل و مدل‌سازی در محیط Python انجام شده و پیاده‌سازی الگوریتم‌ها در محیط توسعه Visual Studio Code (VS Code) صورت گرفته است. ارزیابی عملکرد مدل‌ها با استفاده از شاخص Kling-Gupta Efficiency (KGE) انجام شد که به‌طور هم‌زمان دقت، همبستگی و بایاس شبیه‌سازی را در نظر می‌گیرد. برای ارزیابی کامل‌تر شاخص‌های RMSE و NSE نیز محاسبه شد. این چارچوب امکان مقایسه جامع اثر روش‌های مختلف پیش‌پردازش داده و

جدول ۲ پارامترهای روش برآورد داده گمشده

Table 2 Parameters of the missing data imputation methods

Method	Parameter	Value
MAD	k (n_neighbors)	5
	weights	uniform
	metric	nan_euclidean
Isolation Forest	Length (Window Length)	24
	Thresholdpct	90%
	Method	Tereshold
	Adjustment	True

جدول ۳ سناریوهای استفاده شده در پژوهش

Table 3 Scenarios used in the study

Scenario	Outlier removal	Missing data estimation
S0	Without preprocessing	Without preprocessing
S1	MAD	kNN
S2	MAD	SSA
S3	Isolation Forest	kNN
S4	Isolation Forest	SSA

دو مدل Random Forest (RF) و Extreme Gradient Boosting (XGBoost) برای شبیه‌سازی و پیش‌بینی متغیر مورد نظر به کار گرفته شدند (Amrehn et al., 2018; Kanani et al., 2023).

دو سال پایانی (۲۰۲۳ و ۲۰۲۴) به‌عنوان مجموعه آزمون مستقل انتخاب شدند تا شرایط واقعی پیش‌بینی شبیه‌سازی شود؛ به‌گونه‌ای که مدل تنها با استفاده از داده‌های گذشته آموزش دیده و بر روی داده‌های آینده که در فرآیند آموزش و تنظیم پارامترها هیچ نقشی نداشته‌اند، ارزیابی شود. همچنین، تمامی مراحل تنظیم پارامترهای مدل تنها بر روی داده‌های آموزشی انجام شد تا از نشت اطلاعات بین داده‌های آموزش و آزمون جلوگیری شود. اگرچه مجموعه آزمون شامل تعداد محدودی از مشاهدات ماهانه است، این نوع تقسیم‌بندی زمانی موجب حفظ ترتیب زمانی داده‌ها و ارائه ارزیابی واقع‌بینانه‌تری از توانایی مدل در پیش‌بینی داده‌های آینده می‌شود.

در این پژوهش، ورودی‌های مدل‌های یادگیری ماشین شامل مجموع بارش ماهانه (میلی‌متر) و شماره ماه بودند. انتخاب این متغیرها با هدف بررسی دقیق‌تر اثر روش‌های اصلاح داده و

آموزش و آزمون جلوگیری می‌کند. معیار انتخاب بهترین پارامتر KGE بوده است.

در این پژوهش، دو مدل Random Forest (RF) و XGBoost مورد استفاده قرار گرفتند و پارامترهای اصلی آن‌ها مطابق [جدول ۴](#) تنظیم شدند.

به‌منظور جلوگیری از بروز بیش‌پدازش (Overfitting)، داده‌ها به دو بخش مستقل آموزش و آزمون تقسیم شدند؛ به‌طوری‌که داده‌های آزمون در هیچ‌یک از مراحل آموزش و تنظیم پارامترهای مدل مورد استفاده قرار نگرفتند. همچنین فرآیند بهینه‌سازی پارامترها با استفاده از روش Grid Search تنها بر روی داده‌های آموزشی انجام شد. نزدیک بودن عملکرد مدل در مجموعه‌های آموزش و آزمون نشان‌دهنده توانایی مناسب مدل در تعمیم‌پذیری و عدم وابستگی بیش از حد به داده‌های آموزشی است.

جدول ۴ پارامترهای تنظیم‌شده مدل‌های یادگیری ماشین با استفاده از روش Grid Search

Table 4 Tuned Parameters of Machine Learning Models Using the Grid Search Method

Model	Parameter	Tested Values	Best Value
RF	n_estimators	100, 200, 300	200
	max_depth	5, 10, 15, 20	10
	min_samples_split	2, 5, 10	2
	min_samples_leaf	1, 2, 4	1
	max_features	sqrt, log2	sqrt
XGBoost	n_estimators	100, 200, 300	300
	learning_rate	0.01, 0.05, 0.1	0.05
	max_depth	3, 5, 7	5
	subsample	0.7, 0.8, 1.0	0.8
	colsample_bytree	0.7, 0.8, 1.0	1.0
Linear Regression	gamma	0, 0.1, 0.2	0.1
	fit_intercept	True, False	True
	positive	True, False	False
Linear Regression	n_jobs	-1, 1	-1

مدل‌های یادگیری ماشین را فراهم می‌کند (Gupta et al., 2009). شاخص‌های NSE، RMSE و KGE از فرمول‌های (۱)، (۲) و (۳) محاسبه می‌شود:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2} \quad (2)$$

$$NSE = 1 - \frac{(O_i - P_i)^2}{(O_i - \bar{O})^2} \quad (3)$$

O_i = مقدار مشاهده‌شده (مشاهدات واقعی بارش) در زمان i
 P_i = مقدار پیش‌بینی‌شده توسط مدل در زمان i و $\bar{O}$ = میانگین مقادیر مشاهده‌شده، n = تعداد کل داده‌ها است.

r ضریب همبستگی پیرسون بین مقادیر مشاهده‌شده و پیش‌بینی‌شده را نشان می‌دهد و میزان تطابق الگوی زمانی داده‌ها را ارزیابی می‌کند. پارامتر α نسبت انحراف معیار مقادیر پیش‌بینی‌شده به مقادیر مشاهده‌شده بوده و توانایی مدل در بازتولید میزان تغییرپذیری داده‌ها را بیان می‌کند. همچنین β نسبت میانگین مقادیر پیش‌بینی‌شده به میانگین مقادیر مشاهده‌شده است و بیانگر وجود یا عدم وجود بایاس سیستماتیک در پیش‌بینی‌ها است. مقدار بهینه KGE برابر با ۱ بوده و هرچه مقدار این شاخص به ۱ نزدیک‌تر باشد، عملکرد مدل مطلوب‌تر ارزیابی می‌شود.

به‌منظور بهینه‌سازی عملکرد مدل‌های یادگیری ماشین، تنظیم پارامترها با استفاده از روش جستجوی شبکه‌ای (Grid Search) انجام شد. در این روش، مجموعه‌ای از مقادیر مختلف برای هر پارامتر تعریف شده و تمامی ترکیبات ممکن مورد ارزیابی قرار گرفتند تا بهترین ترکیب پارامترها بر اساس معیار عملکرد مدل انتخاب شود. پس از تعیین بهترین مجموعه پارامترها، مدل نهایی مجدداً بر روی کل داده‌های آموزشی آموزش داده شد و تنها یک‌بار بر روی مجموعه آزمون مستقل ارزیابی گردید. این رویکرد از نشت اطلاعات بین داده‌های

سطحی و زیرسطحی و تأثیرات قابل توجه تغییرات اقلیمی، یک نمونه بارز برای بررسی الگوهای بارش در مناطق نیمه‌خشک محسوب می‌شود (Ravar et al., 2020).

