Scientific-Promotional (Professional) Journal of Aquifer
Homepage: https://abkhansj.ut.ac.ir

Improvement of Monthly Precipitation Forecasting Accuracy Using a Data Preprocessing Framework and Machine Learning Models

Ali Dalir Gabrabad [1]✉ | Ali Milandarzadeh [2] | Seyed Alireza Sheykhan [3]

1. Corresponding Author, M.Sc., Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources, University of Tehran, Iran. E-mail: alidalir@ut.ac.ir
2. M.Sc., Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources, University of Tehran, Iran.
3. M.Sc., Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources, University of Tehran, Iran. E-mail: alireza.sheykhan@ut.ac.ir

Article Info

Article type: Original article

Article history:

Received: Jan. 23, 2026

Revised: Jun. 02, 2026

Accepted: Jun. 21, 2026

Published online: Jul. 18, 2026

Keywords:
Data Preprocessing,
Machine Learning Models,
Monthly Precipitation Forecasting,
, Lake Namak Basin.

ABSTRACT

This study examined the effect of data preprocessing on the performance of machine learning models in forecasting monthly precipitation in the Lake Namak basin. CHIRPS data were first validated against ground based observations (NSE = 70%), then cleaned using the Z Score method to remove outliers and subsequently reconstructed using Singular Spectrum Analysis (SSA) and linear interpolation. Statistical analyses indicated that preprocessing substantially reduced the skewness and kurtosis of the time series and enhanced data stability without noticeably altering the variance. Performance comparisons among the Random Forest, XGBoost, SVR, and CatBoost models showed that preprocessing significantly improved forecasting accuracy; notably, the Random Forest model achieved the largest enhancement, with MSE decreasing from 28.7 in the original series to approximately 15.7–16.0 in the corrected series, while SVR exhibited the least sensitivity to data adjustments. Both visual and numerical evaluations further confirmed the superior performance of Random Forest and XGBoost in terms of higher correlation and lower bias. Overall, the findings demonstrate that integrating effective preprocessing techniques with machine learning models particularly Random Forest plays a critical role in improving precipitation forecasting accuracy and reducing hydrological uncertainties.

Cite this article: Dalir Gabrabad, A., Milandarzadeh, A., Sheykhan, S., (2026) Improvement of Monthly Precipitation Forecasting Accuracy Using a Data Preprocessing Framework and Machine Learning Models, *Scientific-Promotional Journal of Aquifer*, 21 (1).

Publisher: The University of Tehran Press.

نشریه علمی–ترویجی آبخوان، دوره ۲۱، شماره ۱
Homepage: https://abkhansj.ut.ac.ir

بهبود دقت پیش‌بینی بارش ماهانه با استفاده از چارچوب پیش‌پردازش داده و مدل‌های یادگیری ماشین

علی دلیر گبرآباد[۱] ✉ | علی میلان‌درزاده [۲] | سید علیرضا شیخان [۳]

۱. نویسنده مسئول، دانشجوی کارشناسی ارشد گرایش مدیریت منابع آب، گروه آبیاری و آبادانی، دانشکدگان کشاورزی و منابع طبیعی، کرج، ایران. رایانامه: alidalir@ut.ac.ir
۲. دانشجوی کارشناسی ارشد گرایش مدیریت منابع آب، گروه آبیاری و آبادانی، دانشکدگان کشاورزی و منابع طبیعی، کرج، ایران.
۳. دانشجوی کارشناسی ارشد گرایش آبیاری و زهکشی، گروه آبیاری و آبادانی، دانشکدگان کشاورزی و منابع طبیعی، کرج، ایران. رایانامه: alireza.sheykhan@ut.ac.ir

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۱۱/۰۳
تاریخ بازنگری: ۱۴۰۵/۰۳/۱۲
تاریخ پذیرش: ۱۴۰۵/۰۳/۳۱
تاریخ انتشار: ۱۴۰۵/۰۴/۲۷

واژه‌های کلیدی:
یادگیری ماشین ،
پیش‌بینی بارش ماهانه،
پیش‌پردازش داده،
دریاچه نمک.

چکیده

در این پژوهش، اثر پیش‌پردازش داده‌ها بر عملکرد مدل‌های یادگیری ماشین در پیش‌بینی بارش ماهانه حوضه دریاچه نمک بررسی شد. داده‌های CHIRPS پس از اعتبارسنجی با ایستگاه‌های زمینی NSE = ٪۷۰ با استفاده از روش Z-Score از نقاط پرت پاک‌سازی و سپس به کمک روش‌های SSA و درون‌یابی خطی بازسازی شدند. نتایج آماری نشان داد که پیش‌پردازش موجب کاهش قابل‌توجه چولگی و کشیدگی سری‌های زمانی و افزایش پایداری داده‌ها بدون تغییر محسوس در واریانس شد. مقایسه عملکرد مدل‌های Random Forest، XGBoost، SVR و CatBoost نشان داد که پیش‌پردازش داده‌ها به‌طور محسوسی دقت پیش‌بینی را بهبود می‌دهد؛ به‌طوری‌که مدل Random Forest بیشترین بهبود را تجربه کرد و مقدار MSE آن از ۲۸/۷ در سری اولیه به حدود ۱۵/۷ تا ۱۶/۰ در سری‌های اصلاح‌شده کاهش یافت، در حالی که SVR کمترین حساسیت را به اصلاح داده‌ها نشان داد. تحلیل‌های بصری و آماری نیز برتری Random Forest و XGBoost را از نظر همبستگی بالاتر و سوگیری کمتر تأیید کردند. نتایج این مطالعه نشان می‌دهد که ترکیب پیش‌پردازش مؤثر داده‌ها با مدل‌های یادگیری ماشین، به‌ویژه Random Forest، نقش کلیدی در افزایش دقت پیش‌بینی بارش و کاهش عدم‌قطعیت‌های هیدرولوژیکی دارد.

استناد: دلیر گبرآباد؛ علی، میان‌درزاده؛ علی، شیخان؛ سید علیرضا، (۱۴۰۵) بهبود دقت پیش‌بینی بارش ماهانه با استفاده از چارچوب پیش‌پردازش داده و مدل‌های یادگیری ماشین، *نشریه علمی–ترویجی آبخوان*، ۲۱ (۱).

ناشر: مؤسسه انتشارات دانشگاه تهران.

مقدمه

بارش به‌عنوان یکی از مهم‌ترین ورودی‌های چرخه هیدرولوژیک، نقشی کلیدی در تأمین آب سطحی و زیرزمینی، تولید انرژی برق‌آبی و پایداری کشاورزی دارد (Shiru et al., 2020). پیش‌بینی دقیق بارش نه‌تنها برای برنامه‌ریزی کشاورزی و مدیریت سدها و مخازن ضروری است، بلکه در کاهش خطرات ناشی از رخدادهای حدی مانند سیلاب‌ها و خشکسالی‌ها نیز اهمیت دارد (Zhang et al., 2021). باتوجه به تغییرات اقلیمی و افزایش نوسانات بارش، اتکا به داده‌های دقیق و پالایش‌شده برای پیش‌بینی‌های مطمئن، بیش‌ازپیش حیاتی شده است. پیشرفت‌های فناورانه اخیر توانایی‌های جمع‌آوری داده را متحول کرده و امکان تولید مقادیر عظیمی از داده‌های سری زمانی در حوزه‌های مختلف را فراهم ساخته است (Blazquez-Garcia et al., 2020). در این میان، داده‌های مرتبط با منابع آب به‌ویژه بارش جایگاه ویژه‌ای دارند، زیرا بارش در چرخه هیدرولوژیکی نقش تعیین‌کننده‌ای در تغذیه منابع آب سطحی و زیرزمینی، کشاورزی، تولید انرژی و امنیت غذایی ایفا می‌کند (Shiru et al., 2020). پیش‌بینی دقیق بارش برای مدیریت پایدار منابع آب، طراحی سدها و مخازن، برنامه‌ریزی کشاورزی و کاهش ریسک بلایای طبیعی نظیر سیلاب و خشکسالی اهمیت حیاتی دارد (Zhang et al., 2021). بااین‌حال، داده‌های بارش اغلب با مشکلاتی همچون وجود ناهنجاری‌ها، نقاط پرت و مقادیر گمشده مواجه هستند که می‌تواند بر کیفیت تحلیل‌ها و دقت پیش‌بینی‌ها اثر منفی بگذارد (Saha et al., 2022). بنابراین، شناسایی و اصلاح نقاط پرت و گمشده نه‌تنها یک مرحله پیش‌پردازش، بلکه ضرورتی اساسی برای اطمینان از نتایج قابل‌اعتماد در مدل‌سازی و پیش‌بینی سری‌های زمانی بارش محسوب می‌شود.

