Watershed Engineering and Management
DOI: 10.22092/ijwmse.2025.369325.2115

Volume 17, Issue 4, 2026, Pages 424-447

Applying wavelet-based machine learning and deep learning algorithms for streamflow prediction of the Kurkursar River

Edris Merufinia[1], Ahmad Sharafati[2*], Hirad Abghari[3] and Yousef Hassanzadeh[4]

[1] Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Associate Professor, Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3] Associate Professor, Department of Range and Watershed Management, Urmia University, Urmia, Iran
[4] Professor, Department of Water Engineering, Center of Excellence in Hydroinformatics, Faculty of Civil Engineering, University of Tabriz, and Farazab Co. (Consulting Engineers), Research and Writing Capacity Enhancement Affairs, Tabriz, Iran

Received: 01 May 2025 Accepted:15 September 2025

Extended abstract

Introduction

Accurate streamflow prediction is essential for water resources management and flood control. Due to the complex and nonlinear behavior of streamflow, traditional models are often inadequate. Machine learning and deep learning algorithms offer more robust solutions; however, their accuracy can be affected by sudden climatic fluctuations. Consequently, employing hybrid methods is necessary to improve prediction accuracy. The literature review reveals that, despite the high capabilities of machine learning models, a research gap still exists in managing multi-scale fluctuations in streamflow data. This underscores the necessity of using hybrid approaches to enhance prediction accuracy. The innovation of this study is a hybrid framework that simultaneously models both long-term patterns and short-term fluctuations by integrating wavelet analysis, used to decompose the streamflow signal, with a powerful deep learning model.

Materials and methods

In this study, to predict the streamflow of the Kurkursar River in Nowshahr, hydrological data including daily precipitation and river discharge over a 20-year period at a daily resolution were utilized. The input variables included daily precipitation (P_t) and streamflow with time lags of one, two, and three days (Q_{t-1}, Q_{t-2}, Q_{t-3}). Before the modeling process, data preprocessing was performed, which included reconstructing missing data, removing anomalous data (outliers), and normalizing the values to improve data quality and enhance their reliability in hydrological analyses. The hydrological data from the watershed were divided into three subsets: training (70%), validation (15%), and testing (15%). Four streamflow prediction scenarios were selected based on Pearson correlation coefficient analysis to identify sensitive variables and determine the model inputs. The river streamflow modeling process was carried out using two algorithms: Random Forest (RF) and the deep learning Long Short-Term Memory (LSTM) recurrent neural network. Furthermore, to enhance the accuracy and improve the generalizability of the models, various wavelet transform methods, including Daubechies 4 (Db4), Haar, and Mexican Hat wavelets, were used to extract multi-scale features and combine them with the input data for the RF and LSTM models. This hybrid approach facilitated the identification of complex spatio-temporal patterns in the hydrological time series. After the final evaluation of the prediction models' performance, the Daubechies 4 (Db4) wavelet transform was employed to optimize their coefficients and structural parameters. Performance evaluation metrics, including the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Percent Bias (PBIAS), Mean Absolute Percentage Error (MAPE), and Kling-Gupta Efficiency (KGE), were used to assess the accuracy of the models' predictions. Ultimately, the optimal models were selected based on a comparative analysis of these quantitative criteria. Additionally, for data analysis and visual presentation of the results, various plots were used, including scatter plots, time series of observed and predicted data, and error distributions such as error histograms, normal density curves, cumulative distribution functions of errors, and quantile-quantile (Q-Q) plots.

* Corresponding author: asharafati@srbiau.ac.ir

DOI: 10.22092/ijwmse.2025.369325.2115

Results and discussion

The results showed that in streamflow prediction, previous time steps (different lags) were the most important variables for predicting all subsequent horizons. The final results regarding the model scenarios indicated that the first scenario (S1), which only used the precipitation variable, was the weakest performer in all cases. Furthermore, the sixth scenario (S6), which utilized all available variables (P_t,Q_{t-1},Q_{t-2},Q_{t-3}), had the best performance in the training and testing phases for both standalone and hybrid models. The research findings indicated that the hybrid Random Forest-Wavelet (RF-Wavelet) model had the best performance in both the training (R^2=0.907, RMSE=0.0192) and testing (R^2=0.942, RMSE=0.0106) phases. Additionally, the standalone Long Short-Term Memory (LSTM) deep learning model had the weakest performance in the training (R^2=0.499, RMSE=1.6) and testing (R^2=0.579, RMSE=1.149) phases. The findings also showed that the Daubechies 4 wavelet , when combined with the Random Forest model, was able to reduce the error of the standalone RF model by approximately 55%. Additionally, the wavelet, when combined with the LSTM model, was able to increase the prediction accuracy by approximately 39%. Furthermore, a comparison of the wavelet-hybrid models showed that the RF-Wavelet model reduced the error by approximately 23% compared to the hybrid LSTM-Wavelet model.

Conclusion

In this research, various wavelet transform models, including Daubechies 4, Haar, and Mexican Hat, were utilized for integration with RF and LSTM algorithms. Quantitative and qualitative analyses showed that the Daubechies 4 wavelet transform had significant superiority in improving streamflow prediction accuracy compared to other wavelet types within both RF and LSTM model frameworks. Therefore, this type of wavelet transform was selected and used as the primary basis for integration with these two prediction models. Examination of the error distribution pattern in the training data indicates a major concentration of error values in regions adjacent to zero. The distribution of errors was observed to be approximately symmetrical and showed considerable consistency with a normal distribution. This pattern signifies the model's satisfactory accuracy in the training and data-fitting process. Ultimately, the present study focused on the development of data-driven models to determine the optimal combination of predictor variables for modeling and predicting river streamflow. This research demonstrated that integrating the Daubechies 4 wavelet transform with the Random Forest (RF) model served as the optimal and superior approach for predicting hydrological streamflow in the present case study. The aforementioned hybrid model, in addition to significantly enhancing performance compared to standalone models by reducing prediction error by up to 55%, showed notable superiority over complex deep learning models, including LSTM and its associated hybrid combinations. This achievement highlights the importance of extracting multi-scale time-frequency features using the wavelet transform and emphasizes its pivotal role in improving the accuracy and generalizability of hydrological streamflow predictions, even in comparison to advanced architectures of deep temporal models.

Keywords: Haar wavelet, Mexican Hat wavelet, Pearson correlation coefficient, Random Forest, Streamflow prediction

Cite this article: Merufinia, E., Sharafati, A., Abghari, H., Hassanzadeh, Y., 2026. Applying wavelet-based machine learning and deep learning algorithms for streamflow prediction of the Kurkursar River. Water. Eng. Manag. 17(4), 424-447.

مهندسی و مدیریت آبخیز
جلد ۱۷، شماره ۴، ۱۴۰۴، صفحات ۴۲۴-۴۴۷

Watershed Engineering and Management
Volume 17, Issue 4, 2026, Pages 424-447
DOI:10.22092/ijwmse.2025.369325.2115

کاربست الگوریتم‌های یادگیری ماشین و عمیق موجکی برای پیش‌بینی جریان رودخانه کورکورسر نوشهر

ادریس معروفی نیا[۱]، احمد شرافتی[۲*]، هیراد عبقری[۳] و یوسف حسن زاده[۴]

[۱] دانشجوی دکتری، گروه مدیریت ساخت و آب، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

[۲] دانشیار، گروه مدیریت ساخت و آب، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

[۳] دانشیار، گروه مرتع و آبخیزداری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران

[۴] استاد، گروه مهندسی آب، قطب علمی هیدروانفورماتیک، دانشکده مهندسی عمران، دانشگاه تبریز و شرکت فرازآب (مهندسین مشاور) امور ارتقای توانمندی تحقیقات و تألیفات، تبریز، ایران

تاریخ دریافت: ۱۴۰۴/۰۲/۱۱ تاریخ پذیرش: ۱۴۰۴/۰۶/۲۴

چکیده مبسوط

مقدمه

پیش‌بینی دقیق جریان رودخانه برای مدیریت آب و کنترل سیلاب ضروری است. به‌دلیل رفتار پیچیده و غیرخطی جریان، مدل‌های سنتی کارایی لازم را ندارند. الگوریتم‌های یادگیری ماشین و یادگیری عمیق راه‌حل‌های پیشرفته‌تری ارائه می دهند، اما دقت آن‌ها تحت تأثیر نوسانات ناگهانی اقلیمی قرار می‌گیرد. از این رو، برای بهبود پیش‌بینی، به‌کارگیری روش‌های ترکیبی ضروری است. مرور پیشینه نشان می‌دهد که با وجود توانمندی بالای مدل‌های یادگیری ماشین، همچنان شکاف تحقیقاتی در زمینه مدیریت نوسانات چندمقیاسی در داده‌های جریان رودخانه وجود دارد. این موضوع، ضرورت به‌کارگیری روش‌های ترکیبی را برای افزایش دقت پیش‌بینی آشکار می‌سازد. نوآوری این پژوهش در ارائه یک چارچوب ترکیبی است که با ادغام تحلیل موجک برای تجزیه سیگنال جریان و یک مدل یادگیری عمیق قدرتمند، به طور همزمان الگوهای بلندمدت و نوسانات کوتاه‌مدت را مدل‌سازی می‌کند.

مواد و روش‌ها

در این پژوهش، به‌منظور پیش‌بینی جریان رودخانه کورکورسر نوشهر، از داده‌های هیدرولوژیکی شامل بارش روزانه و دبی جریان رودخانه در بازه زمانی ۲۰ ساله و سطح روزانه استفاده شد. متغیرهای ورودی شامل بارش روزانه (P_t) و دبی جریان با تاخیرهای زمانی یک، دو و سه روزه (Qt-1, Qt-2, Qt-3) بودند. پیش از انجام فرایند مدل‌سازی، پیش‌پردازش داده‌ها شامل بازسازی داده‌های گمشده، حذف داده‌های پرت (خارج از محدوده نرمال) و نرمال‌سازی مقادیر به‌منظور بهبود کیفیت داده‌ها و افزایش قابلیت اعتماد آن‌ها در تحلیل‌های هیدرولوژیکی انجام پذیرفت. در این پژوهش، داده‌های حاصل از پایش جریان هیدرولوژیکی حوزه آبخیز به سه زیرمجموعه آموزش (۷۰ درصد)، اعتبارسنجی (۱۵ درصد) و آزمون (۱۵ درصد) تفکیک شد. چهار سناریوی پیش‌بینی جریان بر اساس تحلیل ضریب همبستگی پیرسون به‌منظور شناسایی متغیرهای حساس و تعیین ورودی‌های مدل‌ها انتخاب شدند. فرایند مدل‌سازی جریان رودخانه با بهره‌گیری از دو الگوریتم یادگیری ماشین شامل جنگل تصادفی (RF) و شبکه عصبی بازگشتی یادگیری عمیق حافظه طولانی کوتاه مدت (LSTM) انجام شد. همچنین، به‌منظور افزایش دقت و بهبود قابلیت تعمیم‌پذیری مدل‌ها، روش‌های تبدیل موجکی

* مسئول مکاتبات: asharafati@srbiau.ac.ir

متنوعی ازجمله موجک دابشیز نوع ۴ (Dabchiz 4)، موجک هار (Haar) و موجک کلاه مکزیکی (Mexican Hat) جهت استخراج ویژگی‌های چندمقیاسی و ترکیب آن‌ها با داده‌های ورودی مدل‌های RF و LSTM مورد استفاده قرار گرفتند. این رویکرد ترکیبی امکان شناسایی الگوهای زمانی-فضایی پیچیده در سری‌های زمانی هیدرولوژیکی را تسهیل نمود. پس از اتمام فرایند ارزیابی نهایی عملکرد مدل‌های پیش‌بینی، به‌منظور بهینه‌سازی ضرایب و فراسنجه‌های ساختاری آن‌ها، تبدیل موجک دابشیز نوع ۴ (Db4) به کار گرفته شد. شاخص‌های ارزیابی عملکرد شامل ضریب تعیین (R^2)، میانگین قدر مطلق خطا (MAE)، ریشه میانگین مربعات خطا (RMSE)، درصد نااریبی (PBIAS)، میانگین درصد قدر مطلق خطا (MAPE) و ضریب بهره‌وری کلینگ-گوپتا (KGE) جهت سنجش دقیق بودن پیش‌بینی‌های مدل‌ها به‌کاربرده شدند. درنهایت، انتخاب بهینه‌ترین مدل‌ها بر اساس تحلیل مقایسه‌ای این معیارهای کمی انجام گرفت. همچنین، به‌منظور تجزیه و تحلیل داده‌ها و ارائه بصری نتایج، از نمودارهای پراکندگی، سری‌های زمانی داده‌های مشاهده‌شده و پیش‌بینی شده، توزیع خطاها ازجمله هیستوگرام خطا، منحنی چگالی نرمال، تابع توزیع تجمعی خطا و نمودارهای چندک-چندک استفاده شد.

نتایج و بحث

نتایج نشان داد که در پیش‌بینی جریان رودخانه، گام‌های قبلی (تأخیرهای مختلف) مهم‌ترین متغیر در پیش‌بینی جریان تمام افق‌های بعدی است. نتایج نهایی در خصوص سناریوهای مدل نشان داد که سناریوی اول (S_1) که فقط از متغیر بارش (Q_t) استفاده می‌نماید، در تمامی حالات به‌عنوان ضعیف‌ترین عملکرد در پیش‌بینی انتخاب شد. همچنین سناریوی ششم (S6) که از تمامی متغیرهای موجود بهره می‌برد (P_t، Q_{t-1}، Q_{t-2}، Q_{t-3}) دارای بهترین عملکرد در مرحله آموزش و آزمون برای مدل‌های منفرد و ترکیبی بوده است. یافته‌های پژوهش نشان داد که مدل ترکیبی جنگل تصادفی-موجک (RF-Wavelet) در دو حالت آموزش (R^2=0.907، RMSE=0.0192) و آزمون (R^2=0.942، RMSE=0.0106) دارای بهترین عملکرد بوده است. همچنین مدل منفرد یادگیری عمیق حافظه طولانی کوتاه مدت (LSTM) دارای ضعیف‌ترین عملکرد در مرحله آموزش (R^2=0.499، RMSE=1.6) و آزمون (R^2=0.579، RMSE=1.149) بوده است. همچنین یافته‌ها نشان داد که مدل موجک دابشیز ۴ توانسته است با ترکیب با مدل جنگل تصادفی حدود ۵۵ درصد خطای مدل منفرد RF را کاهش دهد. همچنین مدل موجک در ترکیب با مدل LSTM توانسته است حدود ۳۹ دقت پیش‌بینی را افزایش دهد. همچنین مقایسه مدل‌های ترکیبی ترکیب شده با موجک نشان داد که مدل RF-Wavelet حدود ۲۳ درصد نسبت به مدل ترکیبی LSTM-Wavelet خطای مدل را کاهش دهد.