حوضه سیاه‌کوه نیز در مرکز ایران و در نزدیکی حوضه گاوخونی قرار دارد و ویژگی‌های مشابهی از نظر اقلیمی دارد. تفاوت عمده این حوضه در توپوگرافی و پراکندگی بارش است که چالش‌های خاصی برای شبیه‌سازی و پیش‌بینی دقیق بارش ایجاد می‌کند (Ahmadi et al., 1999).

حوضه قره‌قوم در شرق ایران واقع شده و به دلیل قرارگیری در منطقه خشک و نیمه‌خشک شرقی، کمبود شدید منابع آب و نوسانات زمانی بارش، اهمیت ویژه‌ای برای مطالعات مدل‌سازی هیدرولوژیکی دارد. این حوضه نماینده شرایط خشک شرقی کشور است و تحلیل داده‌های آن می‌تواند به درک بهتر اثرات تغییرات اقلیمی بر بارش در مناطق کم‌آب کمک کند (Pourhashemi & Aliabadi, 2025).

تمامی حوضه‌های مطالعه شده دارای بارش کم و پراکندگی نامنظم بارش در طول سال هستند و به همین دلیل برای تحلیل و پیش‌بینی دقیق بارش، استفاده از داده‌های بلندمدت سنجش از دور و مدل‌های یادگیری ماشین ضرورت دارد. طول سری زمانی داده‌های مورد استفاده در این پژوهش ۳۴ سال (۱۹۹۰ تا ۲۰۲۴) بوده و دو سال پایانی برای آزمون مدل‌ها در نظر گرفته شده است. این حوضه‌ها نمونه‌های ایده‌آل برای ارزیابی روش‌های حذف داده‌های پرت، برآورد مقادیر گمشده و کاربرد مدل‌های یادگیری ماشین در مناطق نیمه‌خشک محسوب می‌شوند.

۳- نتایج و بحث

در این پژوهش، تأثیر روش‌های مختلف حذف داده‌های پرت، برآورد مقادیر گمشده و عملکرد مدل‌های یادگیری ماشین در بازتولید یا پیش‌بینی سری زمانی بارش در چند حوضه با ویژگی‌های متفاوت مورد بررسی قرار گرفت. نتایج حاصل از تحلیل شاخص KGE نشان داد که عملکرد مدل‌ها به‌طور

روی داده‌های آموزشی انجام شد. نزدیک بودن عملکرد مدل در مجموعه‌های آموزش و آزمون نشان‌دهنده توانایی مناسب مدل در تعمیم‌پذیری و عدم وابستگی بیش از حد به داده‌های آموزشی است. بنابراین، نتایج بیانگر آن است که مدل‌های توسعه‌یافته از پایداری مطلوب برخوردار بوده و دچار بیش‌پردازش نشده‌اند. در این پژوهش، هدف اصلی دستیابی به بیشترین دقت ممکن پیش‌بینی بارش نبوده، بلکه تمرکز اصلی بر جداسازی و ارزیابی اثر واقعی روش‌های پیش‌پردازش داده بر عملکرد مدل‌های یادگیری ماشین بوده است. به همین دلیل، ورودی‌های مدل به متغیرهای پایه شامل مجموع بارش ماهانه و شماره ماه محدود شدند تا اثر سایر متغیرهای هواشناسی و اقلیمی بر نتایج به حداقل برسد. استفاده از تعداد زیاد متغیرهای کمکی می‌تواند باعث افزایش پیچیدگی مدل، ایجاد وابستگی‌های پنهان و پوشیده شدن اثر واقعی فرآیند اصلاح داده‌ها شود.

در واقع، این پژوهش تلاش کرده است با کنترل پیچیدگی ورودی‌ها، نقش مستقل روش‌های حذف داده پرت و برآورد داده‌های مفقودی را در بهبود عملکرد مدل‌ها بررسی کند. این رویکرد امکان تحلیل شفاف‌تر تعامل میان کیفیت داده و رفتار مدل‌های یادگیری ماشین را فراهم کرده و موجب شده است که تغییرات عملکرد مدل‌ها مستقیماً به فرآیند پیش‌پردازش نسبت داده شود، نه به افزایش تعداد متغیرهای ورودی یا پیچیدگی ساختار مدل.

۲-۶- منطقه مطالعاتی

مطالعه حاضر در سه حوضه آبریز نیمه‌خشک ایران انجام شده است (شکل ۱). حوضه گاوخونی و حوضه سیاه‌کوه در مرکز ایران و حوضه قره‌قوم در شرق کشور. این حوضه‌ها به دلیل شرایط اقلیمی خشک و نیمه‌خشک، تنوع مکانی و نوسانات زمانی بارش، از اهمیت بالایی برای مطالعات پیش‌بینی بارش و مدیریت منابع آب برخوردارند.

حوضه گاوخونی در مرکز ایران واقع شده و بخشی از دشت کویر را شامل می‌شود. این حوضه به دلیل کاهش منابع آب

است. بر اساس نمودار توزیع فراوانی (Box Plot) شاخص KGE، تفاوت معناداری میان حالت بدون پیش‌پردازش (Non/Non) و سایر ترکیب‌های پردازشی مشاهده می‌شود. در حالت عدم اعمال پیش‌پردازش، میانه شاخص KGE در پایین‌ترین سطح خود (حدود ۰/۱۲) قرار دارد و دامنه تغییرات آن بسیار گسترده است که حتی به مقادیر منفی (تا -۰/۲) نیز کشیده شده است. علت اصلی این افت شدید کارایی، وجود نویزهای آماری، داده‌های گمشده و مقادیر پرت غیرفیزیکی در سری‌های زمانی خام است که منجر به انحراف مدل از یافتن روابط بهینه میان متغیرها و در نتیجه افزایش بایاس و کاهش همبستگی در خروجی نهایی می‌شود.

معناداری تحت تاثیر کیفیت پیش‌پردازش داده‌ها قرار دارد و انتخاب نامناسب روش‌های حذف داده پرت و پر کردن مقادیر گمشده می‌تواند منجر به کاهش شدید دقت مدل‌ها شود. به‌طور کلی، یافته‌ها بر اهمیت در نظر گرفتن تعامل میان روش‌های پیش‌پردازش داده، نوع مدل یادگیری ماشین و ویژگی‌های خاص هر حوضه تاکید دارند و نشان می‌دهند که به‌کارگیری یک راهکار واحد برای تمامی شرایط، لزوماً به عملکرد بهینه منجر نخواهد شد.