در چنین شرایطی، احتیاج به رویکردهایی که بتوانند کیفیت داده‌ها را ارتقا داده و اثر آن را بر عملکرد مدل‌های پیش‌بینی ارزیابی کنند، بیش‌ازپیش احساس می‌شود. این پژوهش باهدف پر کردن این خلأ، چارچوبی چندمرحله‌ای را ارائه می‌دهد. در این چارچوب، ابتدا داده‌های بارش ماهانه از نظر وجود نقاط پرت بررسی و اصلاح شدند، سپس داده‌های حذف‌شده با استفاده از روش‌های متنوع بازسازی گردیدند، و در نهایت مجموعه‌داده اصلاح‌شده وارد مدل‌های پیش‌بینی گردید. نتایج نشان داد که ترکیب روش‌های پیش‌پردازش داده با الگوریتم‌های یادگیری ماشین منجر به بهبود چشمگیر عملکرد مدل‌ها می‌شود. ارزشمند باشد.

با گسترش فناوری‌های جمع‌آوری و پردازش داده، امروزه امکان دسترسی به حجم بسیار بالایی از داده‌های سری زمانی در حوزه‌های مختلف فراهم شده است. این تحول، توجه پژوهشگران را به تحلیل دقیق این داده‌ها، به‌ویژه شناسایی نقاط پرت یا ناهنجاری‌ها، جلب کرده است؛ چرا که ناهنجاری‌ها می‌توانند نشان‌دهنده رویدادهای بحرانی، خطاهای سیستم یا رفتارهای غیرمنتظره در داده‌ها باشند که نیازمند واکنش فوری هستند (Schmidl et al., 2022 ؛ Blazquez-Garcia et al., 2020). اهمیت این موضوع در کاربردهای مختلف، از سیستم‌های مالی و صنعتی تا حوزه‌های پزشکی و محیط‌زیستی، کاملاً مشهود است و باعث شده توسعه روش‌های مؤثر و دقیق برای تشخیص و مدیریت نقاط پرت به یک اولویت تحقیقاتی تبدیل شود (Toribio et al., 2025؛ Braei et al., 2020).

تشخیص ناهنجاری‌ها در سری‌های زمانی مستلزم درک پیچیده‌ای از وابستگی‌های ترتیبی، نوسانات طبیعی و الگوهای فصلی داده‌ها است (Jia et al., 2025). پژوهش‌های پیشین روش‌های متنوعی ارائه کرده‌اند که شامل رویکردهای آماری سنتی، مدل‌های مبتنی بر فاصله و همسایگی، روش‌های یادگیری ماشین و یادگیری عمیق می‌شود (Braei et al., 2020).

با افزایش کاربرد داده‌های چندمتغیره در سیستم‌های صنعتی و محیطی، توجه پژوهش‌ها به تشخیص ناهنجاری‌های جمعی جلب شده است؛ جایی که مقادیر منفرد ممکن است طبیعی به نظر برسند، اما روابط بین آن‌ها نشان‌دهنده رفتار غیرعادی سیستم باشد (Xu et al., 2023). همچنین، ظهور سناریوهای جریان داده در زمان واقعی، نیازمند توسعه الگوریتم‌هایی است که قادر به تشخیص سریع نقاط پرت با استفاده از مدل‌های تطبیقی باشند (Blazquez-Garcia et al., 2020).

در زمینه مدیریت داده‌های گمشده و پرت، روش‌های سنتی مانند پر کردن به جلو–عقب، میان‌یابی خطی و جایگزینی با مقدار غالب، ساده و محاسباتی سبک هستند، اما اغلب برای سری‌های زمانی پیچیده و چندمتغیره ناکافی‌اند (Kazijevs & ؛ Kamalov et al., 2021) Samad., 2023. روش‌های پیشرفته‌تر آماری و مبتنی بر رگرسیون مانند Multiple Imputation ، ARIMA و فیلترینگ کالمن، دقت بالاتری دارند؛ اما اغلب با داده‌های پر نویز و وابسته به زمان چالش دارند (Khattab et al., 2023;Magare et al., 2020).

رویکردهای یادگیری ماشین و یادگیری عمیق، شامل الگوریتم‌هایk-NN ، Random Forest، XGBoost، CatBoost و شبکه‌های عصبی LSTM وCNN ، نشان داده‌اند که می‌توانند بادقت بالاتر و انعطاف بیشتر، نقاط پرت و داده‌های گمشده را شناسایی و بازسازی کنند، حتی در شرایط داده‌های پر نویز و پویا (Shiru et al., 2020 ؛ Mallen et al., 2023؛ Baldyga et al., 2024). روش‌های نوین

مانند MLBUI، با تبدیل سری‌های زمانی تک‌متغیره به چندمتغیره و ترکیب پیش‌بینی‌های جلو و عقب، دقت تخمین و شناسایی روندهای واقعی داده‌ها را بهبود می‌بخشند (Khattab et al., 2023).

باتوجه‌به شواهد موجود، روش حاضر با بهره‌گیری از ترکیبی از تکنیک‌های یادگیری ماشین و الگوریتم‌های نوین سری زمانی، قادر است ناهنجاری‌ها و داده‌های گمشده را بادقت بالا شناسایی کند و اثرات جانبی ناشی از نقاط پرت را بر تحلیل‌های پایین‌دستی کاهش دهد. این ویژگی‌ها بیانگر برتری نسبی رویکرد پیشنهادی در مقایسه با روش‌های موجود و توانایی آن در کاربردهای صنعتی و محیطی با داده‌های پر نویز و پویا است.

در این پژوهش سعی بر آن شده‌است که اثر اصلاح داده بر دقت مدل‌های یادگیری ماشین برای پیش‌بینی بارش بارش در حوضه دریاچه نمک بررسی شود.

مواد و روش‌ها

روش‌شناسی پژوهش

در این پژوهش، به‌منظور بهبود دقت پیش‌بینی بارش ماهانه حوضه دریاچه نمک، داده‌های استخراج‌شده از پایگاه CHIRPS پس از شناسایی و حذف داده‌های پرت و بازسازی مقادیر مفقود، در قالب سری‌های زمانی آماده‌سازی شدند. سپس عملکرد مدل‌های یادگیری ماشین شامل Random Forest، CatBoost، Support Vector Regression و XGBoost برای ارزیابی اثر روش‌های پیش‌پردازش بر دقت پیش‌بینی مورد بررسی قرار گرفت. در شکل ۱ منطقه مورد مطالعه نمایش داده‌شده است.