نتیجه‌گیری

در این پژوهش، از مدل‌های مختلف تبدیل موجک شامل موجک دابشیز ۴، موجک هار و موجک کلاه مکزیکی برای تلفیق با الگوریتم‌های یادگیری ماشین RF و شبکه‌های LSTM بهره‌برداری شده است. تحلیل‌های کمی و کیفی نشان داد که تبدیل موجک دابشیز ۴ در بهبود دقت پیش‌بینی جریان رودخانه نسبت به سایر انواع موجک‌ها در هر دو چارچوب مدل RF و LSTM برتری معنی‌داری داشته است. ازاین‌رو، این نوع تبدیل موجک به‌عنوان مبنای اصلی برای ادغام با این دو مدل پیش‌بینی انتخاب و مورد استفاده قرارگرفته است. بررسی الگوی توزیع خطا در داده‌های آموزش نشان‌دهنده تمرکز عمده مقادیر خطا در نواحی مجاور صفر است. توزیع خطاها تقریباً به‌صورت تقارن دوطرفه مشاهده‌شده و سازگاری قابل‌توجهی با توزیع نرمال از خود نشان داده است. این الگو بیانگر دقت مطلوب مدل در فرایند آموزش و برازش داده‌ها است. درنهایت، مطالعه مذکور به توسعه مدل‌های مبتنی برداده به‌منظور تعیین بهینه‌ترین ترکیب متغیرهای پیش‌بینی کننده برای مدل‌سازی و پیش‌بینی جریان رودخانه متمرکز شد. این پژوهش به‌طور دقیقی نشان داد که ادغام تبدیل موجک دابشیز ۴ با مدل جنگل تصادفی RF به‌عنوان رویکرد بهینه و برتر در پیش‌بینی جریان‌های هیدرولوژیکی در مورد مطالعه حاضر عمل کرده است. مدل ترکیبی مذکور، علاوه بر ارتقاء قابل‌توجه عملکرد نسبت به مدل‌های تک‌محوری، با کاهش خطای پیش‌بینی تا حدود ۵۵ درصد، برتری شاخصی نسبت به مدل‌های پیچیده یادگیری عمیق، ازجمله LSTM

و ترکیبات ترکیبی مرتبط نشان داده است. این دستاورد، اهمیت استخراج ویژگی‌های چندمقیاسی و زمانی-فرکانسی با استفاده از تبدیل موجک را برجسته می‌سازد و بر نقش محوری آن در بهبود دقت و قابلیت تعمیم پیش‌بینی جریان‌های هیدرولوژیکی، حتی در مقایسه با معماری‌های پیشرفته مدل‌های زمانی تأکید می‌کند.

واژه‌های کلیدی: پیش‌بینی رودخانه، جنگل تصادفی، ضریب پیرسون، موجک کلاه مکزیکی، موجک هار

مقدمه

جریان رودخانه تحت تأثیر عوامل متعددی نظیر بارندگی، دما، تبخیر و تعرق، و همچنین ویژگی‌های کاربری اراضی و ویژگی‌های حوزه آبخیز قرار دارد (Adnan et al., 2019). پیش‌بینی دقیق و قابل‌اعتماد پویای هیدرولوژیکی جریان رودخانه از اهمیت حیاتی در طراحی مهندسی سازه‌های هیدرولیکی، برنامه‌ریزی تخصیص منابع آب، بهینه‌سازی بهره‌برداری و مدیریت جامع منابع آب در حوزه‌های آبخیز برخوردار است (Roy and Singh, 2020).

در این راستا، مدل‌های پیش‌بینی جریان رودخانه عموماً به دو دسته کلی مدل‌های مبتنی بر اصول فیزیکی و مدل‌های داده‌محور تقسیم‌بندی می‌شوند (Solomatine and Ostfeld, 2008). مدل‌های مفهومی نمونه‌هایی از رویکردهای مبتنی بر اصول بنیادی فیزیکی به شمار می‌روند، درحالی‌که مدل‌های مبتنی بر روش‌های یادگیری ماشین و مدل‌های آماری در زمره مدل‌های تجربی و داده‌محور قرار دارند (Wu et al., 2010).

در چارچوب مدل‌سازی فیزیکی، فرایند واسنجی نیازمند پذیرش و اعمال فرضیه‌های پایه‌ای و تعاریف دقیق فراسنجه‌های فرایندی است تا صحت و اعتبار مدل تضمین شود. مدل‌های فیزیکی به‌منظور شبیه‌سازی فرایندهای هیدرولوژیکی نیازمند دسترسی به مجموعه‌های داده‌ای گسترده، دقیق و پایدار از متغیرهای هیدرولوژیکی هستند که در اغلب حوزه‌های آبخیز، به‌دلیل محدودیت‌های فنی، مالی و زیست‌محیطی، جمع‌آوری این داده‌ها با دشواری‌های فراوان مواجه می‌شود.

این محدودیت‌ها منجر به کاهش قابلیت کاربردی و دقت مدل‌های فیزیکی در شرایط عملی و محیط‌های واقع‌گرایانه می‌شود (Ragettli et al., 2014). مدل‌های یادگیری ماشین (به‌عنوان مدل‌های جعبه سیاه) یک رویکرد جایگزین برای مدل‌سازی فرایندهای پیچیده و غیرخطی محسوب می‌شوند. بنابراین ویژگی اصلی این نوع مدل‌ها، استخراج یک رابطه پایدار بین متغیرهای ورودی و خروجی تاریخی بدون در نظر گرفتن مستقیم فیزیک مسأله است که به‌عنوان یک مزیت بزرگ تلقی می‌شود (Kumar et al., 2016).

بدیهی است که یکی از کاربردهای کلیدی الگوریتم‌های یادگیری ماشین در حوزه مدیریت جامع منابع آب، مدل‌سازی و پیش‌بینی جریان هیدرولوژیکی رودخانه‌ها است، که این امر نقش محوری در برنامه‌ریزی تأمین منابع آب، کنترل فرایندهای سیلابی و بهینه‌سازی سامانه‌های آبیاری هوشمند ایفا می‌نماید.

این سامانه‌های داده‌محور با تحلیل مجموعه داده‌های تاریخی ازجمله فراسنجه‌های اقلیمی مانند بارندگی، دما، رطوبت خاک، تغذیه زیرسطحی و دبی رودخانه، قابلیت استخراج و شناسایی الگوهای نهفته و پیچیده فرایندهای هیدرولوژیکی را دارا هستند. الگوریتم‌های پیشرفته‌ای همچون شبکه‌های عصبی مصنوعی[1]، ماشین بردار پشتیبان[2] و روش‌های یادگیری عمیق[3] ضمن ارائه دقت پیش‌بینی بالا و تعمیم‌پذیری مناسب، به کاهش وابستگی به داده‌های ورودی دقیق و گسترده کمک نموده و به‌تبع آن ضمن کاهش نیازمندی های محاسباتی، باعث افزایش کارایی و اثربخشی عملیاتی مدل‌های هیدرولوژیکی نسبت به مدل‌های فیزیکی-ریاضی سنتی می‌شوند.

مدل‌های یادگیری عمیق توانایی قابل‌توجهی در کاهش عدم قطعیت‌های مرتبط با داده‌های هیدرولوژیکی دارند (Lin et al., 2021). از مزایای عمده این رویکردها، قابلیت استخراج ویژگی‌های پیچیده و

[1] Artificial Neural Networks (ANN)

[2] Support Vector Machines (SVM)

[3] Deep Learning (DL)

چندبعدی از داده‌های هیدرولوژیکی به‌صورت خودکار است که موجب کاهش قابل‌توجه نیاز به مداخله انسانی در انتخاب و مهندسی ویژگی‌ها می‌شود (Xu et al., 2022). به‌خصوص، استفاده از یادگیری عمیق در مدل‌سازی وابستگی‌های زمانی بلندمدت در داده‌های جریان رودخانه، عملکرد متمایزی را به نمایش می گذارد. مدل‌های مبتنی بر شبکه‌های LSTM با حفظ و گسترش اطلاعات مرتبط با رویدادهای گذشته، قادر به شناسایی الگوهای غیرخطی پیچیده و تحلیل نوسانات زمانی داده‌ها با دقت و کارایی بالا هستند.

قابلیت مذکور در پیش‌بینی جریان رودخانه در حوزه‌های آبخیز با شرایط اقلیمی پیچیده از اهمیت بسزایی برخوردار است (Khosravi et al. , 2022؛ Latif and Ahmed, 2023). در این راستا، Kratzert et al., (2019) با به‌کارگیری مدل یادگیری عمیق LSTM، جریان رودخانه را مورد پیش‌بینی قرار دادند. نتایج پژوهش مذکور بیانگر برتری قابل‌توجه این مدل نسبت به روش‌های سنتی هیدرولوژیکی در دقت پیش‌بینی رواناب است.

Obeta et al., (2020) مدل‌های یادگیری عمیق LSTM و GRU را در پیش‌بینی هیدرولوژیک مورد بررسی قرار دادند. نتایج نهایی نشان داد که GRU با پیچیدگی محاسباتی کمتر، عملکردی مشابه LSTM ارائه می‌دهد. Le et al., (2021) به مقایسه روش‌های یادگیری عمیق (FFNN، CNN، GRU و LSTM) برای پیش‌بینی جریان رودخانه سرخ در شمال ویتنام پرداختند. نتایج نهایی نشان داد که مدل‌های مبتنی بر LSTM می‌توانند حتی در حضور سدها و مخازن بالادست به پیش‌بینی‌های چشمگیری دست یابند. Rahimzad et al., (2021) به مقایسه عملکرد یک مدل یادگیری عمیق مبتنی بر LSTM در مقابل الگوریتم‌های یادگیری ماشین سنتی (رگرسیون خطی، پرسپترون چندلایه و ماشین بردار پشتیبان) برای پیش‌بینی جریان رودخانه پرداختند. نتایج نشان داد که شبکه LSTM در پیش‌بینی جریان روزانه با کمترین مقادیر از سایر مدل‌ها بهتر عمل نموده و تحت همه سناریوها این یافته‌ها نشان داد که LSTM یک تکنیک داده‌محور قوی برای توصیف رفتارهای سری زمانی در برنامه‌های مدل‌سازی هیدرولوژیک است.

جنگل تصادفی با سازوکار ترکیب چندین درخت تصمیم، از بروز مشکل بیش‌برازش جلوگیری کرده و دقت پیش‌بینی را به میزان قابل توجهی افزایش می دهد. این مدل به‌ویژه در شرایطی که داده‌های ورودی دارای اغتشاش یا مقادیر پرت باشند، عملکرد بسیار پایداری از خود نشان داده و با بهینه‌سازی تدریجی مدل، توانایی بالایی در شناسایی الگوهای پیچیده جریان رودخانه ارائه می‌نماید (Han et al. 2002؛ Al-Juboori 2019).

همچنین Sahour et al., (2021) با استفاده از مدل‌های جنگل تصادفی و XGBoost به پیش‌بینی جریان رودخانه در کالیفرنیا پرداختند. نتایج حاکی از برتری مدل XGBoost بر RF بوده و مدل‌ها در مدل‌سازی جریان‌های نرمال و کم در مقایسه با جریان‌های بسیار زیاد، عملکرد بهتری دارند.

در تحقیق دیگری Li et al., (2019) به مقایسه مدل‌های یادگیری شدید (ELM)، مدل جنگل تصادفی (RF)، شبکه عصبی پس‌انتشار (BPNN) و ماشین بردار پشتیبان رگرسیونی (SVR) پرداختند. نتایج نشان داد که مدل RF در پیش‌بینی جریان‌های اوج، نسبت به مدل‌های دیگر برتری داشته و مدل ELM بالاترین دقت پیش‌بینی جریان‌های کم را نشان داد. Pham et al., (2020) به ارزیابی مدل‌های RF در برابر مدل‌های رگرسیون خطی چندگانه (MLR) در پیش‌بینی کوتاه‌مدت جریان روزانه در حوزه‌های آبخیز ناشی از بارش و ذوب برف پرداختند. نتایج نشان داد که RF در مقایسه با MLR در حوزه‌های آبخیز ناشی از ذوب برف در مقایسه با حوزه‌های باران‌محور عملکرد بهتری دارد. هدف اصلی این پژوهش، توسعه و ارزیابی یک چارچوب پیشرفته و کارآمد برای پیش‌بینی دقیق جریان رودخانه کورکورسر نوشهر است.

این چارچوب با بهره‌گیری از هم‌افزایی تبدیل موجک دابشیز ۴ و الگوریتم‌های یادگیری ماشین جنگل تصادفی (RF) و یادگیری عمیق حافظه طولانی کوتاه‌مدت (LSTM) طراحی شده تا بر محدودیت‌های مدل‌های سنتی و حتی مدل‌های یادگیری ماشین عمیق معمول در مواجهه با پیچیدگی‌های غیرخطی، ناایستایی‌ها و نوسانات چند مقیاسی موجود در سری‌های زمانی جریان رودخانه غلبه کند. لذا دستیابی

به یک مدل منعطف و پویا که توانایی بالاتری در مدل سازی و پیش‌بینی پدیده‌های پیچیده داشته باشد و به تصمیم‌گیری‌های آگاهانه‌تر در مدیریت منابع آب و کنترل سیلاب کمک کند و بتواند شکاف‌های مطالعاتی را پوشش دهد جزء اهداف دیگر این پژوهش است. نوآوری اصلی این پژوهش در ارائه یک چارچوب ترکیبی با رویکردی نوین برای ادغام تبدیل موجک با مدل‌های یادگیری ماشین است.

بر خلاف بسیاری از مطالعات که از موجک صرفاً به عنوان ابزار پیش‌پردازش برای کاهش اغتشاش استفاده می‌کنند، موجک دابشیز ۴ به‌عنوان ابزاری برای بهینه‌سازی مستقیم ضرایب و ابرفراسنجه‌های درونی مدل‌های RF و LSTM استفاده شد. این رویکرد، یک سطح عمیق‌تر از ترکیبی‌سازی را معرفی نمود.

این بهینه‌سازی عمیق به مدل‌های ترکیبی این امکان را فراهم ساخت تا ویژگی‌های محلی (نوسانات سریع) و سراسری (روندهای کلی) داده‌ها را به‌طور همزمان و با دقت بالاتری شناسایی نماید. همچنین این پژوهش به‌صورت نظام‌مند عملکرد این رویکرد ترکیبی پیشرفته را هم در مدل‌های یادگیری ماشین کلاسیک (RF) و هم در مدل‌های یادگیری عمیق (LSTM) ارزیابی نموده تا کارایی آن در معماری‌های مختلف سنجیده شود.

مواد و روش‌ها

حوزه آبخیز کورکورسر در استان مازندران و شهرستان نوشهر واقع شده است. این حوضه در مجاورت چهار رودخانه واقع‌شده که ازجمله آن‌ها می‌توان به رودخانه کورکورسر در محدوده شهرستان نوشهر و رودخانه خیررود در فاصله شش کیلومتری شرق آن اشاره کرد. از منظر تقسیمات هیدرولوژیکی، این حوزه به‌عنوان زیرحوضه دریای خزر شناخته می‌شود. ارتفاع متوسط منطقه به‌طور تقریباً ۸۶۰ متر از سطح دریای آزاد و میانگین شیب حوضه حدود ۱۲/۳ درصد تعیین‌شده است.