تحلیل آماری خروجی مدل‌های یادگیری ماشین تحت سناریوهای مختلف پیش‌پردازش، بیانگر تاثیر سرنوشت‌ساز مدیریت داده‌های ورودی بر کارایی شبیه‌سازی هیدرولوژیکی

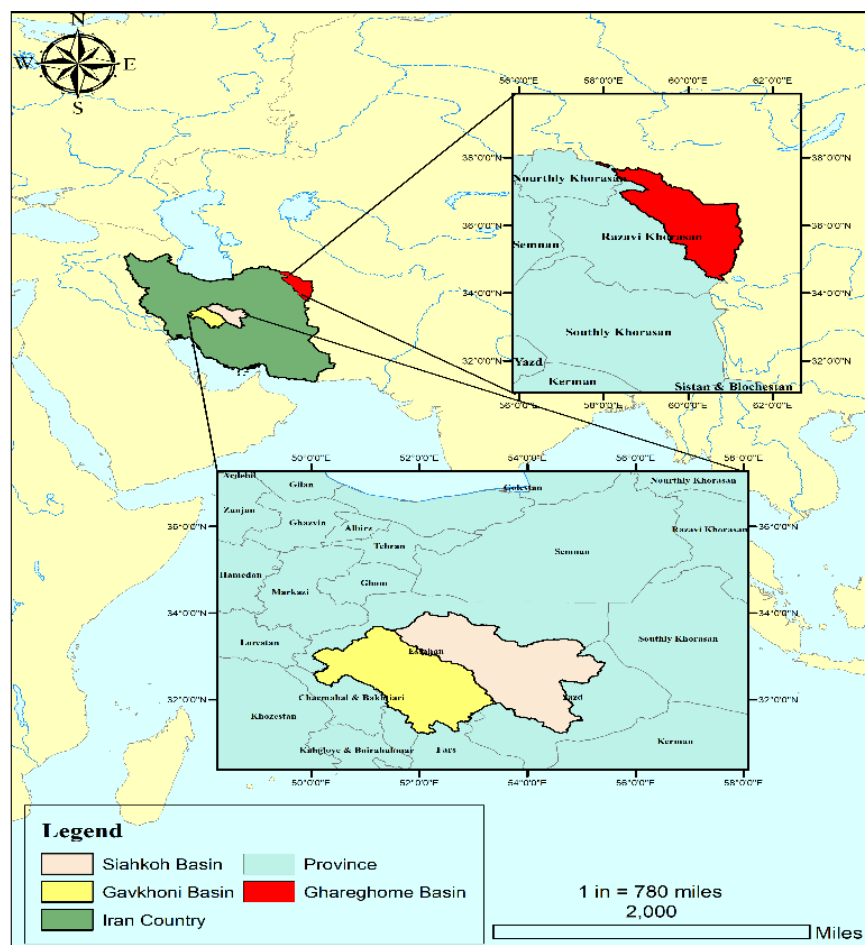


Fig. 1 The Studyarea.

شکل ۱ منطقه مورد مطالعه

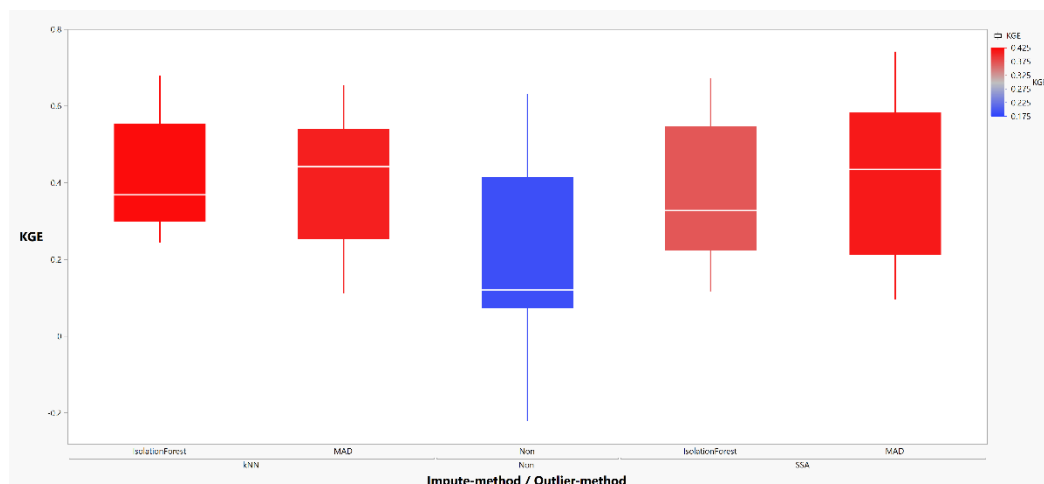


Fig. 2 Box plot of outlier removal methods, missing data imputation methods

شکل ۲ باکس پلات روش‌های حذف داده پرت، روش‌های پر کردن داده مفقودی

L1 و L2 برای مهار بیش‌برازش است. این در حالی است که حوضه سیاه‌کوه به‌دلیل رفتارهای غیرخطی شدید و اقلیم خشک که منجر به تولید داده‌های با انحراف معیار بالا می‌شود، بیشترین حساسیت را به نوع پیش‌پردازش نشان داده و در سناریوهای فاقد پیش‌پردازش، به کمترین سطح کارایی سقوط کرده است. در مجموع، نتایج عددی بر این واقعیت تاکید دارند که اعمال یک راهبرد ترکیبی نظیر kNN/IsolationForest یا SSA/MAD می‌تواند دقت مدل را از نظر شاخص KGE بیش از ۲۰۰ درصد نسبت به داده‌های خام بهبود بخشد. این تحلیل ثابت می‌کند که برای دستیابی به یک مدل هیدرولوژیکی قابل اتکا، علاوه بر انتخاب الگوریتم‌های قدرتمندی نظیر XGBoost، مدیریت هوشمندانه داده‌های پرت و گمشده جهت پالایش سیگنال‌های فیزیکی از نویزهای آماری، امری اجتناب‌ناپذیر است. در بخش برآورد مقادیر گمشده، روش kNN با میانگین KGE حدود ۰/۵۰ عملکرد برتری نسبت به روش SSA با میانگین حدود ۰/۴۹ ثبت کرده است. این نتایج حاکی از آن است که روش kNN به‌دلیل بهره‌گیری از شباهت‌های محلی در فضای ویژگی‌ها، توانسته است مقادیر گمشده را با انطباق بیشتری بر واقعیت فیزیکی حوضه برآورد کند. اگرچه اختلاف عملکرد این دو روش نسبتاً محدود است، اما برتری kNN نشان‌دهنده اهمیت حفظ همبستگی‌های

در مقابل، استفاده از روش‌های جایگذاری داده (Imputation) نظیر kNN و SSA به‌همراه الگوریتم‌های شناسایی داده‌های پرت مانند MAD و IsolationForest، به‌طور چشمگیری منجر به بهبود میانه کارایی و کاهش پراکندگی خطا شده است. ترکیب kNN/MAD با ثبت میانه‌ای در حدود ۰/۴۴ و ترکیب SSA/MAD با میانه‌ای نزدیک به ۰/۴۳، برترین سناریوها در ارتقای شاخص KGE بوده‌اند. علت برتری روش MAD در این توزیع، ماهیت مقاوم (Robust) آن در برابر انحرافات شدید آماری است که با تعدیل داده‌های پرت، واریانس خطا را کنترل کرده و پایداری مدل را افزایش می‌دهد. همچنین، باریک‌تر بودن عرض جعبه‌ها (Interquartile Range - IQR) در سناریوهای مبتنی بر kNN نسبت به حالت خام، نشان‌دهنده افزایش اطمینان (Reliability) مدل در پیش‌بینی‌هاست؛ چرا که kNN با بازسازی داده‌های گمشده بر اساس ویژگی‌های همسایگی فیزیکی، تداوم زمانی سری‌های هیدرولوژیکی را حفظ کرده و از گسستگی الگوهای یادگیری جلوگیری می‌کند.