شکل ۱: منطقه مورد مطالعه

گردآوری و آماده‌سازی داده‌ها (Data Acquisition and Preprocessing)

داده‌های بارش ماهانه حوضه دریاچه نمک برای دوره ۱۳۷۹ تا ۱۴۰۳ (۲۰۰۰ تا ۲۰۲۴) از پایگاه‌داده CHIRPS[1] استخراج شدند. این داده‌ها یکی از معتبرترین منابع برای تحلیل‌های هیدرولوژیکی در مناطق خشک و نیمه‌خشک هستند و در بسیاری از مطالعات اخیر در مدیریت منابع آب و پایش خشکسالی به‌کاررفته‌اند (Dinku et al., 2018؛ Funk et al., 2015). دقت مکانی (Spatial Resolution) داده‌های CHIRPS بادقت ۰/۰۵ درجه (~۵/۳ کیلومتر) ارائه شدند و واحد اندازه‌گیری بارش بر حسب میلی‌متر(mm) است.

به‌این‌ترتیب، یک سری زمانی بارش ماهانه در سطح حوضه، ایجاد شد. داده‌ها به دو بخش تقسیم شدند:

داده آموزشی: سال‌های ۲۰۰۰ تا ۲۰۱۹

داده آزمایشی: سال‌های ۲۰۲۰ تا ۲۰۲۴

این تقسیم‌بندی مشابه مطالعات دیگر در پیش‌بینی هیدرولوژیکی است که برای ارزیابی تعمیم‌پذیری مدل‌ها از داده‌های اخیر استفاده می‌کنند (Dotse et al., 2024). برای اطمینان از نمایندگی و پراکندگی فضایی مناسب در سرتاسر حوضه، چهار ایستگاه منتخب هیدرومتری و اقلیم‌شناسی که دارای سوابق بلندمدت و قابل اعتماد هستند، انتخاب شدند. این ایستگاه‌ها شامل موارد زیر هستند:

ایستگاه آغچه (استان اصفهان)

ایستگاه ارتش آباد (استان قزوین)

ایستگاه‌های آراقلعه (استان مرکزی)

احمدآباد (استان مرکزی)

این توزیع مکانی باعث می‌شود که نتایج ارزیابی، تنوع اقلیمی و توپوگرافی حوضه را پوشش داده و قابلیت تعمیم بالاتری داشته باشند. داده‌های بارش ماهانه از این چهار ایستگاه زمینی به‌عنوان مقادیر معیار و مشاهده‌شده (Observed)، در یک دوره ده‌ساله که از سال ۲۰۰۷ تا ۲۰۱۷ امتداد دارد، جمع‌آوری و با داده‌های ماهواره‌ای متناظر مقایسه شدند. انتخاب این بازه زمانی طولانی‌مدت برای کاهش اثر نویزهای کوتاه‌مدت و دربرگرفتن طیف کاملی از نوسانات اقلیمی (مانند سال‌های مرطوب و خشک) حائز اهمیت است.

شناسایی و حذف داده‌های پرت (Outlier Detection and Removal)

روش Z-Score

روش Z-Score یک روش آماری رایج برای شناسایی داده‌های پرت است. در این روش، ابتدا برای هر مشاهده x، مقدار Z-score محاسبه می‌شود:

$$Z = \frac{X - \mu}{\sigma}$$ رابطهٔ ۱)

که در آن μ میانگین و σ انحراف معیار کل سری داده‌ها است. سپس داده‌هایی که قدرمطلق Z-score آن‌ها بزرگ‌تر از یک آستانه مشخص باشد معمولاً ($|Z| > 2$) به‌عنوان داده پرت شناسایی و حذف می‌شوند (Barnett & Lewis, 1994).

پرکردن داده‌های حذف شده

پر کردن داده‌های حذف شده به روش SSA (Singular Spectrum Analysis)

SSA یکی از روش‌های پرکاربرد در تجزیه و بازسازی سری‌های زمانی است که با تفکیک داده‌ها به مؤلفه‌های روند، الگوهای فصلی و نویز عمل می‌کند. این قابلیت موجب می‌شود که داده‌های گمشده یا پرت شده در سری زمانی با استفاده از الگوهای درونی داده‌ها بازسازی شوند (Golyandina & Korobeynikov, 2014; Hassani, 2007).

پرکردن داده‌های حذف شده به روش درون‌یابی خطی (Linear Interpolation)

روش‌های درون‌یابی یکی از متداول‌ترین تکنیک‌ها برای بازسازی داده‌های گمشده در سری‌های زمانی هستند. این روش‌ها بر اساس نقاط مشاهده‌شده اطراف داده مفقود، مقدار جدید را تخمین می‌زنند و فرض می‌کنند که تغییرات داده به طور نسبی نرم و پیوسته است. رایج‌ترین

[1] Climate Hazards GroupInfraRed Precipitation with Station data

انواع آن شامل درون‌یابی خطی، مکعبی، اسپلاین طبیعی و PCHIP هستند (Shumway & Stoffer, 2017; Hyndman & Athanasopoulos, 2021). در سری‌های زمانی بارش، این روش‌ها به دلیل سادگی و سرعت اجرا بسیار مورداستفاده قرار می‌گیرند، زیرا امکان بازسازی سریع داده‌های گمشده در مقیاس‌های بزرگ را فراهم می‌کنند.

آماده‌سازی سری‌های زمانی (Time Series Preparation)

پس از شناسایی و حذف داده‌های پرت و بازسازی مقادیر حذف‌شده با استفاده از روش‌های مختلف SSA، و درون‌یابی، داده‌ها مجدداً به ساختار سری زمانی بازنویسی شدند. در این مرحله، هدف ایجاد مجموعه‌ای از سری‌های زمانی آماده برای پیش‌بینی بود. به طور مشخص، سه سری زمانی اصلاح‌شده حاصل گردید:

سری اولیه بدون هیچ بازسازی و پیش‌پردازش.

حذف داده‌های پرت با Z-Score سپس سری بازسازی‌شده با SSA.

حذف داده‌های پرت با Z-Score سپس سری بازسازی‌شده با Interpolation.

به‌منظور ارزیابی دقیق عملکرد، علاوه بر این سه سری اصلاح‌شده، سری اولیه بدون هیچ‌گونه اصلاح نیز حفظ گردید تا به‌عنوان مبنای مقایسه (Baseline) عمل کند. این رویکرد امکان بررسی میزان تأثیر هر روش حذف و بازسازی داده‌های پرت بر کیفیت پیش‌بینی‌های آتی را فراهم ساخت.

در نهایت، همه سری‌های زمانی آماده شدند تا ورودی مدل‌های یادگیری ماشین قرار گیرند. برای جلوگیری از بایاس و تضمین قابلیت تعمیم‌پذیری، داده‌ها به دو بخش آموزشی (۲۰۰۰–۲۰۲۰) و آزمون (۲۰۲۱–۲۰۲۴) این تقسیم‌بندی بر اساس توصیه‌های مرسوم در تحلیل سری‌های زمانی انجام شد که پیشنهاد می‌کنند حداقل ۷۰ تا ۸۰ درصد داده‌ها برای آموزش و بخش باقی‌مانده برای آزمون استفاده شود (Hyndman & Athanasopoulos, 2021).

پیشبینی سری زمانی

پس از شناسایی داده‌های پرت و پر کردن مقادیر گمشده، برای ارزیابی اثرگذاری روش ارائه شده، پیش‌بینی سری‌های زمانی انجام شد. در این مطالعه از چهار مدل پیش‌بینی استفاده شد:

- Random Forest (RF)
- CatBoost
- Support Vector Regression (SVR)
- XGBoost

این مدل‌ها به‌منظور بررسی دقت پیش‌بینی پس از اصلاح داده‌ها به کار گرفته شدند. برای هر مدل، سه ستون از سری‌های زمانی به‌عنوان ورودی در نظر گرفته شد و ستون اول با داده‌های نویزی‌تر و ستون‌های دوم و سوم با داده‌های باکیفیت بالاتر مورد پیش‌بینی قرار گرفتند.