رودخانه کورکورسر با طول تقریبی بیست کیلومتر به دریای خزر می‌ریزد. محدوده مطالعاتی این حوضه از شمال به دریای خزر، از شرق به زیر حوزه ماشالک، از جنوب به حوضه رودخانه هاریسک و از غرب به حوزه آبخیز چالوس محدود می‌شود (Khosravi et al., 2021). مساحت حوزه آبخیز کورکورسر نوشهر در بالادست ایستگاه هیدرومتری نوشهر تقریباً برابر با ۷۹ کیلومترمربع است. ارتفاع متوسط حوزه در حدود ۸۶۰ متر از سطح دریا و بیشینه ارتفاع آن به ۱۹۷۲ متر و کمینه ارتفاع به ۱۰- متر می‌رسد.

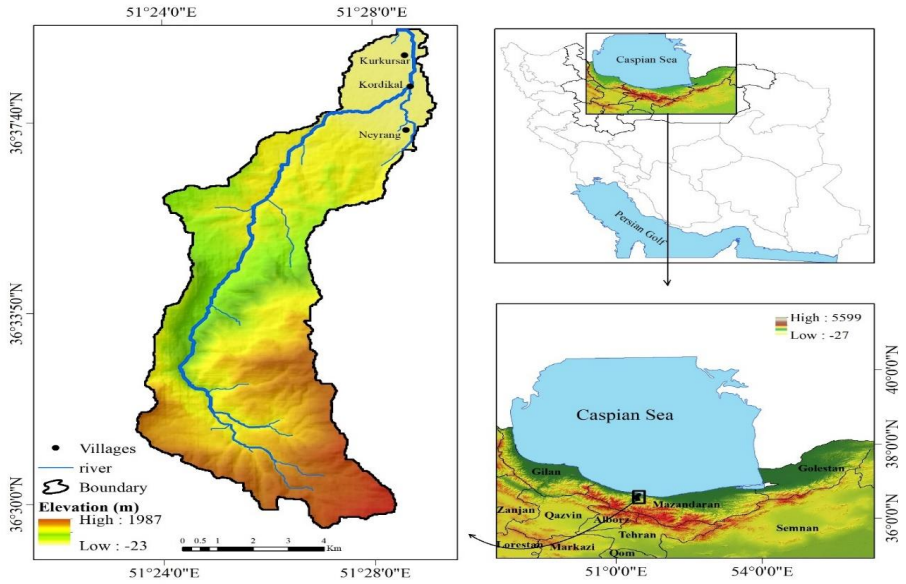
از ویژگی‌های بارز هیدرولوژیکی و ریخت‌سنجی این حوضه می‌توان به وجود مخروط افکنه‌های آبرفتی گسترده، سدهای رسوبی (بایاسدها)، شبکه پیچیده کانال‌های دره‌ای و پهنه‌های دشت‌های سیلابی اشاره نمود. اقلیم حوضه مذکور بر اساس طبقه‌بندی اقلیمی مدیترانه‌ای است و میانگین سالانه بارش آن حدود ۹۰۰ میلی‌متر برآورد شده است.

میانگین جریان حجمی رودخانه اصلی در محدوده حوزه آبخیز کورکورسر برابر با ۱/۲ مترمکعب بر ثانیه محاسبه‌شده است و بیشینه جریان حجمی در بازه‌ای متغیر از ۰/۰۰۲ تا ۷۸ مترمکعب بر ثانیه گزارش شده است (Difi et al., 2023). شکل ۱، موقعیت مکانی جغرافیایی حوزه آبخیز کورکورسر را به‌صورت دقیق نشان می‌دهد.

روش‌شناسی پژوهش: این پژوهش، با هدف پیش‌بینی دقیق جریان رودخانه کورکورسر، از رویکردهای ترکیبی یادگیری ماشین و عمیق با تبدیل موجک بهره می‌برد. پس از اخذ داده‌های بیست ساله بارش و دبی و انجام پیش‌پردازش‌های لازم (مانند مدیریت داده‌های گمشده و نرمال‌سازی)، داده‌ها برای مدل‌سازی تقسیم‌بندی شدند. برای بهبود عملکرد مدل‌های پایه، از تبدیل موجک به‌عنوان یک ابزار قدرتمند پردازش سیگنال استفاده شد.

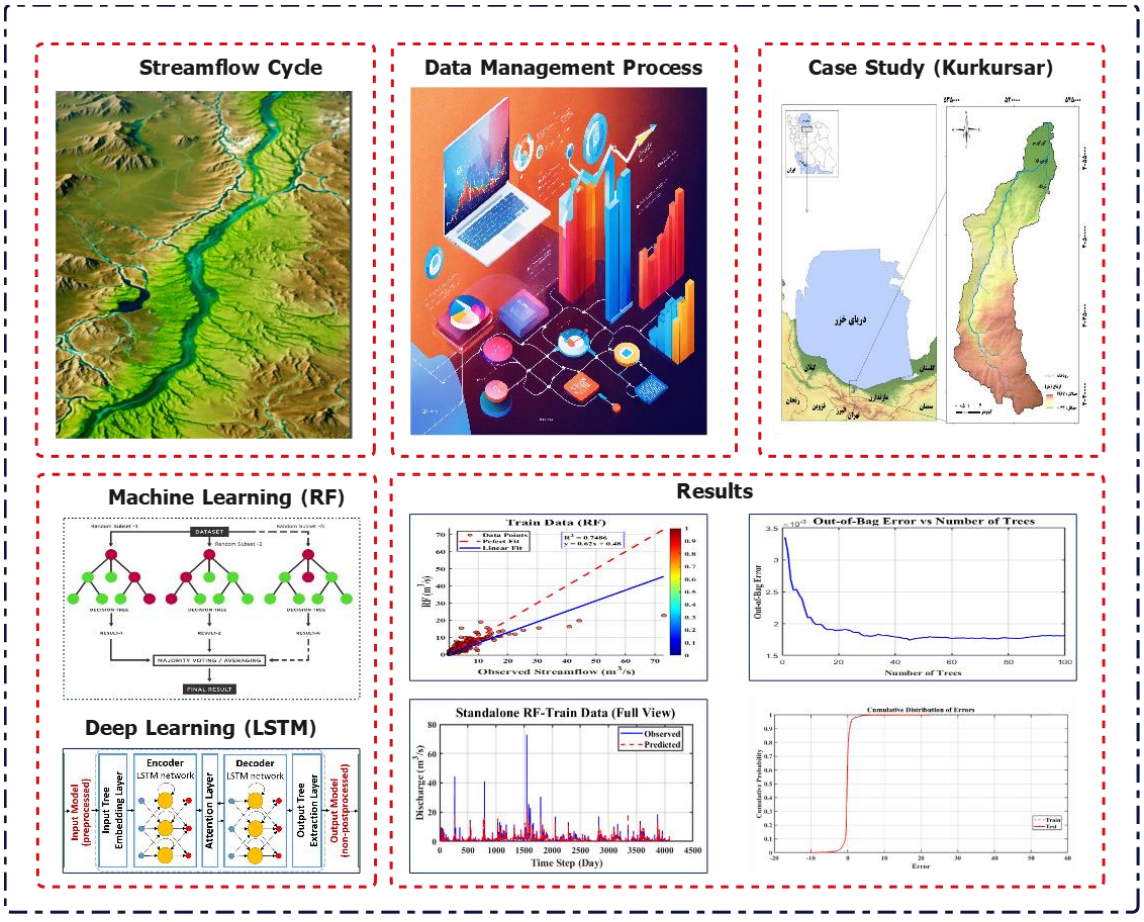
این تکنیک، سری زمانی پیچیده و ناایستای جریان را به مولفه‌های بسامدی ساده‌تر تجزیه کرده و با جداسازی اغتشاش، به الگوریتم‌های یادگیری ماشین امکان می‌دهد تا الگوهای پنهان داده‌ها را بهتر شناسایی کنند. این رویکرد ترکیبی که در آن موجک به‌عنوان یک پیش‌پردازشگر هوشمند عمل می‌کند، منجر به پیش‌بینی‌های دقیق‌تر و قابل اعتمادتری می‌شود. در نهایت، عملکرد مدل‌های بهینه با معیارهای آماری و نمودارهای مختلف ارزیابی شد (Singh et al., 2025).

شکل ۲، روندنمای فرایند کلی انجام پژوهش را نشان می‌دهد.



شکل ۱- موقعیت حوزه آبخیز کورکورسر (Merufinia et al. 2023)

Fig. 1. Location of the Kurkursar Watershed (Merufinia et al. 2023)



شکل ۲- روندنمای فرایند کلی و مراحل انجام پژوهش

Fig. 2. Flowchart of the overall process and stages of the research

مدل جنگل تصادفی (RF): مدل جنگل تصادفی که توسط بریمن (Breiman, 2001) ارائه شد، یک الگوریتم نظارت‌شده و نافراسنجه‌ای از خانواده درختان تصمیم است که از مجموعه‌ای از درختان ناهمبسته تشکیل شده تا پیش‌بینی‌هایی برای مسائل طبقه‌بندی و رگرسیون تولید کند. روش‌های نافراسنجه‌ای مانند جنگل تصادفی خانواده خاصی را برای توزیع داده‌ها در نظر نمی‌گیرند (Altman and Bland, 1999). از آنجایی که یک درخت تصمیم منفرد می‌تواند واریانس بالایی ایجاد کند و مستعد اغتشاش است، جنگل تصادفی این محدودیت را با تولید درختان متعدد، با هر درخت بر روی نمونه خودراه‌انداز داده‌های آموزشی، برطرف می کند (James et al., 2013).

هر بار که یک تقسیم دودویی در یک درخت ساخته می‌شود (که به‌عنوان گره تقسیم نیز شناخته می‌شود)، یک زیر مجموعه تصادفی از پیش‌بینی کننده‌ها (بدون جایگزینی) از مجموعه کامل متغیرهای پیش‌بینی در نظر گرفته می‌شود. یک پیش‌بینی‌کننده از این نامزدها برای ایجاد تقسیم استفاده می‌شود که در آن واریانس‌های مجموع مورد انتظار متغیر پاسخ در دو گره به کمینه رسیده است. فرایند تصادفی‌سازی در تولید زیرمجموعه ویژگی‌ها از انتخاب مکرر یک یا چند پیش بینی کننده قوی در هر تقسیم جلوگیری می‌کند که منجر به درختان بسیار همبسته می‌شود.

پس از رشد همه درختان، جنگل‌ها در مورد درخت جدید پیش‌بینی می‌کنند با اجرای همه درختان از طریق پیش‌بینی کننده‌ها، نقطه داده است. در پایان، جنگل‌ها بیشینه رای بر روی یک طبقه برچسب برای کار طبقه‌بندی می‌دهند یا با میانگین‌گیری همه پیش‌بینی‌ها، مقداری برای کار رگرسیون تولید می‌کنند (Hastie et al., 2009). پیش‌بینی نهایی مدل، میانگین پیش‌بینی‌های تمام درختان است که برای مسائل رگرسیون از رابطه (۱) به‌دست می‌آید.

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (۱)$$

که در این رابطه، $\hat{y}$ مقدار پیش‌بینی‌شده برای نمونه ورودی x، همچنین T تعداد درختان در جنگل و $f_t(x)$ پیش‌بینی درخت tام برای نمونه x است. هر درخت با استفاده از معیار کاهش ناخالصی ساخته می‌شود. برای طبقه‌بندی معمولاً از شاخص جینی یا آنتروپی (بی نظمی) استفاده می‌شود و از رابطه (۲) استفاده می‌شود.

$$Gini=1-\sum_{k=1}^{K} p_k^2 \qquad (۲)$$

که در این رابطه، K تعداد طبقه‌ها، P_k نسبت نمونه‌های طبقه k در یک گره است.

$$Entropy=-\sum_{k=1}^{K} p_k log(p_k) \qquad (۳)$$

$$Var=\frac{1}{N}\sum_{i=1}^{N} (y_i-\bar{y})^2 \qquad (۴)$$

که در این رابطه، N تعداد نمونه‌ها در گره و $\bar{y}$ میانگین مقادیر y_i در گره است. اهمیت هر ویژگی بر اساس میانگین کاهش ناخالصی (یا کاهش واریانس) در تمام درختان از رابطه (۵) محاسبه می‌شود.

$$Importance_j=\frac{1}{T}\sum_{t=1}^{T} \Delta\ Impurity_{j,t} \qquad (۵)$$

که در این رابطه، $\Delta\ Impurity_{j,t}$ کاهش ناخالصی ناشی از ویژگی j در درخت t است. فراسنجه‌های کلی‌دی مدل جنگل تصادفی نیز شامل تعداد درختان (T)، تعداد ویژگی‌های تصادفی برای هر تقسیم (m) حداکثر عمق درخت (Dmax) برای جلوگیری از بیش‌برازش است.

مدل حافظه کوتاه‌مدت طولانی (LSTM): مدل LSTM، به‌عنوان یکی از پیشرفته‌ترین ساختارهای یادگیری عمیق در زمینه تحلیل سری‌های زمانی، به‌طور گسترده‌ای در پیش‌بینی و مدل‌سازی پویای هیدرولوژیکی مورداستفاده قرارگرفته است. این مدل با توانایی ویژه در به‌دست آوردن و حفظ وابستگی‌های زمانی بلندمدت و کوتاه‌مدت در داده‌های متوالی، امکان استخراج الگوهای پیچیده و پیش‌بینی دقیق مقادیر جریان رودخانه را فراهم می‌آورد (Cheng et al., 2020؛ Hunt et al., 2022).

در پیش‌بینی هیدرودینامیکی جریان رودخانه، یکی از قابلیت‌های برجسته شبکه‌های LSTM، مدیریت کارآمد داده‌های اغتشاشی و ناقص است. ساختار مبتنی بر دروازه‌های کنترل‌شده این معماری شامل سه دروازه کلیدی دروازه فراموشی، دروازه ورودی و دروازه خروجی است که امکان پالایش دقیق و انتخابی جریان اطلاعات را به مدل می‌دهد.

به‌واسطه این سازوکار، مدل LSTM قادر است اطلاعات مرتبط و حیاتی را در حافظه بلندمدت خود نگه دارد و داده‌های نامربوط یا دارای اغتشاش را از پردازش حذف نماید، که این امر بهره‌وری مدل را در

مواجهه با داده‌های حقیقی و پراکندگی‌های غیرمطلوب بهبود می‌بخشد (Muhammad et al., 2019). در مدل LSTM، دروازه فراموشی از رابطه (۶) به‌دست می‌آید.

$$f_t=\sigma\left(W_f\cdot[h_{t-1},x_t]+b_f\right) \quad (۶)$$

که در این رابطه، f_t ترتیب خروجی دروازه فراموشی، W_f ماتریس وزن‌های دروازه فراموشی، h_{t-1} حالت پنهان در زمان t-1، x_t ورودی در زمان t، b_f بایاس دروازه فراموشی و σ تابع سیگموئید است. دروازه ورودی نیز به‌ترتیب از رابطه‌های (۷) و (۸) به‌صورت زیر به‌دست می‌آید.