از منظر مدل‌سازی، برتری قاطع XGBoost با ثبت بیشترین مقدار کارایی (NSE=0.81 و KGE=0.741) در حوضه‌های گاوخونی و قره‌قوم، ناشی از معماری پیشرفته آن در بهینه‌سازی توابع هدف پیچیده و استفاده از تنظیم‌کننده‌های

حوضه‌ها و راهبرد پیش‌پردازش داده‌ها نشان می‌دهد. تحلیل جامع شاخص‌های آماری ضریب ناش-ساتکلیف ((NSE)، شاخص کارایی کلینگ-گوپتا (KGE) و ریشه میانگین مربعات خطا (RMSE) نشان داد که مدل غیرخطی و جمعی XGBoost در اکثر سناریوها با ثبت بالاترین دقت مانند NSE معادل ۰/۸۱ در حوضه قره‌قوم و KGE معادل ۰/۷۴ در حوضه گاوخونی برترین عملکرد را داشته است. علت اصلی این برتری، ساختار درخت تصمیم تقویت‌یافته Gradient Boosting در الگوریتم XGBoost است که با اعمال توابع جریمه تصحیح‌کننده (Regularization L1 & L2) از بیش‌برازش (Overfitting) جلوگیری کرده و به‌طور متوالی خطاهای مدل‌های قبلی را به حداقل می‌رساند؛ این ویژگی به مدل اجازه می‌دهد تا الگوهای به‌شدت غیرخطی، پیچیده و پویای چرخه آب در حوضه را با انعطاف‌پذیری بالایی استخراج کند. در مرتبه دوم، مدل جنگل تصادفی (RF) به‌دلیل ماهیت میانگین‌گیری از درخت‌های تصمیم مستقل (Bagging) پایداری مناسبی نشان داد، اما به علت ناتوانی ذاتی در برون‌یابی مقادیر خارج از محدوده داده‌های آموزش و تمایل به بیش‌تخمین مقادیر کم و کم‌تخمین مقادیر بیشینه، در رتبه‌ای پایین‌تر از XGBoost قرار گرفت. در مقابل، رگرسیون خطی (LR) با ثبت مقادیر منفی یا نزدیک به صفر در شاخص‌های NSE و KGE ضعیف‌ترین عملکرد را داشت؛ علت این ضعف شدید، فرض صلب همبستگی خطی میان متغیرهای ورودی و خروجی در مدل LR است، در حالی که فرآیندهای هیدرولوژیکی مانند بارش یا دبی رودخانه تحت تاثیر رفتارهای تاخیری، غیرخطی و آستانه‌ای زمین و اتمسفر هستند که مدل‌های خطی اساساً فاقد ظرفیت ریاضی برای بازنمایی آن‌ها می‌باشند.

تحلیل حوضه‌محور نشان می‌دهد که پاسخ مدل‌ها به‌شدت به ویژگی‌های فیزیکی و فیزیوگرافی هر حوضه وابسته است؛ به‌طوری‌که حوضه گاوخونی پیش‌بینی‌پذیرترین رفتار را با کمترین میزان RMSE نشان داد. علت این امر را می‌توان در وسعت، شیب ملایم‌تر، رفتار هیدرولوژیکی منظم‌تر و احتمالاً

نزدیک در بازسازی سری‌های زمانی هیدرولوژیکی است. همچنین، تحلیل نمودار توزیع فراوانی (Box Plot) نشان داد که در غیاب این پیش‌پردازش‌ها (حالت Non)، میانه کارایی به سطح بحرانی ۰/۱۲ سقوط می‌کند که تأکیدی بر ضرورت غیرقابل‌انکار پالایش داده‌ها پیش از ورود به مرحله آموزش مدل است.

مقایسه مدل‌های یادگیری ماشین در مقیاس میانگین تمامی سناریوها نشان داد که مدل RF با میانگین KGE حدود ۰/۵۱ عملکرد پایدارتری نسبت به مدل XGBoost با میانگین حدود ۰/۴۸ داشته است. این برتری میانگین نشان می‌دهد که مدل جنگل تصادفی (RF) در مواجهه با عدم قطعیت‌ها، نویزهای باقی‌مانده و ناهمگنی موجود در داده‌های هیدرولوژیکی سه حوضه، مقاومت و پایداری بیشتری از خود نشان داده است. با این حال، باید توجه داشت که مدل XGBoost در حوضه‌هایی با داده‌های باکیفیت‌تر (نظیر قره‌قوم با $NSE = 0.81$)، پتانسیل بالاتری برای ثبت رکوردهای دقت نقطه‌ای دارد، اما در نگاه کلی و میانگین حوضه‌ها، مدل RF به‌دلیل خاصیت میانگین‌گیری در فرآیند Bagging، در کنترل واریانس خطا موفق‌تر عمل کرده است. در مقابل، مدل رگرسیون خطی (LR) به‌دلیل فرض صلب خطی بودن روابط، در تمامی شاخص‌ها ضعیف‌ترین عملکرد را داشت که نشان‌دهنده ناتوانی آن در بازنمایی رفتارهای آستانه‌ای و غیرخطی حوضه‌های خشک و نیمه‌خشک نظیر سیاه‌کوه است. در نهایت، تفاوت عملکرد حوضه‌ها (به‌ویژه خطای بالای RMSE در سیاه‌کوه تا ۰/۹۹) تأکید می‌کند که پیچیدگی‌های اقلیمی و هیدرولوژیکی، مانند وقوع سیلاب‌های ناگهانی، انتخاب ترکیب بهینه از روش‌های پیش‌پردازش و مدل‌سازی را به یک ضرورت حیاتی برای دستیابی به پیش‌بینی‌های قابل اتکا در مدیریت منابع آب تبدیل کرده است.