برای افزایش دقت پیش‌بینی، از روش GridSearch برای پیداکردن بهترین مقادیر پارامترهای مدل‌ها استفاده شد. تعداد درخت‌ها، عمق درخت، نرخ یادگیری و سایر پارامترهای مرتبط به‌گونه‌ای تنظیم شدند که مدل بیشترین دقت ارائه دهد. در جدول ۱ پارامترهای انتخابی GridSearch برای هر مدل نشان‌داده‌شده است.

عملکرد مدل‌ها با استفاده از شاخص‌های استاندارد شامل MSE، MAE و NSE (Nash–Sutcliffe Efficiency) سنجیده شد. این شاخص‌ها برای هر ستون محاسبه شد تا تأثیر روش پر کردن داده‌ها بر دقت پیش‌بینی مشخص شود. نتایج

جدول ۱: پارامترهای انتخابی روش *grid search* برای هر مدل

Model	Original Series	Z-Score–SSA	Z-Score–Interpolation
RF	max_depth: 5, max_features: sqrt, n_estimators: 100	max_depth: 10, max_features: sqrt, n_estimators: 200	max_depth:10, max_features:sqrt, n_estimators: 100
SVR	C: 10, epsilon: 0.01, gamma: auto	C: 10, epsilon: 0.01, gamma: auto	C: 10, epsilon: 0.01, gamma: auto
XG Boost	learning_rate: 0.05, max_depth: 3, n_estimators: 100, subsample: 0.8	learning_rate: 0.05, max_depth: 3, n_estimators: 100, subsample: 0.8	learning_rate: 0.05, max_depth: 3, n_estimators: 100, subsample: 0.8
Cat Boost	depth: 4, iterations: 200, l2_leaf_reg: 1, learning_rate: 0.1	depth: 4, iterations: 150, l2_leaf_reg: 5, learning_rate: 0.05	depth: 5, iterations: 150, l2_leaf_reg: 5, learning_rate: 0.05

نتایج

پس از جمع‌آوری داده، نتایج کمی ارزیابی دقت ماهواره‌ای با استفاده از شاخص‌های آماری استاندارد محاسبه شد. این نتایج نشان داد که محصول ماهواره‌ای CHIRPS دارای دقت قابل‌قبولی است؛ با دستیابی به ضریب کارایی NSE معادل ۷۰/۰۰ درصد، این داده‌ها از نظر هیدرولوژیکی در محدوده عملکرد خوب تا بسیار خوب ارزیابی شده و با ضریب تعیین R^2 برابر ۶۸/۷۱ درصد، همبستگی خطی قوی و مثبتی با داده‌های زمینی نشان دادند. همچنین، مقدار سوگیری (Bias) بسیار ناچیز (۰/۰۹-)، حاکی از حداقل سوگیری سیستماتیک در این داده‌ها است و اعتبار لازم برای استفاده از آن‌ها در مرحله بعدی پژوهش را اثبات می‌کند. در جدول ۲ نتایج ارزیابی و دقت ماهواره CHIRPS را نشان می‌دهد.

جدول ۲: دقت ماهواره *CHIRPS*

ماهواره	NSE (%)	MSE (%)	MAE (%)	R^2	BAIAS (%)
CHIRPS	۷۰/۰۰	۱۹/۴۲	۳۱/۳۶	۶۸/۷۱	-۰/۰۹

نتایج اصلاح و پیش پردازش سری‌ها

به‌منظور بررسی تأثیر روش‌های مختلف اصلاح داده بر ویژگی‌های آماری سری زمانی بارش، شاخص‌های آماری اصلی شامل میانگین، انحراف معیار، واریانس، چولگی و کشیدگی، پیش و پس از اعمال روش‌های اصلاح محاسبه شدند. هدف از این تحلیل آن است که مشخص شود اصلاح داده‌ها تا چه اندازه موجب تغییر در ساختار آماری سری زمانی شده و آیا ویژگی‌های اصلی داده‌ها حفظ شده‌اند یا خیر. جدول زیر مقایسه‌ای بین داده‌های اولیه (Observed) و داده‌های اصلاح‌شده با روش‌های مختلف ارائه می‌دهد. در جدول شماره ۳ متغیرهای آماری سری‌ها نشان‌داده‌شده است.

جدول ۳: متغیرهای آماری سری‌ها

Variable	mean	std	min	25%	50%	75%	max	variance	skewness	kurtosis
Origional Series	۰/۹۴۴	۰/۸۵۳	۰/۰۳۰	۰/۱۱۵	۰/۷۸۵	۱/۵۵۰	۴/۲۸۴	۰/۷۲۷	۰/۷۹۷	۰/۱۴۳
Z score-SSA	۰/۸۸۹	۰/۷۶۰	۰/۰۳۰	۰/۱۱۵	۰/۷۸۵	۱/۴۸۰	۲/۶۰	۰/۵۷۸	۰/۴۹۶	-۰/۸۹۲
ZScore-Interpolation	۰/۹۰۲	۰/۷۷۱	۰/۰۳۰	۰/۱۱۵	۰/۷۸۵	۱/۵۲۲	۲/۶۰	۰/۵۹۴	۰/۴۶۶	-۰/۹۸۴

پس از اعمال روش‌های مختلف در اصلاح داده‌های بارش، تغییرات قابل‌توجهی در شاخص‌های آماری مشاهده شد. میانگین داده‌ها در

تمامی روش‌ها نسبت به داده‌ی اولیه (Observed) اندکی کاهش پیدا کرده است (از ۰/۹۴ به حدود ۰/۸۸ در برخی روش‌ها). این موضوع نشان می‌دهد که فرایند اصلاح، سبب بهبود میانگین بارش و نزدیک‌تر شدن مقادیر به سطح واقعی شده است.

انحراف معیار (Std) نیز تغییرات جزئی داشته و در محدوده ۰/۷۱ تا ۰/۸۵ قرار گرفته است. این موضوع بیانگر آن است که روش‌های اصلاح داده، پراکندگی کلی داده‌ها را به‌طور چشمگیری و قابل توجهی تغییر نداده‌اند و نوسانات اصلی داده‌ها همچنان حفظ شده است.

از نظر شاخص‌های توصیفی مانند چارک‌ها و بیشینه، تغییرات در حد جزئی بوده و نشان می‌دهد که ساختار کلی توزیع داده‌ها حفظ شده است. با این حال، مقادیر واریانس و انحراف از مرکز (Skewness) اندکی کاهش یافته است که می‌تواند نشان‌دهنده کاهش نامتقارن بودن داده‌ها و هموارسازی خطاها باشد.

شاخص کشیدگی (Kurtosis) نیز در تمامی روش‌ها کاهش یافته است (از حدود ۰/۱۴ در داده اولیه به حدود ۰/۹۸- در داده اصلاح‌شده). این تغییر نشان می‌دهد که داده‌های اصلاح‌شده نسبت به داده‌های اولیه، دارای توزیع مسطح‌تر و با احتمال کمتر برای وجود داده‌های پرت (Outlier)هستند.

به‌طورکلی نتایج بیانگر آن است که اصلاح داده‌ها ضمن حفظ ویژگی‌های اصلی، سبب بهبود میانگین، کاهش چولگی و کاهش کشیدگی توزیع شده است. در نتیجه، داده‌های اصلاح‌شده پایدارتر، هموارتر و مناسب‌تر برای مدل‌سازی‌های بعدی خواهند بود.

نتایج پیشبینی مدل‌ها

در این پژوهش، برای ارزیابی تأثیر روش‌های مختلف حذف و پرکردن داده‌های پرت و گمشده بر دقت پیش‌بینی، چهار مدل یادگیری ماشین شامل Random Forest (RF)، Support Vector Regression (SVR)، XGBoost، و CatBoost به کار گرفته شدند. عملکرد هر مدل بر روی سه سری زمانی مختلف (شامل داده‌های اصلی بدون پیش‌پردازش و داده‌های اصلاح‌شده) مورد مقایسه قرار گرفت.