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \quad (۷)$$

$$\widetilde{C}_t=tanh(W_C\cdot[h_{t-1},x_t]+b_C) \quad (۸)$$

که در این رابطه، i_t خروجی دروازه ورودی و $\widetilde{C}_t$ نامزد جدید برای حالت حافظه است. همچنین به‌روزرسانی حالت حافظه مطابق رابطه (۹) انجام گرفت.

$$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}_t \quad (۹)$$

که در این رابطه، C_t حالت حافظه جدید و $\odot$ ضرب عنصر به عنصر است. دروازه خروجی نیز از رابطه‌های (۱۰) و (۱۱) به‌دست می‌آید.

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \quad (۱۰)$$

$$h_t=o_t\odot tanh(C_t) \quad (۱۱)$$

که در این رابطه‌ها، o_t خروجی دروازه و h_t حالت پنهان جدید است.

تئوری موجک: تحلیل موجک به‌عنوان یک ابزار قدرتمند در پردازش نشانک‌های غیرایستا و دارای تغییرات زمانی-بسامدی شناخته می‌شود. این روش با ارائه توانایی تجزیه نشانک به مؤلفه‌های زمانی و بسامدی، مزیت قابل توجهی نسبت به روش‌های سنتی مانند تبدیل فوریه دارد. در میان انواع مختلف تبدیل‌های موجک، موجک دابشیز به‌دلیل ویژگی‌های منحصر به فردش در تحلیل داده‌های هیدرولوژیک مورد توجه قرار گرفته است.

موجک دابشیز ۴ (Db4) به‌طور خاص، با داشتن چهار ضریب پالایش و تعادل مناسب بین نرم‌بودن و فشردگی زمانی، برای مدل‌سازی سری‌های زمانی جریان رودخانه بسیار مناسب است (Peng et al., 2017). مزیت اصلی موجک Db4 در پیش‌بینی جریان رودخانه، توانایی آن در استخراج همزمان ویژگی‌های زمانی و بسامدی از داده‌های هیدرولوژیک است. این ویژگی به‌ویژه در مواجهه با داده‌های جریان رودخانه که معمولاً دارای مؤلفه‌های فصلی، روندهای بلندمدت و نوسانات کوتاه‌مدت هستند، بسیار ارزشمند است.

لذا تبدیل موجک می‌تواند این مؤلفه‌ها را به‌طور مؤثر از هم جدا کند و زمینه را برای مدل‌سازی دقیق‌تر هر یک از این اجزا فراهم نماید. در مقایسه با روش‌های سنتی که ممکن است در شناسایی الگوهای پیچیده موجود در داده‌های جریان رودخانه با مشکل مواجه شوند، موجک Db4 با حفظ اطلاعات زمانی و مکانی، امکان تحلیل جامع‌تری را فراهم می‌کند (Hadi and Tombul 2018؛ Liu et al. 2014).

جمع‌آوری و پیش پردازش داده‌ها: جمع‌آوری داده، پیش‌پردازش و مدیریت داده‌ها از مراحل اساسی در توسعه مدل‌های پیش‌بینی جریان رودخانه محسوب می شوند که تأثیر مستقیمی بر دقت و قابلیت اطمینان نتایج مدل‌سازی دارند. لذا توصیف دقیق روش‌های جمع‌آوری داده، پیش‌پردازش و مدیریت داده‌ها به خوانندگان کمک می‌کند تا درک بهتری از فرایند تحقیق داشته باشند و نتایج به‌دست آمده را با اطمینان بیشتری تفسیر کنند.

در گام نخست، جمع‌آوری داده‌ها و آمار ۲۰ ساله بارش و دبی رودخانه در مقیاس روزانه از شرکت آب منطقه‌ای استان مازندران انجام گرفت. داده‌های مورد نیاز برای مدل‌سازی شامل دبی روزانه رودخانه، بارندگی، دمای هوا، تبخیر و تعرق، رطوبت خاک و فراسنجه‌های مکان‌نگاری هستند که نظر به این که منطقه مورد نظر دارای بارندگی مناسبی بوده و تبخیر و تعرق در آن بسیار ناچیز بوده و در آن معنی‌دار نیست.

از سایر فراسنجه‌ها نیز به علت اثرات ناچیز و تأثیر کم در مدل‌سازی صرف نظر شید و صرفاً آمار بارش (Pt) و دبی رودخانه تا سه روز تأخیر (Qt-1، Qt-2 و Qt-3) مورد استفاده قرار گرفت. پس از جمع‌آوری داده‌ها، مرحله پیش‌پردازش آغاز شد که شامل شناسایی و مدیریت مقادیر گمشده، حذف داده‌های پرت و اغتشاش دار، و نرمال‌سازی داده‌ها است.

مقادیر گمشده و مفقوده از طریق الگوریتم نزدیک ترین همسایگی (KNN) بازسازی شد. داده‌های پرت و خارج از محدوده بازه نیز که ممکن است ناشی از خطاهای ابزاری یا وقایع غیرمعمول باشند، با استفاده از

روش آزمون Z-score و دامنه بین چارکی (IQR) شناسایی و اصلاح شد.

علاوه بر این، نرمال‌سازی داده‌ها برای تبدیل مقادیر به یک بازه یکسان (۱- و ۱) انجام گرفت تا از تأثیر نامتوازن متغیرهای با مقیاس بزرگ بر مدل جلوگیری شود. گام بعدی تقسیم‌بندی داده‌های برای فرایند آموزش، اعتبارسنجی و آزمون است که به‌ترتیب از ۷۰ درصد داده‌ها برای فرایند آموزش، ۲۰ درصد داده‌ها برای اعتبارسنجی و ۱۰ درصد داده‌ها در فرایند آزمون جهت تنظیم فراسنجه‌ها و برای ارزیابی نهایی مدل استفاده شد. در نهایت، پس از اتمام مراحل پیش‌پردازش و مدیریت داده‌ها، داده‌های آماده شده به مدل‌های مختلف یادگیری ماشین مورد استفاده قرار گرفت.

معیارهای ارزیابی مدل: در این پژوهش از معیارهای ارزیابی ضریب تبیین (R^2)، میانگین قدر مطلق خطا (MAE)، میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، ضریب بهره وری نش-ساتکلیف (NSE) و خطای میانگین درصد قدر مطلق خطا (MAPE) استفاده شده است که رابطه‌ها و محدوده عددی این معیارها در جدول ۱ خلاصه شده است.

در رابطه‌های زیر به‌ترتیب y_i مقدار واقعی (مشاهداتی) جریان رودخانه برای داده iام، $\hat{y}_i$ مقدار پیش‌بینی‌شده توسط مدل برای داده iام، n تعداد کل نمونه‌ها (داده‌ها) در مجموعه ارزیابی و $\bar{y}$ میانگین مقادیر واقعی (مشاهداتی) در کل مجموعه داده‌ها است.

جدول ۱- معیارهای ارزیابی مدل

Table 1. Model evaluation criteria

Row	Criteria name	Mathematical relationship	Evaluation scope	Optimal value
1	R^2	$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$	to 1∞-	1
2	MAE	$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	∞0 to	0
3	RMSE	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	∞0 to	0
4	MAPE	$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$		A value of less than 10% is desirable.
5	KGE	$KGE=1-\sqrt{(r-1)^2+(\beta-1)^2+(\gamma-1)^2}$ $\beta=\frac{\mu_s}{\mu_o}$,$\gamma=\frac{\sigma_s/\mu_s}{\sigma_o/\mu_o}$	to 1∞-	1

فراسنجه‌های آماری: در تحلیل داده‌ها و مدل‌سازی های آماری، فراسنجه‌ها به‌عنوان مقادیری مطرح می شوند که شاخص‌های کمی ویژگی‌های کل جامعه آماری را توصیف می‌کنند. فراسنجه‌های آماری مرتبط با متغیرهای مطالعه، شامل بیشینه، کمینه، میانه، میانگین، واریانس، انحراف معیار، ضریب تغییرات، ضریب کشیدگی و چولگی، به‌طور خلاصه و نظام‌مند در جدول ۲، ارائه شده است.

جدول۲- مجموعه داده‌های فراسنجه‌های آماری برای منطقه مورد مطالعه

Table 2. Dataset of statistical parameters for the study area

		Statistical parameters of the studied area					
Parameter	Phase	Minimum	Maximum	Average	Standard deviation	Skewness	Kurtosis
Rainfall (mm)	All Data	0	149	3.549	11.220	6.088	49.542
Discharge (M^3/S)	All Data	0.002	73.1	1.260	2.116	13.025	317.687

انتخاب سناریو و ترکیب مدل بهینه: سوابق پژوهشی و بررسی مطالعات مرتبط حاکی از آن است که در فرایند مدل‌سازی جریان رودخانه، فراسنجه‌های بارش و دبی نقش کلیدی و تأثیرگذاری دارند. همچنین، به‌کارگیری داده‌های دبی با تأخیر زمانی تا سه روز، به‌طورمعمول منجر به بهبود دقت پیش‌بینی‌های مدل می‌شود. بر این مبنا، در این تحقیق، داده‌های آماری بارندگی (Pt) و دبی رودخانه با تأخیرهای یک، دو و سه‌روزه (Qt-1, Qt-2, Qt-3) به‌عنوان ورودی‌های مدل و دبی جریان در زمان جاری (Qt) به‌عنوان متغیر خروجی لحاظ شدند.

در این پژوهش، از ضریب همبستگی پیرسون به‌عنوان شاخص آماری اصلی برای سنجش میزان و جهت رابطه خطی میان دو متغیر کمی استفاده شد تا تعامل بین متغیرهای ورودی و خروجی مدل به‌طور دقیق ارزیابی شود. بر این اساس، مقادیر این ضریب به‌عنوان معیار کلیدی در انتخاب سناریوهای مدل لحاظ شد. جدول ۳، نشان‌دهنده میزان روابط بین متغیرهای ورودی و میزان جریان خروجی مدل بر اساس ضریب همبستگی پیرسون است.

در فرایند انتخاب ترکیب‌های مدل، ضریب همبستگی پیرسون مبنای مرتب‌سازی قرار گرفت، به‌گونه‌ای که متغیرهایی با بالاترین میزان همبستگی در ردیف‌های بالایی و متغیرهای با ضریب همبستگی کمتر به‌ترتیب در ردیف‌های پایین‌تر قرار گرفتند. نهایتاً، ترکیب نهایی مدل مشتمل بر کلیه متغیرهای موردمطالعه تدوین شد. همچنین، جدول ۴، ترکیبات مدل و سناریوهای مربوط به هر یک را ارائه می‌کند.

جدول ۳- ارتباط متغیرهای ورودی و خروجی جهت پیش‌بینی جریان رودخانه بر اساس ضریب پیرسون

Table 3. Correlation between input and output variables for river flow prediction based on Pearson's coefficient

P(t)	Q(t-1)	Q(t-2)	Q(t-3)	Q(t)
0.563	0.463	0.297	0.251	1

جدول ۴- انتخاب سناریو و ترکیبات ورودی مختلف برای پیش‌بینی جریان رودخانه

Table 4. Selection of scenarios and various input combinations for streamflow prediction

Scenario number	Input combinations	Output
1	P(t)	Q(t)
2	P(t), Q(t-1)	Q(t)
3	P(t), Q(t-1), Q(t-2)	Q(t)
4	P(t), Q(t-1), Q(t-2), Q(t-3)	Q(t)

تعیین ضرایب و مقادیر فراسنجه‌های مدل‌های مورد استفاده: انتخاب بهینه فراسنجه‌های کنترلی مدل‌های یادگیری ماشین، نقش بسیار مهمی در بهبود دقت پیش‌بینی، کاهش مشکل بیش برازش و افزایش توان تعمیم‌پذیری مدل‌ها دارد. تنظیم دقیق این فراسنجه‌ها به‌صورت مستقیم عملکرد نهایی الگوریتم را تحت تأثیر قرار داده و می‌تواند مرز میان مدل‌هایی با کارایی متوسط و مدل‌هایی با بازدهی بالا را مشخص کند.

در مقابل، انتخاب نامناسب این فراسنجه‌ها نه‌تنها کیفیت نتایج را به‌شدت کاهش می‌دهد بلکه ممکن است باعث مصرف بی‌مورد منابع محاسباتی و کاهش اعتماد به پیش‌بینی‌ها شود. در این زمینه، فراسنجه‌های کلیدی و ابرفراسنجه‌های حیاتی مدل‌های استفاده‌شده شامل RF و LSTM، محدوده‌های انتخاب (شامل مقادیر مشخص و بازه‌های جست‌وجو) و همچنین مقادیر پیشنهادی و به‌کار رفته در این مطالعه در جدول ۵، به‌صورت خلاصه ارائه‌شده است (Nguyen et al., 2025؛ Kratzert et al., 2019).

جدول ۵- مقادیر فراسنجه‌های مدل‌های مورد استفاده پژوهش

Table 5. Values of the parameters for the models used in the study

Row	Parameter or variable	Selection range	Selected value
	Random Forest (RF)		
1	Number of trees (n_estimators)	50-500	100
2	Maximum tree depth (max_depth)	3-20	10
3	Minimum samples to split a node (min_samples_split)	2-10	2
4	Minimum samples in leaves (min_samples_leaf)	1-5	1
5	Number of features for each split (max_features)	log2 / Sqrt	sqrt
	Long short-term memory (LSTM)		

ادامه جدول ۵.

Table 5. Continued

Row	Parameter or variable	Selection range	Selected value
	Random Forest (RF)		
1	Number of hidden layer neurons (units)	32-256	120
2	Dropout	0.2-0.5	0.3
3	Batch size	16-128	32
4	Epochs	50-200	100
5	Learning rate	0.001-0.01	0.001
6	Number of LSTM layers (num_layers)	1-3	1
7	Activation Function	tanh -sigmoid - ReLU	Tanh - ReLU

نتایج و بحث

فرایند مدل‌سازی در دو حالت مدل‌های منفرد و ترکیبی انجام گرفت. نتایج فرایند مدل‌سازی در جدول ۶، خلاصه شد. نتایج نشان داد که در مدل منفرد RF به‌ترتیب در میان سناریوهای موجود سناریوی اول (S_1) در مرحله آموزش (R^2=0.425؛ RMSE=1.714 m^3/s) و آزمون (R^2=0.397؛ RMSE=1.422 m^3/s) دارای عملکرد ضعیف‌تری نسبت به سایر سناریوها بوده است. همچنین سناریوی چهارم (S^4) دارای بهترین عملکرد در مرحله آموزش (R^2=0.748, RMSE=1.181 m^3/s) و آزمون (R^2=0.684, RMSE=0.988 m^3/s) در میان مدل‌های منفرد RF بوده است. همچنین نتایج نشان داد که سناریوی S_1 که صرفا دارای متغیر بارش بوده دارای عملکرد ضعیف‌تری نسبت به سناریوهای دارای دبی پایه است. همچنین نتایج نشان داد که مدل موجک دابشیز ۴ تا حد بسیار قابل قبولی به کاهش خطای مدل کمک نموده است.