ارزیابی عملکرد مدل‌های مختلف یادگیری ماشین در شبیه‌سازی متغیرهای هیدرولوژیکی سه حوضه آبریز مورد مطالعه، تفاوت‌های معناداری را تحت تاثیر مستقیم نوع معماری الگوریتم، ویژگی‌های ژئومورفولوژیکی و اقلیمی

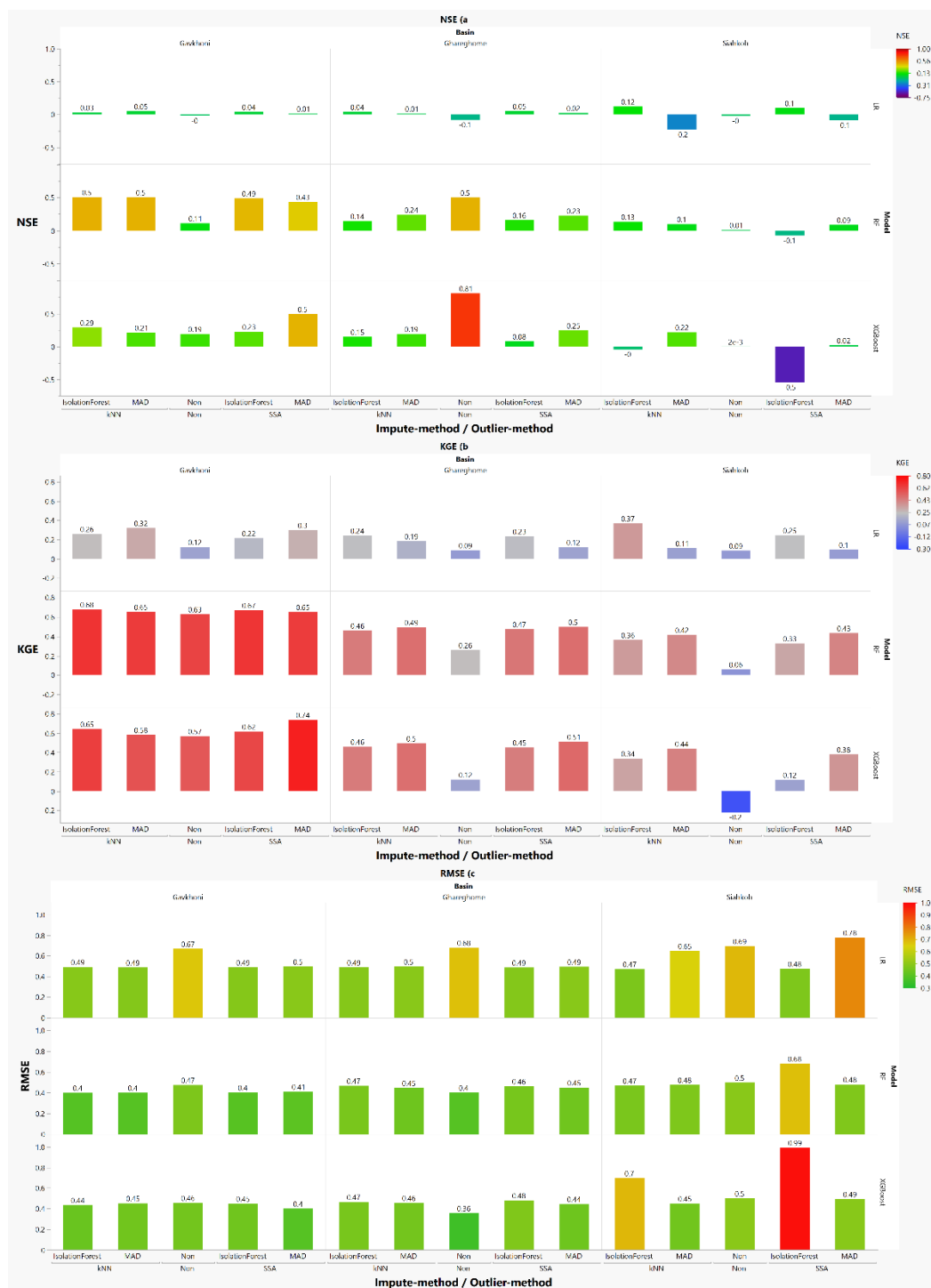


Fig. 3 Performance of corrected series based on the a) NSE, b) KGE, and c) RMSE indexes

شکل ۳ عملکرد سری‌های اصلاح شده با شاخص (آ) NSE، (ب) KGE، و (پ) RMSE

مانع از گسستگی زمانی یا ایجاد بایاس مصنوعی در ساختار داده‌ها می‌شود. در نهایت، تفاوت رفتاری شاخص‌ها (برای مثال ثبت KGE مناسب هم‌زمان با RMSE متوسط) نشان می‌دهد که RMSE به شدت تحت تاثیر خطاهای بزرگ در مقادیر فرین (به دلیل به توان دو رسیدن خطاها) قرار دارد، در حالی که KGE به صورت متوازن همبستگی، بایاس خطی و تغییرپذیری را می‌سنجد؛ از این رو، ارزیابی چندشاخصه برای جلوگیری از تفسیر نادرست عملکرد مدل‌ها و تضمین قابلیت تعمیم آن‌ها در کاربردهای عملی مدیریت منابع آب امری حیاتی است.

شکلهای ۲ و ۳ نشان می‌دهد که عملکرد مدل‌های یادگیری ماشین به شدت تحت تاثیر ترکیب روش حذف داده پرت و روش برآورد داده‌های مفقودی قرار دارد. به‌طور کلی، ترکیب روش MAD برای حذف داده‌های پرت با روش‌های برآورد داده، عملکرد پایدارتری نسبت به سایر روش‌ها نشان داده است. همچنین در اغلب ترکیبات، مدل Random Forest (RF) مقادیر بالاتری از شاخص KGE را نسبت به XGBoost ارائه کرده که بیانگر توانایی بیشتر این مدل در بازتولید الگوی زمانی بارش است. بررسی حوضه‌های آبریز مختلف نیز نشان داد که حوضه گاوخونی (Gavkhoni) بیشترین مقادیر KGE را در اکثر ترکیبات مدل‌سازی داشته است، در حالی که عملکرد مدل‌ها در حوضه سیاهکوه (Siahkoh) نسبتاً ضعیف‌تر بوده است. این موضوع می‌تواند ناشی از پیچیدگی بیشتر الگوی بارش و نوسانات زمانی در این حوضه باشد. همچنین نتایج بیانگر آن است که انتخاب روش مناسب برای اصلاح داده‌ها، تاثیر مستقیمی بر دقت مدل‌های یادگیری ماشین دارد؛ به‌گونه‌ای که در برخی موارد، تغییر روش حذف داده پرت یا روش برآورد داده مفقودی منجر به تغییر قابل توجه در مقدار شاخص KGE شده است. بنابراین، مرحله پیش‌پردازش داده‌ها نقش کلیدی در بهبود عملکرد مدل‌های پیش‌بینی ایفا می‌کند و نمی‌توان آن را مستقل از فرآیند مدل‌سازی در نظر گرفت.