به‌منظور ارزیابی عملکرد پیش‌بینی مدل‌های مختلف، ضریب همبستگی پیرسون و اسپیرمن بین مقادیر پیش‌بینی شده و سری داده‌های آزمون (ستون هدف) محاسبه شد. این محاسبات امکان سنجش میزان تطابق مدل‌ها با داده‌های واقعی را فراهم می‌کند. برای هر ستون، هم مقادیر همبستگی خطی (Pearson r) و هم همبستگی رتبه‌ای (Spearman rho) به همراه مقادیر p-value مربوط به معناداری آماری ارائه شده است. تمامی ستون‌ها در قالب جدول ۴ آورده شده‌اند تا مقایسه روش‌های پیش‌بینی و مدل‌ها به وضوح قابل مشاهده باشد.

جدول ٤: ضریب همبستگی نتایج مدل‌ها

Model	Series	Pearson_r	Pearson_p	Spearman_rho	Spearman_p
	Original Series	0.70	3.2769E-08	0.74	2.42488E-09
XG Boost	Z-Score–SSA	0.79	5.19366E-11	0.79	2.91951E-11
	Z-Score–Interpolation	0.76	6.65878E-10	0.78	7.78086E-11
	Original Series	0.78	1.61096E-10	0.79	5.76775E-11
SVR	Z-Score–SSA	0.78	1.6148E-10	0.79	5.76775E-11
	Z-Score–Interpolation	0.77	1.63807E-10	0.79	6.28974E-11
	Original Series	0.68	1.68701E-07	0.73	5.46795E-09
RF	Z-Score–SSA	0.79	2.54935E-11	0.78	9.0472E-11
	Z-Score–Interpolation	0.79	5.75785E-11	0.79	6.22213E-11
	Original Series	0.81	4.03527E-12	0.80	1.90785E-11
CAT Boost	Z-Score–SSA	0.63	1.94106E-06	0.73	5.00022E-09
	Z-Score–Interpolation	0.69	1.0088E-07	0.75	1.67789E-09

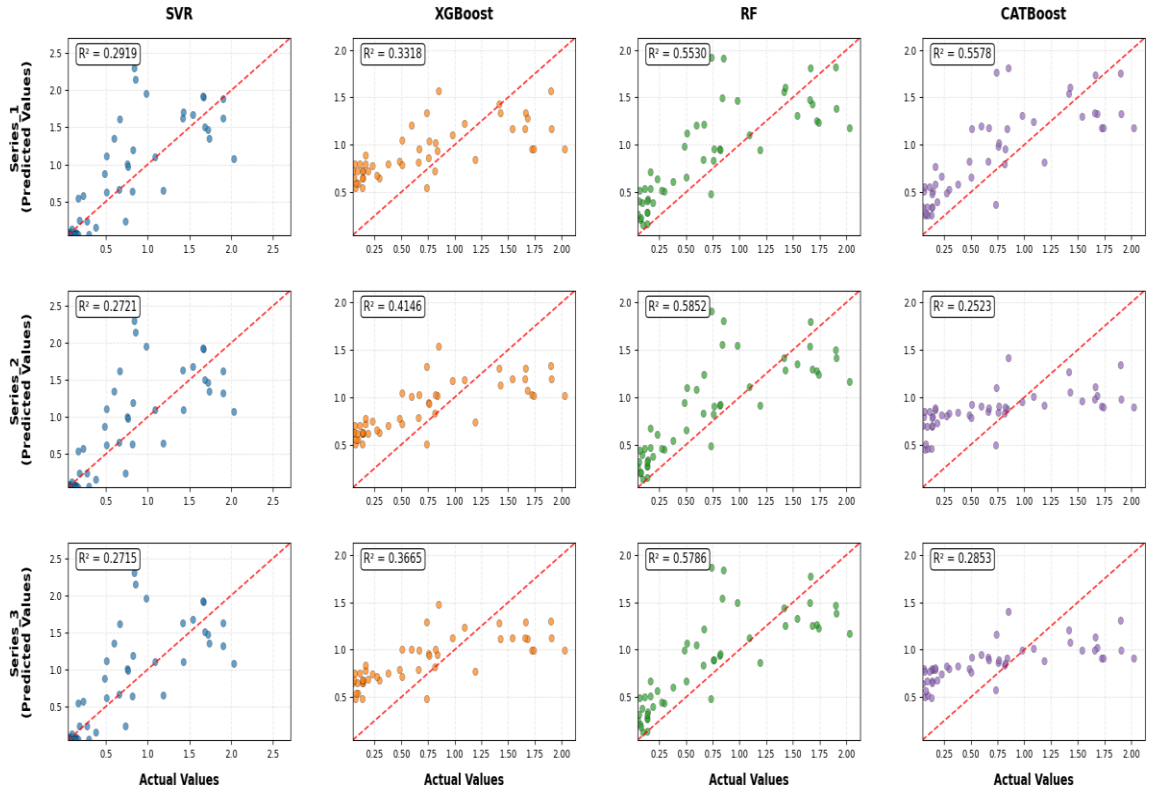
نتایج جدول نشان می‌دهد که اکثر پیش‌بینی‌های مدل‌ها با سری داده‌های آزمون دارای همبستگی مثبت و قابل‌توجه هستند، که نشان‌دهنده عملکرد مناسب مدل‌ها در پیش‌بینی مقادیر واقعی است.

ضریب پیرسون (Pearson r) برای اکثر ستون‌ها بین ۰/۷ تا ۰/۸۴ قرار دارد و رابطه خطی قوی بین مقادیر پیش‌بینی و داده‌های واقعی را نشان می‌دهد. ستون‌هایی با r بالاتر، دقت پیش‌بینی بیشتری دارند.
ضریب اسپیرمن (Spearman rho) نیز بین ۰/۷ تا ۰/۸۲ است، که بیانگر حفظ ترتیب رتبه‌ای مقادیر پیش‌بینی نسبت به داده واقعی می‌باشد و نشان می‌دهد حتی در مواردی که خطاهای جزئی وجود دارد، روند کلی پیش‌بینی با داده واقعی همخوانی دارد.
تمامی مقادیر p-value برای هر دو ضریب همبستگی بسیار کوچک هستند (معمولا ۰/۰۵>)، بنابراین این روابط از نظر آماری معنادار هستند و می‌توان به نتایج اعتماد کرد.
مقایسه مدل‌ها بر اساس میانگین ضریب همبستگی نشان می‌دهد که مدل‌های RFو XGBoost در مجموع بهترین تطابق با داده‌های واقعی را دارند، در حالی که مدل‌های دیگر مانند CatBoost یا ستون‌های خاصی از مدل چهارم اختلاف بیشتری با داده‌های آزمون دارند. نتایج کلی نشان می‌دهد که مدل‌ها توانایی پیش‌بینی مناسبی دارند و می‌توان از مقادیر پیش‌بینی شده برای تحلیل‌های بعدی یا ارزیابی عملکرد مدل‌ها استفاده کرد. علاوه بر این، همبستگی‌های اسپیرمن و پیرسون ترکیبی نشان می‌دهند که پیش‌بینی‌ها هم از نظر خطی و هم از نظر رتبه‌ای با داده‌های واقعی هماهنگی دارند که نشانه پایداری و اعتمادپذیری مدل‌هاست.
برای ارزیابی، از شاخص‌های میانگین مربعات خطا (MSE) و میانگین قدر مطلق خطا (MAE) استفاده شد. نتایج در قالب نمودارهای مقایسه‌ای ارائه گردیده‌اند.