نتایج مدل نشان داد که سناریوی چهارم نیز در مدل ترکیبی دارای عملکرد غالب و برتر در هر دو مرحله آموزش (R^2=0.907, RMSE=0.0192 m^3/s) و آزمون (R^2=0.942, RMSE=0.016 m^3/s) است. همچنین در این مدل ترکیبی عملکرد موجک باعث افزایش دقت قابل قبولی شده است (R^2=0.425, R^2=0.942). لذا دقت حدود حدود ۵۵ درصد نسبت به مدل منفرد افزایش یافته است که عملکرد بسیار خوب مدل را نشان می‌دهد. در خصوص چرایی و علت بهبود نتایج مدل منفرد نیز می‌توان گفت که این امر ناشی از مکانیسم‌های متعددی ازجمله استخراج ویژگی‌های چندمقیاسه، کاهش اغتشاش، بهبود تشخیص روابط غیرخطی و افزایش تعمیم‌پذیری است.

این یافته‌ها با نتایج مطالعات پیشین همسو و هم راستا است و نشان می‌دهد که ادغام روش‌های پردازش سیگنال با یادگیری ماشین می‌تواند راه‌حلی مؤثر برای مسائل پیچیده باشد. همچنین می‌توان گفت که تبدیل موجک، قادر است داده‌ها را هم در حوزه زمان و هم در حوزه بسامدی تحلیل کند. این ویژگی منحصر به فرد، امکان شناسایی الگوهای موقتی و محلی را فراهم می سازد که در داده‌های خام یا تبدیل‌های سنتی پنهان می‌مانند. همچنین Db4 به‌دلیل داشتن خاصیت فشردگی مناسب و تعادل بین زمان و بسامد، برای پردازش نشانک‌های غیرساکن ایده‌آل است.

همچنین Db4 به‌عنوان یک موجک متعامد، در حذف اغتشاش‌های اضافی عملکرد مناسبی دارد. این تبدیل با اعمال آستانه‌گذاری بر روی ضرایب موجک، مؤلفه‌های کم‌اهمیت را حذف می‌کند و در عین حال، ساختار اصلی داده را حفظ می‌نماید. در نتیجه، مدل RF با داده‌هایی تمیزتر و با نسبت نشانک به اغتشاش بالاتر مواجه می‌شود که این امر منجر به کاهش واریانس پیش‌بینی و جلوگیری از بیش‌برازش می‌شود.

لذا تبدیل Db4 با پالایش نمودن اطلاعات و تقویت نشانک‌های مفید، به مدل RF کمک نموده تا تعمیم‌پذیری بهتری داشته باشد. در خصوص ضرایب و ابرفراسنجه‌های RF نیز به روش دستی نیز سعی و خطا صورت گرفت. نتایج نشان داد که در خصوص فراسنجه تعداد درخت‌های تصمیم در جنگل (در محدوده ۱۰ تا ۱۰۰۰) بهترین عملکرد مدل در تعداد ۱۰۰ درخت حاصل می‌شود. همچنین نتایج نشان داد که مقادیر بالاتر از ۲۰۰ معمولاً بهبود ناچیزی دارند. در این خصوص نیز می‌توان گفت که پس از یک آستانه بهینه (معمولاً ۱۰۰–۲۰۰ درخت)، افزایش تعداد درخت‌ها بهبود ناچیزی در دقت ایجاد می‌کند و پس از ۲۰۰ درخت، منحنی دقت به حالت اشباع می‌رسد و با افزایش تعداد درخت یا بهبود خاصی ملاحظه نمی‌شود و یا این

بهبود و تغییرات معنی‌دار نخواهد بود و صرفا اجرای برنامه را طولانی‌تر می‌سازد.

بیشینه عمق هر درختان نیز بین ۲۰-۳ مورد ارزیابی قرار گرفت که نتایج برای این پژوهش نشان داد که بین محدوده ۱۰-۵ بهترین عملکرد برای مدل حاصل می‌شود که در این پژوهش مقدار پنج جواب بهینه‌تری داشته است. همچنین بررسی منابع نشان می دهد که عمق زیاد (بیش از ۱۵) منجر به پدیده بیش برازش می‌شود که با یافته‌های این پژوهش همخوانی دارد. در خصوص معیار تقسیم (جینی یا آنتروپی) نظر به مقیاس روزانه و حجیم بودن داده از معیار آنتروپی استفاده شد که با مطالعات مختلف انجام گرفته همسو است.

همچنین مقایسه نتایج دیگر نشان داد که عملکرد این دو معیار تقسیم در خروجی نتایج کمتر از دو درصد در بهبود نتایج می‌باشد. شکل‌های ۳ و ۴ نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی مدل منفرد و ترکیبی جنگل تصادفی در مراحل آموزش (R^2=0.748, Y=0.62x+0.48) و آزمون (R^2=0.684, Y=0.71x+0.39) را نشان می‌دهد. محور افقی مقادیر داده‌های مشاهداتی و محور عمودی مقادیر پیش‌بینی مدل (مترمکعب بر ثانیه) نشان می‌دهد. شکل ۵، نیز نمودار پراکندگی بین داده‌های مشاهداتی و پیش‌بینی جریان رودخانه مدل منفرد RF در مراحل آموزش و آزمون را نشان می‌دهد. محور افقی گام زمانی (روز) و محور عمودی نیز دبی (مترمکعب بر ثانیه) است.

همچنین برای نمایش بهتر یک بزرگنمایی از مراحل آموزش و آزمون نشان داده شد. شکل ۶، نمودار تحلیل خطای OOB و تعداد درختان و نمودار خطای پیش بینی برای مدل منفرد RF را نشان می‌دهد. هدف اصلی از ترسیم نمودار خطای OOB، مشخص کردن تعداد بهینه درخت‌ها است؛ یعنی جایی که با افزودن درخت‌های بیشتر، بهبود چشمگیری در دقت حاصل نمی‌شود و مدل به حالت پایدار می‌رسد.

به این ترتیب، می‌توان از ایجاد مدل‌های بیش از حد بزرگ و زمان‌بر جلوگیری کرد. در مدل منفرد RF، برای آموزش هر درخت، یک نمونه‌گیری با بازگشت از داده‌های آموزش انجام گرفته و به‌طور متوسط، حدود ۶۰ درصد داده‌ها در هر خودراه‌انداز و ۴۰ درصد باقی مانده به‌عنوان داده‌های OOB در نظر گرفته شد. خطای OOB از میانگین خطای پیش‌بینی روی همین داده‌های دیده‌نشده محاسبه می‌شود و عملکرد مدل را بدون نیاز به مجموعه تست مجزا تخمین می‌زند.

در این نمودار محور افقی تعداد درخت‌های تصمیم‌گیری که در مدل RF (مقدار یک تا ۱۰۰) و محور عمودی خطای OOB را نشان می‌دهد. محور افقی نشان می‌دهد که چگونه اضافه شدن درخت‌های بیشتر بر خطای مدل تأثیر می‌گذارد. یافته‌ها و تحلیل نمودار مذکور نشان داد که کاهش سریع خطا در مراحل اولیه مشاهده شد که با افزایش تعداد درخت‌ها (از یک تا حدود ۲۰)، خطای OOB به‌طور چشم‌گیری کاهش یافته است. این نشان‌دهنده یادگیری مؤثر و تجمیع نتایج از درخت‌های متعدد است. پس از رسیدن به حدود ۴۰ درخت، نرخ کاهش خطا کند می‌شود و تقریباً به یک مقدار ثابت می‌رسد. این بدان معنی است که افزودن درخت‌های بیشتر به مدل، بهبود قابل توجهی در عملکرد ایجاد نمی‌کند.

همچنین خطا در انتهای نمودار دچار نوسانات بسیار جزئی شده که ممکن است ناشی از اغتشاش تصادفی در نمونه‌گیری خودراه‌انداز باشد، اما تغییرات آن بسیار کم و قابل چشم‌پوشی است. نمودار خطای پیش‌بینی به‌صورت پراکنش مقادیر واقعی (داده‌های مشاهداتی) در محور افقی و مقدار باقیمانده‌ها (تفاضل داده‌های پیش بینی از مشاهداتی) در محور عمودی ترسیم شده است. نقاط آبی نشان‌دهنده داده‌های مربوط به مجموعه آموزش، نقاط نارنجی مجموعه آزمون و خط قرمز افقی معرف سطح خطای صفر است.

بررسی توزیع باقیمانده‌ها نشان می‌دهد که اکثر داده‌ها در محدوده نزدیک به محور افقی (خط صفر) متمرکز شده‌اند، که این امر حاکی از عملکرد نسبتاً مناسب مدل در پیش‌بینی داده‌ها است. تمرکز نسبی باقیمانده‌ها در بازه‌ی ۵- و پنج برای بیشتر مقادیر مشاهده‌شده، نشان‌دهنده آن است که خطاها دارای میانگین نزدیک به صفر بوده و فاقد اریبی هستند. مقایسه نقاط آبی و نارنجی نشان می‌دهد که رفتار مدل در هر دو مجموعه آموزش و آزمون تقریباً مشابه است. این همخوانی نشان‌دهنده آن است که مدل دچار بیش‌برازش یا کم‌برازش شدید نشده است. با این حال،

برخی نقاط دورافتاده در هر دو مجموعه مشاهده می شود که مدل نتوانسته آن‌ها را با دقت کافی پیش‌بینی کند.

همچنین در این نمودار چندین نقطه به‌وضوح از سایر نقاط فاصله دارند و دارای باقیمانده‌های بزرگ مثبتی هستند. این نقاط ممکن است ناشی از خطاهای اندازه‌گیری، نمونه‌های نادر یا پدیده‌هایی با ماهیت پویایی متفاوت باشند. شناسایی و تحلیل این نقاط از طریق حذف و با استفاده از الگوریتم‌های مقاوم در برابر داده‌های پرت می‌تواند به بهبود کیفیت مدل کمک شایانی نماید، شکل ۷، نمودار خطای پیش‌بینی نیز شامل دو هیستوگرام خطا برای دو مجموعه داده آموزش و آزمون است. محور افقی، مقدار خطا بین مقدار پیش‌بینی‌شده و مقدار مشاهداتی (واقعی) و محور عمودی نیز چگالی احتمال است که نشان می‌دهد که با چه احتمالی هر مقدار خطا رخ داده است.

بررسی شکل توزیع خطا در داده‌های آموزشی نشان داد که اکثر خطاها در اطراف مقدار صفر متمرکز شده‌اند. توزیع خطا تا حد زیادی متقارن و زنگوله‌ای شکل است که با توزیع نرمال هم‌راستا است. این پیکربندی نشان می‌دهد که مدل در داده‌های آموزش دقت خوبی داشته اما برخی داده‌ها ممکن است به‌درستی یادگیری نشده باشند. شکل ۷، نمودار توزیع تجمعی خطا (نمودار سمت راست) و نمودار چندک-چندک خطا (نمودار چپ) برای مدل منفرد جنگل تصادفی را نشان می‌دهد. نمودار توزیع تجمعی خطاها، به‌منظور تحلیل دقیق‌تر عملکرد مدل RF در پیش‌بینی متغیر وابسته مورد استفاده قرار گرفت.

این نمودار به‌صورت تابع پله‌ای نمایانگر توزیع تجمعی احتمال خطا برای مجموعه‌های داده آموزش و آزمون است. محور افقی مقدار خطا و محور عمودی احتمال تجمعی وقوع خطا را نشان می‌دهد. دو منحنی مجزا برای مجموعه داده آموزش (رنگ آبی) و آزمون (رنگ نارنجی) نشان داده است که مقایسه آن‌ها امکان ارزیابی دقیق تعمیم‌پذیری مدل را فراهم می‌کند. نتایج نشان داد که هر دو منحنی دارای شیب بسیار تند در حوالی خطای صفر هستند، به‌گونه‌ای که بیش از ۹۵ درصد از خطاها در بازه بسیار باریکی حول مقدار صفر قرار گرفته‌اند.

این رفتار نمایانگر دقت بالای مدل در پیش‌بینی مقادیر خروجی است. به بیان دیگر، توزیع خطاها به‌شدت متراکم در اطراف صفر بوده و بخش عمده‌ای از خطاها از نوع خطاهای کوچک هستند که این موضوع نشانه‌ای از برازش مناسب مدل جنگل تصادفی به داده‌ها است. همچنین سایر یافته‌ها نشان می‌دهد که هم‌پوشانی زیاد منحنی‌های آموزش و آزمون دلالت بر آن دارد که مدل در مواجهه با داده‌های نادیده (آزمون) رفتاری مشابه با داده‌های دیده‌شده (آموزش) از خود نشان داده است.

لذا می‌توان گفت که مدل توانسته است الگوهای موجود در داده‌های آموزش را به‌خوبی استخراج کرده و بدون وابستگی بیش از حد به داده‌های خاص، آن‌ها را به مجموعه داده آزمون تعمیم دهد. با این حال، در بخش‌های انتهایی نمودار (سمت راست)، مقادیر خطای نسبتاً بزرگی دیده می‌شود که هرچند از لحاظ فراوانی بسیار اندک هستند، اما وجود آن‌ها حاکی از آن است که مدل در برخی نمونه‌های خاص دچار خطای زیاد شده است.

این نقاط (پرت) ناشی از پیچیدگی زیاد داده‌های ورودی، اغتشاش موجود در اندازه‌گیری‌ها یا حتی محدودیت‌های ساختاری مدل باشند. شکل ۷، نمودار چندک- چندک خطا (نمودار چپ) را نشان می‌دهد. محور افقی در هر دو نمودار مربوط به چارک (چندک) توزیع نرمال استاندارد و محور عمودی مربوط به چارک (چندک) نمونه‌ای خطاهای مدل است. اگر خطاهای مدل از توزیع نرمال تبعیت کنند، نقاط روی یک خط قطری (قرمز رنگ در نمودار) قرار می‌گیرند یا به آن نزدیک خواهند بود. انحراف از این خط نشانه‌ای از نرمال نبودن توزیع خطاها است که می‌تواند اطلاعات زیادی درباره عملکرد مدل، وجود داده‌های پرت، یا مشکلات احتمالی در تنظیم مدل ارائه دهد.

در نمودار مربوط به داده‌های آموزشی، خطاها در بیشتر قسمت‌ها با نرمال بودن فاصله دارند، اما در بخش مرکزی نمودار، نقاط به خوبی روی خط قطری قرار گرفته‌اند که نشان‌دهنده آن است که بخش عمده‌ای از خطاها (یعنی خطاهایی با شدت کم) به توزیع نرمال نزدیک هستند. با این حال، در انتهای نمودار (چندک‌های بالا و پایین)، نقاط از خط قطری فاصله می

گیرند و انحراف قابل توجهی را نشان می‌دهند. این مسأله نشان‌دهنده وجود داده‌های پرت یا رفتار غیرخطی در مدل است که نتوانسته به‌خوبی برخی مشاهدات خاص را پیش‌بینی کند. به‌ویژه در انتهای سمت راست که مربوط به بزرگ‌ترین خطاها است، خطاهایی مشاهده می‌شود که به مراتب بیشتر از آن چیزی هستند که در توزیع نرمال انتظار می‌رود.