نویز کمتر در داده‌های ثبت‌شده این حوضه جستجو کرد که به الگوریتم‌ها فرصت می‌دهد الگوهای فصلی و چرخه‌ای را به‌خوبی یاد بگیرند. در سوی دیگر، حوضه سیاهکوه به‌عنوان چالش‌برانگیزترین حوضه با خطای RMSE بالا (تا ۰/۹۹) و شاخص‌های کارایی منفی ارزیابی شد. علت فیزیکی و آماری این ناپایداری شدید در سیاهکوه، به ماهیت اقلیمی خشک یا نیمه‌خشک حوضه، وقوع سیلاب‌های ناگهانی و لحظه‌ای، نرخ بالای تبخیر و تعرق و رفتارهای هیدرولوژیکی غیرمنظم و آستانه‌ای مربوط می‌شود؛ در این شرایط، سری‌های زمانی حاوی مقادیر صفر طولانی‌مدت به همراه پیک‌های ناگهانی بسیار شدید هستند که این گسستگی و عدم تقارن در توزیع آماری داده‌ها، فرآیند آموزش مدل‌ها را دچار اختلال شدید کرده و منجر به خطا در برآورد مقادیر فرین می‌شود.

بررسی اثر استراتژی‌های پیش‌پردازش داده‌ها، فرضیات سنتی هیدرولوژی محاسباتی را به چالش کشید؛ چرا که در برخی سناریوها مانند مدل XGBoost در حوضه قره‌قوم، حالت بدون پیش‌پردازش (Non) دقیق‌ترین نتایج را ارائه داد. علت علمی این پدیده آن است که در سامانه‌های طبیعی، متغیرها از نظر آماری ممکن است به عنوان داده پرت توسط الگوریتم‌هایی نظیر MAD یا IsolationForest شناسایی و حذف یا تعدیل شوند، در حالی که این داده‌ها حاوی اطلاعات فیزیکی واقعی و حیاتی از رفتار پاسخ حوضه به بارش‌های شدید هستند و حذف آن‌ها عملاً مدل را از یادگیری رفتار حوضه در شرایط بحرانی محروم می‌سازد. با این حال، در شرایطی که داده‌ها آمیخته به نویزهای اندازه‌گیری یا خطاهای انسانی باشند، روش‌های MAD و IsolationForest با حذف این نوسانات کاذب، به پایداری مدل کمک کرده‌اند. همچنین عملکرد موفق جایگذاری داده‌های گمشده با روش kNN به این علت است که این روش با تکیه بر ویژگی‌های نمونه‌های همسایه در فضای چندبعدی متغیرها، مقادیر گمشده را بر اساس نزدیک‌ترین وضعیت‌های فیزیکی مشابه بازسازی می‌کند و

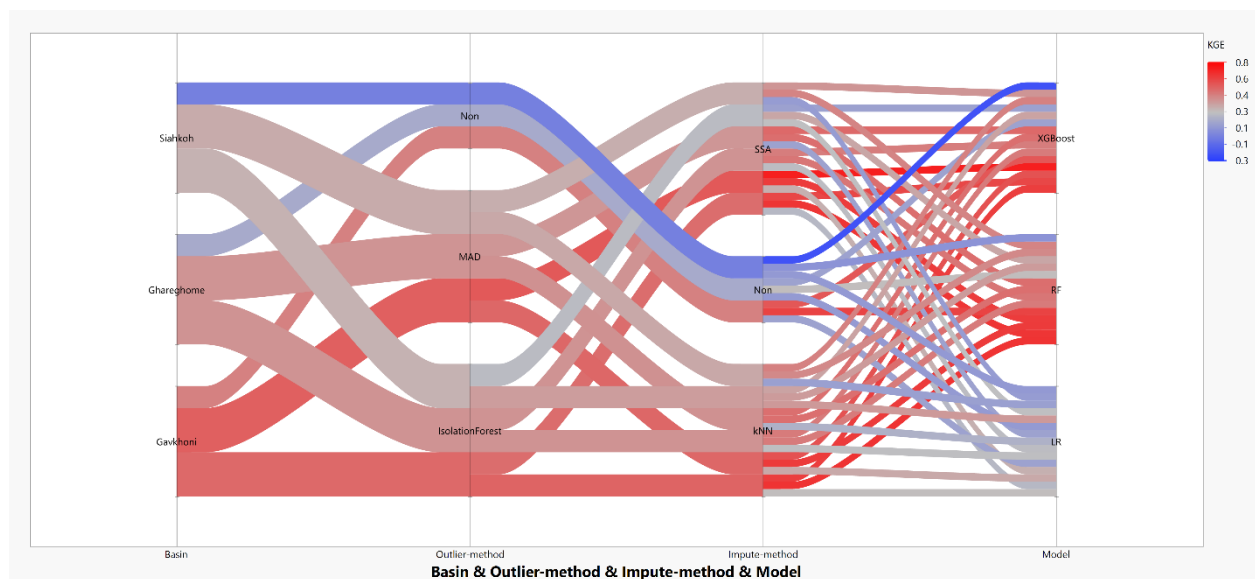


Fig. 4 Sankey diagram illustrating the relationships among outlier detection methods, missing data imputation methods, machine learning models, and basins based on the KGE index

شکل ۴ نمایش جریان ترکیبی روش‌های حذف داده پرت، روش‌های برآورد داده مفقودی، مدل‌های یادگیری ماشین و حوضه‌های آبریز بر اساس شاخص KGE



Fig. 5 Performance of combined correction methods

شکل ۵ عملکرد ترکیبی روش‌های اصلاح

مدل RF، ترکیب Isolation Forest + kNN با مقدار KGE حدود ۰/۶۷۹ و Isolation Forest + SSA با مقدار حدود ۰/۶۷۱ بهترین عملکرد را نشان می‌دهند. در مقابل،

در حوضه گاوخونی، بالاترین مقادیر KGE در میان تمامی حوضه‌ها مشاهده می‌شود که بیانگر کیفیت مناسب داده‌ها و پاسخ‌پذیری بالای مدل‌ها به روش‌های پیش‌پردازش است. در

داده‌ها همخوانی دارد. (Bojang et al. (2020). نشان دادند که ترکیب روش تحلیل طیف منفرد (SSA) با مدل‌های یادگیری ماشین، به‌ویژه Random Forest، می‌تواند عملکرد پیش‌بینی بارش ماهانه را نسبت به مدل‌های بدون پیش‌پردازش بهبود دهد. این یافته با نتایج مطالعه حاضر هم‌راستا است و نشان می‌دهد که کیفیت و ساختار داده‌های ورودی می‌تواند نقش مهمی در عملکرد مدل‌های یادگیری ماشین داشته باشد. با این حال، مطالعه حاضر دامنه گسترده‌تری از روش‌های پیش‌پردازش را بررسی کرده و اثر روش‌های مختلف تشخیص داده‌های پرت و تکمیل داده‌های مفقود را به‌صورت مقایسه‌ای ارزیابی می‌کند.

از سوی دیگر، مطالعات مختلف نشان داده‌اند که عملکرد مدل‌های یادگیری ماشین در پیش‌بینی بارش به ساختار زمانی داده‌ها، متغیرهای ورودی و روش مدل‌سازی وابسته است. (Oliveira et al. (2023 در یک مرور جامع از کاربرد یادگیری ماشین در پیش‌بینی بارش نشان دادند که اگرچه مدل‌های یادگیری عمیق و شبکه‌های عصبی سهم قابل توجهی در این حوزه دارند، مدل‌های درختی از جمله Random Forest و روش‌های مبتنی بر boosting نیز به‌طور گسترده مورد استفاده قرار گرفته‌اند. این موضوع با انتخاب Random Forest و XGBoost به‌عنوان مدل‌های اصلی در مطالعه حاضر سازگار است.