شکل ۲: میانگین مربعات خطا و میانگین قدر مطلق خطا

همان‌طور که در نمودارهای مقایسه خطاها (Error Metrics Comparison) مشاهده می‌شود، نوع پیش‌پردازش داده‌ها نقش تعیین‌کننده‌ای در دقت پیش‌بینی مدل‌ها داشته است. در مدل Random Forest بیشترین بهبود مشاهده شد؛ به‌طوری‌که در سری‌های Z-Score–SSA مقدار MSE به حدود ۱۵/۷ تا ۱۶/۰ کاهش یافت، در حالی‌که این مقدار در سری اصلی حدود ۲۸/۷ بود. این نشان می‌دهد که RF نسبت به داده‌های پرت بسیار حساس بوده و با حذف و پرکردن مناسب، دقت پیش‌بینی آن به شکل چشمگیری افزایش می‌یابد.
مدل CatBoost نیز عملکرد مطلوبی از خود نشان داد که نسبت به داده‌های اصلی (MAE = 33.1) بهبود قابل‌توجهی محسوب می‌شود. البته، در برخی روش‌ها مانند Z-Score–Interpolation دقت کاهش یافت که نشان‌دهنده اهمیت انتخاب درست روش پیش‌پردازش است. در مدل XGBoost تغییرات خطا نسبتاً محدود بوده و این مدل پایداری بیشتری از خود نشان داده است. هرچند دقت آن به اندازه RFنبود، اما همچنان نشان‌دهنده بهبود نسبت به سری اصلی است. در مقابل، مدل SVR کمترین حساسیت به پیش‌پردازش داده‌ها را نشان داد. مقادیر خطای آن در بازه‌ای نسبتاً ثابت (MSE ≈ 26–27) قرار گرفتند و تفاوت چندانی بین سری‌ها مشاهده نشد.

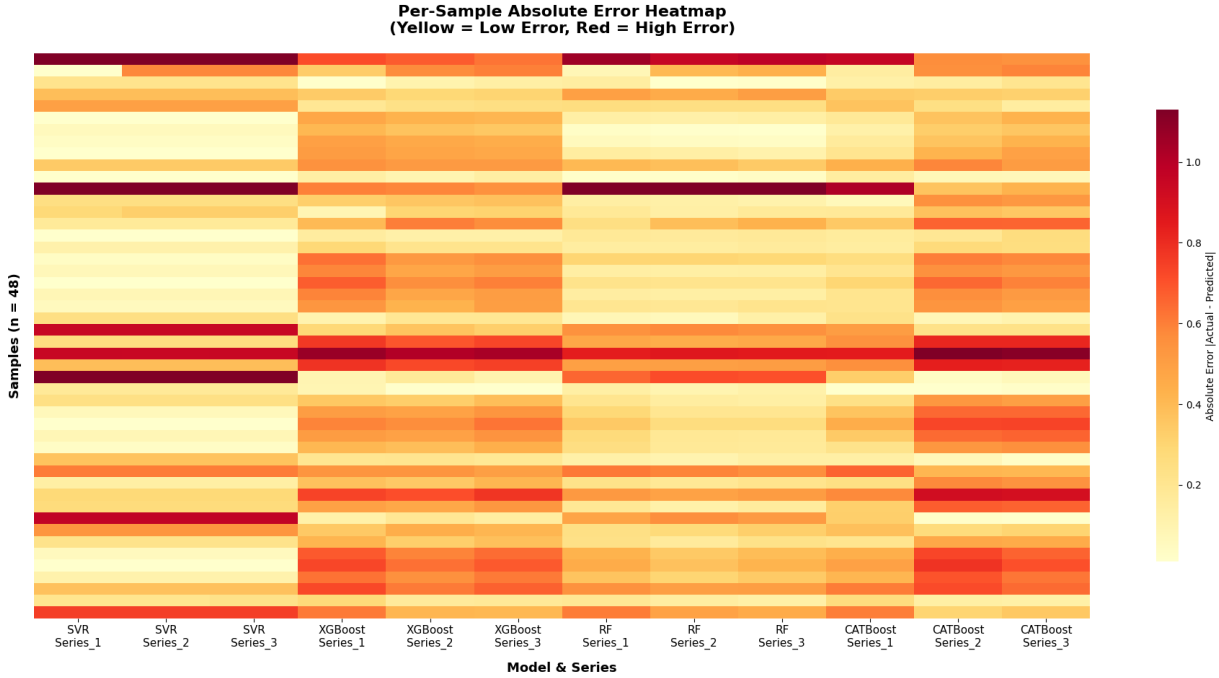
برای ارزیابی بصری عملکرد مدل‌ها و بررسی میزان انطباق مقادیر پیش‌بینی شده با داده‌های واقعی، نمودارهای پراکندگی (Scatter Plot) ترسیم شده است. این نمودارها میزان پراکندگی داده‌ها را نسبت به خط ۱:۱ (خط قرمز منقطع) نشان می‌دهند. در این شکل، هر نقطه نشان‌دهنده یک مقدار پیش‌بینی شده در برابر مقدار واقعی متناظر آن است و هر ستون، یک مدل یادگیری ماشین (SVR, XGBoost, RF, CatBoost) را نشان می‌دهد. ردیف‌ها نیز سه سری زمانی مورد استفاده (سری اصلی، Z-Score–SSA و Z-Score–Interpolation) را به ترتیب نشان می‌دهند. همچنین، ضریب تعیین (R^2) برای هر زیرنمودار ارائه شده که بیانگر قدرت مدل در توضیح واریانس داده‌ها است. شکل ۳. مقایسه مقادیر پیش‌بینی شده در مقابل مقادیر واقعی برای هر مدل و سری داده‌ها را نشان می‌دهد.



شکل ۳: نمودار اسکتر پلات داده‌های پیشبینی شده

همان‌طور که در شکل ۲ مشاهده می‌شود، نزدیک‌تر بودن نقاط به خط ۱:۱ (خط قرمز) نشان‌دهنده دقت بالاتر مدل در پیش‌بینی است. ضریب R^2 که میزان توان تبیین مدل را نشان می‌دهد، به‌وضوح بیانگر عملکرد برتر مدل Random Forest (RF) است که در سری‌های داده اصلاح‌شده (ردیف‌های دوم و سوم) به مقادیر R^2 بالاتر از ۰/۵۸ دست یافته است. این نتیجه تأیید می‌کند که پیش‌پردازش داده‌ها به شکل چشمگیری دقت پیش‌بینی مدل RF را بهبود بخشیده است. در مقابل، مدل‌های SVR و XGBoost مقادیر R^2 پایین‌تری را نشان می‌دهند (حدود ۰/۲۷ تا ۰/۴۱)، که نشان‌دهنده پراکندگی بیشتر و دقت کمتر آن‌ها در پیش‌بینی مقادیر واقعی است. به طور کلی، نمودار پراکندگی عملکرد بهتر مدل RF نسبت به سایر مدل‌ها را به‌ویژه پس از اعمال روش‌های پیش‌پردازش داده، تأیید می‌کند.

برای بررسی نحوه توزیع خطاها در هر نمونه (ماه) از داده‌های آزمون و شناسایی الگوهای خطای سیستماتیک، یک نقشه حرارتی از خطای مطلق برای هر ۴۸ نمونه ترسیم شد. این نقشه حرارتی امکان مقایسه بصری خطای هر مدل را در سری‌های مختلف داده‌ها فراهم می‌کند. همان‌طور که در شکل ۴ مشاهده می‌شود، خطاهای مطلق با طیفی از رنگ‌ها نمایش داده شده‌اند که رنگ زرد روشن نشان‌دهنده کمترین خطا و رنگ قرمز تیره نشان‌دهنده بیشترین خطای پیش‌بینی در آن نمونه خاص است.