شکل ۹، نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی مدل ترکیبی جنگل تصادفی-موجک را نشان می‌دهد. نتایج مدل منفرد LSTM نیز نشان داد که سناری اول (S_1) که صرفا دارای متغیر بارش است، دارای ضعیف‌ترین عملکرد در مرحله آموزش (R^2=0.310, RMSE=1.872) و آزمون (, R^2=0.373 RMSE=1.406) است. همچنین سناریوی چهارم (S_4) که دارای ترکیب سناریوهای بارش و دبی با تأخیر است، دارای بهترین عملکرد در مرحله آموزش (,R^2=0. 499 RMSE=1.60) و آزمون (R^2=0.579, RMSE=1.149) بوده است.

همچنین نتایج مدل ترکیبی LSTM-Wavelet نیز نشان داد که سناریوی چهارم در مرحله آموزش (R^2=0.794, RMSE=1.149) و آزمون (,R^2=0.724 RMSE=0.962) دارای بهترین عملکرد بوده است. هچنین مدل ترکیبی LSTM-Wavelet توانست حدود ۳۹ درصد عملکرد مدل منفرد LSTM را بهبود بخشد. شکل ۱۰، نمودار پراکندگی داده‌های مشاهداتی و پیش بینی مدل منفرد LSTM در مراحل آموزش و آزمون را نشان می‌دهد. همچنین شکل‌های ۱۱ و ۱۲ نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی را نشان می‌دهد. برتری سناریوی S_4 بر S_1 تأیید می‌کند که افزودن مقادیر دبی گذشته (حافظه سیستم) در کنار بارش، برای پیش‌بینی دقیق جریان ضروری است.

بهبود چشمگیر عملکرد مدل RF-Wavelet ناشی از توانایی تبدیل موجک در تجزیه سیگنال پیچیده جریان به مؤلفه‌های فرکانسی ساده‌تر است. این فرایند با کاهش اغتشاش و آشکارسازی الگوهای پنهان، به مدل RF اجازه می‌دهد تا روابط غیرخطی را با دقت بالاتری یاد بگیرد. این یافته که مدل‌های مبتنی بر جنگل تصادفی، به‌ویژه در حالت ترکیبی، عملکرد قدرتمندی دارند، با نتایج Khosravi et al., (2022)، Sharma et al., (2024)، که بر برتری مدل RF در شرایط مشابه تأکید داشتند، همخوانی دارد. همچنین تحلیل ابرفراسنجه‌های مدل RF نیز نشان داد که تعداد بهینه درختان حدود ۱۰۰ عدد است و پس از آن منحنی دقت به حالت اشباع می‌رسد (شکل ۶).

این موضوع از بزرگ شدن بی‌دلیل مدل و افزایش زمان محاسبات جلوگیری می‌کند. همچنین، عمق بهینه درختان بین پنج تا ۱۰ تعیین شد که با یافته‌های پیشین مبنی بر خطر بیش‌برازش در عمق‌های زیاد، مطابقت دارد. مدل منفرد LSTM ضعیف‌ترین عملکرد کلی را در این پژوهش داشت. بهترین نتایج این مدل در سناریوی چهارم (S_4) در مرحله آموزش (,R^2=0.499 RMSE=1.60 m^3/s) و آزمون (,R^2=0.579 RMSE=1.149 m^3/s) ثبت شد.

با این حال، پس از ترکیب با تبدیل موجک، مدل ترکیبی LSTM-Wavelet بهبود قابل توجهی نشان داد و در سناریوی چهارم به نتایج (R^2=0.794) در مرحله آموزش و (R^2=0.724) در مرحله آزمون دست یافت. این رویکرد ترکیبی توانست دقت مدل منفرد LSTM را حدود ۳۹ درصد افزایش دهد.

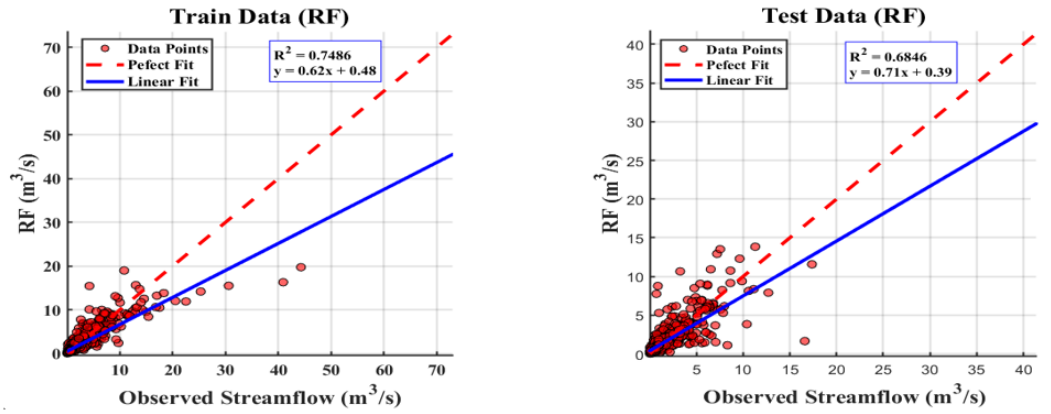
اگرچه مدل LSTM ذاتاً برای داده‌های سری زمانی طراحی شده است، اما نتایج نشان می‌دهد که پیش‌پردازش با موجک همچنان می‌تواند با کاهش اغتشاش و ساده‌سازی الگوها، به بهبود عملکرد آن کمک کند. با این وجود، در مقایسه مستقیم مدل‌های ترکیبی، مدل RF-Wavelet حدود ۲۳ درصد خطای کمتری نسبت به مدل LSTM-Wavelet داشت.

این موضوع ممکن است نشان دهد که برای این مجموعه داده خاص، ساختار مبتنی بر درخت تصمیم RF در ترکیب با ویژگی‌های استخراج شده توسط موجک، انطباق بهتری نسبت به معماری حافظه-محور LSTM داشته است. در حالی که در این پژوهش مدل RF-Wavelet برتر بود، برتری کلی مدل‌های ترکیبی مبتنی بر موجک با LSTM با نتایج Arathy and Adarsh, (2024) که بهبود ۱۰ تا ۱۲ درصدی را گزارش کردند، همسو است.

جدول ۶- نتایج معیارهای ارزیابی مدل‌های RF و LSTM در دو مرحله آموزش و آزمون

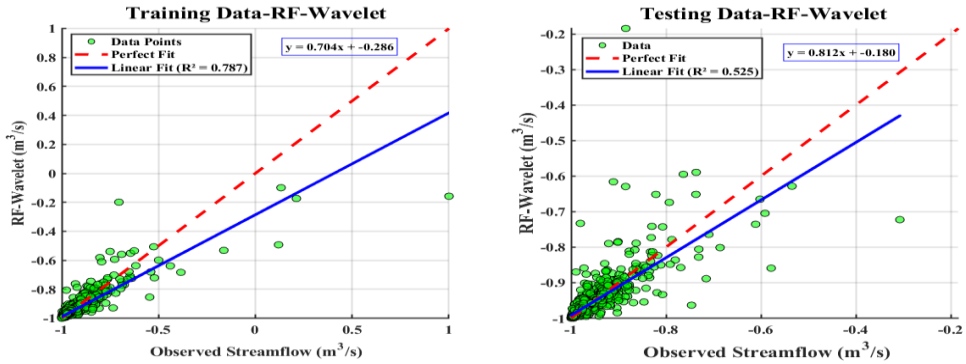
Table 6. Evaluation metrics results of the RF and LSTM models in the training and testing phases

Scenario number	Model name	Phase	R^2	MAE	RMSE	MAPE (%)	PBIAS (%)	KGE
S_1	RF	Train	0.425	0.845	1.714	234.74	0.258	0.458
		Test	0.394	0.826	1.422	251.92	-7.346	0.588
S_2		Train	0.720	0.313	1.227	43.918	0.232	0.678
		Test	0.659	0.355	0.031	56.645	-3.0696	0.768
S_3		Train	0.741	0.259	1.182	24.314	0.328	0.696
		Test	0.659	0.333	1.038	38.462	-7.074	0.783
S_4		Train	0.748	0.253	1.181	25.647	0.594	0.682
		Test	0.684	0.319	0.98	38.989	-3.181	0.773
S_1	RF-Wavelet	Train	0.901	0.0053	0.0197	0.688	-0.015	0.803
		Test	0.917	0.0072	0.012	0.793	-0.011	0.868
S_2		Train	0.900	0.0035	0.019	0.509	-0.027	0.817
		Test	0.942	0.0048	0.0105	0.548	0.023	0.926
S_3		Train	0.891	0.0038	0.0207	0.543	-0.009	0.810
		Test	0.938	0.0050	0.0109	0.574	0.017	0.912
S_4		Train	0.907	0.0036	0.0192	0.496	-0.016	0.820
		Test	0.942	0.0049	0.0106	0.562	0.0165	0.902
S_1	LSTM	Train	0.310	0.895	1.872	248.56	-4.724	0.344
		Test	0.373	0.834	1.406	267.24	-11.56	0.495
S_2		Train	0.472	0.611	1.641	132.33	3.211	0.514
		Test	0.559	0.564	1.167	146.17	-0.869	0.657
S_3		Train	0.485	0.597	1.622	138.2	-4.650	0.524
		Test	0.572	0.554	1.155	153.6	-9.374	0.661
S_4		Train	0.499	0.567	1.6	124.66	-5.532	0.539
		Test	0.579	0.529	1.149	139.44	-10.583	0.679
S_1	LSTM-Wavelet	Train	0.789	0.641	1.264	172.79	-0.869	0.538
		Test	0.669	0.646	1.082	205.13	-410.6	0.552
S_2		Train	0.842	0.296	0.980	34.323	-1.020	0.713
		Test	0.836	0.288	0.740	45.065	-2.991	0.774
S_3		Train	0.811	0.412	1.102	61.306	-1.648	0.646
		Test	0.758	0.391	0.902	71.6	-4.268	0.690
S_4		Train	0.794	0.468	1.149	80.768	-1.843	0.627
		Test	0.724	0.474	0.962	96.81	-6.275	0.656



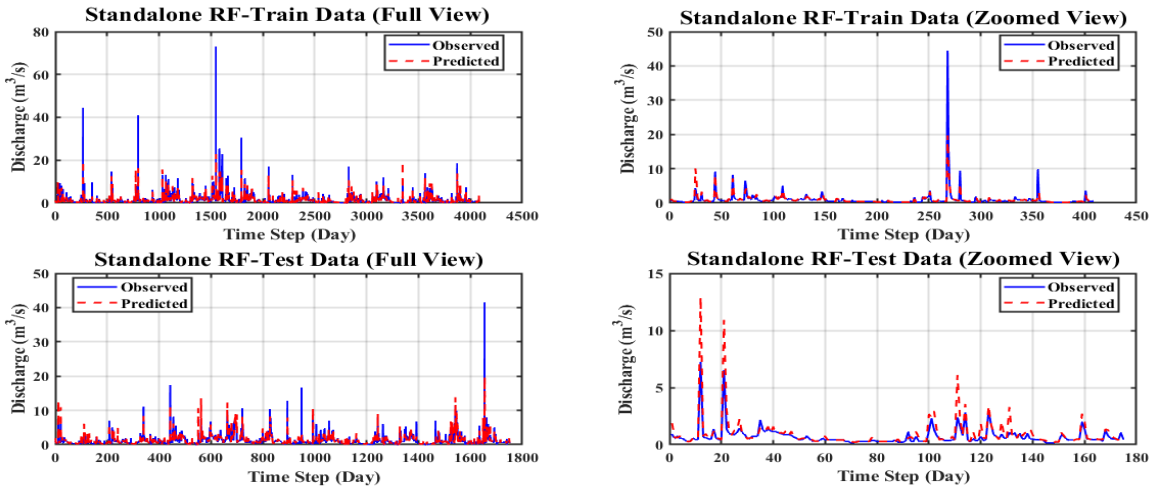
شکل ۳- نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی مدل منفرد RF در مراحل آموزش و آزمون

Fig. 3. Time series plot of observational data and predictions from the single RF model during the training and testing phases



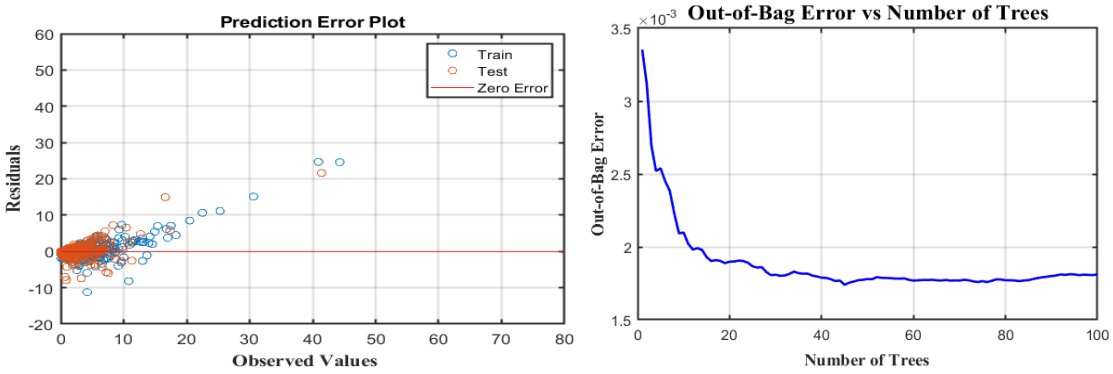
شکل ۴- نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی مدل ترکیبی RF-Wavelet در مراحل آموزش و آزمون

Fig. 4. Time series plot of observational data and predictions from the hybrid RF-Wavelet model during the training and testing phases



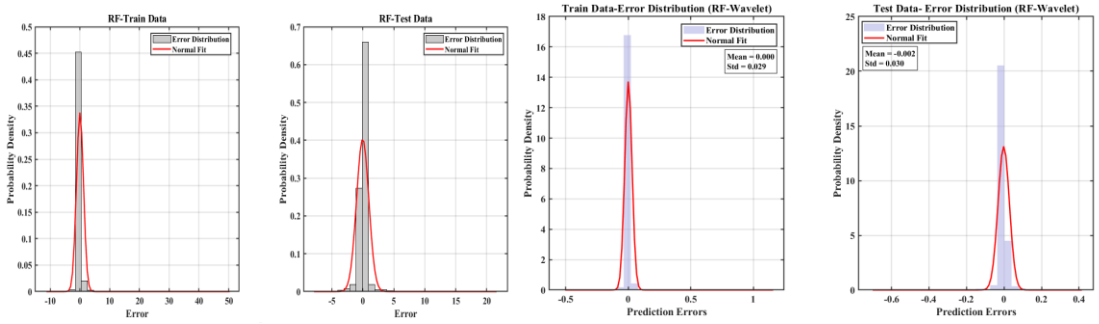
شکل ۵- نمودار پراکندگی بین داده‌های مشاهداتی و پیش‌بینی جریان رودخانه مدل منفرد جنگل تصادفی در مراحل آموزش و آزمون