همچنین، Zhang and Ye (2021) در بررسی گسترده روش‌های یادگیری ماشین برای پس‌پردازش پیش‌بینی‌های بارش نشان دادند که تفاوت در ساختار مدل‌های یادگیری ماشین می‌تواند تاثیر قابل توجهی بر کیفیت پیش‌بینی داشته باشد و انتخاب مناسب ساختار مدل و ورودی‌های زمانی نقش مهمی در دستیابی به عملکرد بهتر دارد. نتایج این مطالعه از اهمیت بررسی سیستماتیک مدل‌ها و شرایط ورودی در مسائل پیش‌بینی بارش حمایت می‌کند و با رویکرد مقایسه‌ای مطالعه حاضر همخوانی دارد.

در مجموع، مقایسه با مطالعات پیشین نشان می‌دهد که بهبود

ترکیب‌های مبتنی بر MAD در این مدل اندکی عملکرد پایین‌تری (حدود ۰/۶۵) دارند. در مدل XGBoost، بیشترین مقدار KGE مربوط به ترکیب MAD + SSA با مقدار حدود ۰/۷۴۱ است که بالاترین عملکرد در کل [شکل ۳](#) محسوب می‌شود. این نتیجه نشان می‌دهد که در این حوضه، مدل XGBoost نسبت به انتخاب روش حذف داده پرت و روش برآورد مقادیر گم‌شده حساسیت بالایی دارد.

در حوضه قره‌قوم، اختلاف عملکرد میان ترکیب‌های مختلف کمتر از گاوخونی است، با این حال اثر روش‌های پیش‌پردازش همچنان قابل توجه است. در مدل RF، ترکیب‌های مبتنی بر MAD چه با kNN و چه با SSA مقادیر KGE بالاتری (حدود ۰/۴۹۵ تا ۰/۴۹۸) نسبت به Isolation Forest ارائه می‌دهند.

در مدل XGBoost نیز ترکیب MAD + SSA با مقدار KGE حدود ۰/۵۱۳ بهترین عملکرد را دارد، در حالی که ترکیب Isolation Forest + SSA کمترین مقدار را در این مدل ثبت کرده است. این نتایج نشان می‌دهد که در این حوضه، MAD یک روش پایدارتر برای حذف داده‌های پرت محسوب می‌شود. در حوضه سیاه‌کوه، بیشترین حساسیت به انتخاب روش‌های پیش‌پردازش مشاهده می‌شود.

در مدل RF، استفاده از MAD در ترکیب با هر دو روش kNN و SSA منجر به عملکرد بهتری KGE (حدود ۰/۴۱۶ تا ۰/۴۳۴) نسبت به Isolation Forest شده است.

در مدل XGBoost، ترکیب Isolation Forest + SSA ضعیف‌ترین عملکرد کل [شکل ۳](#) را با مقدار KGE حدود ۰/۱۱۶ نشان می‌دهد، در حالی که ترکیب MAD + kNN بهترین نتیجه را در این حوضه برای این مدل فراهم کرده است KGE (حدود ۰/۴۴۱). این اختلاف شدید بیانگر حساسیت بالای مدل XGBoost به کیفیت داده‌ها و روش‌های پیش‌پردازش در این حوضه است.

یافته‌های این پژوهش با مطالعات پیشین در زمینه پیش‌بینی بارش با استفاده از روش‌های یادگیری ماشین و پیش‌پردازش

هیدرولوژیکی در تمامی سناریوها ارائه داد.

این مطالعه با ارائه یک چارچوب ارزیابی جامع نشان داد که بر خلاف رویکردهای سنتی تمرکز بر یک مدل منفرد، کل زنجیره آماده‌سازی داده‌ها بر خروجی نهایی اثرگذار است. نتایج اثبات کرد که هیچ ترکیب پیش‌فرضی برای همه حوضه‌ها بهینه نیست و انتخاب روش‌ها باید متناسب با ویژگی‌های فیزیکی و اقلیمی حوضه صورت گیرد.

محدودیت اصلی این پژوهش، عدم قطعیت ناشی از انتخاب ابرپارامترهای بهینه پیش‌پردازش در اقلیم‌های به‌شدت خشک و سیلابی بود. پیشنهاد می‌شود در مطالعات آتی، اثر روش‌های پیش‌پردازش چندگانه بر روی سری‌های زمانی با گام‌های زمانی روزانه و ساعتی بررسی شده و از الگوریتم‌های بهینه‌سازی هوشمند (نظیر Optuna) برای تنظیم هم‌زمان پارامترهای پیش‌پردازش و مدل‌های یادگیری ماشین استفاده شود.

یکی از محدودیت‌های این پژوهش، محدود بودن مجموعه آزمون به دو سال پایانی دوره مطالعه است. افزایش طول دوره آزمون یا استفاده از روش‌هایی مانند اعتبارسنجی گام‌به‌گام (Walk-forward Validation) و اعتبارسنجی متقاطع ویژه سری‌های زمانی (Time Series CrossValidation) در مطالعات آینده می‌تواند به ارزیابی دقیق‌تر پایداری و قابلیت تعمیم مدل‌های پیشنهادی کمک کند.

5- List of symbols

r (-)	Pearson correlation coefficient
α (-)	Variability ratio
β (-)	Bias ratio
O_i (mm)	Observed precipitation
P_i (mm)	Predicted precipitation
$\bar{O}$ (mm)	Mean observed precipitation
σ_o (mm)	Standard deviation of observed precipitation
n (-)	Number of observations

Acknowledgments

The authors would like to thank all individuals and organizations who contributed to the completion of

عملکرد مدل‌های یادگیری ماشین تنها به انتخاب الگوریتم پیش‌بینی وابسته نیست، بلکه کیفیت، ساختار و نحوه پیش‌پردازش داده‌های ورودی نیز می‌تواند نقش تعیین‌کننده‌ای داشته باشد. بنابراین، نتایج مطالعه حاضر بر اهمیت بررسی هم‌زمان مرحله پیش‌پردازش و مرحله مدل‌سازی تاکید می‌کند.

۴- نتیجه‌گیری

این پژوهش به ارزیابی نظام‌مند زنجیره پیش‌پردازش داده‌ها و مدل‌سازی هیدرولوژیکی در حوضه‌های آبریز با کیفیت داده‌های متفاوت پرداخت. نادیده گرفتن مرحله پیش‌پردازش داده‌ها (حالت خام) کارایی شبیه‌سازی را به‌شدت کاهش می‌دهد؛ به‌طوری که اعمال پیش‌پردازش مناسب، میانه شاخص KGE را از مقدار بحرانی ۰/۱۲ به بیش از ۰/۴۴ ارتقا داد. این امر نقش حیاتی پیش‌پردازش را در حوضه‌های با نوسان شدید هیدرولوژیکی (نظیر سیاه‌کوه) به عنوان یک شرط لازم اثبات می‌کند. در میان راهبردهای پیش‌پردازش، ترکیب روش انحراف معیار میانه (MAD) برای حذف داده‌های پرت و الگوریتم kNN برای جایگذاری داده‌های گمشده، پایدارترین عملکرد را نشان دادند. در بخش مدل‌سازی، هرچند مدل XGBoost به دقت‌های نقطه‌ای بالاتری دست یافت، اما مدل Random Forest (RF) با میانگین KGE حدود ۰/۵۱، عملکردی پایدارتر و مقاوم‌تر در برابر ناهمگنی داده‌های

this research.