شکل ٤: هیت مپ خطاها برای تماس سری‌ها

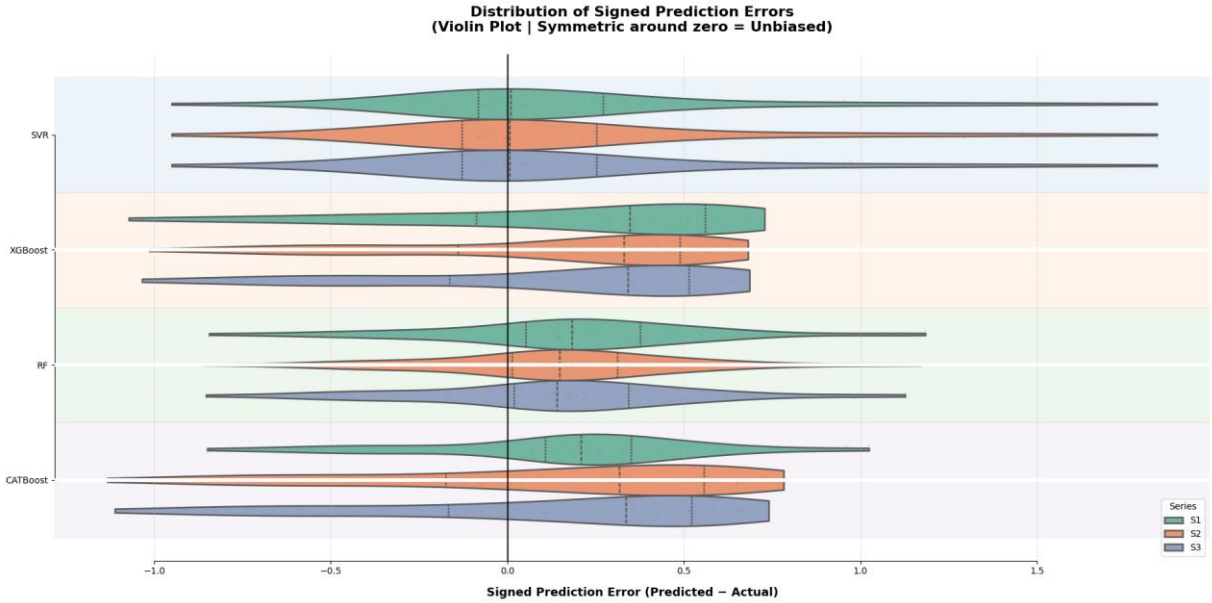
شکل ۳ الگوی خطای مدل‌ها در طول دوره آزمون را مشخص می‌کند. الگوهای افقی قرمز رنگ در برخی ردیف‌ها (نمونه‌ها) نشان‌دهنده وجود نقاط پرت یا رویدادهای حدی در داده‌های واقعی هستند که مدل‌ها در پیش‌بینی آن‌ها دچار خطای بزرگی شده‌اند. به طور خاص، الگوهای قرمز عمودی (ستون‌ها) در مدل SVR و CatBoost نشان می‌دهند که این مدل‌ها در نمونه‌های خاصی دارای خطای سیستماتیک و بزرگ هستند. همچنین، این شکل تأثیر مثبت پیش‌پردازش را در مدل‌هایی مانند RF و XGBoost نشان می‌دهد، به‌طوری که سری‌های اصلاح‌شده Z-Score–SSA و Z-Score–Interpolationدر مقایسه با سری اصلی، رنگ‌های زرد بیشتری را نشان می‌دهند که حاکی از خطای پایین‌تر نمونه به نمونه است. این تحلیل تأیید می‌کند که پیش‌پردازش داده‌ها می‌تواند پایداری پیش‌بینی مدل‌ها را در برابر نوسانات نمونه‌های تکی بهبود بخشد.

برای بررسی سوگیری (Bias) در پیش‌بینی مدل‌ها و توزیع کلی خطاها، نمودار ویولونی از خطای پیش‌بینی علامت‌دار (Predicted – Actual) ترسیم گردید. نمودار ویولونی چگالی و دامنه توزیع خطاها را نشان می‌دهد و اگر توزیع خطاها متقارن حول خط صفر باشد، مدل بدون سوگیری (Unbiased) تلقی می‌شود، به این معنی که مدل به همان اندازه که پیش‌بینی بیشتر دارد، پیش‌بینی کمتر نیز دارد. شکل ۵ توزیع خطاهای پیش‌بینی علامت‌دار (Signed Prediction Errors) به صورت نمودار ویولونی را نشان می‌دهد.

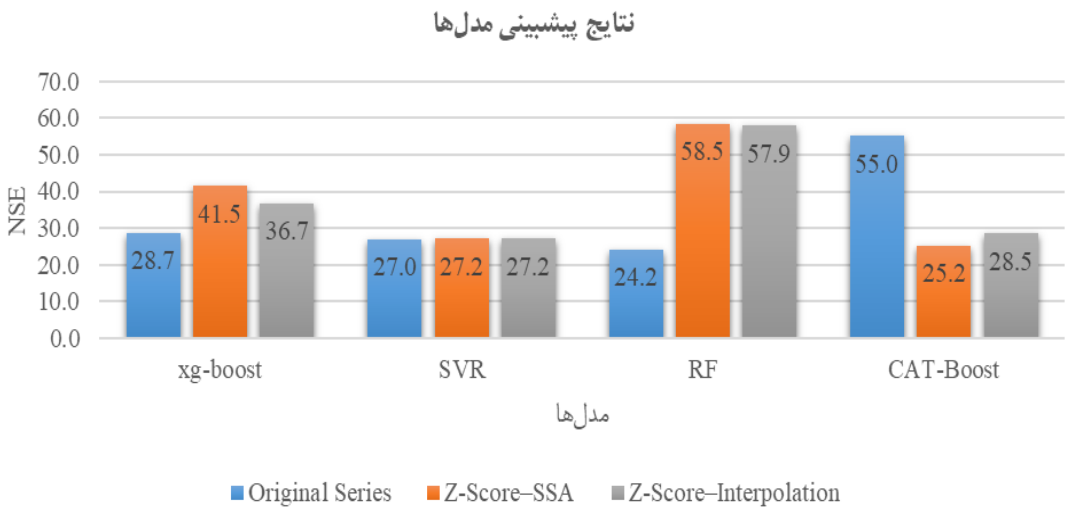
شکل ۵ توزیع خطاهای علامت‌دار را نشان می‌دهد. خط عمودی در صفر، خطای مطلوب را نشان می‌دهد. در اینجا مشاهده می‌شود که: مدل SVR دارای توزیع خطای گسترده‌تری است، به‌ویژه در سری اصلی، که حاکی از پراکندگی بالای خطاها است. مدل‌های XGBoost و RF توزیع خطای باریک‌تر و متمرکزتری را نشان می‌دهند که نشان‌دهنده دقت بالاتر و واریانس خطای کم‌تر است. در مدل RF و به ویژه در سری‌های اصلاح‌شده ، توزیع خطاهای علامت‌دار به شکل قابل توجهی حول صفر متقارن است.

مدل CatBoost دارای توزیع خطای شیفت‌یافته به سمت مقادیر مثبت است (Predicted - Actual > 0) ، به این معنی که تمایل به پیش‌بینی‌های بیشتر (Overestimation) در مقایسه با مقادیر واقعی دارد. به طور خلاصه، این شکل تأکید می‌کند که مدل‌های RF و XGBoost در سری‌های اصلاح‌شده، علاوه بر کاهش دامنه خطا، کمترین سوگیری (Bias) را نیز از خود نشان داده‌اند، که نشان‌دهنده پیش‌بینی‌های پایدار و قابل اطمینان‌تر است.

برای ارزیابی دقت مدل‌های پیش‌بینی بارش از شاخص Nash–Sutcliffe Efficiency (NSE) استفاده شده است که مقدار نزدیک به ۲ نشان‌دهنده دقت بالا و مقدار نزدیک به صفر یا منفی نشان‌دهنده دقت پایین است. در ادامه شکل ۶ شاخص NSE را نشان می‌دهد.