Fig. 5. Scatter plot between observed data and river flow predictions of the single random forest model during the training and testing phases



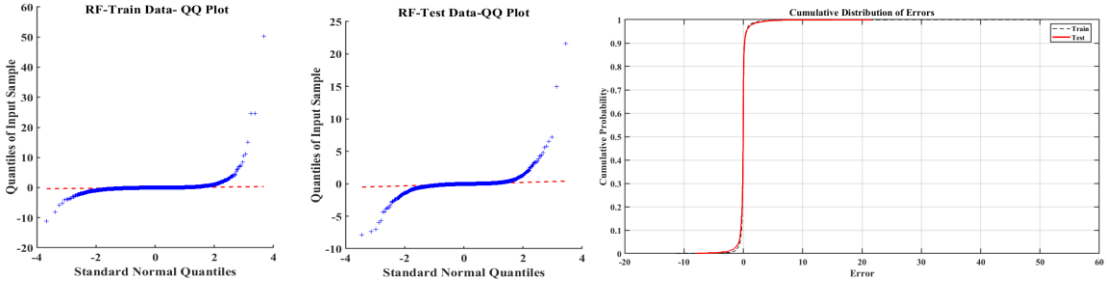
شکل ۶- نمودار تحلیل خطای OOB و تعداد درختان و نمودار خطای پیش بینی برای مدل منفرد جنگل تصادفی

Fig. 6. Out-of-bag error analysis plot and number of trees, and prediction error plot for the single random forest model



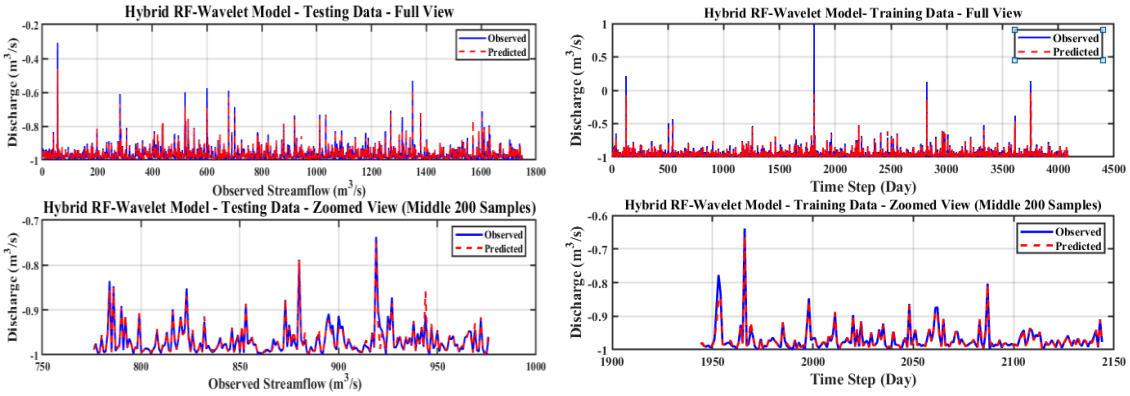
شکل ۷- نمودار هیستوگرام خطا و توزیع نرمال در دو مرحله آموزش و آزمون برای مدل منفرد جنگل تصادفی

Fig. 7. Histogram of error and normal distribution during the training and testing phases for the single random forest model



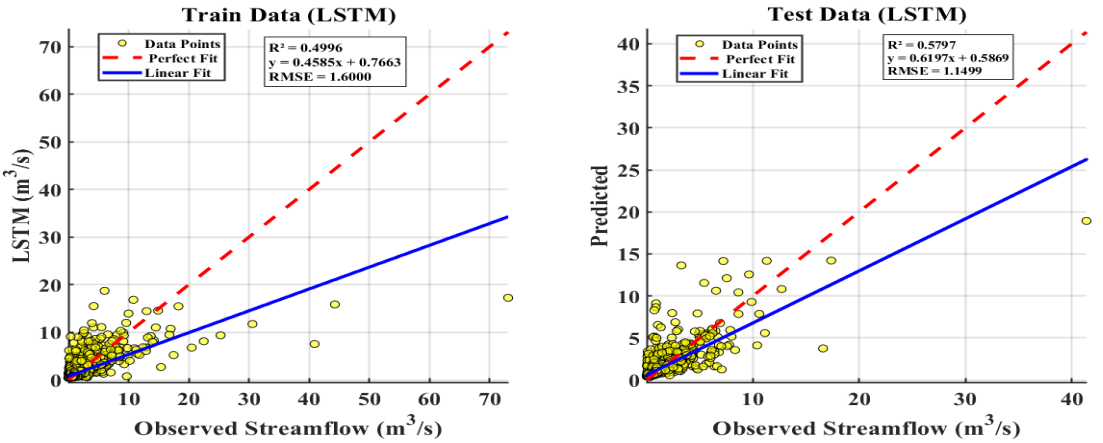
شکل ۸- نمودار توزیع تجمعی خطا و نمودار چندک-چندک خطا برای مدل منفرد جنگل تصادفی

Fig. 8. Cumulative error distribution plot and quantile-quantile error plot for the single random forest model



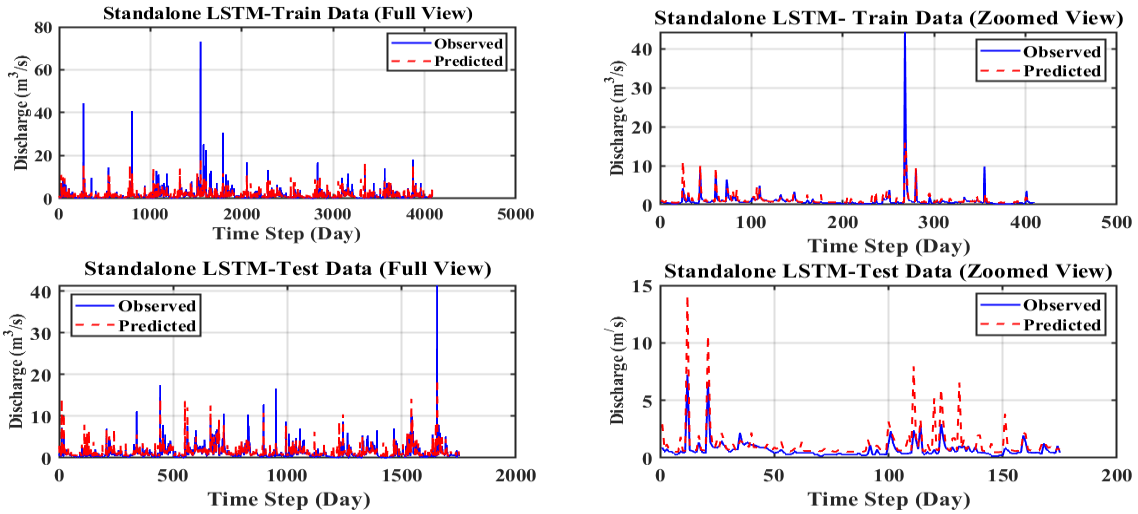
شکل ۹- نمودار سری زمانی داده‌های مشاهداتی و پیش‌بینی مدل ترکیبی جنگل تصادفی-موجک

Fig. 9. Time series plot of observed data and predictions from the hybrid Random Forest–Wavelet model



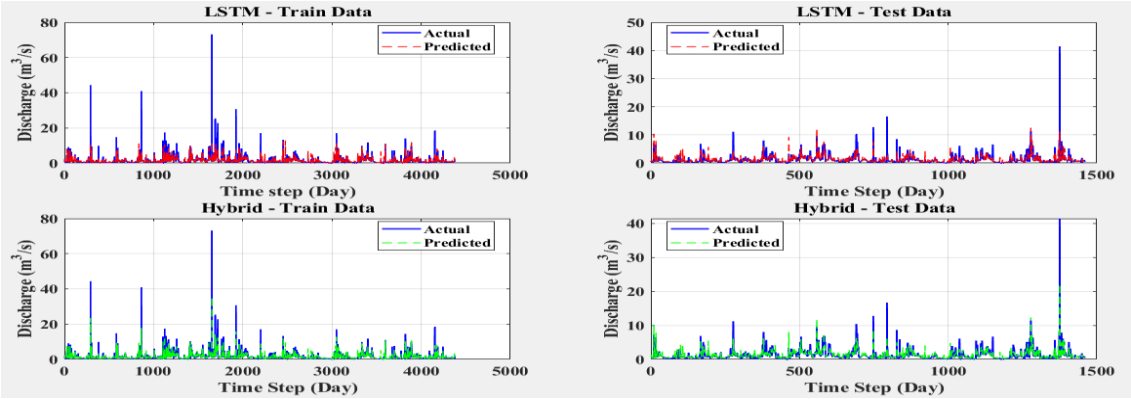
شکل ۱۰- نمودار پراکندگی داده‌های مشاهداتی و پیش‌بینی مدل منفرد LSTM در مراحل آموزش و آزمون

Fig. 10. Scatter plot of observed data and predictions of the single LSTM model during the training and testing phases



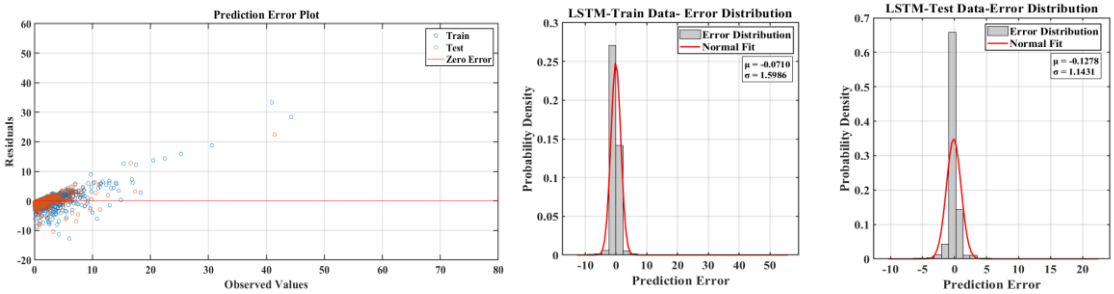
شکل ۱۱- نمودار سری زمانی بین داده‌های مشاهداتی و پیش‌بینی جریان رودخانه مدل منفرد LSTM در مراحل آموزش و آزمون

Fig. 11. Time series plot comparing observed data and river flow predictions from the single LSTM model during the training and testing phases



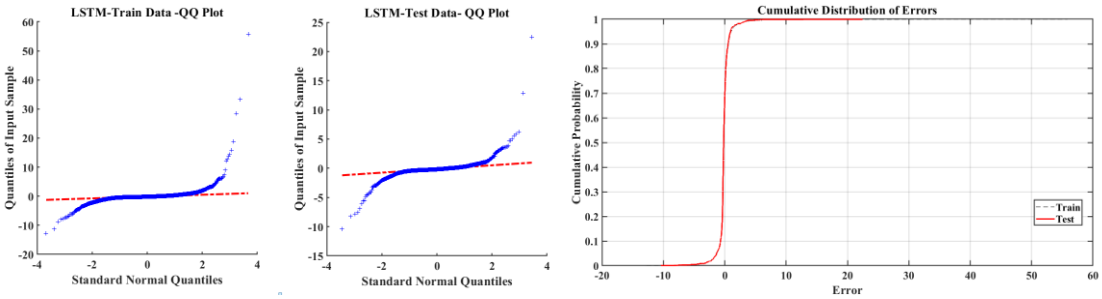
شکل ۱۲- نمودار سری زمانی بین داده‌های مشاهداتی و پیش‌بینی جریان رودخانه مدل منفرد LSTM در مراحل آموزش و آزمون

Fig. 12. Time series plot comparing observational data and river flow predictions from the single LSTM model during the training and testing phases



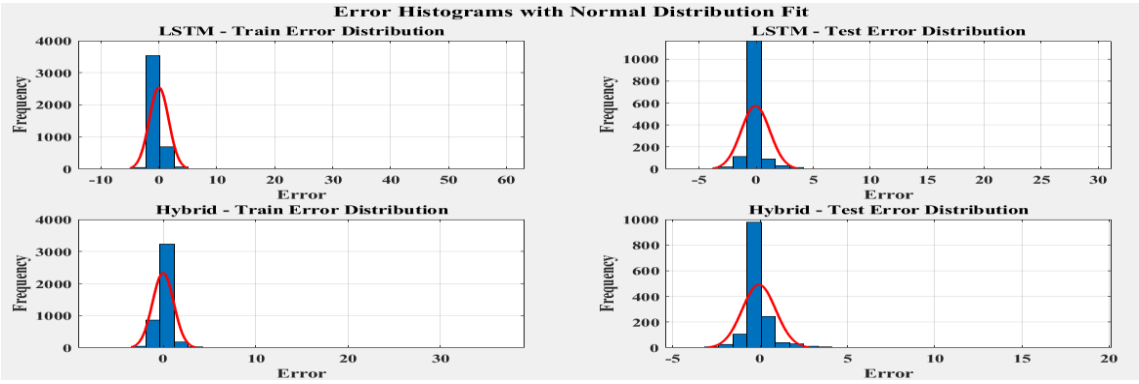
شکل ۱۳- نمودار هیستوگرام خطا و توزیع نرمال و نمودار خطای پیش‌بینی برای مدل منفرد LSTM در دو مرحله آموزش و آزمون

Fig. 13. Histogram of error and normal distribution, and prediction error plot for the single LSTM model during both the training and testing phases



شکل ۱۴- نمودار توزیع تجمعی و نمودار چندک- چندک خطا برای مدل منفرد LSTM

Fig. 14. Cumulative distribution plot and quantile-quantile error plot for the single LSTM model



شکل ۱۵-نمودار هیستوگرام خطا و تابع توزیع نرمال برای مدل منفرد و ترکیبی LSTM

Fig. 15. Error histogram and normal distribution function for the single and hybrid LSTM models

در تحقیقی Danesh et al., (2025)، سه مدل یادگیری ماشین جنگل تصادفی (RF)، درخت تصمیم (DT)، و نزدیک‌ترین همسایه (KNN) و سه مدل یادگیری عمیق شامل شبکه‌های عصبی پیچشی (CNN)، حافظه طولانی کوتاه مدت (LSTM) و یک مدل ترکیبی CNN-LSTM را در دو ایستگاه اندازه‌گیری در رودخانه مک‌کنزی در ایالات متحده مورد بررسی قرار دادند. نتایج نشان داد که مدل‌های یادگیری عمیق به‌طور مداوم در پیش‌بینی جریان رودخانه، بهتر از روش‌های یادگیری ماشین عمل کردند. مدل ترکیبی CNN-LSTM دقیق‌ترین پیش‌بینی‌ها را ارائه داد.