Credit Authorship Contribution Statement

Conceptualization, A.D.G., M.A. and J.S.; Methodology, A.D.G. and M.A.; Software, A.D.G.; Data Curation, A.D.G.; Data Analysis, A.D.G.; Investigation, A.D.G., M.A. and J.S.; Writing—Original Draft Preparation, A.D.G.; Writing—Review and Editing, M.A. and J.S.; Visualization, A.D.G.; Supervision, M.A. and J.S.

All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data used in this study are publicly available and can be accessed upon reasonable request from the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

- Adam, A. Y. F., Gribovszki, Z., & Kalicz, P. (2025). *Enhancing Rainfall Measurement Using Remote Sensing Data in Sparse Rain Gauge Networks: A Case Study in White Nile State, Sudan* The 1st International Conference on Advanced Remote Sensing & Shaping Sustainable Global Landscapes (ICARS 2025),
- Ahmadi, H., Jafari, M., & Mashhadi, N. (1999). Relationship between geomorphology units, soil and groundwater table in Siah-kooch playa.
- Amrehn, M., Mualla, F., Angelopoulou, E., Steidl, S., & Maier, A. (2018). The random forest classifier in WEKA: Discussion and new developments for imbalanced data. *arXiv preprint arXiv:1812.08102*.
- Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., Huffman, G. J., Adler, R. F., & Wood, E. F. (2019). Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrology and Earth System Sciences*, 23(1), 207–224.
- Bojang, P. O., Yang, T. C., Pham, Q. B., & Yu, P. S. (2020). Linking Singular Spectrum Analysis and Machine Learning for Monthly Rainfall Forecasting. *Applied Sciences*, 10(9), 3224. <https://doi.org/10.3390/app10093224>.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21–27
- Dalir Gabrabad, A., Abdollahi, M. A., & Ashrafzadeh, A. (2026). Monthly rainfall simulation using hybrid machine learning models: Salt Lake Basin. *Journal of Water and Soil*, Advance online publication. <https://doi.org/10.22067/jsw.2026.97595.1522>.
- Ghosh, S., Gourisaria, M. K., Sahoo, B., & Das, H. (2023). A pragmatic ensemble learning approach for rainfall prediction. *Discover Internet of Things*, 3(1). <https://doi.org/10.1007/s43926-023-00044-3>
- Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. CRC press.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2), 80–91.
- Hajirad, I., Paria Pourmohammad, P. P., & Pourgholam Amiji, M. (2026). Optimizing Water Productivity in the Face of Climate Change: The Central Role of Machine Learning Approaches. *Iranian Journal of Soil and Water Research*, 56(11), 2929–2949.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730), 1999–2049.
- Jabari, M., Hajirad, I., & Jabbari, M. (2026). Sustainable Irrigation Water Management in Horticultural Crops of Arid and Semi-Arid Regions: Approaches and Challenges. *Applied Research in Water Engineering*, e733343.
- Jumadi, J., Danardono, D., Roziaty, E., Ulinuha, A., Supari, S., Choy, L. K., Sattar, F., & Nawaz, M. (2025). AI-Driven Ensemble Learning for Spatio-Temporal Rainfall Prediction in the Bengawan Solo River Watershed, Indonesia. *Sustainability*, 17(20). <https://doi.org/10.3390/su17209281>
- Kanani, S., Patel, S., Gupta, R. K., Jain, A., & Lin, J. C. (2023). An AI-Enabled ensemble method for rainfall forecasting using Long-Short term memory. *Math Biosci Eng*, 20(5), 8975–9002. <https://doi.org/10.3934/mbe.2023394>
- Kundu, S., Biswas, S. K., Tripathi, D., Karmakar, R., Majumdar, S., & Mandal, S. (2023). A review on rainfall forecasting using ensemble learning techniques. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 6. <https://doi.org/10.1016/j.prime.2023.100296>
- 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. *Journal of experimental social psychology*, 49(4), 764–766.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. 2008 eighth IEEE international conference on data mining.
- Le et al. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms

- for time-series forecasting. *Machine Learning with Applications*, 7, 100204. <https://doi.org/10.1016/j.mlwa.2021.100204>.
- Moghaddam N., Hadizadeh H. & Ashrafzadeh A. Data-driven analysis for estimating and predicting well salinity using artificial intelligence algorithms. *Discov ApplSci* (2025). <https://doi.org/10.1007/s42452-025-08021-x>
- Omary, A., Wedyan, A., Zghoul, A., Banihani, A., & Alsmadi, I. (2012). An interactive predictive system for weather forecasting. 2012 International Conference on Computer, Information and Telecommunication Systems (CITS),
- Oliveira, E. C. L. de, et al. (2023). Precipitation forecasting: from geophysical aspects to machine learning applications. *Frontiers in Climate*, 5, 1250201. <https://doi.org/10.3389/fclim.2023.1250201>
- Patel, A., Keriwala, N., Soni, N., Goel, U., Bhoj, R., Adhyaru, Y., & Yadav, S. M. (2023). Rainfall Prediction using Machine Learning Techniques for Sabarmati River Basin, Gujarat, India. *Journal of Engineering Science and Technology Review*, 16(1), 101–108. <https://doi.org/10.25103/jestr.161.13>
- Pourhashemi, S., & Aliabadi, K. (2025). Flood Susceptibility Mapping in Ghare Ghom watershed. *Journal of Arid Regions Geographic Studies*, 16(60), 188–203.
- Ravar, Z., Zahraie, B., Sharifinejad, A., Gozini, H., & Jafari, S. (2020). System dynamics modeling for assessment of water–food–energy resources security and nexus in Gavkhuni basin in Iran. *Ecological Indicators*, 108, 105682.
- Sesethu, F., & Piotr, W. (2022). Framework for implementation of the Pitman-WR2012 model in seasonal hydrological forecasting: a case study of Kraai River, South Africa. *Water SA*, 48(1 January).
- <https://doi.org/10.17159/wsa/2022.v48.i1.3891>
- Zhang, Y., & Ye, A. (2021). Machine Learning for Precipitation Forecasts Postprocessing: Multimodel Comparison and Experimental Investigation. *Journal of Hydrometeorology*, 22(11), 3065–3085. <https://doi.org/10.1175/JHM-D-21-0096.1>
- Tuysuzoglu, G., Birant, K. U., & Birant, D. (2023). Rainfall Prediction Using an Ensemble Machine Learning Model Based on K-Stars. *Sustainability*, 15(7). <https://doi.org/10.3390/su15075889>.