شکل ۵: نمودار ویلونی خطاها



شکل ٦: دقت مدل‌ها

مدل‌ها با حساسیت متفاوت نسبت به پیش‌پردازش، نشان می‌دهند که قبل از آموزش مدل‌ها باید تحلیل داده‌ها و انتخاب روش پیش‌پردازش به‌دقت انجام شود.

بحث و نتیجه‌گیری

دقت و قابلیت اطمینان در پیش‌بینی بارش، به دلیل نقش حیاتی این متغیر در چرخه هیدرولوژیک و مدیریت منابع آب، همواره یکی از چالش‌های اصلی در حوزه مهندسی آب بوده است؛ به‌ویژه در مناطق خشک و نیمه‌خشک نظیر حوضه موردمطالعه که تأثیرات تغییرات اقلیمی، نوسانات شدید بارش و رخدادهای حدی مانند سیلاب‌ها و خشکسالی‌ها، ضرورت استفاده از مدل‌های پیشرفته و داده‌های پالایش‌شده را دوچندان می‌کند. این پژوهش باهدف توسعه یک چارچوب مطمئن برای پیش‌بینی بارش ماهانه، ابتدا اعتبار داده‌های ماهواره‌ای CHIRPS را تأیید کرد و سپس تأثیر حیاتی فرایند پیش‌پردازش داده‌ها بر عملکرد مدل‌های یادگیری ماشین را به‌صورت کمی نشان داد. مهم‌ترین دستاورد، اثبات این فرضیه بود که کیفیت داده‌های ورودی، تعیین‌کننده نهایی دقت مدل‌های یادگیری ماشین است؛ به طوری که به‌کارگیری اقدامات پیش‌پردازشی مانند حذف نقاط پرت با Z-Score و بازسازی با روش‌هایی نظیر SSA و Interpolation، ساختار آماری سری زمانی را به شکل مطلوبی تثبیت نمود. نتایج نشان داد که این بهینه‌سازی داده، منجر به جهشی چشمگیر در عملکرد مدل Random Forest (RF) شد، به‌گونه‌ای که در مقایسه با سایر الگوریتم‌های مورد بررسی (XGBoost، SVR و CatBoost)، این مدل

با داده‌های پیش پردازش‌شده توانست بالاترین ضریب تعیین (R2) و کمترین خطای میانگین مربعات (MSE) را به ثبت برساند و کمترین سوگیری سیستماتیک را در پیش‌بینی‌ها از خود نشان دهد. این نوع پژوهش‌ها در حوزه مهندسی آب اهمیت مضاعفی دارد؛ چرا که ارائه یک روش‌شناسی معتبر و اثبات‌شده برای پیش‌بینی بارش، مستقیماً به کاهش عدم قطعیت‌های هیدرولوژیکی منجر می‌شود و به مدیران منابع آب این امکان را می‌دهد که با اطمینان بیشتری نسبت به برنامه‌ریزی عملیاتی سدها و مخازن، هشدار سیلاب و مدیریت بهینه تخصیص آب برای مصارف کشاورزی اقدام نمایند. به طور خلاصه، پژوهش حاضر، مدل Random Forest را از بین الگوریتم‌های CATBoost, SVR, XG Boost و Random Forest به‌عنوان برترین الگوریتم برای پیش‌بینی بارش، مشروط به اعمال روش‌های پیش‌پردازش بر داده‌های ورودی، معرفی می‌کند و بر لزوم ادغام مراحل کنترل کیفی داده در هر زنجیره مدل‌سازی تأکید می‌ورزد. در نهایت، برای تکمیل این پژوهش و افزایش گستره کاربرد آن، پیشنهاداتی برای مطالعات آتی ارائه می‌گردد که شامل: اولاً، ارزیابی سایر تکنیک‌های پیشرفته کاهش نویز و تحلیل سری زمانی مانند استفاده از تبدیل ویولت (Wavelet) و یا روش‌های EEMD/CEEMDAN برای مقایسه با SSA؛ دوماً، توسعه مدل‌های پیش‌بینی چندمتغیره با ادغام عوامل اقلیمی دیگر مانند دما، تبخیر و رطوبت؛ سوماً، مقایسه روش بهینه‌شده فعلی با مدل‌های یادگیری عمیق (Deep Learning) مانند شبکه‌های LSTM و GRU برای بررسی پتانسیل آن‌ها در دریافت مستقیم داده‌های پالایش نشده؛ و چهارماً، تکرار این روش‌شناسی در حوضه‌های با ویژگی‌های جغرافیایی و اقلیمی متفاوت برای تأیید قابلیت تعمیم و پایداری چارچوب پیشنهادی است.

REFERENCE

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons. https://doi.org/10.1002/bimj.4710370219

Blazquez-Garcia, A., Conde, A., Mori, U., & Lozano, J. A. (2020). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3), 1-33. https://doi.org/10.1145/3444690

Dinku, T., et al. (2018). *Validation of CHIRPS satellite rainfall estimates over eastern Africa.* Journal of Hydrometeorology, 19(4), 791–810. https://doi.org/10.1002/qj.3244

Dotse, S. Q., Larbi, I., Limantol, A. M., & De Silva, L. C. (2024). A review of the application of hybrid machine learning models to improve rainfall prediction. Modeling Earth Systems and Environment, 10(1), 19-44. https://doi.org/10.1007/s40808-023-01835-x

Funk, C., et al. (2015). *The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes.* Scientific Data, 2, 150066. https://doi.org/10.1038/sdata.2015.66

Golyandina, N., & Korobeynikov, A. (2014). *Basic singular spectrum analysis and forecasting with R*. Computational Statistics & Data Analysis, 71, 934–954. https://doi.org/10.48550/arXiv.1206.6910

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC. https://doi.org/10.1201/9780367801687

Hassani, H. (2021). *Singular spectrum analysis: Methodology and comparison*. Journal of Data Science, 5(2), 239–257. https:/doi.org/10.6339/JDS.2007.05(2).396

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

Jia, X., Xun, P., Peng, W., Zhao, B., Li, H., & Shen, C. (2025). Deep anomaly detection for time series: A survey. Computer Science Review, 58, 100787. DOI: 10.1016/j.cosrev.2025.100787

Leys, C., et al. (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. https://doi.org/10.1016/j.jesp.2013.03.013

Kazijevs, M., & Samad, M. D. (2023). Deep imputation of missing values in time series health data: A review with benchmarking. *Journal of Biomedical Informatics*, *144*, 104440. https://doi.org/https://doi.org/10.1016/j.jbi.2023.104440

Park, J., Müller, J., Arora, B. *et al.* Long-term missing value imputation for time series data using deep neural networks. *Neural Comput & Applic* 35, 9071–9091 (2023). https://doi.org/10.1007/s00521-022-08165-6.

Alejo-Sanchez, L. E., Márquez-Grajales, A., Salas-Martínez, F., Franco-Arcega, A., López-Morales, V., Acevedo-Sandoval, O. A., González-Ramírez, C. A., & Villegas-Vega, R. (2025). Missing data imputation of climate time series: A review. *MethodsX*, *15*, 103455. https://doi.org/https://doi.org/10.1016/j.mex.2025.103455

Shiru, M. S., Shahid, S., Dewan, A., Chung, E. S., Alias, N., Ahmed, K., & Hassan, Q. K. (2020). Projection of Meteorological Droughts in Nigeria during Growing Seasons under Climate Change Scenarios. Scientific Reports, 10, Article No. 10107. https://doi.org/10.1038/s41598-020-67146-8

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. doi:10.14778/3538598.3538602

Zhang, W., Furtado, K., Wu, P., Zhou, T., Chadwick, R., Marzin, C., Rostron, J., & Sexton, D. (2021). Increasing precipitation variability on daily-to-multiyear time scales in a warmer world. Science Advances, 7(31), eabf8021. https://doi.org/doi:10.1126/sciadv.abf8021