بخشی از نتایج با این پژوهش هم راستا است (Danesh et al., 2025). Khosravi et al., (2021)، از پنج مدل درختی RF، Bagging، M5P، RC و RS برای پیش‌بینی جریان رودخانه کورکورسر استفاده کردند. همچنین پنج مدل ترکیبی دیگر نیز مورد استفاده قرار گرفت. نتایج نشان داد که مدل‌های مبتنی بر جنگل تصادفی و ترکیبی عملکرد بهتری نسبت به سایر مدل های درختی دارا هستند که با نتایج این پژوهش هم راستا است (Khosravi et al., 2021). Arathy and Adarsh, (2024) مدل ترکیبی RF-LSTM برای پیش‌بینی جریان روزانه حوضه رودخانه بزرگ پامبا پرداختند. در پژوهش مذکور از متغیرهای روزانه جریان رودخانه، بیشینه و کمینه دما، بارش، از ۱ تا ۷ روزه تاخیر و برای یک دوره ۲۵ ساله استفاده شد.

نتایج نشان داد که این رویکرد ترکیبی نسبت به حالت‌ها منفرد بین ۱۰ و ۱۲ درصد دقت مدل را افزایش می‌دهد که نتایج آن از لحاظ برتری مدل‌های ترکیبی با نتایج ما در یک راستا است (Arathy and Adarsh, 2024).

Adnan et al., (2021)، جهت پیش‌بینی دقیق جریان رودخانه از مدل های LSTM، ELM و RF استفاده کردند. نتایج حاصل از مدل‌های اعمال شده نشان داد که LSTM دقت بالاتری نسبت به روش‌های ELM و RF ارائه می‌کند. نتایج این پژوهش حاکی از برتری مدل LSTM بر RF بوده که با نتایج این پژوهش هم‌راستا نیست اما در هر دو مدل به‌کار رفته مدل ترکیبی عملکرد به مراتب بهتری نسبت به مدل منفرد دارا بوده است (Adnan et al., 2021). Sharma et al., (2023)، به پیش‌بینی جریان رودخانه با استفاده از مدل‌های یادگیری ماشین (مدل‌های SVM، RF، LSTM و MARS) در رودخانه‌های جنوب هند پرداختند. یافته‌های پژوهش مذکور نشان داد که مدل RF برای جریان روزانه و MARS برای پیش‌بینی جریان ماهانه بهتر از سایر مدل‌ها عمل کردند. این نتایج حاکی از برتری مدل RF بر LSTM است که با نتایج این پژوهش هم‌راستا است (Sharma et al., 2024).

نتیجه‌گیری

این پژوهش با هدف توسعه و ارزیابی یک چارچوب ترکیبی پیشرفته برای پیش‌بینی جریان روزانه رودخانه کورکورسر نوشهر، به مقایسه عملکرد مدل‌های یادگیری ماشین جنگل تصادفی (RF) و یادگیری عمیق حافظه طولانی کوتاه‌مدت (LSTM) در دو حالت منفرد و ترکیبی با تبدیل موجک پرداخت. تحلیل نتایج نشان داد که ادغام تبدیل موجک، به‌ویژه موجک دابشیز ۴ (Db4)، با الگوریتم‌های یادگیری ماشین، رویکردی بسیار مؤثر برای بهبود دقت پیش‌بینی جریان‌های هیدرولوژیکی است. مدل ترکیبی جنگل تصادفی-موجک (RF-Wavelet) با عملکردی برجسته، به‌عنوان بهینه‌ترین مدل در این پژوهش شناخته شد.

این مدل در مرحله آزمون (R^2=0.942؛ RMSE=0.0106 m^3/s) دست یافت. این موفقیت نشان‌دهنده توانایی بالای موجک Db4 در استخراج ویژگی‌های چندمقیاسی و کاهش اغتشاش از نشانک پیچیده جریان رودخانه است که به مدل RF امکان می‌دهد الگوهای غیرخطی را با دقت بیشتری شناسایی کند. ترکیب مدل RF با موجک Db4 توانست خطای پیش‌بینی را در مقایسه با مدل منفرد RF حدود ۵۵ درصد کاهش دهد.

همچنین، مدل ترکیبی LSTM-Wavelet دقت پیش‌بینی را نسبت به مدل منفرد LSTM حدود ۳۹ درصد افزایش داد. این نتایج، نقش محوری پیش‌پردازش با موجک را در افزایش دقت و قابلیت تعمیم‌پذیری مدل‌های پیش‌بینی، حتی در مقایسه با معماری‌های پیشرفته یادگیری عمیق، برجسته می سازد. در مقایسه مستقیم مدل‌های ترکیبی، RF-

Wavelet با کاهش ۲۳ درصدی خطا نسبت به LSTM-Wavelet، برتری خود را به اثبات رساند. از سوی دیگر، مدل منفرد LSTM ضعیف‌ترین عملکرد را در میان تمامی مدل‌ها به ثبت رساند که نشان می‌دهد برای این مجموعه داده خاص، ساختار مبتنی بر درخت تصمیم RF انطباق بهتری نسبت به معماری حافظه-محور LSTM داشته است.

نتایج سناریوهای مختلف به‌طور واضح نشان داد که سناریوی چهارم (S4) که از ترکیب داده‌های بارش (Pt) و دبی با تأخیرهای زمانی یک، دو و سه‌روزه(Qt−1, Qt−2, Qt−3) استفاده می‌کرد، در تمامی مدل‌ها بهترین عملکرد را داشت. این موضوع تأیید می‌کند که افزودن حافظه سیستم (مقادیر دبی گذشته) در کنار داده‌های بارش برای پیش‌بینی دقیق جریان ضروری است. اگرچه مدل‌های توسعه‌یافته در این پژوهش عملکرد بسیار مطلوبی داشتند، اما رویکرد داده‌محور ذاتاً با عدم قطعیت‌هایی ناشی از کیفیت داده‌ها و عدم لحاظ صریح تغییرات فیزیکی حوزه آبخیز مواجه است.

با این حال، نتایج این تحقیق پیامدهای عملی مهمی برای مدیریت منابع آب دارد. پیش‌بینی‌های دقیق ارائه‌شده توسط مدل RF-Wavelet می‌تواند به‌طور مستقیم در سیستم‌های هشدار سیلاب، بهینه‌سازی تخصیص آب برای مصارف کشاورزی، شرب و مدیریت بهره‌برداری از نیروگاه‌های برق‌آبی به کار گرفته شود. تحقیقات آتی، پیشنهاد می‌شود سایر الگوریتم‌های یادگیری ماشین و عمیق در ترکیب با مدل‌های موجکی و الگوریتم‌های فرابتکاری برای بهینه‌سازی ابرفراسنجه‌ها مورد ارزیابی قرار گیرند. همچنین، ادغام داده‌های سنجش از دور به‌منظور لحاظ کردن تغییرات کاربری اراضی و پوشش گیاهی می‌تواند به افزایش دقت و واقع‌گرایی فیزیکی مدل‌ها کمک نماید.

تشکر و قدردانی

این مقاله برگرفته از رساله دکتری گروه مدیریت آب و ساخت دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران است. نویسندگان این پژوهش مراتب تشکر و قدردانی خود را از آن دانشگاه اعلام می‌دارند.

تعارض منافع

در این مقاله تضاد منافعی وجود ندارد و این مسأله مورد تأیید همه نویسندگان است.

منابع مورد استفاده

Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., Kisi, O., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577, 123981.

Adnan, R.M., Zounemat-Kermani, M., Kuriqi, A., Kisi, O., 2020. Machine learning method in prediction streamflow considering periodicity component. In Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation: Theory and Practice of Hazard Mitigation, 383-403. Singapore: Springer Singapore.

Al-Juboori, A.M., 2019. Generating monthly stream flow using nearest river data: Assessing different trees models. Water Resour. Manage. 33(9), 3257-3270.

Altman, D.G., Bland, J.M., 1999. Statistics notes variables and parameters. Bmj, 318(7199), 1667.

Arathy, N.G., Adarsh, S., 2024. A hybrid RF-LSTM model for daily streamflow prediction of greater Pamba River Basin, kerala incorporating dominant hydro-climatic drivers. In Recent Advances in Civil Engineering, 169-174. CRC Press.

Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.

Cheng, M., Fang, F., Kinouchi, T., Navon, I.M., Pain, C.C., 2020. Long lead-time daily and monthly streamflow forecasting using machine learning methods. J. Hydrol. 590, 125376.

Danesh, M., Gharehbaghi, A., Mehdizadeh, S., Danesh, A., 2025. A comparative assessment of machine learning and deep learning models for the daily river streamflow forecasting. Water Resour. Manage. 39(4), 1911-1930.

Difi, S., Elmeddahi, Y., Hebal, A., Singh, V.P., Heddam, S., Kim, S., Kisi, O., 2023. Monthly streamflow prediction using hybrid extreme learning machine optimized by bat algorithm: a case study of Cheliff watershed, Algeria. Hydrol. Sci. J. 68(2), 189-208.

Hadi, S.J., Tombul, M., 2018. Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study. Water Resour. Manage. 32(14), 4661-

4679.
Han, D., Cluckie, I.D., Karbassioun, D., Lawry, J., Krauskopf, B., 2002. River flow modelling using fuzzy decision trees. Water Resour. Manage. 16(6), 431-445.
Hastie, T., 2009. The elements of statistical learning: data mining, inference, and prediction.
Hunt, K.M., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrol. Earth Sys. Sci. 26(21), 5449-5472. https://doi.org/10.5194/hess-26-5449-2022
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning: with applications in R. 103, New York: Springer.
Khosravi, K., Golkarian, A., Tiefenbacher, J.P., 2022. Using optimized deep learning to predict daily streamflow: A comparison to common machine learning algorithms. Water Resour. Manage. 36(2), 699-716. https://doi.org/10.1007/s11269-021-03051-7
Khosravi, K., Miraki, S., Saco, P.M., Farmani, R., 2021. Short-term River streamflow modeling using Ensemble-based additive learner approach. J. Hydro-Environ. Res. 39, 81-91.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Sys. Sci. 23(12), 5089-5110. https://doi.org/10.5194/hess-23-5089-2019
Kumar, P.S., Praveen, T.V., Prasad, M.A., 2016. Artificial neural network model for rainfall-runoff-A case study. Int. J. Hybrid Info. Technol. 9(3), 263-272. http://dx.doi.org/10.14257/ijhit.2016.9.3.24
Latif, S.D., Ahmed, A.N., 2023. Streamflow prediction utilizing deep learning and machine learning algorithms for sustainable water supply management. Water Resour. Manage. 37(8), 3227-3241. https://doi.org/10.1007/s11269-023-03499-9
Le, X.H., Nguyen, D.H., Jung, S., Yeon, M., Lee, G., 2021. Comparison of deep learning techniques for river streamflow forecasting. IEEE Access. 9, 71805-71820.
Li, X., Sha, J., Wang, Z.L., 2019. Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. Hydrol. Sci. J. 64(15), 1857-1866. https://doi.org/10.1080/02626667.2019.1680846
Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., Dai, Y., 2021. A hybrid deep learning algorithm and its application to streamflow prediction. J. Hydrol. 601, 126636.
Liu, Z., Zhou, P., Chen, G., Guo, L., 2014. Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting. J. Hydrol. 519, 2822-2831. https://doi.org/10.1016/j.jhydrol.2014.06.050
Merufinia, E., Sharafati, A., Abghari, H., Hassanzadeh, Y., 2023. On the simulation of streamflow using hybrid tree-based machine learning models: A case study of Kurkursar basin, Iran. Arabian J. Geosci. 16(1), 28. https://doi.org/10.1007/s12517-022-11045-x
Muhammad, A.U., Li, X., Feng, J., 2019. Using LSTM GRU and hybrid models for streamflow forecasting. In International Conference on Machine Learning and Intelligent Communications, 510-524. Cham: Springer International Publishing. https://doi.org/10.2166/hydro.2025.276
Nguyen, N.Y., Kha, D.D., Ninh, L.V., Anh, V.T., Anh, T.N., 2025. Streamflow prediction using Long Short-Term Memory networks: a case study at the Kratie Hydrological Station, Mekong River Basin. J. Hydroinform. 27(2), 275-298. https://doi.org/10.2166/hydro.2025.276
Obeta, S., Grisan, E., Kalu, C.V., 2020. A comparative study of long short-term memory and gated recurrent unit. Available at SSRN 4442677.
Peng, T., Zhou, J., Zhang, C., Fu, W., 2017. Streamflow forecasting using empirical wavelet transform and artificial neural networks. Water 9(6), 406. https://doi.org/10.3390/w9060406
Pham, L.T., Luo, L., Finley, A.O., 2020. Evaluation of Random Forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds. Hydrol. Earth Sys. Sci. Discu. 1-33. https://doi.org/10.5194/hess-25-2997-2021.
Ragettli, S., Cortés, G., McPhee, J., Pellicciotti, F., 2014. An evaluation of approaches for modelling hydrological processes in high- elevation, glacierized Andean watersheds. Hydrol. Process. 28(23), 5674-5695. https://doi.org/10.1002/hyp.10055
Rahimzad, M., Moghaddam Nia, A., Zolfonoon, H., Soltani, J., Danandeh Mehr, A., Kwon, H.H., 2021. Performance comparison of an LSTM-based deep learning model versus conventional machine learning algorithms for streamflow forecasting. Water Resour. Manage. 35(12), 4167-4187. https://doi.org/10.1007/s11269-021-02937-w
Roy, B., Singh, M.P., 2020. An empirical-based rainfall-runoff modelling using optimization technique. Int. J. River Basin Manage. 18(1), 49-67. https://doi.org/10.1080/15715124.2019.1680557
Sahour, H., Gholami, V., Torkaman, J., Vazifedan, M., Saeedi, S., 2021. Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. Environ. Earth Sci. 80(22), 747. https://doi.org/10.1007/s12665-021-10054-5

Sharma, R.K., Kumar, S., Padmalal, D., Roy, A., 2024. Streamflow prediction using machine learning models in selected rivers of Southern India. Int. Journal River Basin Manage. 22(4), 529-555. https://doi.org/10.1080/15715124.2023.2196635

Singh, A., Saranya Das, K., Vijay, A., Nath, A.R., Chithra, N.R., 2025. Comparative study of different wavelet-machine learning models for agricultural drought prediction. Acta Geophysica 1-22. https://doi.org/10.1007/s11600-025-01660-z

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: some past experiences and new approaches. J. Hydroinform. 10(1), 3-22. https://doi.org/10.2166/hydro.2008.015

Wu, C.L., Chau, K.W., Fan, C., 2010. Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. J. Hydrol. 389(1-2), 146-167. https://doi.org/10.1016/j.jhydrol.2010.05.040

Xu, W., Chen, J., Zhang, X. J., 2022. Scale effects of the monthly streamflow prediction using a state-of-the-art deep learning model. Water Resour. Manage. 36(10), 3609-3625. https://doi.org/10.1007/s11269-022-03216-y